

Handbook of Statistical Genomics

Handbook of Statistical Genomics

Volume 1

Edited by

David J. Balding

University of Melbourne, Australia

Ida Moltke

University of Copenhagen, Denmark

John Marioni

University of Cambridge, United Kingdom

Fourth Edition

Founding Editors

Chris Cannings

Martin Bishop

WILEY

This fourth edition first published 2019

© 2019 John Wiley & Sons Ltd

Edition History

John Wiley & Sons, Ltd (1e, 2001), John Wiley & Sons, Ltd (2e, 2003), John Wiley & Sons, Ltd (3e, 2007)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law.

Advice on how to obtain permission to reuse material from this title is available at
<http://www.wiley.com/go/permissions>.

The right of David J. Balding, Ida Moltke and John Marioni to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Balding, D. J., editor. | Moltke, Ida, editor. | Marioni, John, editor.

Title: Handbook of statistical genomics / edited by David J. Balding (University of Melbourne, Australia),
Ida Moltke (University of Copenhagen, Denmark), and John Marioni (University of Cambridge, United Kingdom).

Other titles: Handbook of statistical genetics. | Handbook of statistical genetics.

Description: Fourth edition. | Hoboken, NJ : Wiley, 2019– | Previous title: Handbook of statistical genetics. | Includes bibliographical references and indexes. |

Identifiers: LCCN 2018060346 (print) | LCCN 2019003813 (ebook) | ISBN 9781119429227 (Adobe PDF) | ISBN 9781119429258 (ePub) | ISBN 9781119429142 (hardcover)

Subjects: LCSH: Genetics—Statistical methods—Handbooks, manuals, etc.

Classification: LCC QH438.4.S73 (ebook) | LCC QH438.4.S73 H36 2019 (print) | DDC 572.8/60727–dc23
LC record available at <https://lccn.loc.gov/2018060346>

Cover Design: Wiley

Cover Image: © Zita / Shutterstock

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

Contents

Volume 1

- List of Contributors xxiii**
Editors' Preface to the Fourth Edition xxvii
Glossary xxix
Abbreviations and Acronyms xxxix

1	Statistical Modeling and Inference in Genetics 1
	<i>Daniel Wegmann and Christoph Leuenberger</i>
1.1	Statistical Models and Inference 1
1.1.1	Statistical Models 2
1.1.2	Inference Methods and Algorithms 4
1.2	Maximum Likelihood Inference 4
1.2.1	Properties of Maximum Likelihood Estimators 6
1.2.2	Quantifying Confidence: the Fisher Information Matrix 8
1.2.3	Newton's Method 9
1.2.4	Latent Variable Problems: the EM Algorithm 11
1.2.5	Approximate Techniques 16
1.3	Bayesian Inference 20
1.3.1	Choice of Prior Distributions 21
1.3.2	Bayesian Point Estimates and Confidence Intervals 22
1.3.3	Markov Chain Monte Carlo 23
1.3.4	Empirical Bayes for Latent Variable Problems 30
1.3.5	Approximate Bayesian Computation 31
1.4	Model Selection 37
1.4.1	Likelihood Ratio Statistic 37
1.4.2	Bayesian Model Choice 38
1.5	Hidden Markov Models 40
1.5.1	Bayesian Inference of Hidden States Using Forward-Backward Algorithm 42
1.5.2	Finding the Most Likely Hidden Path (Viterbi Algorithm) 43
1.5.3	MLE Inference of Hierarchical Parameters (Baum–Welch Algorithm) 44
	Acknowledgements 46
	References 47

2	Linkage Disequilibrium, Recombination and Haplotype Structure	51
	<i>Gil McVean and Jerome Kelleher</i>	
2.1	What Is Linkage Disequilibrium?	51
2.2	Measuring Linkage Disequilibrium	53
	2.2.1 Single-Number Summaries of LD	54
	2.2.2 The Spatial Distribution of LD	56
	2.2.3 Various Extensions of Two-Locus LD Measures	60
2.3	Modelling Linkage Disequilibrium and Genealogical History	60
	2.3.1 A Historical Perspective	60
	2.3.2 Coalescent Modelling	62
	2.3.3 Relating Genealogical History to LD	67
2.4	Data Analysis	69
	2.4.1 Estimating Recombination Rates	69
	2.4.2 Methods Exploiting Haplotype Structure	72
2.5	Prospects	75
	Acknowledgements	75
	References	76
3	Haplotype Estimation and Genotype Imputation	87
	<i>Jonathan Marchini</i>	
3.1	Haplotype Estimation	87
	3.1.1 A Simple Haplotype Frequency Model	88
	3.1.2 Hidden Markov Models for Phasing	89
	3.1.3 Phasing in Related Samples	93
	3.1.4 Phasing Using Sequencing Data	94
	3.1.5 Phasing from a Reference Panel	95
	3.1.6 Measuring Phasing Performance	96
3.2	Genotype Imputation	97
	3.2.1 Uses of Imputation in GWASs	98
	3.2.2 Haploid Imputation	99
	3.2.3 Imputation Methods	100
	3.2.4 Testing Imputed Genotypes for Association	103
	3.2.5 Summary Statistic Imputation	104
	3.2.6 Factors Affecting Accuracy	104
	3.2.7 Quality Control for Imputed Data	107
3.3	Future Directions	109
	References	109
4	Mathematical Models in Population Genetics	115
	<i>Nick Barton and Alison Etheridge</i>	
4.1	Introduction	115
4.2	Single-Locus Models	116
	4.2.1 Random Drift and the Kingman Coalescent	117
	4.2.2 Diffusion Approximations	120
	4.2.3 Spatially Structured Populations	126
4.3	Multiple Loci	130
	4.3.1 Linkage Equilibrium	131
	4.3.2 Beyond Linkage Equilibrium	134
4.4	Outlook	140
	References	140

5	Coalescent Theory	145
	<i>Magnus Nordborg</i>	
5.1	Introduction	145
5.2	The Coalescent	146
	5.2.1 The Fundamental Insights	146
	5.2.2 The Coalescent Approximation	148
5.3	Generalizing the Coalescent	151
	5.3.1 Robustness and Scaling	151
	5.3.2 Variable Population Size	152
	5.3.3 Population Structure on Different Time-Scales	153
5.4	Geographical Structure	155
	5.4.1 The Structured Coalescent	155
	5.4.2 The Strong-Migration Limit	156
5.5	Diploidy and Segregation	157
	5.5.1 Hermaphrodites	157
	5.5.2 Males and Females	159
5.6	Recombination	159
	5.6.1 The Ancestral Recombination Graph	160
	5.6.2 Properties and Effects of Recombination	163
5.7	Selection	164
	5.7.1 Balancing Selection	165
	5.7.2 Selective Sweeps	166
	5.7.3 Background Selection	168
5.8	Neutral Mutations	168
5.9	Concluding Remarks	169
	5.9.1 The Coalescent and ‘Classical’ Population Genetics	169
	5.9.2 The Coalescent and Phylogenetics	169
	5.9.3 Prospects	171
	Acknowledgements	171
	References	171
6	Phylogeny Estimation Using Likelihood-Based Methods	177
	<i>John P. Huelsenbeck</i>	
6.1	Introduction	177
	6.1.1 Statistical Phylogenetics	178
	6.1.2 Chapter Outline	178
6.2	Maximum Likelihood and Bayesian Estimation	179
	6.2.1 Maximum Likelihood	179
	6.2.2 Bayesian Inference	180
6.3	Choosing among Models Using Likelihood Ratio Tests and Bayes Factors	184
6.4	Calculating the Likelihood for a Phylogenetic Model	186
	6.4.1 Character Matrices and Alignments	186
	6.4.2 The Phylogenetic Model	186
	6.4.3 Calculating the Probability of a Character History	187
	6.4.4 Continuous-Time Markov Model	188
	6.4.5 Marginalizing over Character Histories	189
6.5	The Mechanics of Maximum Likelihood and Bayesian Inference	192
	6.5.1 Maximum Likelihood	192
	6.5.2 Bayesian Inference and Markov Chain Monte Carlo	193

6.6	Applications of Likelihood-Based Methods in Molecular Evolution	199
6.6.1	A Taxonomy of Commonly Used Substitution Models	199
6.6.2	Expanding the Model around Groups of Sites	202
6.6.3	Rate Variation across Sites	204
6.6.4	Divergence Time Estimation	206
6.7	Conclusions	212
	References	213
7	The Multispecies Coalescent	219
	<i>Laura Kubatko</i>	
7.1	Introduction	219
7.2	Probability Distributions under the Multispecies Coalescent	221
7.2.1	Gene Tree Probabilities	221
7.2.2	Site Pattern Probabilities	227
7.2.3	Species Tree Likelihoods under the Multispecies Coalescent	229
7.2.4	Model Assumptions and Violations	230
7.3	Species Tree Inference under the Multispecies Coalescent	231
7.3.1	Summary Statistics Methods	231
7.3.2	Bayesian Full-Data Methods	234
7.3.3	Site Pattern-Based Methods	235
7.3.4	Multilocus versus SNP Data	236
7.3.5	Empirical Examples	237
7.4	Coalescent-Based Estimation of Parameters at the Population and Species Levels	239
7.4.1	Speciation Times and Population Sizes	239
7.4.2	Hybridization and Gene Flow	240
7.4.3	Species Delimitation	241
7.4.4	Future Prospects	242
	Acknowledgements	242
	References	242
8	Population Structure, Demography and Recent Admixture	247
	<i>G. Hellenthal</i>	
8.1	Introduction	247
8.1.1	'Admixture' versus 'Background' Linkage Disequilibrium	248
8.2	Spatial Summaries of Genetic Variation Using Principal Components Analysis	249
8.3	Clustering Algorithms	251
8.3.1	Defining 'Populations'	251
8.3.2	Clustering Based on Allele Frequency Patterns	252
8.3.3	Incorporating Admixture	253
8.3.4	Incorporating Admixture Linkage Disequilibrium	254
8.3.5	Incorporating Background Linkage Disequilibrium: Using Haplotypes to Improve Inference	255
8.3.6	Interpreting Genetic Clusters	258
8.4	Inferring Population Size Changes and Split Times	259
8.4.1	Allele Frequency Spectrum Approaches	260
8.4.2	Approaches Using Whole-Genome Sequencing	261
8.5	Identifying/Dating Admixture Events	262
8.5.1	Inferring DNA Segments Inherited from Different Sources	263
8.5.2	Measuring Decay of Linkage Disequilibrium	265

8.6	Conclusion 267	
	Acknowledgements 268	
	References 268	
9	Statistical Methods to Detect Archaic Admixture and Identify Introgressed Sequences 275	
	<i>Liming Li and Joshua M. Akey</i>	
9.1	Introduction 275	
9.2	Methods to Test Hypotheses of Archaic Admixture and Infer Admixture Proportions 277	
	9.2.1 Genetic Drift and Allele Frequency Divergence in Genetically Structured Populations 277	
	9.2.2 Three-Population Test 277	
	9.2.3 D-Statistic 279	
	9.2.4 F_4 -Statistic 282	
9.3	Methods to Identify Introgressed Sequences 283	
	9.3.1 S^* -Statistic 284	
	9.3.2 Hidden Markov and Conditional Random Field Models 287	
	9.3.3 Relative Advantages and Disadvantages of Approaches to Detect Introgressed Sequences 289	
9.4	Summary and Perspective 289	
	References 290	
10	Population Genomic Analyses of DNA from Ancient Remains 295	
	<i>Torsten Günther and Mattias Jakobsson</i>	
10.1	Introduction 295	
10.2	Challenges of Working with and Analyzing Ancient DNA Data 296	
	10.2.1 Sequence Degradation 296	
	10.2.2 Contamination 297	
	10.2.3 Handling Sequence Data from Ancient Material 300	
	10.2.4 Different Sequencing Approaches and the Limitations in their Resulting Data 301	
	10.2.5 Effects of Limited Amounts of Data on Downstream Analysis 301	
10.3	Opportunities of Ancient DNA 302	
	10.3.1 Population Differentiation in Time and Space 303	
	10.3.2 Continuity 306	
	10.3.3 Migration and Admixture over Time 307	
	10.3.4 Demographic Inference Based on High-Coverage Ancient Genomes 308	
	10.3.5 Allele Frequency Trajectories 308	
10.4	Some Examples of How Genetic Studies of Ancient Remains Have Contributed to a New Understanding of the Human Past 310	
	10.4.1 Archaic Genomes and the Admixture with Modern Humans 310	
	10.4.2 Neolithic Revolution in Europe and the Bronze Age Migrations 311	
10.5	Summary and Perspective 313	
	Acknowledgements 313	
	References 314	
11	Sequence Covariation Analysis in Biological Polymers 325	
	<i>William R. Taylor, Shaun Kandathil, and David T. Jones</i>	
11.1	Introduction 325	

11.2	Methods 326
11.2.1	DCA Method 326
11.2.2	PSICOV 327
11.2.3	plmDCA, GREMLIN and CCMpred 327
11.3	Applications 328
11.3.1	Globular Protein Fold Prediction 328
11.3.2	Transmembrane Protein Prediction 328
11.3.3	RNA Structure Prediction 328
11.3.4	Protein Disordered Regions 329
11.3.5	Protein–Protein Interactions 329
11.3.6	Allostery and Dynamics 330
11.3.7	CASP 330
11.4	New Developments 332
11.4.1	Sequence Alignment 332
11.4.2	Comparison to Known Structures 333
11.4.3	Segment Parsing 334
11.4.4	Machine Learning 335
11.4.5	Deep Learning Methods 336
11.4.6	Sequence Pairing 338
11.4.7	Phylogeny Constraints 339
11.5	Outlook 340
	Acknowledgements 341
	References 342
12	Probabilistic Models for the Study of Protein Evolution 347
	<i>Umberto Pernon, Iain H. Moal, Jeffrey L. Thorne, and Nick Goldman</i>
12.1	Introduction 347
12.2	Empirically Derived Models of Amino Acid Replacement 348
12.2.1	The Dayhoff and Eck Model 348
12.2.2	Descendants of the Dayhoff Model 350
12.3	Heterogeneity of Replacement Rates among Sites 351
12.4	Protein Structural Environments 351
12.5	Variation of Preferred Residues among Sites 353
12.6	Models with a Physicochemical Basis 355
12.7	Codon-Based Models 355
12.8	Dependence among Positions 357
12.9	Stochastic Models of Structural Evolution 359
12.10	Conclusion 360
	Acknowledgements 361
	References 361
13	Adaptive Molecular Evolution 369
	<i>Ziheng Yang</i>
13.1	Introduction 369
13.2	Markov Model of Codon Substitution 371
13.3	Estimation of Synonymous and Non-synonymous Substitution Rates between Two Sequences and Test of Selection on the Protein 372
13.3.1	Heuristic Estimation Methods 372
13.3.2	Maximum Likelihood Estimation 374

13.3.3	Bayesian Estimation	377
13.3.4	A Numerical Example	377
13.4	Likelihood Calculation on a Phylogeny	379
13.5	Detecting Adaptive Evolution along Lineages	380
13.5.1	Likelihood Calculation under Models of Variable ω Ratios among Lineages	380
13.5.2	Adaptive Evolution in the Primate Lysozyme	381
13.5.3	Comparison with Methods Based on Reconstructed Ancestral Sequences	382
13.6	Inferring Amino Acid Sites under Positive Selection	384
13.6.1	Likelihood Ratio Test under Models of Variable ω Ratios among Sites	384
13.6.2	Methods that Test One Site at a Time	386
13.6.3	Positive Selection in the HIV-1 <i>vif</i> Genes	386
13.7	Testing Positive Selection Affecting Particular Sites and Lineages	388
13.7.1	Branch-Site Test of Positive Selection	388
13.7.2	Clade Models and Other Variants	389
13.8	Limitations of Current Methods	390
13.9	Computer Software	391
	References	391
14	Detecting Natural Selection	397
	<i>Aaron J. Stern and Rasmus Nielsen</i>	
14.1	Introduction	397
14.2	Types of Selection	398
14.2.1	Directional Selection	398
14.2.2	Balancing Selection	399
14.2.3	Polygenic Selection	399
14.3	The Signature of Selection in the Genome	399
14.3.1	The Signature of Positive Directional Selection	400
14.3.2	Balancing Selection	403
14.3.3	Polygenic Selection	403
14.3.4	Confounders	404
14.4	Methods for Detecting Selection	405
14.4.1	Substitution-Based Methods	405
14.4.2	Methods Comparing Substitutions and Diversity	406
14.4.3	Methods Using the Frequency Spectrum	407
14.4.4	Methods Using Genetic Differentiation	408
14.4.5	Methods Using Haplotype Structure	410
14.4.6	Why Full-Likelihood Methods Are Intractable for Population Samples	412
14.4.7	Composite Likelihood Methods	412
14.4.8	Approximate Bayesian Computation	413
14.4.9	Machine Learning Methods	413
14.5	Discussion	414
	References	415
15	Evolutionary Quantitative Genetics	421
	<i>Bruce Walsh and Michael B. Morrissey</i>	
15.1	Introduction	421

15.2	Resemblances, Variances, and Additive Genetic Values	422
15.2.1	Fisher's Genetic Decomposition	422
15.2.2	Additive Genetic Variances and Covariances	423
15.3	Parent–Offspring Regressions and the Response to Selection	423
15.3.1	Single-Trait Parent–Offspring Regressions	424
15.3.2	Selection Differentials and the Breeder's Equation	424
15.3.3	Multiple-Trait Parent–Offspring Regressions	425
15.3.4	The Genetic and Phenotypic Covariance Matrices	425
15.3.5	The Multivariate Breeder's Equation	425
15.4	The Infinitesimal Model	426
15.4.1	Linearity of Parent–Offspring Regressions under the Infinitesimal Model	426
15.4.2	Allele Frequency Changes under the Infinitesimal Model	426
15.4.3	Changes in Variances	427
15.4.4	The Equilibrium Additive Genetic Variance	429
15.5	Inference of σ_A^2 and \mathbf{G}	430
15.6	Fitness	432
15.6.1	Individual Fitness	432
15.6.2	Episodes of Selection	433
15.7	The Robertson–Price Identity, and Theorems of Selection	434
15.7.1	Description of the Theorems	435
15.7.2	Empirical Operationalization of the Theorems	436
15.8	The Opportunity for Selection	437
15.9	Selection Coefficients	438
15.9.1	Measures of Selection on the Mean	439
15.9.2	Measures of Selection on the Variance	439
15.10	Fitness Functions and the Characterization of Selection	441
15.10.1	Individual and Mean Fitness Functions	441
15.10.2	Gradients and the Local Geometry of Fitness Surfaces	442
15.11	Multivariate Selection	444
15.11.1	Short-Term Changes in Means: The Multivariate Breeder's Equation	444
15.11.2	The Effects of Genetic Correlations: Direct and Correlated Responses	444
15.11.3	Selection Gradients and Understanding which Traits Affect Fitness	446
15.12	Inference of Selection Gradients	447
15.12.1	Ordinary Least Squares Analysis	448
15.12.2	Flexible Inference of Fitness Functions with Associated Selection Gradient Estimates	449
15.12.3	Normality and Selection Gradients	450
15.13	Summary	451
	References	452
16	Conservation Genetics	457
	<i>Mark Beaumont and Jinliang Wang</i>	
16.1	Introduction	457
16.2	Estimating Effective Population Size	458
16.2.1	Methods Based on Heterozygosity Excess	459
16.2.2	Methods Based on Linkage Disequilibrium	460
16.2.3	Methods Based on Relatedness	463
16.2.4	Methods Based on Temporal Changes in Allele Frequency	466

16.3	Estimating Census Size by the Genotype Capture–Recapture Approach	470
16.3.1	Methods Based on Multilocus Genotype Mismatches	472
16.3.2	Methods Based on Pairwise Relatedness	472
16.3.3	Methods Based on Pairwise Relationships	473
16.3.4	Methods Based on Pedigree Reconstruction	474
16.4	Inferring Genetic Structure	475
16.4.1	Measuring Genetic Differentiation	475
16.4.2	Population Assignment	477
16.4.3	Population Clustering and Inference of Ancestry Proportions	479
16.4.4	Inferring Levels of Recent Gene Flow	483
16.4.5	Landscape Genetics	486
16.5	Deintrogression Strategies	487
16.6	Genetic Species Delimitation	489
16.7	Conclusions and Outlook	491
	Acknowledgements	492
	References	492
17	Statistical Methods for Plant Breeding	501
	<i>Ian Mackay, Hans-Peter Piepho, and Antonio Augusto Franco Garcia</i>	
17.1	Introduction	501
17.2	Heritability and the Breeder's Equation in Plant Breeding	502
17.3	The Breeding System of Plants	504
17.4	Polyploidy in Plants and Its Genetic Consequences	505
17.5	Genomic Rearrangements in Plants	509
17.6	Genetic Architecture of Traits in Plants	510
17.7	Response to the Environment and Plasticity	511
17.8	Genomic Selection	514
17.8.1	Genotype–Environment Interaction	514
17.8.2	Quantitative Trait Loci and Major Genes	516
17.8.3	Genomic Selection and Cross Prediction	517
17.8.4	Genomic Selection and Phenotyping Cost	517
17.8.5	Mate Selection	517
17.8.6	Sequential Selection	518
17.8.7	Genomic Prediction of Hybrid Performance and Heterosis	518
17.8.8	Marker Imputation	519
17.9	Experimental Design and Analysis	519
17.10	Conclusions	521
	References	521
18	Forensic Genetics	531
	<i>B.S. Weir</i>	
18.1	Introduction	531
18.2	Principles of Interpretation	532
18.3	Profile Probabilities	534
18.3.1	Genetic Models for Allele Frequencies	535
18.3.2	Y-STR Profiles	539
18.4	Mixtures	542
18.4.1	Combined Probabilities of Inclusion and Exclusion	542
18.4.2	Likelihood Ratios	542
18.5	Behavior of Likelihood Ratio	546

18.6	Single Nucleotide Polymorphism, Sequence and Omic Data	547
	References	548

Volume 2

List of Contributors	<i>xxiii</i>
Editors' Preface to the Fourth Edition	<i>xxvii</i>
Glossary	<i>xxix</i>
Abbreviations and Acronyms	<i>xxxix</i>

19	Ethical Issues in Statistical Genetics	551
	<i>Susan E. Wallace and Richard Ashcroft</i>	
19.1	Introduction	551
	19.1.1 What Is Ethics?	552
	19.1.2 Models for Analysing the Ethics of Population Genetic Research	553
19.2	Ethics and Governance in Population Genetics Research: Two Case Studies	554
	19.2.1 'Healthy Volunteer' Longitudinal Cohort Studies: UK Biobank	555
	19.2.2 Precision Medicine Approaches: 100,000 Genomes Project	556
	19.2.3 The Scientific and Clinical Value of the Research	556
	19.2.4 Recruitment of Participants	558
	19.2.5 Consent	559
	19.2.6 Returning Individual Genetic Research Results	563
	19.2.7 Confidentiality and Security	564
19.3	Stewardship and Wider Social Issues	565
	19.3.1 Benefit Sharing	566
	19.3.2 Community Involvement and Public Engagement	567
	19.3.3 Race, Ethnicity and Genetics	567
19.4	Conclusion	568
	Acknowledgements	568
	References	568
20	Descent-Based Gene Mapping in Pedigrees and Populations	573
	<i>E.A. Thompson</i>	
20.1	Introduction to Genetic Mapping and Genome Descent	573
	20.1.1 Genetic Mapping: The Goal and the Data	573
	20.1.2 The Process of Meiosis and the Descent of DNA	574
	20.1.3 Genetic Linkage Mapping: Association or Descent?	576
20.2	Inference of Local IBD Sharing from Genetic Marker Data	577
	20.2.1 Identity by Descent at a Locus	577
	20.2.2 Probabilities of Marker Data Given IBD Pattern	579
	20.2.3 Modeling the Probabilities of Patterns of IBD	580
	20.2.4 Inferring Local IBD from Marker Data	581
20.3	IBD-Based Detection of Associations between Markers and Traits	583
	20.3.1 Trait Data Probabilities for Major Gene Models	583
	20.3.2 Quantitative Trait Data Probabilities under Random Effects Models	584
	20.3.3 IBD-Based Linkage Likelihoods for Major Gene Models	585
	20.3.4 IBD-Based Linkage Likelihoods for Random-Effects Models	587
20.4	Other Forms of IBD-Based Genetic Mapping	589

20.4.1	IBD-Based Case–Control Studies	589
20.4.2	Patterns of IBD in Affected Relatives	590
20.5	Summary	592
	Acknowledgements	592
	References	593
21	Genome-Wide Association Studies	597
	<i>Andrew P. Morris and Lon R. Cardon</i>	
21.1	Introduction	597
21.2	GWAS Design Concepts	599
21.2.1	Phenotype Definition	599
21.2.2	Structure of Common Genetic Variation and Design of GWAS Genotyping Technology	599
21.2.3	Sample Size Considerations	601
21.2.4	Genome-Wide Significance and Correction for Multiple Testing	601
21.2.5	Replication	602
21.3	GWAS Quality Control	602
21.3.1	SNP Quality Control Procedures	603
21.3.2	Sample Quality Control Procedures	604
21.3.3	Software	606
21.4	Single SNP Association Analysis	606
21.4.1	Generalised Linear Modelling Framework	606
21.4.2	Accounting for Confounding Factors as Covariates	606
21.4.3	Coding of SNP Genotypes	607
21.4.4	Imputed Genotypes	609
21.4.5	Visualisation of Results of Single SNP GWAS Analyses	609
21.4.6	Interactions with Non-Genetic Risk Factors	609
21.4.7	Bayesian Methods	611
21.4.8	Software	611
21.5	Detecting and Accounting for Genetic Structure in GWASs	611
21.5.1	Identification of Related Individuals	612
21.5.2	Multivariate Approaches to Identify Ethnic Outliers and Account for Population Stratification	613
21.5.3	Mixed Modelling Approaches to Account for Genetic Structure	614
21.5.4	Software	615
21.6	Multiple SNP Association Analysis	616
21.6.1	Haplotype-Based Analyses	616
21.6.2	SNP–SNP Interaction Analyses	617
21.6.3	Gene-Based Analyses	619
21.6.4	Software	619
21.7	Discussion	620
	References	623
22	Replication and Meta-analysis of Genome-Wide Association Studies	631
	<i>Frank Dudbridge and Paul Newcombe</i>	
22.1	Introduction	631
22.2	Replication	632
22.2.1	Motivation	632
22.2.2	Different Forms of Replication	632

22.2.3	Two-Stage Genome-Wide Association Studies	634
22.2.4	Significance Thresholds for Replication	634
22.2.5	A Key Challenge: Heterogeneity	635
22.3	Winner's Curse	635
22.3.1	Description of the Problem	635
22.3.2	Methods for Correcting for Winner's Curse	637
22.3.3	Applicability of These Methods	639
22.4	Meta-analysis	640
22.4.1	Motivation	640
22.4.2	An Illustrative Example	640
22.4.3	Fixed Effect Meta-analysis	641
22.4.4	Chi-Square Test for Heterogeneity in Effect	641
22.4.5	Random Effects Meta-analysis	642
22.4.6	Interpretation and Significance Testing of Meta-analysis Estimates	643
22.4.7	Using Funnel Plots to Investigate Small Study Bias in Meta-analysis	644
22.4.8	Improving Analyses via Meta-analysis Consortia and Publicly Available Data	645
22.5	Summary	647
	References	647
23	Inferring Causal Relationships between Risk Factors and Outcomes Using Genetic Variation	651
	<i>Stephen Burgess, Christopher N. Foley, and Verena Zuber</i>	
23.1	Background	651
23.1.1	Correlation and Causation	651
23.1.2	Chapter Outline	652
23.2	Introduction to Mendelian Randomization and Motivating Example	652
23.2.1	Instrumental Variable Assumptions	653
23.2.2	Assessing the Instrumental Variable Assumptions	654
23.2.3	Two-Sample Mendelian Randomization and Summarized Data	655
23.3	Monogenic Mendelian Randomization Analyses: The Easy Case	655
23.4	Polygenic Mendelian Randomization Analyses: The Difficult Case	656
23.4.1	Example: Low-Density Lipoprotein Cholesterol and Coronary Heart Disease Risk	656
23.4.2	More Complex Examples	657
23.4.3	Two-Stage Least Squares and Inverse-Variance Weighted Methods	659
23.5	Robust Approaches for Polygenic Mendelian Randomization Analyses	660
23.5.1	Median Estimation Methods	660
23.5.2	Modal Estimation Methods	660
23.5.3	Regularization Methods	660
23.5.4	Other Outlier-Robust Methods	661
23.5.5	MR-Egger Method	661
23.5.6	Multivariable Methods	663
23.5.7	Interactions and Subsetting	663
23.5.8	Practical Advice	664
23.6	Alternative Approaches for Causal Inference with Genetic Data	665
23.6.1	Fine-Mapping and Colocalization	665
23.6.2	LD Score Regression	666
23.7	Causal Estimation in Mendelian Randomization	667

23.7.1	Relevance of Causal Estimate	667
23.7.2	Heterogeneity and Pleiotropy	668
23.7.3	Weak Instrument Bias and Sample Overlap	668
23.7.4	Time-Dependent Causal Effects	669
23.7.5	Collider Bias	670
23.8	Conclusion	670
	References	671
24	Improving Genetic Association Analysis through Integration of Functional Annotations of the Human Genome	679
	<i>Qiongshi Lu and Hongyu Zhao</i>	
24.1	Introduction	679
24.2	Types of Functional Annotation Data in GWAS Applications	680
24.2.1	Transcriptomic Annotation Data	680
24.2.2	Epigenetic Annotation Data	681
24.2.3	DNA Conservation	681
24.3	Methods to Synthesize Annotation Data	682
24.3.1	Genome Browsers and Annotator Software	682
24.3.2	Supervised Learning Methods	682
24.3.3	Unsupervised Learning Methods	684
24.3.4	Improving Specificity of Computational Annotations	685
24.4	Methods to Integrate Functional Annotations in Genetic Association Analysis	685
24.4.1	Partitioning Heritability and Genetic Covariance	685
24.4.2	Imputation-Based Gene-Level Association Analysis	688
24.4.3	Other Applications and Future Directions	690
	Acknowledgements	690
	References	691
25	Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci	697
	<i>Solveig K. Sieberts and Eric E. Schadt</i>	
25.1	Introduction	697
25.1.1	An Overview of Transcription as a Complex Process	700
25.1.2	Modeling Approaches for Biological Processes	702
25.1.3	Human versus Experimental Models	704
25.2	Modeling for eQTL Detection and Causal Inference	705
25.2.1	Heritability of Expression Traits	705
25.2.2	Single-Trait eQTL Mapping	706
25.2.3	Joint eQTL Mapping	706
25.2.4	eQTL and Clinical Trait Linkage Mapping to Infer Causal Associations	708
25.3	Inferring Gene Regulatory Networks	714
25.3.1	From Assessing Causal Relationships among Trait Pairs to Predictive Gene Networks	714
25.3.2	Building from the Bottom Up or Top Down?	714
25.3.3	Using eQTL Data to Reconstruct Coexpression Networks	715
25.3.4	An Integrative Genomics Approach to Constructing Predictive Network Models	718

25.3.5	Integrating Genetic Data as a Structure Prior to Enhance Causal Inference in the Bayesian Network Reconstruction Process	720
25.3.6	Incorporating Other Omics Data as Network Priors in the Bayesian Network Reconstruction Process	721
25.3.7	Illustrating the Construction of Predictive Bayesian Networks with an Example	722
25.4	Conclusions	723
25.5	Software	724
	References	725
26	Statistical Methods for Single-Cell RNA-Sequencing	735
	<i>Tallulah S. Andrews, Vladimir Yu. Kiselev, and Martin Hemberg</i>	
26.1	Introduction	735
26.2	Overview of scRNA-Seq Experimental Platforms and Low-Level Analysis	736
26.2.1	Low-Throughput Methods	736
26.2.2	High-Throughput Methods	737
26.2.3	Computational Analysis	739
26.3	Novel Statistical Challenges Posed by scRNA-Seq	739
26.3.1	Estimating Transcript Levels	739
26.3.2	Analysis of the Expression Matrix	747
	References	753
27	Variant Interpretation and Genomic Medicine	761
	<i>K. Carss, D. Goldstein, V. Aggarwal, and S. Petrovski</i>	
27.1	Introduction and Current Challenges	761
27.2	Understanding the Effect of a Variant	765
27.3	Understanding Genomic Variation Context through Large Human Reference Cohorts	771
27.4	Functional Assays of Genetic Variation	777
27.5	Leveraging Existing Information about Gene Function Including Human and Model Phenotype Resources	779
27.6	Holistic Variant Interpretation	782
27.7	Future Challenges and Closing Remarks	783
27.8	Web Resources	786
	References	788
28	Prediction of Phenotype from DNA Variants	799
	<i>M.E. Goddard, T.H.E. Meuwissen, and H.D. Daetwyler</i>	
28.1	Introduction	799
28.2	Genetic Variation Affecting Phenotype	800
28.3	Data on DNA Polymorphisms Used for Prediction of Genetic Effects	801
28.4	Prediction of Additive Genetic Values	802
28.4.1	An Equivalent Model	804
28.4.2	Single-Step BLUP	804
28.4.3	Multiple Traits	805
28.4.4	Gene Expression	806
28.4.5	Using External Information	806
28.5	Factors Affecting Accuracy of Prediction	806
28.6	Other Uses of the Bayesian Genomic Selection Models	808

28.7	Examples of Genomic Prediction	809
28.7.1	Cattle	809
28.7.2	Humans	809
28.8	Conclusions	810
	References	810
29	Disease Risk Models	815
	<i>Allison Meisner and Nilanjan Chatterjee</i>	
29.1	Introduction and Background	815
29.1.1	Disease Risk Models and Their Applications	815
29.1.2	Examples of Available Disease Risk Models	817
29.1.3	Incorporating Genetic Factors	817
29.2	Absolute Risk Model	818
29.2.1	General Software for Building Absolute Risk Models	820
29.3	Building a Polygenic Risk Score	821
29.3.1	Expected Performance	821
29.3.2	Standard Approach to Constructing a PRS: LD Clumping and <i>p</i> -Value Thresholding	823
29.3.3	Advanced Approaches to Constructing a PRS	825
29.4	Combining PRS and Epidemiologic Factors	826
29.5	Model Validation	827
29.6	Evaluating Clinical Utility	828
29.7	Example: Breast Cancer	829
29.8	Discussion	832
29.8.1	Future Directions	832
29.8.2	Challenges	832
	References	833
30	Bayesian Methods for Gene Expression Analysis	843
	<i>Alex Lewin, Leonardo Bottolo, and Sylvia Richardson</i>	
30.1	Introduction	843
30.2	Modelling Microarray Data	845
30.2.1	Modelling Intensities	845
30.2.2	Gene Variability	845
30.2.3	Normalization	846
30.3	Modelling RNA-Sequencing Reads	846
30.3.1	Alignments for RNA-Sequencing Data	846
30.3.2	Likelihood for Read-Level Data	847
30.3.3	Likelihood for Transcript-Level Read Counts	848
30.3.4	Likelihood for Gene-Level Read Counts	850
30.4	Priors for Differential Expression Analysis	851
30.4.1	Differential Expression from Microarray Data	851
30.4.2	Differential Expression from RNA-Sequencing Data	856
30.5	Multivariate Gene Selection Models	857
30.5.1	Variable Selection Approach	857
30.5.2	Bayesian Shrinkage with Sparsity Priors	861
30.6	Quantitative Trait Loci	863
30.6.1	Single-Response Models	863
30.6.2	Multiple-Response Models	864

Acknowledgements	868
References	869
31 Modelling Gene Expression Dynamics with Gaussian Process Inference	879
<i>Magnus Rattray, Jing Yang, Sumon Ahmed, and Alexis Boukouvalas</i>	
31.1 Introduction	879
31.1.1 Covariance Function	880
31.1.2 Inference	882
31.2 Applications to Bulk Time Series Expression Data	883
31.2.1 Identifying Differential Expression in Time	884
31.2.2 Identifying Changes between Two Time Course Experiments	885
31.2.3 Hierarchical Models of Replicates and Clusters	887
31.2.4 Differential Equation Models of Production and Degradation	888
31.3 Modelling Single-Cell Data	889
31.3.1 Modelling Single-Cell Trajectory Data	889
31.3.2 Dimensionality Reduction and Pseudotime Inference	890
31.3.3 Modelling Branching Dynamics with Single-Cell RNA-Sequencing Data	892
31.4 Conclusion	893
Acknowledgements	894
References	894
32 Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models	899
<i>Marco Grzegorczyk and Dirk Husmeier</i>	
32.1 Introduction	899
32.2 Methodology	901
32.2.1 Dynamic Bayesian Networks (DBN)	901
32.2.2 Bayesian Linear Regression	902
32.2.3 Bayesian Piecewise Linear Regression (NH-DBN)	905
32.2.4 Bayesian Piecewise Linear Regression with Coupled Regression Coefficients (Coupled NH-DBNs)	908
32.2.5 NH-DBNs with More Flexible Allocation Schemes	915
32.2.6 NH-DBNs with Time-Varying Network Structures	916
32.2.7 Dynamic Bayesian Network Modelling	918
32.2.8 Computational Complexity	920
32.3 Application Examples	921
32.3.1 Morphogenesis in Drosophila	921
32.3.2 Synthetic Biology in Yeast	922
32.4 Summary	927
Appendix A: Coupling Schemes	927
A.1 Hard Information Coupling Based on an Exponential Prior	928
A.2 Hard Information Coupling Based on a Binomial Prior	928
A.3 Soft Information Coupling Based on a Binomial Prior	929
References	929
33 DNA Methylation	933
<i>Kasper D. Hansen, Kimberly D. Siegmund, and Shili Lin</i>	
33.1 A Brief Introduction	933
33.2 Measuring DNA Methylation	934

33.3	Differential DNA Methylation	936
33.3.1	Differential Methylation with Bisulfite-Sequencing Data	936
33.3.2	Differential Methylation with Capture-Sequence Data	939
33.3.3	Differential Methylation with HumanMethylation Array Data	940
33.4	Other Topics of Interest	941
	References	942
34	Statistical Methods in Metabolomics	949
	<i>Timothy M.D. Ebbels, Maria De Iorio, and David A. Stephens</i>	
34.1	Introduction	949
34.2	Preprocessing and Deconvolution	950
34.2.1	Nuclear Magnetic Resonance Spectroscopy	950
34.2.2	Liquid Chromatography – Mass Spectrometry	952
34.3	Univariate Methods	954
34.3.1	Metabolome-Wide Significance Levels	956
34.3.2	Sample Size and Power	957
34.4	Multivariate Methods and Chemometrics Techniques	958
34.4.1	Linear Regression Methods	959
34.4.2	Shrinkage Methods	960
34.5	Orthogonal Projection Methods	961
34.5.1	Principal Components Analysis	962
34.5.2	Partial Least Squares	964
34.5.3	Orthogonal Projection onto Latent Structures	965
34.6	Network Analysis	966
34.7	Metabolite Identification and Pathway Analysis	969
34.7.1	Statistical Correlation Spectroscopy	969
34.7.2	Pathway and Metabolite Set Analysis	971
34.8	Conclusion	972
	References	972
35	Statistical and Computational Methods in Microbiome and Metagenomics	977
	<i>Hongzhe Li</i>	
35.1	Microbiome in Human Health and Disease	977
35.2	Estimation of Microbiome Features from 16S rRNA and Shotgun Metagenomic Sequencing Data	980
35.2.1	Estimation of Microbiome Features in 16S rRNA Data	980
35.2.2	Estimation of Microbial Composition in Shotgun Metagenomic Data	981
35.2.3	Estimation of Microbial Gene/Pathway Abundance in Shotgun Metagenomic Data	982
35.2.4	Quantification of Bacterial Growth Dynamics	982
35.2.5	Microbial Diversity Index	983
35.3	Methods for Analysis of Microbiome as an Outcome of an Intervention or Exposure	983
35.3.1	Modeling Multivariate Sparse Count Data as the Response Variable	984
35.3.2	Modeling High-Dimensional Compositional Response Data in Microbiome Studies	984
35.4	Methods for Analysis of Microbiome as a Covariate	985
35.4.1	Regression Analysis with Compositional Covariates	985
35.4.2	Kernel-Based Regression in Microbiome Studies	986

35.5	Methods for Analysis of Microbiome as a Mediator	987
35.6	Integrative Analysis of Microbiome, Small Molecules and Metabolomics Data	989
35.6.1	Computational Analysis of Small Molecules from the Human Microbiota	989
35.6.2	Metabolic Modeling in Microbiome	990
35.7	Discussion and Future Directions	991
	Acknowledgements	991
	References	992
36	Bacterial Population Genomics	997
	<i>Jukka Corander, Nicholas J. Croucher, Simon R. Harris, John A. Lees, and Gerry Tonkin-Hill</i>	
36.1	Introduction	997
36.2	Genetic Population Structure and Clustering of Genotypes	998
36.2.1	Background	998
36.2.2	Model-Based Clustering	998
36.2.3	Linkage Disequilibrium	1000
36.2.4	Distance-Based Methods	1000
36.3	Phylogenetics and Dating Analysis	1001
36.4	Transmission Modeling	1004
36.4.1	Challenges	1004
36.5	Genome-Wide Association Studies in Bacteria	1008
36.5.1	Background	1008
36.5.2	Phylogenetic Methods	1009
36.5.3	Regression-Based Methods	1011
36.6	Genome-Wide Epistasis Analysis	1012
36.7	Gene Content Analysis	1013
	References	1014
	Reference Author Index	1021
	Subject Index	1109

List of Contributors

V. Aggarwal

Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Sumon Ahmed

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Joshua M. Akey

Department of Ecology and Evolutionary Biology and Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA

Tallulah S. Andrews

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Richard Ashcroft

School of Law, Queen Mary University of London, London, UK

Nick Barton

Institute of Science and Technology Austria, Klosterneuburg, Austria

Mark Beaumont

School of Biological Sciences, Bristol University, Bristol, UK

Leonardo Bottolo

Department of Medical Genetics, University of Cambridge, Cambridge, UK, The Alan Turing Institute, London, UK, and MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Alexis Boukouvalas

Prowler.io, Cambridge, UK

Stephen Burgess

MRC Biostatistics Unit, University of Cambridge and Cardiovascular Epidemiology Unit, University of Cambridge, UK

Lon R. Cardon

BioMarin Pharmaceutical, Novato, CA, USA

K. Cars

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Nilanjan Chatterjee

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Jukka Corander

Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, Department of Biostatistics, University of Oslo, Oslo, Norway, and Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Nicholas J. Croucher

Department of Infectious Disease Epidemiology, Imperial College London, London, UK

H.D. Daetwyler

Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia, and School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia

Maria De Iorio

Department of Statistical Science, University College London, London, UK

Frank Dudbridge

Department of Health Sciences, University of Leicester, Leicester, UK

Timothy M.D. Ebbels

Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, UK

Alison Etheridge

University of Oxford, UK

Christopher N. Foley

MRC Biostatistics Unit, University of Cambridge, UK

Antonio Augusto Franco Garcia

University of São Paulo, Piracicaba, Brazil

M.E. Goddard

Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Victoria, Australia, and Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia

Nick Goldman

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, UK

D. Goldstein

Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Marco Grzegorczyk

Bernoulli Institute (BI), Faculty of Science and Engineering, Rijksuniversiteit Groningen, Groningen, Netherlands

Torsten Günther

Human Evolution, Department of Organismal Biology, Uppsala University, Sweden

Kasper D. Hansen

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Simon R. Harris

Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Garrett Hellenthal

University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, UK

Martin Hemberg

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

John P. Huelsenbeck

Department of Integrative Biology, University of California, Berkeley, CA, USA

Dirk Husmeier

School of Mathematics & Statistics, University of Glasgow, Glasgow, UK

Mattias Jakobsson

Human Evolution, Department of Organismal Biology, Uppsala University, Sweden

David T. Jones

University College London, London, UK

Shaun Kandathil

University College London, London, UK

Jerome Kelleher

University of Oxford, UK

Vladimir Yu. Kiselev

Wellcome Sanger Institute, Hinxton,
Cambridgeshire, UK

Laura Kubatko

Ohio State University, Columbus, OH, USA

John A. Lees

Department of Microbiology, School of
Medicine, New York University, New York,
USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Christoph Leuenberger

University of Fribourg, Switzerland

Hongzhe Li

Department of Biostatistics, Epidemiology
and Informatics, Perelman School of
Medicine, University of Pennsylvania,
Philadelphia, USA

Liming Li

Department of Ecology and Evolutionary
Biology, Princeton University, Princeton,
NJ, USA

Shili Lin

Department of Statistics, Ohio State
University, Columbus, OH, USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Qiongshi Lu

Department of Biostatistics and Medical
Informatics, University of Madison-
Wisconsin, Madison, WI, USA

Jonathan Marchini

Regeneron Genetics Center, Tarrytown,
NY, USA

Gil McVean

University of Oxford, UK

Allison Meisner

Department of Biostatistics, Johns Hopkins
Bloomberg School of Public Health,
Baltimore, MD, USA

Ian Mackay

IMplant Consultancy Ltd, Chelmsford, UK

T.H.E. Meuwissen

Norwegian University of Life Sciences,
Ås, Norway

Iain H. Moul

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

Andrew P. Morris

Department of Biostatistics, University of
Liverpool, Liverpool, UK

Michael B. Morrissey

School of Biology, University of St Andrews,
St Andrews, UK

Paul Newcombe

MRC Biostatistics Unit, University of
Cambridge, Cambridge, UK

Rasmus Nielsen

Department of Integrative Biology and
Department of Statistics, University of
California, Berkeley, CA, USA

Magnus Nordborg

Gregor Mendel Institute, Austrian
Academy of Sciences, Vienna BioCenter,
Vienna, Austria

Umberto Peron

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

S. Petrovski

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Hans-Peter Piepho

University of Hohenheim, Stuttgart, Germany

Magnus Rattray

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Sylvia Richardson

MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, and The Alan Turing Institute, London, UK

Eric E. Schadt

Sema4, Stamford, CT, USA, and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Solveig K. Sieberts

Sage Bionetworks, Seattle, WA, USA

Kimberly D. Siegmund

Department of Preventive Medicine, Keck School of Medicine of USC, Los Angeles, USA

David A. Stephens

Department of Mathematics and Statistics, McGill University, Montreal, Canada

Aaron J. Stern

Graduate Group in Computational Biology, University of California, Berkeley, CA

William R. Taylor

The Francis Crick Institute, London, UK

E.A. Thompson

Department of Statistics, University of Washington, Seattle, WA, USA

Jeffrey L. Thorne

Department of Statistics & Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Gerry Tonkin-Hill

Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Susan E. Wallace

Department of Health Sciences, University of Leicester, Leicester, UK

Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Jinliang Wang

Institute of Zoology, Zoological Society of London, UK

Daniel Wegmann

University of Fribourg, Switzerland

B.S. Weir

Department of Biostatistics, University of Washington, Seattle, WA, USA

Jing Yang

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Ziheng Yang

Department of Genetics, Evolution and Environment, University College London, London, UK

Hongyu Zhao

Department of Biostatistics, Yale University, New Haven, CT, USA

Verena Zuber

MRC Biostatistics Unit, University of Cambridge, UK and School of Public Health, Imperial College London

Editors' Preface to the Fourth Edition

After a break of more than 10 years since the third edition, we are pleased to present the fourth edition of the *Handbook of Statistical Genomics*. Genomics has moved on enormously during this period, and so has the *Handbook*: almost everything is new or much revised, with only a small amount of material carried forward from previous editions. Two new editors have joined, Ida Moltke from Copenhagen and John Marioni from Cambridge. With sadness we note the death of founding editor Professor Chris Cannings during 2018. He first saw the need and had the vision for the *Handbook*, one of his many contributions to mathematical and statistical genetics. We also acknowledge the fundamental contribution of the other founding editor, Professor Martin Bishop.

While the content has changed, the mission has not: the *Handbook* is intended as an introduction suitable for advanced graduate students and early-career researchers who have achieved at least a good first-year undergraduate level in both statistics and genetics, and preferably more in at least one of those fields. The chapters are not thorough literature reviews, but focus on explaining the key ideas, methods and algorithms, citing key recent and historic literature for further details and references.

The change of title (from *Genetics* to *Genomics*) is not intended to indicate a substantial change of focus, but reflects both changes in the field, with increased emphasis on transcriptomics and epigenetics for example, and changes in usage. We interpret 'genomics' broadly, to include studies of whole genomes and epigenomes, near-genome processes such as transcription and metabolomics, as well as genomic mechanisms underlying whole-organism outcomes related to selection, adaptation and disease. We also interpret 'statistics' broadly, to include for example relevant aspects of data science and bioinformatics.

The 36 chapters are intended to be largely independent, so that to benefit from the *Handbook* it is not necessary to read every chapter, or to read chapters in order. This structure necessitates some duplication of material, which we have tried to minimize but not always eliminate. Alternative approaches to the same topic by different authors can be beneficial. The extensive subject and author indexes allow easy reference to topics arising in different chapters.

For those with minimal genetics background the glossary has been newly updated. Thanks to Keren Carss for contributing many new terms, in particular those relevant to genomic medicine. Gerry Tonkin-Hill and John Lees also contributed some advanced statistical terminology, but the glossary is predominantly of genetic terms. For those with limited background in statistical modeling and inference we have added an initial chapter that covers these topics, ranging from basic concepts to state-of-art models and methods.

We thank the many commentators on previous editions who were generous in their praise and helpful feedback. No doubt many more improvements will be possible for future editions and we welcome comments e-mailed to any of the editors. We are grateful to all of our authors for taking the time to write and update their chapters with care, and we would like to express our appreciation to all the professional staff working with and for Wiley who helped us to bring this project to fruition.

Glossary

Many of the updates were prepared by Keren Carss (Chapter 27) and Gerry Tonkin-Hill and John Lees (Chapter 36)

NB: Some of the definitions below assume that the organism of interest is diploid.

Adenine (A): purine base that forms a pair with thymine in DNA and uracil in RNA.

Admixture: arises when two previously isolated populations begin interbreeding.

Allele: one of the possible forms of a gene at a given locus. Depending on the technology used to type the gene, it may be that not all DNA sequence variants are recognized as distinct alleles.

Allele frequency: often used to mean the relative frequency (i.e. proportion) of an allele in a sample or population.

Allelic heterogeneity: different alleles in the same gene cause different phenotypes.

Alpha helix: a helical (usually right-handed) arrangement that can be adopted by a polypeptide chain; a common type of protein secondary structure.

Amino acid: the basic building block of proteins. There are 20 naturally occurring amino acids in animals which when linked by peptide bonds form polypeptide chains.

Aneuploid cells: do not have the normal number of chromosomes.

Antisense strand: the DNA strand complementary to the coding strand, determined by the covalent bonding of adenine with thymine and cytosine with guanine.

Approximate Bayesian computation (ABC): methods that approximate the likelihood function by comparing simulations with the observed data.

Ascertainment: the strategy by which study subjects are identified, selected, and recruited for participation in a study.

Autosome: a chromosome other than the sex chromosomes. Humans have 22 pairs of autosomes plus 2 sex chromosomes.

Autozygosity: regions of the genome that are identical due to inheritance from a common ancestor.

Backcross: a linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed back to one of the inbred parental strains.

Bacterial artificial chromosome (BAC): a vector used to clone a large segment of DNA (100–200 kb) in bacteria resulting in many copies.

Base: (abbreviated term for a purine or pyrimidine in the context of nucleic acids), a cyclic chemical compound containing nitrogen that is linked to either a deoxyribose (DNA) or a ribose (RNA).

Base pair (bp): a pair of bases that occur opposite each other (one in each strand) in double stranded DNA/RNA. In DNA adenine base pairs with thymine and cytosine with guanine. RNA is the same except that uracil takes the place of thymine.

Bayesian: a statistical school of thought that, in contrast with the frequentist school, holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes' theorem allows one to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

Beta-sheet: a (hydrogen-bonded) sheet arrangement which can be adopted by a polypeptide chain; a common type of protein secondary structure.

Cell line: an established, immortalized cell culture with a uniform genetic background.

Centimorgan (cM): measure of genetic distance. Two loci separated by 1 cM have an average of one recombination between them every 100 meioses. Because of the variability in recombination rates, genetic distance differs from physical distance, measured in base pairs. Genetic distance differs between male and female meioses; an average over the sexes is usually used.

Centromere: the region where the two sister chromatids join, separating the short (p) arm of the chromosome from the long (q) arm.

Chiasma: the visible structure formed between paired homologous chromosomes (non-sister chromatids) in meiosis.

Chromatid: a single strand of the (duplicated) chromosome, containing a double-stranded DNA molecule.

Chromatin: the material composed of DNA and chromosomal proteins that makes up chromosomes. Comes in two types, euchromatin and heterochromatin.

Chromosome: the self-replicating threadlike structure found in cells. Chromosomes, which at certain stages of meiosis and mitosis consist of two identical sister chromatids, joined at the centromere, carry the genetic information encoded in the DNA sequence.

cis-acting: regulatory elements and expression quantitative trait locus whose DNA sequence directly influences transcription. The physical locations for *cis*-acting elements are in or near the gene or genes they regulate. Contrast *trans*.

Clones: genetically engineered identical cells/sequences.

Co-dominance: both alleles contribute to the phenotype, in contrast with recessive or dominant alleles.

Codon: a nucleotide triplet that encodes an amino acid or a termination signal.

Common disease common variant (CDCV) hypothesis: the hypothesis that many genetic variants underlying complex diseases are common, and hence susceptible to detection using SNP-based association studies.

complementary DNA (cDNA): DNA that is synthesized from a messenger RNA template using the reverse transcriptase enzyme.

Consanguinous mating: mating between closely-related individuals.

Conservation: the similarity of a sequence across species. High conservation indicates selective pressure against mutations, and is suggestive of functional importance.

Constraint: measured by the degree of similarity of a sequence between individuals of the same species, high constraint indicates selective pressure against mutations within that species, which is suggestive of functional importance. See also *intolerance*.

Contig: a group of contiguous, overlapping, cloned DNA sequences.

Core genome: the set of genes shared by all isolates in a collection of related bacterial genomes.

Cytosine (C): pyrimidine base that forms a pair with guanosine in DNA.

Degrees of freedom (df): this term is used in different senses in statistics and in other fields. It can often be interpreted as the number of values that can be defined arbitrarily in the

specification of a system; for example, the number of coefficients in a regression model. Frequently it suffices to regard a df as a parameter used to define certain probability distributions.

De novo variant: a variant in an index case not inherited from either parent.

Deoxyribonucleic acid (DNA): polymer made up of deoxyribonucleotides linked together by phosphodiester bonds.

Deoxyribose: the sugar compound found in DNA.

Diagnostic yield: the fraction of patients for whom a genetic diagnosis is made.

Diploid: has two versions of each autosome, one inherited from the father and one from the mother. Compare with *haploid*.

Dizygotic (DZ) twins: twins derived from different pairs of egg and sperm. DZ twins are genetically equivalent to full sibs.

DNA methylation: the addition of a methyl group to DNA. In mammals this occurs at the C-5 position of cytosine, most often at CpG dinucleotides.

DNA microarray: small slide or 'chip' used to simultaneously measure the quantity of large numbers of different mRNA gene transcripts present in cell or tissue samples.

Dominant allele: results in the same phenotype irrespective of the other allele at the locus.

Dominant negative allele: changes the function of a gene product to be antagonistic to the reference allele.

Effective population size: the size of a theoretical population that best approximates a given natural population under an assumed model. The criterion for assessing the 'best' approximation can vary, but is often some measure of total genetic variation.

Endogamy: mating restricted to a small gene pool.

Enzyme: a protein that controls the rate of a biochemical reaction.

Epigenome: the set of chemical alterations (such as methylation) to a genome's DNA and histones.

Epistasis: the physiological interaction between different genes such that one gene alters the effects of other genes.

Epitope: the part of an antigen that the antibody interacts with.

Eukaryote: organism whose cells include a membrane-bound nucleus. Compare with *prokaryote*.

Exons: parts of a gene that are transcribed into RNA and remain in the mature RNA product after splicing. An exon may code for a specific part of the final protein.

Expression quantitative trait locus (eQTL): a locus influencing the expression of one or more genes.

Fixation: occurs when a locus which was previously polymorphic in a population becomes monomorphic because all but one allele has been lost through genetic drift.

Frequentist: the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more 'extreme' data), given the hypothesis or value. Contrast with *Bayesian*.

Gain-of-function: a new molecular function or increase in the normal function/expression of a gene product.

Gamete: a sex cell, sperm in males, egg in females. Two haploid gametes fuse to form a diploid zygote.

Gene: a segment (not necessarily contiguous) of DNA that codes for a protein or functional RNA.

Gene expression: the process by which coding DNA sequences are converted into functional elements in a cell.

Genealogy: the ancestral relationships among a sample of homologous genes drawn from different individuals, which can be represented by a tree. Also sometimes used in place of pedigree, the ancestral relationships among a set of individuals, which can be represented by a graph.

Genetic drift: the changes in allele frequencies that occur over time due to the randomness inherent in reproductive success.

Genome: all the genetic material of an organism.

Genotype: the (unordered) allele pair(s) carried by an individual at one or more loci. A multilocus genotype is equivalent to the individual's two haplotypes without the phase information.

Germline: sex cells (egg or sperm) that transmit DNA over generations.

Guanine (G): purine base that forms a pair with cytosine in DNA.

Haemoglobin: the red oxygen-carrying pigment of the blood, made up of two pairs of polypeptide chains called globins (2α and 2β subunits).

Haploid: has a single version of each chromosome.

Haploinsufficiency: a dose-sensitive gene such that two working alleles are required to maintain a healthy organism.

Haplotype: the alleles at different loci on a chromosome. An individual's two haplotypes imply the genotype; the converse is not true, but in the presence of strong linkage disequilibrium haplotypes can be accurately inferred from genotype.

Hardy–Weinberg disequilibrium: the non-independence within a population of an individual's two alleles at a locus; can arise due to inbreeding or selection for example. Compare with *linkage disequilibrium*.

Heritability: the proportion of the phenotypic variation in the population that can be attributed to genetic variation.

Heterozygosity: the proportion of individuals in a population that are heterozygotes at a locus. Also sometimes used as shorthand for expected heterozygosity under random mating, which equals the probability that two homologous genes drawn at random from a population are different alleles.

Heterozygote: a single-locus genotype consisting of two different alleles.

Hidden Markov model (HMM): a probabilistic model in which a sequence of unobserved states is a finite Markov chain and a sequence of observed states are random variables that depend on the corresponding hidden state only.

Hierarchical Dirichlet process: a nonparametric generalization of the latent finite mixture model where the number of populations K can be unbounded and is learnt from the data.

Homology: similarities between sequences that arise because of shared evolutionary history (descent from a common ancestral sequence). Homology of different genes within a genome is called paralogous, while that between the genomes of different species is called orthologous.

Homozygote: a single-locus genotype consisting of two versions of the same allele.

Human immunodeficiency virus (HIV): a virus that causes acquired immune deficiency syndrome (AIDS) which destroys the body's ability to fight infection.

Hybrid: the offspring of a cross between parents of different genetic types or different species.

Hybridization: the base pairing of a single stranded DNA or RNA sequence to its complementary sequence.

Hypomorphic: reduces the activity but does not eliminate gene function.

Identity by descent (IBD): two genes are IBD if they have descended without alteration from an ancestral gene.

Inbred lines: derived and maintained by intensive inbreeding, the individuals have almost identical genomes.

Inbreeding: a system of mating in which mates are on average more closely related than under random mating. Inbreeding reduces genetic variation and increases homozygosity and hence the prevalence of recessive traits.

Intercross (or F2 design): a linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed or selfed.

Intron: non-coding DNA sequence separating the exons of a gene. Introns are initially transcribed into messenger RNA but are subsequently spliced out.

Karyotype: the number and structure of an individual's chromosomes.

Kilobase (kb): 1000 base pairs.

Latent finite mixture model: a model which assumes that observed data points belong to an unobserved (latent) set of subpopulations.

Linear mixed model: an extension of linear models that allows for both fixed and random effects, the latter used to model dependence within data sets.

Linkage: two genes are said to be linked if they are located close together on the same chromosome. The alleles at linked genes tend to be co-inherited more often than those at unlinked genes because of the reduced opportunity for an intervening recombination.

Linkage disequilibrium (LD): the non-independence within a population of a gamete's alleles at different loci; can arise due to linkage, population stratification, or selection. The term is misleading and 'gametic phase disequilibrium' is sometimes preferred. Various measures of linkage disequilibrium are available, including D' and r^2 .

Locus (pl. loci): the position of a gene on a chromosome.

Locus heterogeneity: the degree to which variants in different genes can cause similar phenotype(s).

LOD score: a likelihood ratio statistic used to infer whether two loci are genetically linked (have recombination fraction less than 0.5).

Loss-of-function variant: the mutant allele abolishes the function of a gene product.

Metabolome: the complement of all metabolites in a sample.

Marker gene: a polymorphic gene of known location which can be readily typed; used, for example, in genetic mapping.

Markov chain (first-order): a sequence of random variables such that, given the current state of the sequence, the probability distribution of the next state is independent of all previous states.

Markov chain Monte Carlo: a stochastic algorithm used to sample from a target probability distribution. It works by constructing a Markov chain that has the target distribution as its equilibrium distribution, so that over many steps of the chain the fraction of time spent in any region of the state space converges to the probability of that region under the target distribution.

Megabase (Mb): 1000 kilobases = 1,000,000 base pairs.

Meiosis: the process by which (haploid) gametes are formed from (diploid) somatic cells.

messenger RNA (mRNA): the RNA sequence that acts as the template for protein synthesis.

Microarray: see *DNA microarray*.

Microbiome: the complement of all microbes in a sample.

Microsatellite DNA: small stretches of DNA (usually 1–6 bp) tandemly repeated.

Microsatellite loci are often highly polymorphic, and alleles can be distinguished by length, making them useful as marker loci.

Mitochondrial DNA (mtDNA): the genetic material of the mitochondria which consists of a circular DNA duplex inherited maternally.

Mitosis: the process by which a somatic cell is replaced by two daughter somatic cells.

Modifier variant: a variant that modifies the phenotype caused by another variant.

Monomorphic: a locus at which only one allele arises in the sample or population.

Monozygotic twins: genetically identical individuals derived from a single fertilized egg.

Morgan: unit of genetic distance = 100 centimorgans.

Morpholino: a synthetic oligonucleotide used experimentally to modify gene expression.

Mosaicism: two or more populations of cells arising from the same fertilized egg. Can result from a somatic mutation.

Mutation: a process that changes an allele.

Negative selection: removal of deleterious mutations by natural selection. Also known as ‘purifying selection’.

Neutral: not subject to selection.

Neutral evolution: evolution of alleles with nearly zero selective coefficient. When $|Ns| \ll 1$, where N is the population size and s is the selective coefficient, the fate of the allele is mainly determined by random genetic drift rather than natural selection.

Non-coding RNA (ncRNA): RNA transcripts not translated into protein product. There are many classes of ncRNA, some of which can affect the rate of transcription or transcript degradation.

Nonsense mediated decay: the cellular pathway by which transcripts that contain premature stop codons are degraded to prevent translation into aberrant proteins.

Nonsynonymous substitution: nucleotide substitution in a protein-coding gene that alters the encoded amino acid.

Nucleoside: a base attached to a sugar, either ribose or deoxyribose.

Nucleotide: the structural units with which DNA and RNA are formed. Nucleotides consist of a base attached to a five-carbon sugar and mono-, di-, or triphosphate.

Nucleotide substitution: the replacement of one nucleotide by another during evolution. Substitution is generally considered to be the product of both mutation and selection.

Oligonucleotide: a short sequence of single-stranded DNA or RNA, often used as a probe for detecting the complementary DNA or RNA.

Open reading frame (ORF): a long sequence of DNA with an initiation codon at the 5' end and no termination codon except for one at the 3' end.

Pangenome: the totality of genes identified in a collection of related bacterial genomes.

Pathogenicity: the degree to which a variant increases an individual's predisposition to a genetic disease.

Pedigree: a diagram showing the relationship of each family member and the heredity of a particular trait through several generations of a family.

Penetrance: the probability that a particular phenotype is observed in an individual with a given genotype. Penetrance can vary with environment and the alleles at other loci.

Peptide bond: linkages between amino acids occur through a covalent peptide bond joining the C terminal of one amino acid to the N terminal of the next (with loss of a water molecule).

Phase (of linked markers): the relationship (either coupling or repulsion) between alleles at two linked loci. The two alleles at the linked loci are said to be ‘in coupling’ if they are present on the same physical chromosome or ‘in repulsion’ if they are present on different parental homologs.

Phenotype: the observed characteristic under study, may be quantitative (i.e. continuous) such as height, or binary (e.g. disease/no disease), or ordered categorical (e.g. mild/moderate/severe).

Phenotypic heterogeneity: the degree to which different phenotypes are associated with the same variant or gene. Also known as ‘variable expressivity’.

Phenotypic spectrum: the range of different phenotypes associated with the same variant or gene.

Pleiotropy: the effect of a gene on several different traits.

Polygenic traits: affected by multiple genes.

Polymerase chain reaction (PCR): a laboratory process by which a specific, short, DNA sequence is amplified many times.

Polymorphic: a locus that is not monomorphic. Usually a stricter criterion is imposed: a locus is polymorphic if no allele has frequency greater than 0.99.

Polynucleotide: a polymer of either DNA or RNA nucleotides.

Polypeptide: a long chain of amino acids joined together by peptide bonds.

Polypeptide chain: a series of amino acids linked by peptide bonds. Short chains are sometimes referred to as ‘oligopeptides’ or simply ‘peptides’.

Polytene: refers to the giant chromosomes that are generated by the successive replication of chromosome pairs without the nuclear division, thus several chromosome sets are joined together.

Population stratification (or population structure): refers to a situation in which the population of interest can be divided into strata such that an individual tends to be more closely related to others within the same stratum than to other individuals.

Positive selection: fixation, by natural selection, of an advantageous allele with a positive selective coefficient. Also known as ‘Darwinian selection’.

Precision medicine: tailoring healthcare management, including targeted treatment, to individual patients considering variability in genes, environment and lifestyle.

Primary cells: cells taken from a tissue sample, which have undergone few *in vitro* mitoses and have a limited lifespan.

Proband: an individual through whom a family is ascertained, typically by their phenotype.

Product-partition model: models that partition a data set into distinct clusters. They assume that, given a partition of objects in a data set, objects belonging to the same set are exchangeable and objects in different sets are independent.

Prokaryote: a unicellular organism with no nucleus.

Promoter: located upstream of the gene, the promoter allows the binding of RNA polymerase which initiates transcription of the gene.

Protein: a large, complex, molecule made up of one or more chains of amino acids.

Protein-truncating variant: a variant predicted to truncate a gene product. They usually, but not always, are loss-of-function variants.

Proteome: the complement of all protein molecules in a sample.

Pseudogene: a DNA sequence that is either an imperfect, non-functioning, copy of a gene, or a former gene which no longer functions due to mutations.

Purine: adenine or guanine; particular kinds of nitrogen-containing heterocyclic rings.

Pyrimidine: cytosine, thymine, or uracil; particular kinds of nitrogen-containing heterocyclic rings.

Quantitative trait locus (QTL): a locus influencing a continuously varying phenotype.

Radiation hybrid: a cell line, usually rodent, that has incorporated fragments of foreign chromosomes that have been broken by irradiation. They are used in physical mapping.

Recessive allele: has no effect on phenotype except when present in homozygote form.

Recombination: the formation of new haplotypes by physical exchange between two homologous chromosomes during meiosis.

Restriction enzyme: recognizes specific nucleotide sequences in double-stranded DNA and cuts at a specified position with respect to the sequence.

Restriction site: a 4–8 bp DNA sequence (usually palindromic) that is recognized by a restriction enzyme.

Retrovirus: an RNA virus whose replication depends on a reverse transcriptase function, allowing the formation of a cDNA copy that can be stably inserted into the host chromosome.

Ribonucleic acid (RNA): polymer made up of ribonucleotides that are linked together by phosphodiester bonds.

Ribosome: a cytoplasmic organelle, consisting of RNA and protein, that is involved in the translation of messenger RNA into proteins.

Ribosomal RNA (rRNA): the RNA molecules contained in ribosomes.

Selection: a process such that expected allele frequencies do not remain constant, in contrast with genetic drift. Alleles that convey an advantage to the organism in its current environment tend to become more frequent in the population (positive, or adaptive, selection), while deleterious alleles become less frequent. Under stabilizing (or balancing) selection, allele frequencies tend towards a stable, intermediate value.

Sense strand: the DNA strand in the direction of coding.

Sex-linked: a trait influenced by a gene located on a sex (X or Y) chromosome.

Single nucleotide polymorphism (SNP): a polymorphism consisting of a single nucleotide.

Sister chromatids: two chromatids that are copies of the same chromosome. Non-sister chromatids are different but homologous.

Somatic cell: a non-sex cell.

Somatic variant: a variant that arose from mutation in a somatic cell and so is not transmitted to descendants.

Structural variant: genomic rearrangements including deletions (in excess of 50 bp), insertions, tandem duplications, dispersed duplications, inversions, translocations, and complex structural variants.

Synonymous substitution: nucleotide substitution in a protein-coding gene that does not alter the encoded amino acid.

TATA box: a conserved sequence (TATAAAA) found about 25–30 bp upstream from the start of transcription site in most but not all genes.

Thymine (T): pyrimidine base that forms a pair with adenine in DNA.

trans-acting: eQTL whose DNA sequence influences gene expression through its gene product. These regulatory elements are often coded for at loci far from or unlinked to the genes they regulate. Contrast *cis*.

Transcription: the synthesis of a single-stranded RNA version of a DNA sequence.

Transcriptome: the complement of all RNA molecules that can be transcribed from a genome.

Transition: a mutation that changes either one purine base to the other, or one pyrimidine base to the other.

Translation: the process whereby messenger RNA is ‘read’ by transfer RNA and its corresponding polypeptide chain synthesized.

Transposon: a genetic element that can move over generations from one genomic location to another.

Transversion: a mutation that changes a purine base to a pyrimidine, or vice versa.

Triplosensitivity: dose-sensitivity of a genomic region such that excess copies are pathogenic.

Ultra-rare variant: a variant found in an index sample that is unobserved across all available reference cohorts; this may be due to an under-represented genetic ancestry group.

Uracil (U): pyrimidine base in RNA that takes the place of thymine in DNA, also forming a pair with adenine.

Wild-type: the common, or standard, allele/genotype/phenotype in a population.

Yeast artificial chromosome (YAC): a cloning vector able to carry large (e.g. 1 megabase) inserts of DNA and replicate in yeast cells.

Zygote: an egg cell that has been fertilized by a sperm cell.

Abbreviations and Acronyms

ABC	Approximate Bayesian Computation
aDNA	Ancient DNA
AIC	Akaike's Information Criterion
ARG	Ancestral Recombination Graph
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Predictor
BMI	Body Mass Index
bp	Base Pairs
cDNA	Complementary DNA
CHD	Coronary Heart Disease
ChIP	Chromatin Immunoprecipitation
CI	Confidence Interval
DAG	Directed Acyclic Graph
DCA	Direct Coupling Analysis
df	Degrees of Freedom
DNA	Deoxyribonucleic Acid
EHH	Extended Haplotype Homozygosity
EM	Expectation Maximization
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
GBLUP	Genomic Best Linear Unbiased Prediction
GC	Gas Chromatography
GC	Guanine and Cytosine
GLM	Generalized Linear Model
GTR	General Time Reversible
GWAS	Genome-Wide Association Study
HMM	Hidden Markov Model
HPD	Highest Probability Density
HWE	Hardy–Weinberg Equilibrium
IBD	Identical by Descent
IBS	Identical by State
kb	kilobase
KDEs	Kernel Density Estimators
kya	Thousand Years Ago
LD	Linkage Disequilibrium
MAF	Minor Allele Fraction
MAP	Maximum <i>A Posteriori</i>

MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis–Hastings
MHC	Major Histocompatibility Complex
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate/Estimator
MR	Mendelian Randomization
MRCA	Most Recent Common Ancestor
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial DNA
NGS	Next-Generation Sequencing
NMR	Nuclear Magnetic Resonance
ODE	Ordinary Differential Equation
OLS	Ordinary Least Squares
OR	Odds Ratio
OU	Ornstein–Uhlenbeck
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
PPI	Protein–Protein Interaction
PRS	Polygenic Risk Score
QTLs	Quantitative Trait Loci
RCT	Randomized Controlled Trial
SFS	Site Frequency Spectrum
SMC	Sequentially Markov Coalescent
SNP	Single Nucleotide Polymorphism
SVD	Singular Value Decomposition
TMRCA	Time to the Most Recent Common Ancestor
WTCCC	Wellcome Trust Case Control Consortium

Handbook of Statistical Genomics

Handbook of Statistical Genomics

Volume 2

Edited by

David J. Balding

University of Melbourne, Australia

Ida Moltke

University of Copenhagen, Denmark

John Marioni

University of Cambridge, United Kingdom

Fourth Edition

Founding Editors

Chris Cannings

Martin Bishop

WILEY

This fourth edition first published 2019

© 2019 John Wiley & Sons Ltd

Edition History

John Wiley & Sons, Ltd (1e, 2001), John Wiley & Sons, Ltd (2e, 2003), John Wiley & Sons, Ltd (3e, 2007)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law.

Advice on how to obtain permission to reuse material from this title is available at
<http://www.wiley.com/go/permissions>.

The right of David J. Balding, Ida Moltke and John Marioni to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Balding, D. J., editor. | Moltke, Ida, editor. | Marioni, John, editor.

Title: Handbook of statistical genomics / edited by David J. Balding (University of Melbourne, Australia),
Ida Moltke (University of Copenhagen, Denmark), and John Marioni (University of Cambridge, United Kingdom).

Other titles: Handbook of statistical genetics. | Handbook of statistical genetics.

Description: Fourth edition. | Hoboken, NJ : Wiley, 2019– | Previous title: Handbook of statistical genetics. | Includes bibliographical references and indexes. |

Identifiers: LCCN 2018060346 (print) | LCCN 2019003813 (ebook) | ISBN 9781119429227 (Adobe PDF) | ISBN 9781119429258 (ePub) | ISBN 9781119429142 (hardcover)

Subjects: LCSH: Genetics—Statistical methods—Handbooks, manuals, etc.

Classification: LCC QH438.4.S73 (ebook) | LCC QH438.4.S73 H36 2019 (print) | DDC 572.8/60727–dc23
LC record available at <https://lccn.loc.gov/2018060346>

Cover Design: Wiley

Cover Image: © Zita / Shutterstock

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

Contents

Volume 1

- List of Contributors xxiii**
Editors' Preface to the Fourth Edition xxvii
Glossary xxix
Abbreviations and Acronyms xxxix

1	Statistical Modeling and Inference in Genetics 1
	<i>Daniel Wegmann and Christoph Leuenberger</i>
1.1	Statistical Models and Inference 1
1.1.1	Statistical Models 2
1.1.2	Inference Methods and Algorithms 4
1.2	Maximum Likelihood Inference 4
1.2.1	Properties of Maximum Likelihood Estimators 6
1.2.2	Quantifying Confidence: the Fisher Information Matrix 8
1.2.3	Newton's Method 9
1.2.4	Latent Variable Problems: the EM Algorithm 11
1.2.5	Approximate Techniques 16
1.3	Bayesian Inference 20
1.3.1	Choice of Prior Distributions 21
1.3.2	Bayesian Point Estimates and Confidence Intervals 22
1.3.3	Markov Chain Monte Carlo 23
1.3.4	Empirical Bayes for Latent Variable Problems 30
1.3.5	Approximate Bayesian Computation 31
1.4	Model Selection 37
1.4.1	Likelihood Ratio Statistic 37
1.4.2	Bayesian Model Choice 38
1.5	Hidden Markov Models 40
1.5.1	Bayesian Inference of Hidden States Using Forward-Backward Algorithm 42
1.5.2	Finding the Most Likely Hidden Path (Viterbi Algorithm) 43
1.5.3	MLE Inference of Hierarchical Parameters (Baum–Welch Algorithm) 44
	Acknowledgements 46
	References 47

2	Linkage Disequilibrium, Recombination and Haplotype Structure	51
	<i>Gil McVean and Jerome Kelleher</i>	
2.1	What Is Linkage Disequilibrium?	51
2.2	Measuring Linkage Disequilibrium	53
	2.2.1 Single-Number Summaries of LD	54
	2.2.2 The Spatial Distribution of LD	56
	2.2.3 Various Extensions of Two-Locus LD Measures	60
2.3	Modelling Linkage Disequilibrium and Genealogical History	60
	2.3.1 A Historical Perspective	60
	2.3.2 Coalescent Modelling	62
	2.3.3 Relating Genealogical History to LD	67
2.4	Data Analysis	69
	2.4.1 Estimating Recombination Rates	69
	2.4.2 Methods Exploiting Haplotype Structure	72
2.5	Prospects	75
	Acknowledgements	75
	References	76
3	Haplotype Estimation and Genotype Imputation	87
	<i>Jonathan Marchini</i>	
3.1	Haplotype Estimation	87
	3.1.1 A Simple Haplotype Frequency Model	88
	3.1.2 Hidden Markov Models for Phasing	89
	3.1.3 Phasing in Related Samples	93
	3.1.4 Phasing Using Sequencing Data	94
	3.1.5 Phasing from a Reference Panel	95
	3.1.6 Measuring Phasing Performance	96
3.2	Genotype Imputation	97
	3.2.1 Uses of Imputation in GWASs	98
	3.2.2 Haploid Imputation	99
	3.2.3 Imputation Methods	100
	3.2.4 Testing Imputed Genotypes for Association	103
	3.2.5 Summary Statistic Imputation	104
	3.2.6 Factors Affecting Accuracy	104
	3.2.7 Quality Control for Imputed Data	107
3.3	Future Directions	109
	References	109
4	Mathematical Models in Population Genetics	115
	<i>Nick Barton and Alison Etheridge</i>	
4.1	Introduction	115
4.2	Single-Locus Models	116
	4.2.1 Random Drift and the Kingman Coalescent	117
	4.2.2 Diffusion Approximations	120
	4.2.3 Spatially Structured Populations	126
4.3	Multiple Loci	130
	4.3.1 Linkage Equilibrium	131
	4.3.2 Beyond Linkage Equilibrium	134
4.4	Outlook	140
	References	140

5	Coalescent Theory	145
	<i>Magnus Nordborg</i>	
5.1	Introduction	145
5.2	The Coalescent	146
	5.2.1 The Fundamental Insights	146
	5.2.2 The Coalescent Approximation	148
5.3	Generalizing the Coalescent	151
	5.3.1 Robustness and Scaling	151
	5.3.2 Variable Population Size	152
	5.3.3 Population Structure on Different Time-Scales	153
5.4	Geographical Structure	155
	5.4.1 The Structured Coalescent	155
	5.4.2 The Strong-Migration Limit	156
5.5	Diploidy and Segregation	157
	5.5.1 Hermaphrodites	157
	5.5.2 Males and Females	159
5.6	Recombination	159
	5.6.1 The Ancestral Recombination Graph	160
	5.6.2 Properties and Effects of Recombination	163
5.7	Selection	164
	5.7.1 Balancing Selection	165
	5.7.2 Selective Sweeps	166
	5.7.3 Background Selection	168
5.8	Neutral Mutations	168
5.9	Concluding Remarks	169
	5.9.1 The Coalescent and ‘Classical’ Population Genetics	169
	5.9.2 The Coalescent and Phylogenetics	169
	5.9.3 Prospects	171
	Acknowledgements	171
	References	171
6	Phylogeny Estimation Using Likelihood-Based Methods	177
	<i>John P. Huelsenbeck</i>	
6.1	Introduction	177
	6.1.1 Statistical Phylogenetics	178
	6.1.2 Chapter Outline	178
6.2	Maximum Likelihood and Bayesian Estimation	179
	6.2.1 Maximum Likelihood	179
	6.2.2 Bayesian Inference	180
6.3	Choosing among Models Using Likelihood Ratio Tests and Bayes Factors	184
6.4	Calculating the Likelihood for a Phylogenetic Model	186
	6.4.1 Character Matrices and Alignments	186
	6.4.2 The Phylogenetic Model	186
	6.4.3 Calculating the Probability of a Character History	187
	6.4.4 Continuous-Time Markov Model	188
	6.4.5 Marginalizing over Character Histories	189
6.5	The Mechanics of Maximum Likelihood and Bayesian Inference	192
	6.5.1 Maximum Likelihood	192
	6.5.2 Bayesian Inference and Markov Chain Monte Carlo	193

6.6	Applications of Likelihood-Based Methods in Molecular Evolution	199
6.6.1	A Taxonomy of Commonly Used Substitution Models	199
6.6.2	Expanding the Model around Groups of Sites	202
6.6.3	Rate Variation across Sites	204
6.6.4	Divergence Time Estimation	206
6.7	Conclusions	212
	References	213
7	The Multispecies Coalescent	219
	<i>Laura Kubatko</i>	
7.1	Introduction	219
7.2	Probability Distributions under the Multispecies Coalescent	221
7.2.1	Gene Tree Probabilities	221
7.2.2	Site Pattern Probabilities	227
7.2.3	Species Tree Likelihoods under the Multispecies Coalescent	229
7.2.4	Model Assumptions and Violations	230
7.3	Species Tree Inference under the Multispecies Coalescent	231
7.3.1	Summary Statistics Methods	231
7.3.2	Bayesian Full-Data Methods	234
7.3.3	Site Pattern-Based Methods	235
7.3.4	Multilocus versus SNP Data	236
7.3.5	Empirical Examples	237
7.4	Coalescent-Based Estimation of Parameters at the Population and Species Levels	239
7.4.1	Speciation Times and Population Sizes	239
7.4.2	Hybridization and Gene Flow	240
7.4.3	Species Delimitation	241
7.4.4	Future Prospects	242
	Acknowledgements	242
	References	242
8	Population Structure, Demography and Recent Admixture	247
	<i>G. Hellenthal</i>	
8.1	Introduction	247
8.1.1	'Admixture' versus 'Background' Linkage Disequilibrium	248
8.2	Spatial Summaries of Genetic Variation Using Principal Components Analysis	249
8.3	Clustering Algorithms	251
8.3.1	Defining 'Populations'	251
8.3.2	Clustering Based on Allele Frequency Patterns	252
8.3.3	Incorporating Admixture	253
8.3.4	Incorporating Admixture Linkage Disequilibrium	254
8.3.5	Incorporating Background Linkage Disequilibrium: Using Haplotypes to Improve Inference	255
8.3.6	Interpreting Genetic Clusters	258
8.4	Inferring Population Size Changes and Split Times	259
8.4.1	Allele Frequency Spectrum Approaches	260
8.4.2	Approaches Using Whole-Genome Sequencing	261
8.5	Identifying/Dating Admixture Events	262
8.5.1	Inferring DNA Segments Inherited from Different Sources	263
8.5.2	Measuring Decay of Linkage Disequilibrium	265

8.6	Conclusion 267	
	Acknowledgements 268	
	References 268	
9	Statistical Methods to Detect Archaic Admixture and Identify Introgressed Sequences 275	
	<i>Liming Li and Joshua M. Akey</i>	
9.1	Introduction 275	
9.2	Methods to Test Hypotheses of Archaic Admixture and Infer Admixture Proportions 277	
	9.2.1 Genetic Drift and Allele Frequency Divergence in Genetically Structured Populations 277	
	9.2.2 Three-Population Test 277	
	9.2.3 D-Statistic 279	
	9.2.4 F_4 -Statistic 282	
9.3	Methods to Identify Introgressed Sequences 283	
	9.3.1 S^* -Statistic 284	
	9.3.2 Hidden Markov and Conditional Random Field Models 287	
	9.3.3 Relative Advantages and Disadvantages of Approaches to Detect Introgressed Sequences 289	
9.4	Summary and Perspective 289	
	References 290	
10	Population Genomic Analyses of DNA from Ancient Remains 295	
	<i>Torsten Günther and Mattias Jakobsson</i>	
10.1	Introduction 295	
10.2	Challenges of Working with and Analyzing Ancient DNA Data 296	
	10.2.1 Sequence Degradation 296	
	10.2.2 Contamination 297	
	10.2.3 Handling Sequence Data from Ancient Material 300	
	10.2.4 Different Sequencing Approaches and the Limitations in their Resulting Data 301	
	10.2.5 Effects of Limited Amounts of Data on Downstream Analysis 301	
10.3	Opportunities of Ancient DNA 302	
	10.3.1 Population Differentiation in Time and Space 303	
	10.3.2 Continuity 306	
	10.3.3 Migration and Admixture over Time 307	
	10.3.4 Demographic Inference Based on High-Coverage Ancient Genomes 308	
	10.3.5 Allele Frequency Trajectories 308	
10.4	Some Examples of How Genetic Studies of Ancient Remains Have Contributed to a New Understanding of the Human Past 310	
	10.4.1 Archaic Genomes and the Admixture with Modern Humans 310	
	10.4.2 Neolithic Revolution in Europe and the Bronze Age Migrations 311	
10.5	Summary and Perspective 313	
	Acknowledgements 313	
	References 314	
11	Sequence Covariation Analysis in Biological Polymers 325	
	<i>William R. Taylor, Shaun Kandathil, and David T. Jones</i>	
11.1	Introduction 325	

11.2	Methods 326
11.2.1	DCA Method 326
11.2.2	PSICOV 327
11.2.3	plmDCA, GREMLIN and CCMpred 327
11.3	Applications 328
11.3.1	Globular Protein Fold Prediction 328
11.3.2	Transmembrane Protein Prediction 328
11.3.3	RNA Structure Prediction 328
11.3.4	Protein Disordered Regions 329
11.3.5	Protein–Protein Interactions 329
11.3.6	Allostery and Dynamics 330
11.3.7	CASP 330
11.4	New Developments 332
11.4.1	Sequence Alignment 332
11.4.2	Comparison to Known Structures 333
11.4.3	Segment Parsing 334
11.4.4	Machine Learning 335
11.4.5	Deep Learning Methods 336
11.4.6	Sequence Pairing 338
11.4.7	Phylogeny Constraints 339
11.5	Outlook 340
	Acknowledgements 341
	References 342
12	Probabilistic Models for the Study of Protein Evolution 347
	<i>Umberto Pernon, Iain H. Moal, Jeffrey L. Thorne, and Nick Goldman</i>
12.1	Introduction 347
12.2	Empirically Derived Models of Amino Acid Replacement 348
12.2.1	The Dayhoff and Eck Model 348
12.2.2	Descendants of the Dayhoff Model 350
12.3	Heterogeneity of Replacement Rates among Sites 351
12.4	Protein Structural Environments 351
12.5	Variation of Preferred Residues among Sites 353
12.6	Models with a Physicochemical Basis 355
12.7	Codon-Based Models 355
12.8	Dependence among Positions 357
12.9	Stochastic Models of Structural Evolution 359
12.10	Conclusion 360
	Acknowledgements 361
	References 361
13	Adaptive Molecular Evolution 369
	<i>Ziheng Yang</i>
13.1	Introduction 369
13.2	Markov Model of Codon Substitution 371
13.3	Estimation of Synonymous and Non-synonymous Substitution Rates between Two Sequences and Test of Selection on the Protein 372
13.3.1	Heuristic Estimation Methods 372
13.3.2	Maximum Likelihood Estimation 374

13.3.3	Bayesian Estimation	377
13.3.4	A Numerical Example	377
13.4	Likelihood Calculation on a Phylogeny	379
13.5	Detecting Adaptive Evolution along Lineages	380
13.5.1	Likelihood Calculation under Models of Variable ω Ratios among Lineages	380
13.5.2	Adaptive Evolution in the Primate Lysozyme	381
13.5.3	Comparison with Methods Based on Reconstructed Ancestral Sequences	382
13.6	Inferring Amino Acid Sites under Positive Selection	384
13.6.1	Likelihood Ratio Test under Models of Variable ω Ratios among Sites	384
13.6.2	Methods that Test One Site at a Time	386
13.6.3	Positive Selection in the HIV-1 <i>vif</i> Genes	386
13.7	Testing Positive Selection Affecting Particular Sites and Lineages	388
13.7.1	Branch-Site Test of Positive Selection	388
13.7.2	Clade Models and Other Variants	389
13.8	Limitations of Current Methods	390
13.9	Computer Software	391
	References	391
14	Detecting Natural Selection	397
	<i>Aaron J. Stern and Rasmus Nielsen</i>	
14.1	Introduction	397
14.2	Types of Selection	398
14.2.1	Directional Selection	398
14.2.2	Balancing Selection	399
14.2.3	Polygenic Selection	399
14.3	The Signature of Selection in the Genome	399
14.3.1	The Signature of Positive Directional Selection	400
14.3.2	Balancing Selection	403
14.3.3	Polygenic Selection	403
14.3.4	Confounders	404
14.4	Methods for Detecting Selection	405
14.4.1	Substitution-Based Methods	405
14.4.2	Methods Comparing Substitutions and Diversity	406
14.4.3	Methods Using the Frequency Spectrum	407
14.4.4	Methods Using Genetic Differentiation	408
14.4.5	Methods Using Haplotype Structure	410
14.4.6	Why Full-Likelihood Methods Are Intractable for Population Samples	412
14.4.7	Composite Likelihood Methods	412
14.4.8	Approximate Bayesian Computation	413
14.4.9	Machine Learning Methods	413
14.5	Discussion	414
	References	415
15	Evolutionary Quantitative Genetics	421
	<i>Bruce Walsh and Michael B. Morrissey</i>	
15.1	Introduction	421

15.2	Resemblances, Variances, and Additive Genetic Values	422
15.2.1	Fisher's Genetic Decomposition	422
15.2.2	Additive Genetic Variances and Covariances	423
15.3	Parent–Offspring Regressions and the Response to Selection	423
15.3.1	Single-Trait Parent–Offspring Regressions	424
15.3.2	Selection Differentials and the Breeder's Equation	424
15.3.3	Multiple-Trait Parent–Offspring Regressions	425
15.3.4	The Genetic and Phenotypic Covariance Matrices	425
15.3.5	The Multivariate Breeder's Equation	425
15.4	The Infinitesimal Model	426
15.4.1	Linearity of Parent–Offspring Regressions under the Infinitesimal Model	426
15.4.2	Allele Frequency Changes under the Infinitesimal Model	426
15.4.3	Changes in Variances	427
15.4.4	The Equilibrium Additive Genetic Variance	429
15.5	Inference of σ_A^2 and \mathbf{G}	430
15.6	Fitness	432
15.6.1	Individual Fitness	432
15.6.2	Episodes of Selection	433
15.7	The Robertson–Price Identity, and Theorems of Selection	434
15.7.1	Description of the Theorems	435
15.7.2	Empirical Operationalization of the Theorems	436
15.8	The Opportunity for Selection	437
15.9	Selection Coefficients	438
15.9.1	Measures of Selection on the Mean	439
15.9.2	Measures of Selection on the Variance	439
15.10	Fitness Functions and the Characterization of Selection	441
15.10.1	Individual and Mean Fitness Functions	441
15.10.2	Gradients and the Local Geometry of Fitness Surfaces	442
15.11	Multivariate Selection	444
15.11.1	Short-Term Changes in Means: The Multivariate Breeder's Equation	444
15.11.2	The Effects of Genetic Correlations: Direct and Correlated Responses	444
15.11.3	Selection Gradients and Understanding which Traits Affect Fitness	446
15.12	Inference of Selection Gradients	447
15.12.1	Ordinary Least Squares Analysis	448
15.12.2	Flexible Inference of Fitness Functions with Associated Selection Gradient Estimates	449
15.12.3	Normality and Selection Gradients	450
15.13	Summary	451
	References	452
16	Conservation Genetics	457
	<i>Mark Beaumont and Jinliang Wang</i>	
16.1	Introduction	457
16.2	Estimating Effective Population Size	458
16.2.1	Methods Based on Heterozygosity Excess	459
16.2.2	Methods Based on Linkage Disequilibrium	460
16.2.3	Methods Based on Relatedness	463
16.2.4	Methods Based on Temporal Changes in Allele Frequency	466

16.3	Estimating Census Size by the Genotype Capture–Recapture Approach	470
16.3.1	Methods Based on Multilocus Genotype Mismatches	472
16.3.2	Methods Based on Pairwise Relatedness	472
16.3.3	Methods Based on Pairwise Relationships	473
16.3.4	Methods Based on Pedigree Reconstruction	474
16.4	Inferring Genetic Structure	475
16.4.1	Measuring Genetic Differentiation	475
16.4.2	Population Assignment	477
16.4.3	Population Clustering and Inference of Ancestry Proportions	479
16.4.4	Inferring Levels of Recent Gene Flow	483
16.4.5	Landscape Genetics	486
16.5	Deintrogession Strategies	487
16.6	Genetic Species Delimitation	489
16.7	Conclusions and Outlook	491
	Acknowledgements	492
	References	492
17	Statistical Methods for Plant Breeding	501
	<i>Ian Mackay, Hans-Peter Piepho, and Antonio Augusto Franco Garcia</i>	
17.1	Introduction	501
17.2	Heritability and the Breeder's Equation in Plant Breeding	502
17.3	The Breeding System of Plants	504
17.4	Polyploidy in Plants and Its Genetic Consequences	505
17.5	Genomic Rearrangements in Plants	509
17.6	Genetic Architecture of Traits in Plants	510
17.7	Response to the Environment and Plasticity	511
17.8	Genomic Selection	514
17.8.1	Genotype–Environment Interaction	514
17.8.2	Quantitative Trait Loci and Major Genes	516
17.8.3	Genomic Selection and Cross Prediction	517
17.8.4	Genomic Selection and Phenotyping Cost	517
17.8.5	Mate Selection	517
17.8.6	Sequential Selection	518
17.8.7	Genomic Prediction of Hybrid Performance and Heterosis	518
17.8.8	Marker Imputation	519
17.9	Experimental Design and Analysis	519
17.10	Conclusions	521
	References	521
18	Forensic Genetics	531
	<i>B.S. Weir</i>	
18.1	Introduction	531
18.2	Principles of Interpretation	532
18.3	Profile Probabilities	534
18.3.1	Genetic Models for Allele Frequencies	535
18.3.2	Y-STR Profiles	539
18.4	Mixtures	542
18.4.1	Combined Probabilities of Inclusion and Exclusion	542
18.4.2	Likelihood Ratios	542
18.5	Behavior of Likelihood Ratio	546

18.6	Single Nucleotide Polymorphism, Sequence and Omic Data	547
	References	548

Volume 2

List of Contributors	<i>xxiii</i>
Editors' Preface to the Fourth Edition	<i>xxvii</i>
Glossary	<i>xxix</i>
Abbreviations and Acronyms	<i>xxxix</i>

19	Ethical Issues in Statistical Genetics	551
	<i>Susan E. Wallace and Richard Ashcroft</i>	
19.1	Introduction	551
	19.1.1 What Is Ethics?	552
	19.1.2 Models for Analysing the Ethics of Population Genetic Research	553
19.2	Ethics and Governance in Population Genetics Research: Two Case Studies	554
	19.2.1 'Healthy Volunteer' Longitudinal Cohort Studies: UK Biobank	555
	19.2.2 Precision Medicine Approaches: 100,000 Genomes Project	556
	19.2.3 The Scientific and Clinical Value of the Research	556
	19.2.4 Recruitment of Participants	558
	19.2.5 Consent	559
	19.2.6 Returning Individual Genetic Research Results	563
	19.2.7 Confidentiality and Security	564
19.3	Stewardship and Wider Social Issues	565
	19.3.1 Benefit Sharing	566
	19.3.2 Community Involvement and Public Engagement	567
	19.3.3 Race, Ethnicity and Genetics	567
19.4	Conclusion	568
	Acknowledgements	568
	References	568
20	Descent-Based Gene Mapping in Pedigrees and Populations	573
	<i>E.A. Thompson</i>	
20.1	Introduction to Genetic Mapping and Genome Descent	573
	20.1.1 Genetic Mapping: The Goal and the Data	573
	20.1.2 The Process of Meiosis and the Descent of DNA	574
	20.1.3 Genetic Linkage Mapping: Association or Descent?	576
20.2	Inference of Local IBD Sharing from Genetic Marker Data	577
	20.2.1 Identity by Descent at a Locus	577
	20.2.2 Probabilities of Marker Data Given IBD Pattern	579
	20.2.3 Modeling the Probabilities of Patterns of IBD	580
	20.2.4 Inferring Local IBD from Marker Data	581
20.3	IBD-Based Detection of Associations between Markers and Traits	583
	20.3.1 Trait Data Probabilities for Major Gene Models	583
	20.3.2 Quantitative Trait Data Probabilities under Random Effects Models	584
	20.3.3 IBD-Based Linkage Likelihoods for Major Gene Models	585
	20.3.4 IBD-Based Linkage Likelihoods for Random-Effects Models	587
20.4	Other Forms of IBD-Based Genetic Mapping	589

20.4.1	IBD-Based Case–Control Studies	589
20.4.2	Patterns of IBD in Affected Relatives	590
20.5	Summary	592
	Acknowledgements	592
	References	593
21	Genome-Wide Association Studies	597
	<i>Andrew P. Morris and Lon R. Cardon</i>	
21.1	Introduction	597
21.2	GWAS Design Concepts	599
21.2.1	Phenotype Definition	599
21.2.2	Structure of Common Genetic Variation and Design of GWAS Genotyping Technology	599
21.2.3	Sample Size Considerations	601
21.2.4	Genome-Wide Significance and Correction for Multiple Testing	601
21.2.5	Replication	602
21.3	GWAS Quality Control	602
21.3.1	SNP Quality Control Procedures	603
21.3.2	Sample Quality Control Procedures	604
21.3.3	Software	606
21.4	Single SNP Association Analysis	606
21.4.1	Generalised Linear Modelling Framework	606
21.4.2	Accounting for Confounding Factors as Covariates	606
21.4.3	Coding of SNP Genotypes	607
21.4.4	Imputed Genotypes	609
21.4.5	Visualisation of Results of Single SNP GWAS Analyses	609
21.4.6	Interactions with Non-Genetic Risk Factors	609
21.4.7	Bayesian Methods	611
21.4.8	Software	611
21.5	Detecting and Accounting for Genetic Structure in GWASs	611
21.5.1	Identification of Related Individuals	612
21.5.2	Multivariate Approaches to Identify Ethnic Outliers and Account for Population Stratification	613
21.5.3	Mixed Modelling Approaches to Account for Genetic Structure	614
21.5.4	Software	615
21.6	Multiple SNP Association Analysis	616
21.6.1	Haplotype-Based Analyses	616
21.6.2	SNP–SNP Interaction Analyses	617
21.6.3	Gene-Based Analyses	619
21.6.4	Software	619
21.7	Discussion	620
	References	623
22	Replication and Meta-analysis of Genome-Wide Association Studies	631
	<i>Frank Dudbridge and Paul Newcombe</i>	
22.1	Introduction	631
22.2	Replication	632
22.2.1	Motivation	632
22.2.2	Different Forms of Replication	632

22.2.3	Two-Stage Genome-Wide Association Studies	634
22.2.4	Significance Thresholds for Replication	634
22.2.5	A Key Challenge: Heterogeneity	635
22.3	Winner's Curse	635
22.3.1	Description of the Problem	635
22.3.2	Methods for Correcting for Winner's Curse	637
22.3.3	Applicability of These Methods	639
22.4	Meta-analysis	640
22.4.1	Motivation	640
22.4.2	An Illustrative Example	640
22.4.3	Fixed Effect Meta-analysis	641
22.4.4	Chi-Square Test for Heterogeneity in Effect	641
22.4.5	Random Effects Meta-analysis	642
22.4.6	Interpretation and Significance Testing of Meta-analysis Estimates	643
22.4.7	Using Funnel Plots to Investigate Small Study Bias in Meta-analysis	644
22.4.8	Improving Analyses via Meta-analysis Consortia and Publicly Available Data	645
22.5	Summary	647
	References	647
23	Inferring Causal Relationships between Risk Factors and Outcomes Using Genetic Variation	651
	<i>Stephen Burgess, Christopher N. Foley, and Verena Zuber</i>	
23.1	Background	651
23.1.1	Correlation and Causation	651
23.1.2	Chapter Outline	652
23.2	Introduction to Mendelian Randomization and Motivating Example	652
23.2.1	Instrumental Variable Assumptions	653
23.2.2	Assessing the Instrumental Variable Assumptions	654
23.2.3	Two-Sample Mendelian Randomization and Summarized Data	655
23.3	Monogenic Mendelian Randomization Analyses: The Easy Case	655
23.4	Polygenic Mendelian Randomization Analyses: The Difficult Case	656
23.4.1	Example: Low-Density Lipoprotein Cholesterol and Coronary Heart Disease Risk	656
23.4.2	More Complex Examples	657
23.4.3	Two-Stage Least Squares and Inverse-Variance Weighted Methods	659
23.5	Robust Approaches for Polygenic Mendelian Randomization Analyses	660
23.5.1	Median Estimation Methods	660
23.5.2	Modal Estimation Methods	660
23.5.3	Regularization Methods	660
23.5.4	Other Outlier-Robust Methods	661
23.5.5	MR-Egger Method	661
23.5.6	Multivariable Methods	663
23.5.7	Interactions and Subsetting	663
23.5.8	Practical Advice	664
23.6	Alternative Approaches for Causal Inference with Genetic Data	665
23.6.1	Fine-Mapping and Colocalization	665
23.6.2	LD Score Regression	666
23.7	Causal Estimation in Mendelian Randomization	667

23.7.1	Relevance of Causal Estimate	667
23.7.2	Heterogeneity and Pleiotropy	668
23.7.3	Weak Instrument Bias and Sample Overlap	668
23.7.4	Time-Dependent Causal Effects	669
23.7.5	Collider Bias	670
23.8	Conclusion	670
	References	671
24	Improving Genetic Association Analysis through Integration of Functional Annotations of the Human Genome	679
	<i>Qiongshi Lu and Hongyu Zhao</i>	
24.1	Introduction	679
24.2	Types of Functional Annotation Data in GWAS Applications	680
24.2.1	Transcriptomic Annotation Data	680
24.2.2	Epigenetic Annotation Data	681
24.2.3	DNA Conservation	681
24.3	Methods to Synthesize Annotation Data	682
24.3.1	Genome Browsers and Annotator Software	682
24.3.2	Supervised Learning Methods	682
24.3.3	Unsupervised Learning Methods	684
24.3.4	Improving Specificity of Computational Annotations	685
24.4	Methods to Integrate Functional Annotations in Genetic Association Analysis	685
24.4.1	Partitioning Heritability and Genetic Covariance	685
24.4.2	Imputation-Based Gene-Level Association Analysis	688
24.4.3	Other Applications and Future Directions	690
	Acknowledgements	690
	References	691
25	Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci	697
	<i>Solveig K. Sieberts and Eric E. Schadt</i>	
25.1	Introduction	697
25.1.1	An Overview of Transcription as a Complex Process	700
25.1.2	Modeling Approaches for Biological Processes	702
25.1.3	Human versus Experimental Models	704
25.2	Modeling for eQTL Detection and Causal Inference	705
25.2.1	Heritability of Expression Traits	705
25.2.2	Single-Trait eQTL Mapping	706
25.2.3	Joint eQTL Mapping	706
25.2.4	eQTL and Clinical Trait Linkage Mapping to Infer Causal Associations	708
25.3	Inferring Gene Regulatory Networks	714
25.3.1	From Assessing Causal Relationships among Trait Pairs to Predictive Gene Networks	714
25.3.2	Building from the Bottom Up or Top Down?	714
25.3.3	Using eQTL Data to Reconstruct Coexpression Networks	715
25.3.4	An Integrative Genomics Approach to Constructing Predictive Network Models	718

25.3.5	Integrating Genetic Data as a Structure Prior to Enhance Causal Inference in the Bayesian Network Reconstruction Process	720
25.3.6	Incorporating Other Omics Data as Network Priors in the Bayesian Network Reconstruction Process	721
25.3.7	Illustrating the Construction of Predictive Bayesian Networks with an Example	722
25.4	Conclusions	723
25.5	Software	724
	References	725
26	Statistical Methods for Single-Cell RNA-Sequencing	735
	<i>Tallulah S. Andrews, Vladimir Yu. Kiselev, and Martin Hemberg</i>	
26.1	Introduction	735
26.2	Overview of scRNA-Seq Experimental Platforms and Low-Level Analysis	736
26.2.1	Low-Throughput Methods	736
26.2.2	High-Throughput Methods	737
26.2.3	Computational Analysis	739
26.3	Novel Statistical Challenges Posed by scRNA-Seq	739
26.3.1	Estimating Transcript Levels	739
26.3.2	Analysis of the Expression Matrix	747
	References	753
27	Variant Interpretation and Genomic Medicine	761
	<i>K. Carss, D. Goldstein, V. Aggarwal, and S. Petrovski</i>	
27.1	Introduction and Current Challenges	761
27.2	Understanding the Effect of a Variant	765
27.3	Understanding Genomic Variation Context through Large Human Reference Cohorts	771
27.4	Functional Assays of Genetic Variation	777
27.5	Leveraging Existing Information about Gene Function Including Human and Model Phenotype Resources	779
27.6	Holistic Variant Interpretation	782
27.7	Future Challenges and Closing Remarks	783
27.8	Web Resources	786
	References	788
28	Prediction of Phenotype from DNA Variants	799
	<i>M.E. Goddard, T.H.E. Meuwissen, and H.D. Daetwyler</i>	
28.1	Introduction	799
28.2	Genetic Variation Affecting Phenotype	800
28.3	Data on DNA Polymorphisms Used for Prediction of Genetic Effects	801
28.4	Prediction of Additive Genetic Values	802
28.4.1	An Equivalent Model	804
28.4.2	Single-Step BLUP	804
28.4.3	Multiple Traits	805
28.4.4	Gene Expression	806
28.4.5	Using External Information	806
28.5	Factors Affecting Accuracy of Prediction	806
28.6	Other Uses of the Bayesian Genomic Selection Models	808

28.7	Examples of Genomic Prediction	809
28.7.1	Cattle	809
28.7.2	Humans	809
28.8	Conclusions	810
	References	810
29	Disease Risk Models	815
	<i>Allison Meisner and Nilanjan Chatterjee</i>	
29.1	Introduction and Background	815
29.1.1	Disease Risk Models and Their Applications	815
29.1.2	Examples of Available Disease Risk Models	817
29.1.3	Incorporating Genetic Factors	817
29.2	Absolute Risk Model	818
29.2.1	General Software for Building Absolute Risk Models	820
29.3	Building a Polygenic Risk Score	821
29.3.1	Expected Performance	821
29.3.2	Standard Approach to Constructing a PRS: LD Clumping and <i>p</i> -Value Thresholding	823
29.3.3	Advanced Approaches to Constructing a PRS	825
29.4	Combining PRS and Epidemiologic Factors	826
29.5	Model Validation	827
29.6	Evaluating Clinical Utility	828
29.7	Example: Breast Cancer	829
29.8	Discussion	832
29.8.1	Future Directions	832
29.8.2	Challenges	832
	References	833
30	Bayesian Methods for Gene Expression Analysis	843
	<i>Alex Lewin, Leonardo Bottolo, and Sylvia Richardson</i>	
30.1	Introduction	843
30.2	Modelling Microarray Data	845
30.2.1	Modelling Intensities	845
30.2.2	Gene Variability	845
30.2.3	Normalization	846
30.3	Modelling RNA-Sequencing Reads	846
30.3.1	Alignments for RNA-Sequencing Data	846
30.3.2	Likelihood for Read-Level Data	847
30.3.3	Likelihood for Transcript-Level Read Counts	848
30.3.4	Likelihood for Gene-Level Read Counts	850
30.4	Priors for Differential Expression Analysis	851
30.4.1	Differential Expression from Microarray Data	851
30.4.2	Differential Expression from RNA-Sequencing Data	856
30.5	Multivariate Gene Selection Models	857
30.5.1	Variable Selection Approach	857
30.5.2	Bayesian Shrinkage with Sparsity Priors	861
30.6	Quantitative Trait Loci	863
30.6.1	Single-Response Models	863
30.6.2	Multiple-Response Models	864

Acknowledgements	868
References	869
31 Modelling Gene Expression Dynamics with Gaussian Process Inference	879
<i>Magnus Rattray, Jing Yang, Sumon Ahmed, and Alexis Boukouvalas</i>	
31.1 Introduction	879
31.1.1 Covariance Function	880
31.1.2 Inference	882
31.2 Applications to Bulk Time Series Expression Data	883
31.2.1 Identifying Differential Expression in Time	884
31.2.2 Identifying Changes between Two Time Course Experiments	885
31.2.3 Hierarchical Models of Replicates and Clusters	887
31.2.4 Differential Equation Models of Production and Degradation	888
31.3 Modelling Single-Cell Data	889
31.3.1 Modelling Single-Cell Trajectory Data	889
31.3.2 Dimensionality Reduction and Pseudotime Inference	890
31.3.3 Modelling Branching Dynamics with Single-Cell RNA-Sequencing Data	892
31.4 Conclusion	893
Acknowledgements	894
References	894
32 Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models	899
<i>Marco Grzegorczyk and Dirk Husmeier</i>	
32.1 Introduction	899
32.2 Methodology	901
32.2.1 Dynamic Bayesian Networks (DBN)	901
32.2.2 Bayesian Linear Regression	902
32.2.3 Bayesian Piecewise Linear Regression (NH-DBN)	905
32.2.4 Bayesian Piecewise Linear Regression with Coupled Regression Coefficients (Coupled NH-DBNs)	908
32.2.5 NH-DBNs with More Flexible Allocation Schemes	915
32.2.6 NH-DBNs with Time-Varying Network Structures	916
32.2.7 Dynamic Bayesian Network Modelling	918
32.2.8 Computational Complexity	920
32.3 Application Examples	921
32.3.1 Morphogenesis in Drosophila	921
32.3.2 Synthetic Biology in Yeast	922
32.4 Summary	927
Appendix A: Coupling Schemes	927
A.1 Hard Information Coupling Based on an Exponential Prior	928
A.2 Hard Information Coupling Based on a Binomial Prior	928
A.3 Soft Information Coupling Based on a Binomial Prior	929
References	929
33 DNA Methylation	933
<i>Kasper D. Hansen, Kimberly D. Siegmund, and Shili Lin</i>	
33.1 A Brief Introduction	933
33.2 Measuring DNA Methylation	934

33.3	Differential DNA Methylation	936
33.3.1	Differential Methylation with Bisulfite-Sequencing Data	936
33.3.2	Differential Methylation with Capture-Sequence Data	939
33.3.3	Differential Methylation with HumanMethylation Array Data	940
33.4	Other Topics of Interest	941
	References	942
34	Statistical Methods in Metabolomics	949
	<i>Timothy M.D. Ebbels, Maria De Iorio, and David A. Stephens</i>	
34.1	Introduction	949
34.2	Preprocessing and Deconvolution	950
34.2.1	Nuclear Magnetic Resonance Spectroscopy	950
34.2.2	Liquid Chromatography – Mass Spectrometry	952
34.3	Univariate Methods	954
34.3.1	Metabolome-Wide Significance Levels	956
34.3.2	Sample Size and Power	957
34.4	Multivariate Methods and Chemometrics Techniques	958
34.4.1	Linear Regression Methods	959
34.4.2	Shrinkage Methods	960
34.5	Orthogonal Projection Methods	961
34.5.1	Principal Components Analysis	962
34.5.2	Partial Least Squares	964
34.5.3	Orthogonal Projection onto Latent Structures	965
34.6	Network Analysis	966
34.7	Metabolite Identification and Pathway Analysis	969
34.7.1	Statistical Correlation Spectroscopy	969
34.7.2	Pathway and Metabolite Set Analysis	971
34.8	Conclusion	972
	References	972
35	Statistical and Computational Methods in Microbiome and Metagenomics	977
	<i>Hongzhe Li</i>	
35.1	Microbiome in Human Health and Disease	977
35.2	Estimation of Microbiome Features from 16S rRNA and Shotgun Metagenomic Sequencing Data	980
35.2.1	Estimation of Microbiome Features in 16S rRNA Data	980
35.2.2	Estimation of Microbial Composition in Shotgun Metagenomic Data	981
35.2.3	Estimation of Microbial Gene/Pathway Abundance in Shotgun Metagenomic Data	982
35.2.4	Quantification of Bacterial Growth Dynamics	982
35.2.5	Microbial Diversity Index	983
35.3	Methods for Analysis of Microbiome as an Outcome of an Intervention or Exposure	983
35.3.1	Modeling Multivariate Sparse Count Data as the Response Variable	984
35.3.2	Modeling High-Dimensional Compositional Response Data in Microbiome Studies	984
35.4	Methods for Analysis of Microbiome as a Covariate	985
35.4.1	Regression Analysis with Compositional Covariates	985
35.4.2	Kernel-Based Regression in Microbiome Studies	986

35.5	Methods for Analysis of Microbiome as a Mediator	987
35.6	Integrative Analysis of Microbiome, Small Molecules and Metabolomics Data	989
35.6.1	Computational Analysis of Small Molecules from the Human Microbiota	989
35.6.2	Metabolic Modeling in Microbiome	990
35.7	Discussion and Future Directions	991
	Acknowledgements	991
	References	992
36	Bacterial Population Genomics	997
	<i>Jukka Corander, Nicholas J. Croucher, Simon R. Harris, John A. Lees, and Gerry Tonkin-Hill</i>	
36.1	Introduction	997
36.2	Genetic Population Structure and Clustering of Genotypes	998
36.2.1	Background	998
36.2.2	Model-Based Clustering	998
36.2.3	Linkage Disequilibrium	1000
36.2.4	Distance-Based Methods	1000
36.3	Phylogenetics and Dating Analysis	1001
36.4	Transmission Modeling	1004
36.4.1	Challenges	1004
36.5	Genome-Wide Association Studies in Bacteria	1008
36.5.1	Background	1008
36.5.2	Phylogenetic Methods	1009
36.5.3	Regression-Based Methods	1011
36.6	Genome-Wide Epistasis Analysis	1012
36.7	Gene Content Analysis	1013
	References	1014
	Reference Author Index	1021
	Subject Index	1109

List of Contributors

V. Aggarwal

Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Sumon Ahmed

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Joshua M. Akey

Department of Ecology and Evolutionary Biology and Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA

Tallulah S. Andrews

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Richard Ashcroft

School of Law, Queen Mary University of London, London, UK

Nick Barton

Institute of Science and Technology Austria, Klosterneuburg, Austria

Mark Beaumont

School of Biological Sciences, Bristol University, Bristol, UK

Leonardo Bottolo

Department of Medical Genetics, University of Cambridge, Cambridge, UK, The Alan Turing Institute, London, UK, and MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Alexis Boukouvalas

Prowler.io, Cambridge, UK

Stephen Burgess

MRC Biostatistics Unit, University of Cambridge and Cardiovascular Epidemiology Unit, University of Cambridge, UK

Lon R. Cardon

BioMarin Pharmaceutical, Novato, CA, USA

K. Cars

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Nilanjan Chatterjee

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Jukka Corander

Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, Department of Biostatistics, University of Oslo, Oslo, Norway, and Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Nicholas J. Croucher

Department of Infectious Disease Epidemiology, Imperial College London, London, UK

H.D. Daetwyler

Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia, and School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia

Maria De Iorio

Department of Statistical Science, University College London, London, UK

Frank Dudbridge

Department of Health Sciences, University of Leicester, Leicester, UK

Timothy M.D. Ebbels

Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, UK

Alison Etheridge

University of Oxford, UK

Christopher N. Foley

MRC Biostatistics Unit, University of Cambridge, UK

Antonio Augusto Franco Garcia

University of São Paulo, Piracicaba, Brazil

M.E. Goddard

Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Victoria, Australia, and Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia

Nick Goldman

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, UK

D. Goldstein

Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Marco Grzegorczyk

Bernoulli Institute (BI), Faculty of Science and Engineering, Rijksuniversiteit Groningen, Groningen, Netherlands

Torsten Günther

Human Evolution, Department of Organismal Biology, Uppsala University, Sweden

Kasper D. Hansen

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Simon R. Harris

Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Garrett Hellenthal

University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, UK

Martin Hemberg

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

John P. Huelsenbeck

Department of Integrative Biology, University of California, Berkeley, CA, USA

Dirk Husmeier

School of Mathematics & Statistics, University of Glasgow, Glasgow, UK

Mattias Jakobsson

Human Evolution, Department of Organismal Biology, Uppsala University, Sweden

David T. Jones

University College London, London, UK

Shaun Kandathil

University College London, London, UK

Jerome Kelleher

University of Oxford, UK

Vladimir Yu. Kiselev

Wellcome Sanger Institute, Hinxton,
Cambridgeshire, UK

Laura Kubatko

Ohio State University, Columbus, OH, USA

John A. Lees

Department of Microbiology, School of
Medicine, New York University, New York,
USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Christoph Leuenberger

University of Fribourg, Switzerland

Hongzhe Li

Department of Biostatistics, Epidemiology
and Informatics, Perelman School of
Medicine, University of Pennsylvania,
Philadelphia, USA

Liming Li

Department of Ecology and Evolutionary
Biology, Princeton University, Princeton,
NJ, USA

Shili Lin

Department of Statistics, Ohio State
University, Columbus, OH, USA

Alex Lewin

Department of Medical Statistics, London
School of Hygiene and Tropical Medicine,
London, UK

Qiongshi Lu

Department of Biostatistics and Medical
Informatics, University of Madison-
Wisconsin, Madison, WI, USA

Jonathan Marchini

Regeneron Genetics Center, Tarrytown,
NY, USA

Gil McVean

University of Oxford, UK

Allison Meisner

Department of Biostatistics, Johns Hopkins
Bloomberg School of Public Health,
Baltimore, MD, USA

Ian Mackay

IMplant Consultancy Ltd, Chelmsford, UK

T.H.E. Meuwissen

Norwegian University of Life Sciences,
Ås, Norway

Iain H. Moul

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

Andrew P. Morris

Department of Biostatistics, University of
Liverpool, Liverpool, UK

Michael B. Morrissey

School of Biology, University of St Andrews,
St Andrews, UK

Paul Newcombe

MRC Biostatistics Unit, University of
Cambridge, Cambridge, UK

Rasmus Nielsen

Department of Integrative Biology and
Department of Statistics, University of
California, Berkeley, CA, USA

Magnus Nordborg

Gregor Mendel Institute, Austrian
Academy of Sciences, Vienna BioCenter,
Vienna, Austria

Umberto Peron

European Molecular Biology Laboratory,
European Bioinformatics Institute
(EMBL-EBI), Hinxton, Cambridgeshire,
UK

S. Petrovski

Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Hans-Peter Piepho

University of Hohenheim, Stuttgart, Germany

Magnus Rattray

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Sylvia Richardson

MRC Biostatistics Unit, University of Cambridge, Cambridge, UK, and The Alan Turing Institute, London, UK

Eric E. Schadt

Sema4, Stamford, CT, USA, and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Solveig K. Sieberts

Sage Bionetworks, Seattle, WA, USA

Kimberly D. Siegmund

Department of Preventive Medicine, Keck School of Medicine of USC, Los Angeles, USA

David A. Stephens

Department of Mathematics and Statistics, McGill University, Montreal, Canada

Aaron J. Stern

Graduate Group in Computational Biology, University of California, Berkeley, CA

William R. Taylor

The Francis Crick Institute, London, UK

E.A. Thompson

Department of Statistics, University of Washington, Seattle, WA, USA

Jeffrey L. Thorne

Department of Statistics & Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Gerry Tonkin-Hill

Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Susan E. Wallace

Department of Health Sciences, University of Leicester, Leicester, UK

Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Jinliang Wang

Institute of Zoology, Zoological Society of London, UK

Daniel Wegmann

University of Fribourg, Switzerland

B.S. Weir

Department of Biostatistics, University of Washington, Seattle, WA, USA

Jing Yang

Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

Ziheng Yang

Department of Genetics, Evolution and Environment, University College London, London, UK

Hongyu Zhao

Department of Biostatistics, Yale University, New Haven, CT, USA

Verena Zuber

MRC Biostatistics Unit, University of Cambridge, UK and School of Public Health, Imperial College London

Editors' Preface to the Fourth Edition

After a break of more than 10 years since the third edition, we are pleased to present the fourth edition of the *Handbook of Statistical Genomics*. Genomics has moved on enormously during this period, and so has the *Handbook*: almost everything is new or much revised, with only a small amount of material carried forward from previous editions. Two new editors have joined, Ida Moltke from Copenhagen and John Marioni from Cambridge. With sadness we note the death of founding editor Professor Chris Cannings during 2018. He first saw the need and had the vision for the *Handbook*, one of his many contributions to mathematical and statistical genetics. We also acknowledge the fundamental contribution of the other founding editor, Professor Martin Bishop.

While the content has changed, the mission has not: the *Handbook* is intended as an introduction suitable for advanced graduate students and early-career researchers who have achieved at least a good first-year undergraduate level in both statistics and genetics, and preferably more in at least one of those fields. The chapters are not thorough literature reviews, but focus on explaining the key ideas, methods and algorithms, citing key recent and historic literature for further details and references.

The change of title (from *Genetics* to *Genomics*) is not intended to indicate a substantial change of focus, but reflects both changes in the field, with increased emphasis on transcriptomics and epigenetics for example, and changes in usage. We interpret 'genomics' broadly, to include studies of whole genomes and epigenomes, near-genome processes such as transcription and metabolomics, as well as genomic mechanisms underlying whole-organism outcomes related to selection, adaptation and disease. We also interpret 'statistics' broadly, to include for example relevant aspects of data science and bioinformatics.

The 36 chapters are intended to be largely independent, so that to benefit from the *Handbook* it is not necessary to read every chapter, or to read chapters in order. This structure necessitates some duplication of material, which we have tried to minimize but not always eliminate. Alternative approaches to the same topic by different authors can be beneficial. The extensive subject and author indexes allow easy reference to topics arising in different chapters.

For those with minimal genetics background the glossary has been newly updated. Thanks to Keren Carss for contributing many new terms, in particular those relevant to genomic medicine. Gerry Tonkin-Hill and John Lees also contributed some advanced statistical terminology, but the glossary is predominantly of genetic terms. For those with limited background in statistical modeling and inference we have added an initial chapter that covers these topics, ranging from basic concepts to state-of-art models and methods.

We thank the many commentators on previous editions who were generous in their praise and helpful feedback. No doubt many more improvements will be possible for future editions and we welcome comments e-mailed to any of the editors. We are grateful to all of our authors for taking the time to write and update their chapters with care, and we would like to express our appreciation to all the professional staff working with and for Wiley who helped us to bring this project to fruition.

Glossary

Many of the updates were prepared by Keren Carss (Chapter 27) and Gerry Tonkin-Hill and John Lees (Chapter 36)

NB: Some of the definitions below assume that the organism of interest is diploid.

Adenine (A): purine base that forms a pair with thymine in DNA and uracil in RNA.

Admixture: arises when two previously isolated populations begin interbreeding.

Allele: one of the possible forms of a gene at a given locus. Depending on the technology used to type the gene, it may be that not all DNA sequence variants are recognized as distinct alleles.

Allele frequency: often used to mean the relative frequency (i.e. proportion) of an allele in a sample or population.

Allelic heterogeneity: different alleles in the same gene cause different phenotypes.

Alpha helix: a helical (usually right-handed) arrangement that can be adopted by a polypeptide chain; a common type of protein secondary structure.

Amino acid: the basic building block of proteins. There are 20 naturally occurring amino acids in animals which when linked by peptide bonds form polypeptide chains.

Aneuploid cells: do not have the normal number of chromosomes.

Antisense strand: the DNA strand complementary to the coding strand, determined by the covalent bonding of adenine with thymine and cytosine with guanine.

Approximate Bayesian computation (ABC): methods that approximate the likelihood function by comparing simulations with the observed data.

Ascertainment: the strategy by which study subjects are identified, selected, and recruited for participation in a study.

Autosome: a chromosome other than the sex chromosomes. Humans have 22 pairs of autosomes plus 2 sex chromosomes.

Autozygosity: regions of the genome that are identical due to inheritance from a common ancestor.

Backcross: a linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed back to one of the inbred parental strains.

Bacterial artificial chromosome (BAC): a vector used to clone a large segment of DNA (100–200 kb) in bacteria resulting in many copies.

Base: (abbreviated term for a purine or pyrimidine in the context of nucleic acids), a cyclic chemical compound containing nitrogen that is linked to either a deoxyribose (DNA) or a ribose (RNA).

Base pair (bp): a pair of bases that occur opposite each other (one in each strand) in double stranded DNA/RNA. In DNA adenine base pairs with thymine and cytosine with guanine. RNA is the same except that uracil takes the place of thymine.

Bayesian: a statistical school of thought that, in contrast with the frequentist school, holds that inferences about any unknown parameter or hypothesis should be encapsulated in a probability distribution, given the observed data. Bayes' theorem allows one to compute the posterior distribution for an unknown from the observed data and its assumed prior distribution.

Beta-sheet: a (hydrogen-bonded) sheet arrangement which can be adopted by a polypeptide chain; a common type of protein secondary structure.

Cell line: an established, immortalized cell culture with a uniform genetic background.

Centimorgan (cM): measure of genetic distance. Two loci separated by 1 cM have an average of one recombination between them every 100 meioses. Because of the variability in recombination rates, genetic distance differs from physical distance, measured in base pairs. Genetic distance differs between male and female meioses; an average over the sexes is usually used.

Centromere: the region where the two sister chromatids join, separating the short (p) arm of the chromosome from the long (q) arm.

Chiasma: the visible structure formed between paired homologous chromosomes (non-sister chromatids) in meiosis.

Chromatid: a single strand of the (duplicated) chromosome, containing a double-stranded DNA molecule.

Chromatin: the material composed of DNA and chromosomal proteins that makes up chromosomes. Comes in two types, euchromatin and heterochromatin.

Chromosome: the self-replicating threadlike structure found in cells. Chromosomes, which at certain stages of meiosis and mitosis consist of two identical sister chromatids, joined at the centromere, carry the genetic information encoded in the DNA sequence.

cis-acting: regulatory elements and expression quantitative trait locus whose DNA sequence directly influences transcription. The physical locations for *cis*-acting elements are in or near the gene or genes they regulate. Contrast *trans*.

Clones: genetically engineered identical cells/sequences.

Co-dominance: both alleles contribute to the phenotype, in contrast with recessive or dominant alleles.

Codon: a nucleotide triplet that encodes an amino acid or a termination signal.

Common disease common variant (CDCV) hypothesis: the hypothesis that many genetic variants underlying complex diseases are common, and hence susceptible to detection using SNP-based association studies.

complementary DNA (cDNA): DNA that is synthesized from a messenger RNA template using the reverse transcriptase enzyme.

Consanguinous mating: mating between closely-related individuals.

Conservation: the similarity of a sequence across species. High conservation indicates selective pressure against mutations, and is suggestive of functional importance.

Constraint: measured by the degree of similarity of a sequence between individuals of the same species, high constraint indicates selective pressure against mutations within that species, which is suggestive of functional importance. See also *intolerance*.

Contig: a group of contiguous, overlapping, cloned DNA sequences.

Core genome: the set of genes shared by all isolates in a collection of related bacterial genomes.

Cytosine (C): pyrimidine base that forms a pair with guanosine in DNA.

Degrees of freedom (df): this term is used in different senses in statistics and in other fields. It can often be interpreted as the number of values that can be defined arbitrarily in the

specification of a system; for example, the number of coefficients in a regression model. Frequently it suffices to regard a df as a parameter used to define certain probability distributions.

De novo variant: a variant in an index case not inherited from either parent.

Deoxyribonucleic acid (DNA): polymer made up of deoxyribonucleotides linked together by phosphodiester bonds.

Deoxyribose: the sugar compound found in DNA.

Diagnostic yield: the fraction of patients for whom a genetic diagnosis is made.

Diploid: has two versions of each autosome, one inherited from the father and one from the mother. Compare with *haploid*.

Dizygotic (DZ) twins: twins derived from different pairs of egg and sperm. DZ twins are genetically equivalent to full sibs.

DNA methylation: the addition of a methyl group to DNA. In mammals this occurs at the C-5 position of cytosine, most often at CpG dinucleotides.

DNA microarray: small slide or 'chip' used to simultaneously measure the quantity of large numbers of different mRNA gene transcripts present in cell or tissue samples.

Dominant allele: results in the same phenotype irrespective of the other allele at the locus.

Dominant negative allele: changes the function of a gene product to be antagonistic to the reference allele.

Effective population size: the size of a theoretical population that best approximates a given natural population under an assumed model. The criterion for assessing the 'best' approximation can vary, but is often some measure of total genetic variation.

Endogamy: mating restricted to a small gene pool.

Enzyme: a protein that controls the rate of a biochemical reaction.

Epigenome: the set of chemical alterations (such as methylation) to a genome's DNA and histones.

Epistasis: the physiological interaction between different genes such that one gene alters the effects of other genes.

Epitope: the part of an antigen that the antibody interacts with.

Eukaryote: organism whose cells include a membrane-bound nucleus. Compare with *prokaryote*.

Exons: parts of a gene that are transcribed into RNA and remain in the mature RNA product after splicing. An exon may code for a specific part of the final protein.

Expression quantitative trait locus (eQTL): a locus influencing the expression of one or more genes.

Fixation: occurs when a locus which was previously polymorphic in a population becomes monomorphic because all but one allele has been lost through genetic drift.

Frequentist: the school of statistical thought in which support for a hypothesis or parameter value is assessed using the probability of the observed data (or more 'extreme' data), given the hypothesis or value. Contrast with *Bayesian*.

Gain-of-function: a new molecular function or increase in the normal function/expression of a gene product.

Gamete: a sex cell, sperm in males, egg in females. Two haploid gametes fuse to form a diploid zygote.

Gene: a segment (not necessarily contiguous) of DNA that codes for a protein or functional RNA.

Gene expression: the process by which coding DNA sequences are converted into functional elements in a cell.

Genealogy: the ancestral relationships among a sample of homologous genes drawn from different individuals, which can be represented by a tree. Also sometimes used in place of pedigree, the ancestral relationships among a set of individuals, which can be represented by a graph.

Genetic drift: the changes in allele frequencies that occur over time due to the randomness inherent in reproductive success.

Genome: all the genetic material of an organism.

Genotype: the (unordered) allele pair(s) carried by an individual at one or more loci. A multilocus genotype is equivalent to the individual's two haplotypes without the phase information.

Germline: sex cells (egg or sperm) that transmit DNA over generations.

Guanine (G): purine base that forms a pair with cytosine in DNA.

Haemoglobin: the red oxygen-carrying pigment of the blood, made up of two pairs of polypeptide chains called globins (2α and 2β subunits).

Haploid: has a single version of each chromosome.

Haploinsufficiency: a dose-sensitive gene such that two working alleles are required to maintain a healthy organism.

Haplotype: the alleles at different loci on a chromosome. An individual's two haplotypes imply the genotype; the converse is not true, but in the presence of strong linkage disequilibrium haplotypes can be accurately inferred from genotype.

Hardy–Weinberg disequilibrium: the non-independence within a population of an individual's two alleles at a locus; can arise due to inbreeding or selection for example. Compare with *linkage disequilibrium*.

Heritability: the proportion of the phenotypic variation in the population that can be attributed to genetic variation.

Heterozygosity: the proportion of individuals in a population that are heterozygotes at a locus. Also sometimes used as shorthand for expected heterozygosity under random mating, which equals the probability that two homologous genes drawn at random from a population are different alleles.

Heterozygote: a single-locus genotype consisting of two different alleles.

Hidden Markov model (HMM): a probabilistic model in which a sequence of unobserved states is a finite Markov chain and a sequence of observed states are random variables that depend on the corresponding hidden state only.

Hierarchical Dirichlet process: a nonparametric generalization of the latent finite mixture model where the number of populations K can be unbounded and is learnt from the data.

Homology: similarities between sequences that arise because of shared evolutionary history (descent from a common ancestral sequence). Homology of different genes within a genome is called paralogous, while that between the genomes of different species is called orthologous.

Homozygote: a single-locus genotype consisting of two versions of the same allele.

Human immunodeficiency virus (HIV): a virus that causes acquired immune deficiency syndrome (AIDS) which destroys the body's ability to fight infection.

Hybrid: the offspring of a cross between parents of different genetic types or different species.

Hybridization: the base pairing of a single stranded DNA or RNA sequence to its complementary sequence.

Hypomorphic: reduces the activity but does not eliminate gene function.

Identity by descent (IBD): two genes are IBD if they have descended without alteration from an ancestral gene.

Inbred lines: derived and maintained by intensive inbreeding, the individuals have almost identical genomes.

Inbreeding: a system of mating in which mates are on average more closely related than under random mating. Inbreeding reduces genetic variation and increases homozygosity and hence the prevalence of recessive traits.

Intercross (or F2 design): a linkage study design in which the progeny (F1s) of a cross between two inbred lines are crossed or selfed.

Intron: non-coding DNA sequence separating the exons of a gene. Introns are initially transcribed into messenger RNA but are subsequently spliced out.

Karyotype: the number and structure of an individual's chromosomes.

Kilobase (kb): 1000 base pairs.

Latent finite mixture model: a model which assumes that observed data points belong to an unobserved (latent) set of subpopulations.

Linear mixed model: an extension of linear models that allows for both fixed and random effects, the latter used to model dependence within data sets.

Linkage: two genes are said to be linked if they are located close together on the same chromosome. The alleles at linked genes tend to be co-inherited more often than those at unlinked genes because of the reduced opportunity for an intervening recombination.

Linkage disequilibrium (LD): the non-independence within a population of a gamete's alleles at different loci; can arise due to linkage, population stratification, or selection. The term is misleading and 'gametic phase disequilibrium' is sometimes preferred. Various measures of linkage disequilibrium are available, including D' and r^2 .

Locus (pl. loci): the position of a gene on a chromosome.

Locus heterogeneity: the degree to which variants in different genes can cause similar phenotype(s).

LOD score: a likelihood ratio statistic used to infer whether two loci are genetically linked (have recombination fraction less than 0.5).

Loss-of-function variant: the mutant allele abolishes the function of a gene product.

Metabolome: the complement of all metabolites in a sample.

Marker gene: a polymorphic gene of known location which can be readily typed; used, for example, in genetic mapping.

Markov chain (first-order): a sequence of random variables such that, given the current state of the sequence, the probability distribution of the next state is independent of all previous states.

Markov chain Monte Carlo: a stochastic algorithm used to sample from a target probability distribution. It works by constructing a Markov chain that has the target distribution as its equilibrium distribution, so that over many steps of the chain the fraction of time spent in any region of the state space converges to the probability of that region under the target distribution.

Megabase (Mb): 1000 kilobases = 1,000,000 base pairs.

Meiosis: the process by which (haploid) gametes are formed from (diploid) somatic cells.

messenger RNA (mRNA): the RNA sequence that acts as the template for protein synthesis.

Microarray: see *DNA microarray*.

Microbiome: the complement of all microbes in a sample.

Microsatellite DNA: small stretches of DNA (usually 1–6 bp) tandemly repeated.

Microsatellite loci are often highly polymorphic, and alleles can be distinguished by length, making them useful as marker loci.

Mitochondrial DNA (mtDNA): the genetic material of the mitochondria which consists of a circular DNA duplex inherited maternally.

Mitosis: the process by which a somatic cell is replaced by two daughter somatic cells.

Modifier variant: a variant that modifies the phenotype caused by another variant.

Monomorphic: a locus at which only one allele arises in the sample or population.

monozygotic twins: genetically identical individuals derived from a single fertilized egg.

Morgan: unit of genetic distance = 100 centimorgans.

Morpholino: a synthetic oligonucleotide used experimentally to modify gene expression.

Mosaicism: two or more populations of cells arising from the same fertilized egg. Can result from a somatic mutation.

Mutation: a process that changes an allele.

Negative selection: removal of deleterious mutations by natural selection. Also known as ‘purifying selection’.

Neutral: not subject to selection.

Neutral evolution: evolution of alleles with nearly zero selective coefficient. When $|Ns| \ll 1$, where N is the population size and s is the selective coefficient, the fate of the allele is mainly determined by random genetic drift rather than natural selection.

Non-coding RNA (ncRNA): RNA transcripts not translated into protein product. There are many classes of ncRNA, some of which can affect the rate of transcription or transcript degradation.

Nonsense mediated decay: the cellular pathway by which transcripts that contain premature stop codons are degraded to prevent translation into aberrant proteins.

Nonsynonymous substitution: nucleotide substitution in a protein-coding gene that alters the encoded amino acid.

Nucleoside: a base attached to a sugar, either ribose or deoxyribose.

Nucleotide: the structural units with which DNA and RNA are formed. Nucleotides consist of a base attached to a five-carbon sugar and mono-, di-, or triphosphate.

Nucleotide substitution: the replacement of one nucleotide by another during evolution. Substitution is generally considered to be the product of both mutation and selection.

Oligonucleotide: a short sequence of single-stranded DNA or RNA, often used as a probe for detecting the complementary DNA or RNA.

Open reading frame (ORF): a long sequence of DNA with an initiation codon at the 5' end and no termination codon except for one at the 3' end.

Pangenome: the totality of genes identified in a collection of related bacterial genomes.

Pathogenicity: the degree to which a variant increases an individual's predisposition to a genetic disease.

Pedigree: a diagram showing the relationship of each family member and the heredity of a particular trait through several generations of a family.

Penetrance: the probability that a particular phenotype is observed in an individual with a given genotype. Penetrance can vary with environment and the alleles at other loci.

Peptide bond: linkages between amino acids occur through a covalent peptide bond joining the C terminal of one amino acid to the N terminal of the next (with loss of a water molecule).

Phase (of linked markers): the relationship (either coupling or repulsion) between alleles at two linked loci. The two alleles at the linked loci are said to be ‘in coupling’ if they are present on the same physical chromosome or ‘in repulsion’ if they are present on different parental homologs.

Phenotype: the observed characteristic under study, may be quantitative (i.e. continuous) such as height, or binary (e.g. disease/no disease), or ordered categorical (e.g. mild/moderate/severe).

Phenotypic heterogeneity: the degree to which different phenotypes are associated with the same variant or gene. Also known as ‘variable expressivity’.

Phenotypic spectrum: the range of different phenotypes associated with the same variant or gene.

Pleiotropy: the effect of a gene on several different traits.

Polygenic traits: affected by multiple genes.

Polymerase chain reaction (PCR): a laboratory process by which a specific, short, DNA sequence is amplified many times.

Polymorphic: a locus that is not monomorphic. Usually a stricter criterion is imposed: a locus is polymorphic if no allele has frequency greater than 0.99.

Polynucleotide: a polymer of either DNA or RNA nucleotides.

Polypeptide: a long chain of amino acids joined together by peptide bonds.

Polypeptide chain: a series of amino acids linked by peptide bonds. Short chains are sometimes referred to as ‘oligopeptides’ or simply ‘peptides’.

Polytene: refers to the giant chromosomes that are generated by the successive replication of chromosome pairs without the nuclear division, thus several chromosome sets are joined together.

Population stratification (or population structure): refers to a situation in which the population of interest can be divided into strata such that an individual tends to be more closely related to others within the same stratum than to other individuals.

Positive selection: fixation, by natural selection, of an advantageous allele with a positive selective coefficient. Also known as ‘Darwinian selection’.

Precision medicine: tailoring healthcare management, including targeted treatment, to individual patients considering variability in genes, environment and lifestyle.

Primary cells: cells taken from a tissue sample, which have undergone few *in vitro* mitoses and have a limited lifespan.

Proband: an individual through whom a family is ascertained, typically by their phenotype.

Product-partition model: models that partition a data set into distinct clusters. They assume that, given a partition of objects in a data set, objects belonging to the same set are exchangeable and objects in different sets are independent.

Prokaryote: a unicellular organism with no nucleus.

Promoter: located upstream of the gene, the promoter allows the binding of RNA polymerase which initiates transcription of the gene.

Protein: a large, complex, molecule made up of one or more chains of amino acids.

Protein-truncating variant: a variant predicted to truncate a gene product. They usually, but not always, are loss-of-function variants.

Proteome: the complement of all protein molecules in a sample.

Pseudogene: a DNA sequence that is either an imperfect, non-functioning, copy of a gene, or a former gene which no longer functions due to mutations.

Purine: adenine or guanine; particular kinds of nitrogen-containing heterocyclic rings.

Pyrimidine: cytosine, thymine, or uracil; particular kinds of nitrogen-containing heterocyclic rings.

Quantitative trait locus (QTL): a locus influencing a continuously varying phenotype.

Radiation hybrid: a cell line, usually rodent, that has incorporated fragments of foreign chromosomes that have been broken by irradiation. They are used in physical mapping.

Recessive allele: has no effect on phenotype except when present in homozygote form.

Recombination: the formation of new haplotypes by physical exchange between two homologous chromosomes during meiosis.

Restriction enzyme: recognizes specific nucleotide sequences in double-stranded DNA and cuts at a specified position with respect to the sequence.

Restriction site: a 4–8 bp DNA sequence (usually palindromic) that is recognized by a restriction enzyme.

Retrovirus: an RNA virus whose replication depends on a reverse transcriptase function, allowing the formation of a cDNA copy that can be stably inserted into the host chromosome.

Ribonucleic acid (RNA): polymer made up of ribonucleotides that are linked together by phosphodiester bonds.

Ribosome: a cytoplasmic organelle, consisting of RNA and protein, that is involved in the translation of messenger RNA into proteins.

Ribosomal RNA (rRNA): the RNA molecules contained in ribosomes.

Selection: a process such that expected allele frequencies do not remain constant, in contrast with genetic drift. Alleles that convey an advantage to the organism in its current environment tend to become more frequent in the population (positive, or adaptive, selection), while deleterious alleles become less frequent. Under stabilizing (or balancing) selection, allele frequencies tend towards a stable, intermediate value.

Sense strand: the DNA strand in the direction of coding.

Sex-linked: a trait influenced by a gene located on a sex (X or Y) chromosome.

Single nucleotide polymorphism (SNP): a polymorphism consisting of a single nucleotide.

Sister chromatids: two chromatids that are copies of the same chromosome. Non-sister chromatids are different but homologous.

Somatic cell: a non-sex cell.

Somatic variant: a variant that arose from mutation in a somatic cell and so is not transmitted to descendants.

Structural variant: genomic rearrangements including deletions (in excess of 50 bp), insertions, tandem duplications, dispersed duplications, inversions, translocations, and complex structural variants.

Synonymous substitution: nucleotide substitution in a protein-coding gene that does not alter the encoded amino acid.

TATA box: a conserved sequence (TATAAAA) found about 25–30 bp upstream from the start of transcription site in most but not all genes.

Thymine (T): pyrimidine base that forms a pair with adenine in DNA.

trans-acting: eQTL whose DNA sequence influences gene expression through its gene product. These regulatory elements are often coded for at loci far from or unlinked to the genes they regulate. Contrast *cis*.

Transcription: the synthesis of a single-stranded RNA version of a DNA sequence.

Transcriptome: the complement of all RNA molecules that can be transcribed from a genome.

Transition: a mutation that changes either one purine base to the other, or one pyrimidine base to the other.

Translation: the process whereby messenger RNA is ‘read’ by transfer RNA and its corresponding polypeptide chain synthesized.

Transposon: a genetic element that can move over generations from one genomic location to another.

Transversion: a mutation that changes a purine base to a pyrimidine, or vice versa.

Triplosensitivity: dose-sensitivity of a genomic region such that excess copies are pathogenic.

Ultra-rare variant: a variant found in an index sample that is unobserved across all available reference cohorts; this may be due to an under-represented genetic ancestry group.

Uracil (U): pyrimidine base in RNA that takes the place of thymine in DNA, also forming a pair with adenine.

Wild-type: the common, or standard, allele/genotype/phenotype in a population.

Yeast artificial chromosome (YAC): a cloning vector able to carry large (e.g. 1 megabase) inserts of DNA and replicate in yeast cells.

Zygote: an egg cell that has been fertilized by a sperm cell.

Abbreviations and Acronyms

ABC	Approximate Bayesian Computation
aDNA	Ancient DNA
AIC	Akaike's Information Criterion
ARG	Ancestral Recombination Graph
BIC	Bayesian Information Criterion
BLUP	Best Linear Unbiased Predictor
BMI	Body Mass Index
bp	Base Pairs
cDNA	Complementary DNA
CHD	Coronary Heart Disease
ChIP	Chromatin Immunoprecipitation
CI	Confidence Interval
DAG	Directed Acyclic Graph
DCA	Direct Coupling Analysis
df	Degrees of Freedom
DNA	Deoxyribonucleic Acid
EHH	Extended Haplotype Homozygosity
EM	Expectation Maximization
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
GBLUP	Genomic Best Linear Unbiased Prediction
GC	Gas Chromatography
GC	Guanine and Cytosine
GLM	Generalized Linear Model
GTR	General Time Reversible
GWAS	Genome-Wide Association Study
HMM	Hidden Markov Model
HPD	Highest Probability Density
HWE	Hardy–Weinberg Equilibrium
IBD	Identical by Descent
IBS	Identical by State
kb	kilobase
KDEs	Kernel Density Estimators
kya	Thousand Years Ago
LD	Linkage Disequilibrium
MAF	Minor Allele Fraction
MAP	Maximum <i>A Posteriori</i>

MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MH	Metropolis–Hastings
MHC	Major Histocompatibility Complex
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate/Estimator
MR	Mendelian Randomization
MRCA	Most Recent Common Ancestor
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial DNA
NGS	Next-Generation Sequencing
NMR	Nuclear Magnetic Resonance
ODE	Ordinary Differential Equation
OLS	Ordinary Least Squares
OR	Odds Ratio
OU	Ornstein–Uhlenbeck
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PLS	Partial Least Squares
PPI	Protein–Protein Interaction
PRS	Polygenic Risk Score
QTLs	Quantitative Trait Loci
RCT	Randomized Controlled Trial
SFS	Site Frequency Spectrum
SMC	Sequentially Markov Coalescent
SNP	Single Nucleotide Polymorphism
SVD	Singular Value Decomposition
TMRCA	Time to the Most Recent Common Ancestor
WTCCC	Wellcome Trust Case Control Consortium

1

Statistical Modeling and Inference in Genetics

Daniel Wegmann and Christoph Leuenberger

University of Fribourg, Switzerland

Abstract

Given the long mathematical history and tradition in genetics, and particularly in population genetics, it is not surprising that model-based statistical inference has always been an integral part of statistical genetics, and vice versa. Since the big data revolution due to novel sequencing technologies, statistical genetics has further relied heavily on numerical methods for inference. In this chapter we give a brief overview over the foundations of statistical inference, including both the frequentist and Bayesian schools, and introduce analytical and numerical methods commonly applied in statistical genetics. A particular focus is put on recent approximate techniques that now play an important role in several fields of statistical genetics. We conclude by applying several of the algorithms introduced to hidden Markov models, which have been used very successfully to model processes along chromosomes. Throughout we strive for the impossible task of making the material accessible to readers with limited statistical background, while hoping that it will also constitute a worthy refresher for more advanced readers. Readers who already have a solid statistical background may safely skip the first introductory part and jump directly to Section 1.2.3.

1.1 Statistical Models and Inference

Statistical inference offers a formal approach to characterizing a random phenomenon using observations, either by providing a description of a past phenomenon, or by giving some predictions about future phenomena of a similar nature. This is typically done by estimating a vector of parameters θ from a vector of observations or data D , using the formal framework and laws of probability. The interpretation of probabilities is a somewhat contentious issue, with multiple competing interpretations. Specifically, probabilities can be seen as the frequencies with which specific events occur in a repeatable experiment (the *frequentist* interpretation; Lehmann and Casella, 2006), or as reflecting the uncertainty or *degree of belief* about the state of a random variable (the *Bayesian* interpretation; Robert, 2007). In frequentist statistics, only D is thus considered as a random variable, while in Bayesian statistics both D and θ are considered random variables.

The goal of this chapter is not, however, to enter any debate about the validity of the two competing schools of thought. Instead, our aim is to introduce the most commonly used inference methods of both schools. Indeed, most researchers in statistical genetics, including ourselves, choose their approaches pragmatically based on computational considerations rather than

strong philosophical grounds. Yet, the two schools differ slightly in their language. To keep the introduction succinct and consistent, we introduce the basic concepts of statistical modeling first from the Bayesian point of view. The main differences with respect to the frequentist view are then discussed below.

1.1.1 Statistical Models

1.1.1.1 Independence Assumptions

The first step in statistical inference is to specify a statistical model, which consists of identifying all relevant variables \mathcal{D}, θ and formulating the joint probability distribution $\mathbb{P}(\mathcal{D}, \theta)$ of their interaction, usually under some simplifying assumptions.¹ It is hard to overestimate the importance of this step: ignoring a variable makes the strong assumption that this variable is independent or conditionally independent of all variables considered. By focusing on a summary statistic or subset $T(\mathcal{D})$ of the data \mathcal{D} , for instance, it is implied that $T(\mathcal{D})$ contains all information about θ present in \mathcal{D} . Similarly, all variables not included in θ are assumed to be independent of \mathcal{D} conditioned on θ . A third type of assumption that is often made is to consider specific variables to be conditionally independent of each other. That is particularly relevant in *hierarchical models* where the probability distribution of one parameter is dependent on the values of other hierarchical parameters.

Example 1.1 (Allele frequencies). We strive to illustrate all concepts in this chapter through a limited number of compelling scenarios that we revisit frequently. One of these is the problem of inferring the frequency f of the derived allele at a bi-allelic locus from DNA sequence data. While f may denote the frequency of either of the two alleles, we will assume here, without loss of generality, that the two alleles can be polarized into the ancestral and derived allele, where the latter arose from the former through a mutation. Consider now DNA sequence data $\mathbf{d} = \{d_1, \dots, d_n\}$ obtained for n diploid individuals with sequencing errors at rate ϵ . Obviously, f could easily be calculated if all genotypes were known. However, using a statistical model that properly accounts for genotyping uncertainty, a hierarchical parameter such as f can be estimated from much less data (and hence sequencing depth) than would be necessary to accurately infer all n genotypes.

An appropriate statistical model with parameters $\theta = \{f, \epsilon\}$ and data $\mathcal{D} = \mathbf{d}$ might look as follows:

$$\mathbb{P}(\mathbf{d}, f, \epsilon) = \mathbb{P}(\mathbf{d}|f, \epsilon)\mathbb{P}(f, \epsilon) = \left[\prod_{i=1}^n \sum_{g_i} \mathbb{P}(d_i|g_i, \epsilon) \mathbb{P}(g_i|f) \right] \mathbb{P}(f, \epsilon). \quad (1.1)$$

Here, the sum runs over all possible values of the unknown genotypes g_i . ■

The model introduced in Example 1.1 makes the strong assumptions that the only relevant variables are the sequencing data \mathbf{d} , the unknown genotypes $\mathbf{g} = \{g_1, \dots, g_n\}$, the sequencing error rate ϵ and the allele frequency f . In addition, the model makes the conditional independence assumptions $\mathbb{P}(d_i|g_i, \epsilon, f, \mathbf{d}_{-i}) = \mathbb{P}(d_i|g_i, \epsilon)$ that the sequencing data d_i obtained for individual i is independent of f and the sequencing data of all other individuals \mathbf{d}_{-i} when conditioning on a particular genotype g_i .

Variables may also become conditionally dependent, as do, for instance, f and ϵ once specific data is considered in the above model. Undeniably, the data \mathbf{d} constrains ϵ and f : observing around 5% of derived alleles, for instance, is only compatible with $f = 0$ if $\epsilon \approx 0.05$, but not

¹ To keep the notation simple, we will denote by $\mathbb{P}(\cdot)$ the probability of both discrete and continuous variables. Also, we will typically assume the continuous case when describing general concepts and thus use integrals instead of sums.

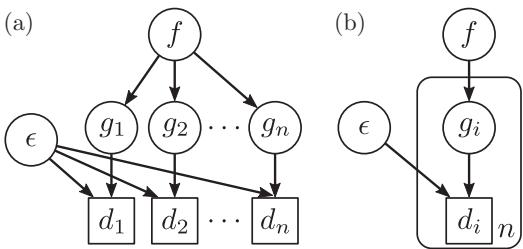


Figure 1.1 (a) Directed acyclic graph (DAG) representing the independence assumptions of Example 1.1 as given in equation (1.1). Observed data is shown as squares; unknown variables as circles. (b) The same DAG in *plate* notation, where a plate replicates the inside quantities as many times as specified in the plate (here n times).

with a much lower ϵ . This also highlights that statistical dependence in a model never implies causality in nature. Indeed, the allele frequency does not causally affect the error rate of the sequencing machine, yet in the model the two variables f and ϵ are dependent as they are connected through the data \mathbf{d} . Importantly, therefore, a statistical model is not a statement about causality, but only about (conditional) independence assumptions. An excellent discussion on this is given in Barber (2012, Ch. 2).

It is often helpful to illustrate the specific independence assumptions of a statistical model graphically using a so-called *directed acyclic graph* (DAG; Barber, 2012; Koller and Friedman, 2009). In a DAG, each variable x_i is a node, and any variable x_j from which a directed edge points to x_i is considered a parental variable of x_i . A DAG for the model of Example 1.1 is given in Figure 1.1, from which the independence assumptions of the model are easily read: (1) Each variable in the DAG is assumed to be independent of any variable not included in the DAG. (2) Each variable is assumed not to be independent of its parental variables. In our case, for instance, we assume that the data d_i of individual i is not independent of the genotype g_i , nor the sequencing error rate ϵ . (3) Each pair of variables a, b connected as $a \rightarrow x \rightarrow b$ or $a \leftarrow x \rightarrow b$ is independent when conditioning on x . In our example, all d_i are independent of f and all other $d_j, j \neq i$, when conditioning on g_i and ϵ . (4) If variables a, b are connected as $a \rightarrow x \leftarrow b$, x is called a *collider*; conditioning on it, a and b become dependent. In our example, ϵ and g_i are thus not independent as soon as specific data d_i is considered. The same holds for ϵ and f , unless we additionally condition on all g_i .

Let us recall at this point that a frequentist would discuss the above concepts with a slightly different vocabulary.

1.1.1.2 Probability Distributions

Once independence assumptions are set, explicit assumptions on the probability distributions have to be made. We note that this is not a requirement for so-called *nonparametric* statistical approaches. However, we will not consider these here because most nonparametric approaches are either restricted to hypothesis testing or only justified when sample sizes are very large, while many problems in genetics have to be solved with limited data. Instead, we will focus on *parametric* modeling and assume that the observations \mathcal{D} were generated from parameterized probability distributions $\mathbb{P}(\mathcal{D}|\theta)$ with unknown parameters θ , but known function \mathbb{P} , which thus need to be specified.

Example 1.2 (Allele frequency). For the model of Example 1.1 given in equation (1.1), two probability functions have to be specified: $\mathbb{P}(d_i|g_i, \epsilon)$ and $\mathbb{P}(g_i|f)$. For the latter, we might be willing to assume that genotypes are in Hardy–Weinberg equilibrium (Hardy, 1908; Weinberg, 1908), such that

$$\mathbb{P}(g_i|f) = \binom{2}{g} f^g (1-f)^{2-g} = \begin{cases} (1-f)^2 & \text{if } g = 0, \\ 2f(1-f) & \text{if } g = 1, \\ f^2 & \text{if } g = 2, \end{cases} \quad (1.2)$$

where g denotes the number of one of the two alleles (e.g. the derived allele).

For the former, usually referred to as *genotype likelihoods* in the genetics literature, we will adopt here a very simple model originally proposed for GATK (Li, 2011):

$$\mathbb{P}(d_i|g, \epsilon) = \prod_{j=1}^{n_i} \mathbb{P}(b_{ij}|g, \epsilon), \quad \mathbb{P}(b_{ij}|g, \epsilon) = \begin{cases} 1 - \epsilon & \text{if } g = 0, b_{ij} = A \text{ or } g = 2, b_{ij} = D, \\ \frac{1}{2} & \text{if } g = 1, \\ \epsilon & \text{if } g = 0, b_{ij} = D \text{ or } g = 2, b_{ij} = A. \end{cases} \quad (1.3)$$

Here, $b_{ij} = A, D$ denotes the bases (alleles) observed in the j th read among the n_i reads covering this site in individual i , under the assumption that only ancestral (A) or derived (D) alleles may be observed. As an example, consider data $d_1 = \{A, D, A\}$. This data is compatible with all three possible genotypes, yet with different likelihoods. If $g_1 = 0$, the allele D must have been the result of a sequencing error, and we get $\mathbb{P}(d_1|g_1 = 0) = (1 - \epsilon)^2\epsilon$. Similarly, $\mathbb{P}(d_1|g_1 = 1) = (1/2)^3$ and $\mathbb{P}(d_1|g_1 = 2) = (1 - \epsilon)\epsilon^2$.

To keep matters simple we assume a single error rate for all bases, although error rates are often considered specific to each base b_{ij} . Genotype likelihood models that more accurately reflect the errors present in the data may increase statistical power. For instance, a genotype likelihood model accounting for post-mortem damage was shown to outperform simpler models for the analysis of ancient DNA (Hofmanová et al., 2016; Kousathanas et al., 2017). However, all calculations presented in this chapter are equally valid under other genotype likelihood models. ■

In order to fully specify equation (1.1), the remaining probability function $\mathbb{P}(f, \epsilon)$ also has to be specified. But as we will outline below, this is only necessary in Bayesian inference, in which case $\mathbb{P}(f, \epsilon)$ is called a *prior distribution*.

1.1.2 Inference Methods and Algorithms

Once the model is specified, the second step in statistical inference is to choose a method of inference, as well as an algorithm to obtain estimates. Many methods for inference have been developed (Lehmann and Romano, 2006), but we will restrict ourselves to two of the most commonly used: the frequentist *maximum likelihood inference* and *Bayesian inference*. Both of these methods are justified by the *likelihood principle*, which states that in parametric statistics, all evidence in D relevant for θ is entirely contained in the *likelihood function* or *likelihood* $\mathcal{L}(\theta) = \mathbb{P}(D|\theta)$ (Davison, 2003, Ch. 11). But as mentioned above, the two methods differ on philosophical grounds. In the frequentist interpretation, θ is not considered as a random variable, and only the likelihood $\mathbb{P}(D|\theta)$ is used for inference. In the Bayesian interpretation, in contrast, the full joint probability $\mathbb{P}(D, \theta) = \mathbb{P}(D|\theta)\mathbb{P}(\theta)$ is considered. This has practical implications for inference, as we will outline in the following sections.

1.2 Maximum Likelihood Inference

The most widely used frequentist inference scheme is maximum likelihood (ML), which was systematically analyzed by Ronald Fisher in his seminal 1922 paper (Fisher, 1922), although it was applied earlier by Gauss and others (see Stigler, 1986). Given some data D following the likelihood function (or simply ‘likelihood’) $\mathcal{L}(\theta) = \mathbb{P}(D|\theta)$, the maximum likelihood approach considers the *maximum likelihood estimator* (often abbreviated as ML estimator or MLE)

$$\hat{\theta} = \arg \max_{\theta} \mathbb{P}(D|\theta).$$

In words, the ML point estimate is the vector $\hat{\theta}$ that maximizes the probability $\mathbb{P}(D|\theta)$ of observing the data.

Example 1.3 (Genotype calling). To illustrate the approach for discrete parameters, consider a statistical model with the likelihood function $\mathbb{P}(d|g)$ of some sequencing data d obtained from a locus at which an individual carries the genotype $g = \text{AA}, \text{AC}, \dots, \text{TT}$. The ML estimate \hat{g} now simply corresponds to the genotype for which $\mathbb{P}(d|g)$ is the highest. Revisiting the data from Example 1.2, the MLE estimate would correspond to $g_1 = 1$ for any error rate $\epsilon < 0.19$. But unless ϵ was extremely low, the difference between $\mathbb{P}(d_1|g_1 = 1)$ and $\mathbb{P}(d_1|g_1 = 0)$ is very small, illustrating that a sequencing depth of 3 is usually not sufficient to accurately estimate or ‘call’ genotypes. ■

For simple models, the maximum of the likelihood function $\mathcal{L}(\theta)$ can be found analytically by setting the *gradient* or *score function* $\nabla \mathcal{L}(\theta)$ equal to zero and solving for θ . In practice this implies taking the partial derivatives of $\mathcal{L}(\theta)$ with respect to each parameter, and then solving the system of equations that results from setting these derivatives equal to zero. This is normally much simplified by optimizing the log-likelihood function $\ell(\theta) = \log \mathcal{L}(\theta)$ instead, as it enables dealing with sums across replicates instead of products. Considering the log-likelihood is justified because the logarithm is a monotonous transformation, that is, $x_1 > x_2$ implies $\log x_1 > \log x_2$, and hence the maximum is at the same location.

Example 1.4 (Pool-seq). A cost-efficient method to estimate population allele frequencies is to sequence pools of DNA extracted from multiple individuals of one population (Pool-seq). For large pools (e.g. large samples from bacterial populations), it is fair to assume that each of the n obtained sequencing reads was derived from a different individual, and hence

$$\mathcal{L}(f) = \mathbb{P}(d|f, \epsilon) = \prod_{i=1}^n \mathbb{P}(d_i|f, \epsilon), \quad \text{where } \mathbb{P}(d_i|f, \epsilon) = \begin{cases} (1-f)(1-\epsilon) + \epsilon f & \text{if } d_i = 0, \\ f(1-\epsilon) + (1-f)\epsilon & \text{if } d_i = 1. \end{cases}$$

Here, f is the derived allele frequency and ϵ the sequencing error rate of the machine. Note that f and ϵ cannot both be estimated, as their effects are indistinguishable (i.e. $\mathbb{P}(d|f, \epsilon) = \mathbb{P}(d|1-f, 1-\epsilon)$), and we will hence assume that ϵ is known. We can then write the likelihood in the form

$$\mathcal{L}(f) = \prod_{i=1}^n \mathbb{P}(0|f, \epsilon)^{1-d_i} (1 - \mathbb{P}(0|f, \epsilon))^{d_i}$$

and hence we obtain the log-likelihood

$$\ell(f) = \sum_{i=1}^n (1 - d_i) \log \mathbb{P}(0|f, \epsilon) + d_i \log(1 - \mathbb{P}(0|f, \epsilon)).$$

To find the value that maximizes $\ell(f)$ (and thus to find the MLE of f) we have to take the derivative with respect to f and solve the equation

$$\frac{d}{df} \ell(f) = \sum_{i=1}^n \left[(1 - d_i) \frac{2\epsilon - 1}{\mathbb{P}(0|f, \epsilon)} - d_i \frac{2\epsilon - 1}{1 - \mathbb{P}(0|f, \epsilon)} \right] = 0. \quad (1.4)$$

This is readily solved for $\mathbb{P}(0|f, \epsilon) = 1 - \frac{1}{n} \sum_i d_i$ (hardly surprising), from which we get the MLE by plugging this solution back into equation (1.4) as

$$\hat{f} = \frac{1}{1 - 2\epsilon} \left(\frac{1}{n} \sum_i d_i - \epsilon \right). \quad (1.5) \quad \blacksquare$$

Note that, using this technique, multiple solutions may be found if the likelihood function is multi-peaked, in which case the true MLE must be distinguished from local maxima

numerically. Note further that ML estimation does not always yield reasonable results (Pawitan, 2001, Section 5.1). In the above example, for instance, \hat{f} will turn out to be negative if all sequence reads contain the ancestral base (i.e. all $d_i = 0$)!

Sometimes the maximum of the likelihood function has to be found under some constraints. An important example from genetics is the inference of genotype frequencies π_g , $g = 0, 1, 2$, which obviously must satisfy the constraint $\pi_0 + \pi_1 + \pi_2 = 1$. As we will show in the following example, accounting for constraints is easily done with the *method of Lagrange multipliers*, which is the standard strategy for finding extrema subject to equality constraints (Lange, 2004, Section 1.4).

Example 1.5 (Genotype frequencies). Assume we observe n genotypes $\mathbf{g} = \{g_1, \dots, g_n\}$. Instead of stipulating that they are in Hardy–Weinberg equilibrium as in equation (1.2), we can be less restrictive and assume that the genotypes are generated with probabilities $\mathbb{P}(g|\boldsymbol{\pi}) = \pi_g$, $g = 0, 1, 2$. As mentioned above, the three parameters must satisfy the constraint $\pi_0 + \pi_1 + \pi_2 = 1$, so that this model actually has a two-dimensional parameter space. Now, if $\mathbf{n} = \{n_0, n_1, n_2\}$ are the observed frequencies (counts) of the genotypes, then the likelihood is given by the *multinomial distribution*

$$\mathbb{P}(\mathbf{g}|\boldsymbol{\pi}) = \frac{n!}{n_0!n_1!n_2!} \pi_0^{n_0} \pi_1^{n_1} \pi_2^{n_2}. \quad (1.6)$$

Ignoring the constant factor of factorials, the log-likelihood is

$$\ell(\pi_0, \pi_1, \pi_2) = \sum_{g=0}^2 n_g \log \pi_g.$$

We cannot, however, just go ahead and maximize because we have to take the constraint into account. This is most easily done by introducing a dummy variable λ (the Lagrange multiplier) and writing down the *Lagrangian function*, which is of the form ‘log-likelihood minus λ times constraint’ in our case

$$L(\pi_0, \pi_1, \pi_2, \lambda) := \ell(\pi_0, \pi_1, \pi_2) - \lambda \cdot \left(\sum_{g=0}^2 \pi_g - 1 \right).$$

The maximum under constraints is now found by solving the system of four equations

$$\begin{aligned} \frac{\partial}{\partial \pi_g} L(\pi_0, \pi_1, \pi_2, \lambda) &= 0, \quad g = 0, 1, 2, \\ \frac{\partial}{\partial \lambda} L(\pi_0, \pi_1, \pi_2, \lambda) &= 0. \end{aligned}$$

A straightforward calculation yields the unsurprising result

$$\hat{\pi}_g = \frac{n_g}{n}, \quad g = 0, 1, 2.$$

■

1.2.1 Properties of Maximum Likelihood Estimators

Maximum likelihood estimators have the agreeable *invariance property* with respect to parameter transformations. Suppose we rewrite the likelihood with respect to some new parameter $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ where the transformation is supposed to be one-to-one. Then the ML estimator of $\boldsymbol{\psi}$ is obtained by simply plugging in the ML estimator of $\boldsymbol{\theta}$: $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}})$. Under mild regularity conditions, ML estimators also have some very neat asymptotic properties (Lehmann and

Casella, 2006; Pawitan, 2001; Held and Sabanés Bové, 2014): First, they are asymptotically *consistent*, which refers to the property that ML estimators converge to the true parameter values $\hat{\theta} \rightarrow \theta^*$ as the amount of data grows to infinity (e.g. sequencing depth increases to infinity in Example 1.4). Second, they are asymptotically normally distributed around the true parameter $\hat{\theta} \sim \mathcal{N}(\theta^*, \mathbf{I}(\theta^*)^{-1})$, with the variance given by the inverse of the *Fisher information* $\mathbf{I}(\theta)$ (see below). Third, they are asymptotically *efficient*, which refers to the property that as data grows to infinity, no other consistent estimator exists having a lower mean squared error $\mathbb{E}[|\hat{\theta} - \theta^*|^2]$. However, these are asymptotic properties that may not hold in applications with finite sample sizes. Indeed, estimators from finite data cannot in general be simultaneously unbiased and have minimal mean squared error, and ML estimators are no exception: while they have low squared error, they tend to be biased. An estimator $\hat{\theta}$ is called *unbiased* if $\mathbb{E}[\hat{\theta}] = \theta$ for all parameter values. Since ML estimators are consistent, they are also asymptotically unbiased, but they may only slowly converge to the true parameter with increasing data.

Example 1.6 (Normal distribution). To illustrate this, let us derive ML estimators for the mean μ and variance σ^2 of a normal distribution from n independently drawn samples $\mathbf{x} = \{x_1, \dots, x_n\}$. The likelihood and log-likelihood functions of this model are given by

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &= \mathbb{P}(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right), \\ \ell(\mu, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,\end{aligned}\tag{1.7}$$

from which we get the pair of equations

$$\begin{aligned}\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0.\end{aligned}$$

These can readily be solved as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i; \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2.\tag{1.8}$$

Applying the invariance property to the transformation $(\mu, \sigma^2) \mapsto (\mu, \sigma)$ yields the MLE for the standard deviation

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}.$$

However, and as mentioned above, MLEs are not, in general, unbiased. In case of the MLE of the variance $\hat{\sigma}$, for instance, we get

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2.$$

But as this example illustrates, MLEs are consistent and hence *asymptotically unbiased*, that is,

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}] = \theta.$$

One can get an unbiased estimator S^2 for σ^2 by setting $S^2 = \frac{n}{n-1}\hat{\sigma}^2$. The unbiasedness property, however, will not survive a nonlinear transformation. In our example, we infer from Jensen's inequality (see, for example, Venkatesh, 2013, XIV.3) that

$$\mathbb{E}(S) = \mathbb{E}[\sqrt{S^2}] < \sqrt{\mathbb{E}[S^2]} = \sigma.$$

■

Usually, estimators with minimal squared errors are preferred over unbiased ones. And as the following example illustrates, unbiased estimators need not even exist for a given model.

Example 1.7 (Allele frequency). Consider the model for genotypes under the Hardy–Weinberg equilibrium given in Example 1.2, equation (1.2). Suppose $T(g)$ is an estimator for the odds $\theta = f/(1-f)$ from observed genotypes g . The expected estimate T given a specific allele frequency f is given by

$$\begin{aligned}\mathbb{E}[T] &= \sum_{g=0}^2 T(g)\mathbb{P}(g|f) = \sum_{g=0}^2 T(g) \binom{2}{g} f^g (1-f)^{2-g} \\ &= T(0) + 2(T(1) - T(0))f + (T(0) + T(2) - 2T(1))f^2,\end{aligned}$$

which is a quadratic polynomial in f . However, using the well-known formula for geometric series, we have $\theta = f/(1-f) = f + f^2 + f^3 + \dots$, which is not a quadratic polynomial. There is thus no estimator T such that $\mathbb{E}[T] = \theta$ for all parameter values. ■

1.2.2 Quantifying Confidence: the Fisher Information Matrix

For correct interpretation, it is crucial to quantify the confidence associated with any estimate. For ML estimators this is usually done using the Fisher information, as we will describe here.

Let $\ell(\theta) = \log \mathbb{P}(\mathcal{D}|\theta)$ be the log-likelihood and $\hat{\theta}$ the corresponding ML estimator. Suppose $\theta = \{\theta_1, \dots, \theta_p\}$ is the unknown true parameter vector. By supposing that the log-likelihood function is sufficiently smooth, we get the following approximation using a Taylor series up to second order:

$$\begin{aligned}\ell(\theta) &\approx \ell(\hat{\theta}) + \sum_{k=1}^p \frac{\partial}{\partial \theta_k} \ell(\hat{\theta})(\theta_k - \hat{\theta}_k) + \frac{1}{2} \sum_{k,l=1}^p \frac{\partial^2}{\partial \theta_k \partial \theta_l} \ell(\hat{\theta})(\theta_k - \hat{\theta}_k)(\theta_l - \hat{\theta}_l) \\ &= \ell(\hat{\theta}) + \frac{1}{2} \sum_{k,l=1}^p \frac{\partial^2}{\partial \theta_k \partial \theta_l} \ell(\hat{\theta})(\theta_k - \hat{\theta}_k)(\theta_l - \hat{\theta}_l),\end{aligned}\tag{1.9}$$

where we used the fact that the partial derivatives of $\ell(\theta)$ vanish at the maximum $\hat{\theta}$. We define the *observed Fisher information matrix* $\mathbf{I}(\hat{\theta})$ to be the symmetric $p \times p$ matrix whose entries are

$$[\mathbf{I}(\hat{\theta})]_{kl} = -\frac{\partial^2}{\partial \theta_k \partial \theta_l} \ell(\hat{\theta}).$$

Using this definition, we can write the approximation (1.9) in matrix notation as

$$\log \frac{\mathcal{L}(\theta)}{\mathcal{L}(\hat{\theta})} \approx -\frac{1}{2}(\theta - \hat{\theta})^T \mathbf{I}(\hat{\theta})(\theta - \hat{\theta}).$$

From the distribution theory of the ML estimator (see, for example, Held and Sabanés Bové, 2014, 4.2.4) it follows that the so-called *Wald statistic* $(\hat{\theta}_k - \theta_k)/\text{se}(\hat{\theta}_k)$ is approximately

standard normally distributed. Here the standard error $\text{se}(\hat{\theta}_k)$ is the square root of the k th diagonal element of the inverted observed Fisher information matrix:

$$\text{se}(\hat{\theta}_k) = ([\mathbf{I}^{-1}(\hat{\theta})]_{kk})^{\frac{1}{2}}.$$

This allows a confidence interval to be estimated that contains the true parameter θ_k with probability $1 - \alpha$:

$$\hat{\theta}_k \pm z_{1-\alpha/2} \text{se}(\hat{\theta}_k). \quad (1.10)$$

Example 1.8 (Normal distribution). In our normal example above, we get for the log-likelihood equation (1.7) the Fisher information matrix

$$\begin{aligned} \mathbf{I}(\mu, \sigma^2) &= - \begin{pmatrix} \frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) & \frac{\partial^2}{\partial \sigma^2 \partial \mu} \ell(\mu, \sigma^2) \\ \frac{\partial^2}{\partial \mu \partial \sigma^2} \ell(\mu, \sigma^2) & \frac{\partial^2}{\partial (\sigma^2)^2} \ell(\mu, \sigma^2) \end{pmatrix} \\ &= -\frac{1}{\sigma^4} \begin{pmatrix} n\sigma^2 & \sum_{i=1}^n (x_i - \mu) \\ \sum_{i=1}^n (x_i - \mu) & \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \end{pmatrix}. \end{aligned}$$

Inserting the ML estimates (1.8), we obtain the observed Fisher information matrix

$$\mathbf{I}(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{(\hat{\sigma}^2)^2} \begin{pmatrix} n\hat{\sigma}^2 & 0 \\ 0 & \frac{n}{2} \end{pmatrix},$$

which in this case turns out to be diagonal. Thus its inverse is obtained by simply inverting the diagonal elements. This immediately yields the Wald confidence intervals

$$\hat{\mu} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{n}}, \quad \hat{\sigma}^2 \pm z_{1-\alpha/2} \sqrt{\frac{2}{n}} \hat{\sigma}^2. \quad \blacksquare$$

The distribution theory of MLEs relies heavily on the notion of the expected Fisher information matrix (see Lehmann and Casella, 2006; Pawitan, 2001; Casella and Berger, 2002).

1.2.3 Newton's Method

As described above, the MLE may be obtained by setting the gradient $\nabla \ell(\theta)$ equal to zero and solving for θ . While this works well for sufficiently regular log-likelihood functions, more often than not the resulting equation allows for no analytical solution and one must have recourse to numerical schemes. As far as speed of convergence is concerned, the gold standard of root-finding algorithms is *Newton's algorithm* (Nocedal and Wright, 1999; Press, 2007; Lange, 2010). We start with an initial guess θ_0 and iteratively improve the current value by linearly projecting the location at which $\nabla \ell(\theta) = 0$ using the second derivative of the log-likelihood (Figure 1.2).

Formally, the idea is to consider a second-order Taylor expansion around the current point θ_k similarly to approximation (1.9):

$$\ell(\theta) \approx \ell(\theta_k) + \nabla \ell(\theta_k)(\theta - \theta_k) - \frac{1}{2}(\theta - \theta_k)^t \mathbf{I}(\theta_k)(\theta - \theta_k).$$

We maximize the right-hand side of this approximation by setting its gradient

$$\nabla \ell(\theta) \approx \nabla \ell(\theta_k) - \mathbf{I}(\theta_k)(\theta - \theta_k)$$

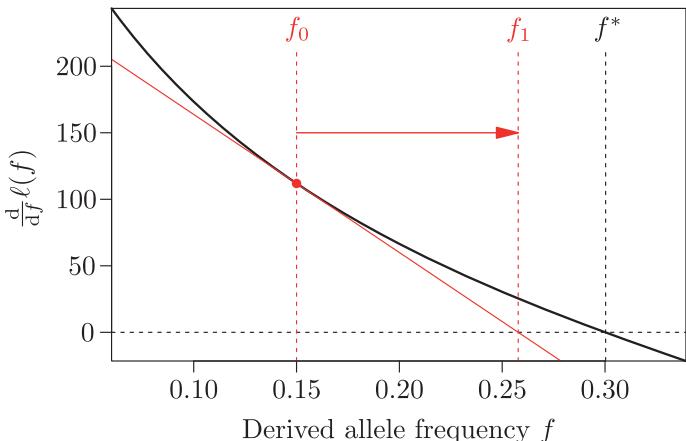


Figure 1.2 Illustration of a Newton update for the model discussed in Example 1.9 for 200 reads simulated with $f = 0.3$ and all $\epsilon_i = 0.1$. The first derivative of $\ell(f)$ (black line) indicates the MLE $f^* = 0.301$. Using the second derivative of $\ell(f)$ (red line), Newton's method updates an initial guess $f_0 = 0.15$ to $f_1 = 0.258$.

equal to zero and solving for θ . This way we obtain the next iterate,

$$\theta_{k+1} = \theta_k + \mathbf{I}(\theta_k)^{-1} \nabla \ell(\theta_k). \quad (1.11)$$

More generally, Newton's method can be used to find zeros of some differentiable function $\mathbf{F} : \mathbb{R}^p \rightarrow \mathbb{R}^p$, that is, to find the values θ^* at which $\mathbf{F}(\theta^*) = \mathbf{0}$. This leads to solving the system of equations

$$F_i(\theta_1, \dots, \theta_p) = 0, \quad i = 1, \dots, p.$$

The corresponding iteration scheme is

$$\theta_{k+1} = \theta_k - \mathbf{J}(\theta_k)^{-1} \mathbf{F}(\theta_k),$$

where \mathbf{J} denotes the Jacobian matrix $[\mathbf{J}(\theta)]_{ij} = \frac{\partial F_i}{\partial \theta_j}(\theta)$. Newton's algorithm, if it converges to the zero θ^* of \mathbf{F} , typically converges very fast: if θ_k is close enough to θ^* one has

$$\|\theta_{k+1} - \theta^*\| \leq M \|\theta_k - \theta^*\|^2$$

for a suitable constant M , which depends on the second derivatives of \mathbf{F} (this is called *quadratic convergence*). For the exact formulation of this convergence result and the necessary conditions on \mathbf{F} , see Nocedal and Wright (1999, Ch. 11).

Newton's method tends to be very unstable in practice because the iteration step easily overshoots the zero location if the initial guess is not close to the root. One way to tame the Newton step is to backtrack it along its direction until the step is acceptable (Press, 2007, Ch. 9.6).

As an optimization scheme like in equation (1.11), Newton's method needs the Fisher information matrix, that is, the second derivatives of the log-likelihood. This is in many cases too expensive to calculate. So-called *quasi-Newton methods* try to approximate the second-order matrix by using information gleaned from the gradient vector in each step. The most popular method in this family is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970); see Murphy (2012, Ch. 8.3.5).

Example 1.9 (Pool-seq). Let us revisit Example 1.4 about estimating the allele frequency from pooled sequencing data, but this time assuming that each of the n read bases has its own error rate ϵ_i such that

$$\mathcal{L}(f) = \mathbb{P}(\mathbf{d}|f, \epsilon) = \prod_{i=1}^n \mathbb{P}(d_i|f, \epsilon_i) \quad \text{where} \quad \mathbb{P}(d_i|f, \epsilon_i) = \begin{cases} (1-f)(1-\epsilon_i) + \epsilon_i f & \text{if } d_i = 0, \\ f(1-\epsilon_i) + (1-f)\epsilon_i & \text{if } d_i = 1. \end{cases}$$

We can proceed exactly as in Example 1.4 and arrive at equation (1.4), except that this time the ϵ s carry an index:

$$\frac{d}{df} \ell(f) = \sum_{i=1}^n \left[(1-d_i) \frac{2\epsilon_i - 1}{\mathbb{P}(0|f, \epsilon_i)} - d_i \frac{2\epsilon_i - 1}{1 - \mathbb{P}(0|f, \epsilon_i)} \right] = 0.$$

In contrast to equation (1.4), this equation cannot be solved analytically for f and we must have recourse to a numerical scheme such as Newton's method given in equation (1.11). The Fisher information in this one-parameter situation is simply minus the second derivative

$$I(f) = -\frac{d^2}{df^2}(f) = \sum_{i=1}^n (2\epsilon_i - 1)^2 \left[\frac{1-d_i}{\mathbb{P}(0|f, \epsilon_i)^2} + \frac{d_i}{(1-\mathbb{P}(0|f, \epsilon_i))^2} \right].$$

Convergence to the MLE will occur via a sequence of Newton steps $f_{k+1} = f_k - \frac{d}{df} \ell(f_k) / \frac{d^2}{df^2} \ell(f_k)$. A graphical illustration of a first such update from $f_0 \rightarrow f_1$ is given in Figure 1.2. ■

1.2.4 Latent Variable Problems: the EM Algorithm

Often it is desirable to integrate out the uncertainty of latent or hidden variables when inferring hierarchical parameters. As an example, consider the model introduced in Example 1.1, where the goal was to infer the allele frequency f while integrating out the uncertainty of the unknown and hidden genotypes \mathbf{g} (Figure 1.1).

Let us denote by \mathbf{x} a set of observed data, \mathbf{z} a set of latent data and $\boldsymbol{\theta}$ the parameters of a general latent variable model. The aim then is to maximize the marginal likelihood of the observed data

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} \mathbb{P}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \sum_{\mathbf{z}} \mathcal{L}_c(\boldsymbol{\theta}), \quad (1.12)$$

where $\mathcal{L}_c(\boldsymbol{\theta}) = \mathbb{P}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ is the *complete-data likelihood* in which the latent variable \mathbf{z} is treated as observed data.

The challenge in deriving ML estimates for such models lies in the tedious summation over all possible values of \mathbf{z} , which often prohibits an analytical optimization. While $\mathcal{L}(\boldsymbol{\theta})$ may sometimes be maximized numerically using Newton's or a similar method, there exists a more elegant solution for latent-variable problems, the so-called *Expectation-Maximization* (EM) algorithm originally introduced by Dempster et al. (1977), which we will derive in the following.

Taking the logarithm on both sides of

$$\mathbb{P}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = \mathbb{P}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \cdot \mathbb{P}(\mathbf{x}|\boldsymbol{\theta})$$

and rearranging, we obtain

$$\log \mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log \mathbb{P}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}).$$

Now we take the expectation of this expression over the latent data \mathbf{z} given the observed data \mathbf{x} and some current parameter estimate $\boldsymbol{\theta}'$. Observing that the left-hand side is independent of \mathbf{z} , this yields

$$\log \mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}') + H(\boldsymbol{\theta}|\boldsymbol{\theta}'), \quad (1.13)$$

where we introduced the so-called Q-function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ and the entropy $H(\boldsymbol{\theta}|\boldsymbol{\theta}')$,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') := \mathbb{E}[\log \mathbb{P}(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}'], \quad H(\boldsymbol{\theta}|\boldsymbol{\theta}') := -\mathbb{E}[\log \mathbb{P}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}'].$$

The information inequality (which itself follows from Jensen's inequality) states that $H(\theta|\theta') - H(\theta'|\theta') \geq 0$ (see, for example, Murphy, Thm. 2.8.1; or Lange, Prop. 2.4.2). Subtracting equation (1.13) with the particular parameter value $\theta = \theta'$ from equation (1.13) with general θ , we get

$$\begin{aligned}\log \mathbb{P}(\mathbf{x}|\theta) - \log \mathbb{P}(\mathbf{x}|\theta') &= Q(\theta|\theta') - Q(\theta'|\theta') + H(\theta|\theta') - H(\theta'|\theta') \\ &\geq Q(\theta|\theta') - Q(\theta'|\theta').\end{aligned}$$

In other words, if we find a parameter value θ with $Q(\theta|\theta') > Q(\theta'|\theta')$ then we also have $\log \mathbb{P}(\mathbf{x}|\theta) > \log \mathbb{P}(\mathbf{x}|\theta')$, thus improving the marginal likelihood. A natural choice is the parameter value θ that maximizes $Q(\theta|\theta')$, which motivates the following algorithm that iterates alternatively between an expectation step (E-step) and a maximization step (M-step):

1. **E-step.** Given a current estimate θ' , calculate the corresponding Q-function

$$Q(\theta|\theta') := \mathbb{E}[l_c(\theta)|\mathbf{x}, \theta']$$

where $l_c(\theta) = \log \mathbb{P}(\mathbf{x}, \mathbf{z}|\theta)$ is the complete-data log-likelihood.

2. **M-step.** Determine the parameter θ maximizing the Q-function, and use it as new current estimate in the next E-step.

This algorithm is guaranteed to increase the likelihood in each iteration until a maximum is reached. However, it is worth noting that the convergence to that maximum may sometimes be rather slow in practice. Several schemes for speeding up the EM algorithm have been proposed, and we particularly highlight the general-purpose SQUAREM method (Varadhan and Roland, 2008).

Example 1.10 (Allele frequencies). We illustrate an application of the EM algorithm with the aid of the model to estimate the allele frequencies f and sequencing error rate ϵ from next-generation sequencing data given by equation (1.1) in Example 1.1 and the specific probability functions specified in Example 1.2. We have the likelihood

$$\mathcal{L}(f, \epsilon) = \mathbb{P}(\mathbf{d}|f, \epsilon) = \prod_{i=1}^n \sum_{g=0}^2 \mathbb{P}(d_i|g, \epsilon) \mathbb{P}(g|f).$$

Now, we treat the unobserved genotypes $\mathbf{g} = \{g_1, \dots, g_n\}$ as the latent data called \mathbf{z} in the general theory above. We get the complete-data likelihood

$$\mathcal{L}_c(f, \epsilon) = \mathbb{P}(\mathbf{d}, \mathbf{g}|f, \epsilon) = \prod_{i=1}^n \mathbb{P}(d_i|g_i, \epsilon) \mathbb{P}(g_i|f)$$

and the complete-data log-likelihood

$$\ell_c(f, \epsilon) = \sum_{i=1}^n (\log \mathbb{P}(d_i|g_i, \epsilon) + \log \mathbb{P}(g_i|f)).$$

E-step. Given a current estimate f', ϵ' , we now determine the Q-function,

$$\begin{aligned}Q(f, \epsilon|f', \epsilon') &= \mathbb{E}[\ell_c(f, \epsilon)| \mathbf{d}, f', \epsilon',] \\ &= \sum_{i=1}^n \sum_{g=0}^2 (\log \mathbb{P}(d_i|g, \epsilon) + \log \mathbb{P}(g|f)) \mathbb{P}(g|d_i, f', \epsilon'),\end{aligned}$$

where with aid of Bayes' theorem we have

$$p_{gi} := \mathbb{P}(g|d_i, f', \epsilon') = \frac{\mathbb{P}(d_i|g, \epsilon')\mathbb{P}(g|f')}{\sum_{h=0}^2 \mathbb{P}(d_i|h, \epsilon')\mathbb{P}(h|f')}.$$
 (1.14)

All the conditional probabilities needed to evaluate equation (1.14) are given in Example 1.2.

M-step. The EM algorithm starts to reveal its charm. The Q-function splits into two separate terms for f and ϵ . In order to optimize the Q-function with respect to f we calculate the derivative

$$\begin{aligned} \frac{\partial}{\partial f} Q(f, \epsilon|f', \epsilon') &= \frac{\partial}{\partial f} \sum_{g=0}^2 \log \mathbb{P}(g|f) \sum_{i=1}^n p_{gi} \\ &= \frac{\partial}{\partial f} \sum_{g=0}^2 (g \log f + (2-g) \log(1-f)) \sum_{i=1}^n p_{gi} \\ &= \sum_{g=0}^2 \left(\frac{g}{f} - \frac{2-g}{1-f} \right) \sum_{i=1}^n p_{gi}. \end{aligned}$$

Setting this equal to zero, we can solve for f :

$$f = \frac{\sum_i (p_{1i} + 2p_{2i})}{2 \sum_i (p_{0i} + p_{1i} + p_{2i})} = \frac{1}{2n} \sum_i (p_{1i} + 2p_{2i}).$$

This will be the new value of f' in the next E-step.

Now let us optimize the Q-function for ϵ . Referring to equation (1.3), we classify the n_i reads at site i as follows:

$$n_i = n_{0i}^A + n_{2i}^D + n_{1i}^A + n_{1i}^D + n_{0i}^D + n_{2i}^A.$$

For the derivative of the Q-function with respect to ϵ we obtain

$$\frac{\partial}{\partial \epsilon} Q(f, \epsilon|f', \epsilon') = \sum_{i=1}^n \left[\frac{n_{0i}^D p_{i0} + n_{2i}^A p_{i2}}{\epsilon} - \frac{n_{0i}^A p_{i0} + n_{2i}^D p_{i2}}{1-\epsilon} \right].$$

We set this equal to zero and solve for ϵ :

$$\epsilon = \frac{\sum_i (n_{0i}^D p_{i0} + n_{2i}^A p_{i2})}{\sum_i (n_{0i}^D p_{i0} + n_{2i}^A p_{i2} + n_{0i}^A p_{i0} + n_{2i}^D p_{i2})}.$$

This will be new the value of ϵ' in the next E-step.

In many applications, including this one, the EM algorithm converges to the ML estimate within a few iterations (less than 10). This is illustrated by example runs of this EM algorithm on simulated data shown in Figure 1.3. Interestingly, the likelihood surface of the model presented here has two peaks of equal height, and hence the two ML estimates $\hat{f}_1, \hat{\epsilon}_1$ and $\hat{f}_2 = 1 - \hat{f}_1, \hat{\epsilon}_2 = 1 - \hat{\epsilon}_1$, even if the solution with $\hat{\epsilon} > 0.5$ may not be considered realistic. This nicely illustrates that external or *a priori* information on parameters (such as that the sequencing machine is expected to be correct more often than incorrect) often contributes crucially to a successful inference. While we will discuss below how such information can be naturally included in Bayesian inference, we note that the likelihood could also be maximized under some constraint (e.g. $\epsilon < 0.5$) in a frequentist setting. ■

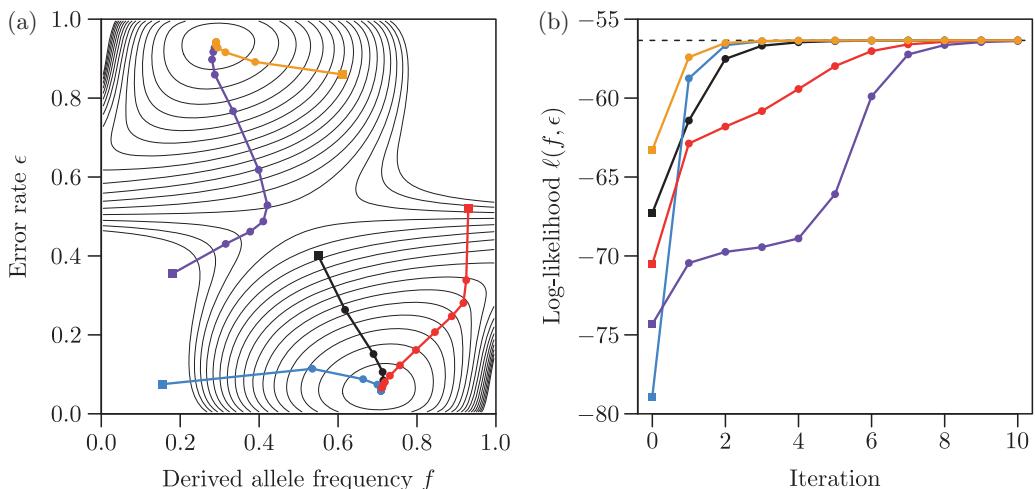


Figure 1.3 (a) Runs of the EM algorithm developed in Example 1.10 on five sequencing reads simulated with sequencing error rate $\epsilon = 0.05$ for each of 20 diploid individuals with genotypes drawn according to Hardy–Weinberg proportions with derived allele frequency $f = 0.7$. Squares and dots indicate the initial and successive parameter estimates, respectively. Contour lines indicate the surface of the likelihood $\ell(f, \epsilon)$. (b) Likelihood values at each iteration for the EM runs shown in (a).

More generally, the example shows that the EM algorithm is only guaranteed to climb a local maximum of the likelihood function, but not to necessarily find the global MLE. This is a direct consequence of the EM algorithm’s property of improving the likelihood in each step, which prohibits it from crossing valleys in the likelihood surface. It is thus advised to run the EM from multiple starting locations if the likelihood surface is multi-peaked, or if the shape of the likelihood surface is unknown. See McLachlan and Krishnan (2007) for an extensive treatment of the EM algorithm and its variants.

1.2.4.1 EM Algorithm with Numerical Optimization

In many applications of the EM algorithm, the maximization in the M-step cannot be done analytically. However, it is usually possible to conduct a numerical optimization, as we will exemplify with the inference of the heterozygosity of an individual. This example is somewhat more substantial than the previous ones but is worth the effort as it beautifully illustrates the EM algorithm and Newton’s root-finding algorithm in interaction.

Example 1.11 (Heterozygosity). The heterozygosity of an individual relates to the fraction of sites at which an individual carries two different alleles. While heterozygosity is readily inferred from accurately called genotypes using a binomial model, uncertainty in genotypes must be accounted for in many applications, for instance by treating the unknown genotypes as latent variables. Here we will use the model proposed by Kousathanas et al. (2017) and implemented in ATLAS (Link et al., 2017). Specifically, let us infer heterozygosity as the rate of substitutions ϑ along the genealogy connecting the two alleles of an individual.

For this, we will adopt the model given by the likelihood

$$\mathcal{L}(\vartheta) = \prod_{l=1}^L \sum_g \mathbb{P}(d_l|g)\mathbb{P}(g|\vartheta),$$

where the sum runs over all possible genotypes $g = AA, AG, \dots, TT$ and $d = \{d_1, \dots, d_L\}$ is the vector of sequencing data obtained for L loci. As in previous examples, the uncertainty

in the genotypes shall be reflected by the genotype likelihoods $\mathbb{P}(d_l|g)$. Finally, let us adopt Felsenstein's substitution model (Felsenstein, 1981), under which the probability of observing a specific genotype $g = rs$ with nucleotides $r, s = A, C, G, T$ is given by

$$\mathbb{P}(g = rs|\vartheta) = \begin{cases} \pi_r(e^{-\vartheta} + \pi_r(1 - e^{-\vartheta})) & \text{if } r = s, \\ \pi_r\pi_s(1 - e^{-\vartheta}) & \text{if } r \neq s. \end{cases}$$

Here, $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ denote the frequencies of the four nucleotides, which we consider to be known to keep derivations simpler.

The complete-data log-likelihood $\ell_c(\vartheta)$, in which the latent variables $\mathbf{g} = \{g_1, \dots, g_L\}$ are considered as known data, is given by

$$\ell_c(\vartheta) = \log \mathbb{P}(\mathbf{d}, \mathbf{g}|\vartheta) = \sum_{l=1}^L [\log \mathbb{P}(d_l|g_l) + \log \mathbb{P}(g_l|\vartheta)].$$

E-step. The expected complete-data log-likelihood is calculated as

$$Q(\vartheta|\vartheta') = \mathbb{E}[l_c(\vartheta) | \mathbf{d}; \vartheta'] = \sum_{l=1}^L \sum_g [\log \mathbb{P}(d_l|g) + \log \mathbb{P}(g|\vartheta)] \mathbb{P}(g|d_l; \vartheta').$$

Only the second part Q_2 of this sum depends on the parameter ϑ . We have

$$Q_2(\vartheta|\vartheta') = \sum_{l=1}^L \sum_g \log \mathbb{P}(g|\vartheta) \mathbb{P}(g|d_l; \vartheta') = \sum_g p_g \log \mathbb{P}(g|\vartheta)$$

where we use the shorthand notation $p_g = \sum_{l=1}^L \mathbb{P}(g|d_l; \vartheta')$. We have by Bayes' theorem

$$p_g = \sum_{l=1}^L \frac{\mathbb{P}(d_l|g)\mathbb{P}(g|\vartheta')}{\sum_h \mathbb{P}(d_l|h)\mathbb{P}(h|\vartheta')} \quad (1.15)$$

Let us write out Q_2 explicitly:

$$Q_2(\vartheta|\vartheta') = \sum_r p_{rr} [\log \pi_r + \log(e^{-\vartheta} + \pi_r(1 - e^{-\vartheta}))] + \sum_r \sum_{s \neq r} p_{rs} [\log \pi_r + \log \pi_s + \log(1 - e^{-\vartheta})].$$

M-step. We have to maximize Q_2 with respect to ϑ . We obtain the derivative

$$\frac{d}{d\vartheta} Q_2(\vartheta|\vartheta') = -e^{-\vartheta} \sum_r \frac{p_{rr}(1 - \pi_r)}{e^{-\vartheta} + \pi_r(1 - e^{-\vartheta})} + \frac{e^{-\vartheta}}{1 - e^{-\vartheta}} \sum_r \sum_{s \neq r} p_{rs}.$$

In order to simplify this expression, we transform to the new variable $\tau = e^{-\vartheta}/(1 - e^{-\vartheta})$. Using the fact that $\sum_g p_g = n$, after a few algebraic manipulations we arrive at

$$\frac{d}{d\vartheta} Q_2(\vartheta|\vartheta') = \tau \cdot \left(n - \sum_r p_{rr} \frac{\tau + 1}{\tau + \pi_r} \right).$$

We have to set this equation to zero and solve for τ . The solution $\tau = 0$ does not correspond to a value of ϑ and thus we have to solve

$$F(\tau) := \left(n - \sum_r p_{rr} \frac{\tau + 1}{\tau + \pi_r} \right) = 0.$$

Since there is no analytical expression for the zero of this function, we will revert to Newton's algorithm. To this end we need the derivative

$$\frac{d}{d\tau} F(\tau) = \sum_r p_{rr} \frac{1 - \pi_r}{\tau + \pi_r}.$$

After a few Newton steps $\tau_{k+1} = \tau_k - F(\tau_k)/\frac{d}{d\tau} F(\tau)$ we will hopefully have converged to a value τ_* satisfying $F(\tau_*) = 0$ and we can then transform back to $\vartheta = -\log(\tau_*/(1 + \tau_*))$. This will be the new value of ϑ' in the next E-step. ■

1.2.5 Approximate Techniques

1.2.5.1 Using Summary Statistics

For some models, the evaluation of the likelihood $\mathcal{L}(\theta) = \mathbb{P}(D|\theta)$ is not possible or is computationally very expensive. In such cases it is sometimes possible to reduce the computational burden by focusing on summary statistics $t = T(D)$ of the full data D and evaluating the approximate likelihood

$$\mathcal{L}(\theta) \approx \hat{\mathcal{L}}(\theta) = \mathbb{P}(t|\theta).$$

The quality of this approximation depends on how much of the information about θ contained in D is also present in t . In the best case the chosen statistic is *sufficient* such that

$$\mathbb{P}(D|t, \theta) = \mathbb{P}(D|t),$$

in which case $\mathbb{P}(D|\theta) \propto \mathbb{P}(t|\theta)$, and hence any inference on θ will be the same when using t or the full data D . If non-sufficient statistics are used, however, there is no such guarantee and the inference will undoubtedly be biased. See Lehmann and Casella (2006, Section 1.6) for a discussion of sufficiency. ■

Example 1.12 (Pool-seq). Let us revisit the MLE of the population allele frequency f from pooled data developed in Example 1.4. As is obvious from equation (1.5), the MLE of f only requires the sum $T(d) = \sum_i d_i$ across each observed base d_i , and hence $T(d)$ is a sufficient statistic for d under this model. ■

1.2.5.2 Monte Carlo Sampling for Intractable Likelihoods

A particular challenge for any inference method are models with *intractable likelihoods*, that is, models for which the likelihood cannot be calculated in reasonable time even for fixed parameters. One typical class of such models are latent variable models where the latent variables are either continuous or take on any of an infinite number of discrete states, and where there is no analytical solution for that integral or sum. In such cases, the integral (or sum) can usually be evaluated numerically using *Monte-Carlo* samples (Robert and Casella, 1999).

Consider a latent variable model of the form

$$\mathcal{L}(\theta) = \mathbb{P}(x|\theta) = \int \mathbb{P}(x|z)\mathbb{P}(z|\theta)dz$$

and suppose that there is no analytical solution for the integral. If samples of the latent variable z_1, \dots, z_k can be drawn from $\mathbb{P}(z|\theta)$ for a fixed value θ' , the integral can be solved numerically as

$$\mathcal{L}(\theta) = \int \mathbb{P}(x|z)\mathbb{P}(z|\theta)dz \approx \hat{\mathcal{L}}_k(\theta) = \frac{1}{k} \sum_{i=1}^k \mathbb{P}(x|z_i), \quad z_i \sim \mathbb{P}(z|\theta).$$

The MLE of θ may then be found numerically, for instance using a grid search. Since such numerical searches rely on the comparisons of likelihood values for different parameter vectors, it is crucial that $\hat{\mathcal{L}}(\theta) < \hat{\mathcal{L}}(\theta')$ for every pair θ, θ' with $\mathcal{L}(\theta) < \mathcal{L}(\theta')$. While $\hat{\mathcal{L}}(\theta) \rightarrow \mathcal{L}(\theta)$ as $k \rightarrow \infty$, a very large number of samples may be required to ensure an acceptable approximation error $|\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)| < \epsilon$. This is particularly true if the variance in $\mathbb{P}(x|z_i)$ is large between different z_i such that the biggest contribution to the integral stems from just a few values z_i .

Example 1.13 (Demographic inference). Perhaps the best-studied example of a model with intractable likelihood in genetics is the problem of inferring demographic parameters θ such as population sizes, split times or migration rates from genetic data D of n individuals. Obviously, the demographic parameters θ do not directly affect the genetic diversity, but shape the evolutionary relationship of the samples characterized by their genealogy \mathcal{G} . The observed genetic diversity is then the result of mutations occurring along the branches of that genealogy, and the likelihood of the model is thus given by (Felsenstein, 1988)

$$\mathcal{L}(\theta, \mu) = \mathbb{P}(D|\theta, \mu) = \int \mathbb{P}(D|\mathcal{G}, \mu) \mathbb{P}(\mathcal{G}|\theta) d\mathcal{G}. \quad (1.16)$$

Here, μ are the parameters of the mutation process and the integral runs across all possible genealogies. There is no general closed-form solution of that integral, not least because the topological space is complex even for small sample sizes (there are, for instance, already more than 10^7 topologies for $n = 10$). Since $\mathbb{P}(D|\mathcal{G}, \mu)$ is readily calculated for any particular genealogy \mathcal{G} , however, the integral in equation (1.16) may be numerically approximated through a large sample of genealogies $\mathcal{G}_i \sim \mathbb{P}(\mathcal{G}|\theta)$, $i = 1, \dots, k$, as

$$\mathcal{L}(\theta, \mu) \approx \hat{\mathcal{L}}_k(\theta, \mu) = \frac{1}{k} \sum_{i=1}^k \mathbb{P}(D|\mathcal{G}_i, \mu).$$

Luckily, sampling genealogies is fast under almost any demographic model thanks to coalescent theory (see Kingman, 1982, and **Chapter 5**). However, the variation in $\mathbb{P}(D|\mathcal{G}_i, \mu)$ may be huge among sampled genealogies even for rather simple models with often just a tiny number of them contributing substantially to the average. Consequently, a very large number of genealogies are often required to accurately approximate the full likelihood $\mathcal{L}(\theta, \mu)$, as these very rare genealogies are otherwise missed.

To demonstrate this, we used `fastsimcoal2` (Excoffier et al., 2013) to simulate data D of a single, fully linked locus of 10 kb with mutation rate $\mu = 10^{-8}$ for 20 haploid samples taken from each of two populations of haploid sizes $N = 10^4$ exchanging migrants at the symmetric rate $m = 10^{-3}$ per generation. We then summarized the data with the allele frequency spectrum $S = T(D)$, which is a highly informative statistic under this model, and generated $k = 5 \cdot 10^8$ genealogies $\mathcal{G}_i \sim \mathbb{P}(\mathcal{G}_i|N, m)$, also with `fastsimcoal2`, to estimate the likelihood for those specific parameters as

$$\mathcal{L}(N, m, \mu) \approx \hat{\mathcal{L}}_k(N, m, \mu) = \frac{1}{k} \sum_{i=1}^k \mathbb{P}(S|\mathcal{G}_i, \mu).$$

As shown in Figure 1.4(a), the variation in $\mathbb{P}(S|\mathcal{G}_i, \mu)$ is huge, with just a tiny number of genealogies resulting in high values. As a consequence, accurate Monte Carlo estimates $\hat{\mathcal{L}}_k(N, m, \mu)$ require an extremely large number of genealogy samples to be accurate, even at the true parameter values (Figure 1.4(b)). ■

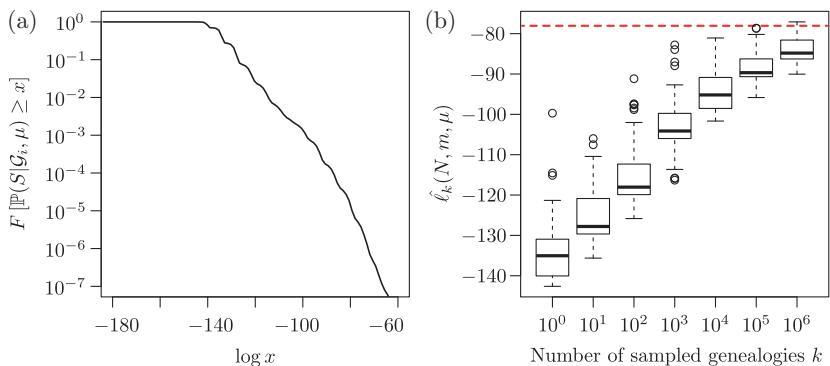


Figure 1.4 (a) Complementary cumulative distribution of $\mathbb{P}(S|G_i, \mu)$, $G_i \sim \mathbb{P}(G_i|N, m)$, for data D of a single locus summarized by the allele frequency spectrum $S = T(D)$ and simulated under a model of two populations exchanging migrants at a symmetric rate, calculated for $5 \cdot 10^8$ genealogies simulated at the true parameters. (b) Distribution of 100 estimates of the log-likelihood $\hat{\ell}_k(N, m, \mu) = \log \hat{L}_k(N, m, \mu)$ from different numbers k of simulated genealogies. The dashed red line indicates a single estimate obtained from $k = 5 \cdot 10^8$ genealogies. (See text for details.)

Importance Sampling The sampling process described above could be massively improved by increasing the probability of generating samples with big contributions. A common technique to achieve this is *importance sampling* (Robert and Casella, 1999, Section 3.3), in which the z_i are samples from an alternative distribution $z_i \sim F$ that preferentially generates samples of high $\mathbb{P}(x|z)$, and to then estimate the integral as

$$\int \mathbb{P}(x|z) \mathbb{P}(z|\theta) dz \approx \frac{1}{k} \sum_{i=1}^k \mathbb{P}(x|z_i) \frac{\mathbb{P}(z|\theta)}{F(z_i)},$$

where the importance weights $\mathbb{P}(z_i|\theta)/F(z_i)$ correct for the oversampling of z_i with high contributions.

1.2.5.3 Composite Likelihood

For many complex hierarchical models it is difficult (or impossible) to evaluate the full likelihood, even with the approximate techniques presented above. Consider, for instance, the vector of data $x = \{x_1, \dots, x_n\}$ and some hierarchical parameters θ . Any factorization of the full likelihood of such a model will involve large number of conditional densities. As an example examine the factorization

$$\mathbb{P}(x|\theta) = \mathbb{P}(x_1|\theta) \mathbb{P}(x_2|x_1, \theta) \mathbb{P}(x_3|x_1, x_2, \theta) \dots \mathbb{P}(x_n|x_{-n}, \theta),$$

where x_{-n} denotes the vector x without its n th element.

If the calculation of these conditional probabilities is computationally challenging, or impossible because the exact nature of the dependence is not specified, one may consider using an approximation of the full likelihood. A popular type of approximation for such cases are *composite likelihoods* (Lindsay, 1988), also called *quasi* or *pseudo likelihoods*, and defined as the product

$$\mathcal{L}_C(\theta) = \prod_{k=1}^K \mathcal{L}_k(\theta)^{w_k}$$

of low-dimensional marginal or conditional likelihood components $L_k(\theta)$ with associated weights w_k . Usually, so-called block composite likelihood components are chosen (Varin et al., 2011), which include bivariate marginal likelihoods

$$\mathcal{L}_{C,pm}(\theta) = \prod_{i=1}^{n-1} \prod_{j=i+1}^n \mathbb{P}(x_i, x_j | \theta)^{w_{ij}},$$

pairwise conditional likelihoods

$$\mathcal{L}_{C,pc}(\theta) = \prod_{k=i}^n \prod_{j \neq i} \mathbb{P}(x_i | x_j, \theta)^{w_{ij}},$$

or simply one-wise independence likelihoods

$$\mathcal{L}_{C,o}(\theta) = \prod_{i=1}^n \mathbb{P}(x_i | \theta)^{w_i}.$$

Inference is then conducted just as with regular likelihoods, for instance through the maximum composite likelihood estimator (MCLE)

$$\hat{\theta}_C = \arg \max_{\theta} \mathcal{L}_C(\theta).$$

What makes composite likelihood approximations so appealing is that MCLEs have well-defined properties. In particular, they are consistent, that is, they converge to the true value $\hat{\theta}_C \rightarrow \theta^*$ as $n \rightarrow \infty$. Also, just like regular MLEs, they are asymptotically normally distributed, but with the variance given by the inverse of the *Godambe information matrix* $\mathbf{G}(\theta)$ (Godambe, 1960), rather than the Fisher information matrix:

$$\hat{\theta}_C \stackrel{\text{d}}{\sim} \mathcal{N}(\theta^*, \mathbf{G}(\theta)^{-1}).$$

Importantly, this allows the calculation of confidence intervals, analogous to expression (1.10).

Example 1.14 (Demographic inference). In population genetics, MCLEs are frequently used as they allow all loci to be treated as independent, while in fact they are often linked. Consider, for instance, the problem of inferring demographic parameters as outlined in Example 1.13, but now with data from L loci $\mathbf{d} = \{d_1, \dots, d_L\}$. Due to linkage, a factorization of the likelihood of the full model would involve complicated terms to reflect the interdependence of loci, the specification of which would require additional parameters. Under a one-wise independence composite likelihood assumption, we get

$$\mathcal{L}_C(\theta, \mu) = \mathbb{P}(\mathbf{d} | \theta, \mu) = \prod_{l=1}^L \int \mathbb{P}(d_l | \mathcal{G}, \mu) \mathbb{P}(\mathcal{G} | \theta) d\mathcal{G},$$

which is much simpler and similar to the one-locus case in equation (1.16). Thanks to the property of consistency, an MCLE based on equation (1.14) will converge to the true parameters (θ^*, μ^*) as $L \rightarrow \infty$, which forms the basis of popular inference methods such as `fastsimcoal2` (Excoffier et al., 2013) or `daadi` (Gutenkunst et al., 2009). ■

1.3 Bayesian Inference

At the heart of statistical inference lies the concept of the statistical model, which describes the uncertainty about how the data was produced. The ultimate aim is to obtain information about the unknown parameter θ given the data D . As discussed above, frequentist statisticians treat the parameter as a fixed but unknown quantity. Bayesian statisticians, on the other hand, prefer to use a fully probabilistic model and to treat the parameter as a random quantity as well. To do so, one has to choose an appropriate *prior distribution* $\mathbb{P}(\theta)$, which reflects the knowledge (i.e. uncertainty) about θ prior to the experiment. The goal is then to update this knowledge (i.e. reduce the uncertainty) given the information contained in the data D . This updated knowledge is encapsulated in the *posterior distribution* $\mathbb{P}(\theta|D)$, which is calculated via Bayes' theorem:

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\mathbb{P}(\theta)}{\mathbb{P}(D)}. \quad (1.17)$$

This theorem was first stated by Thomas Bayes, a Presbyterian minister, in *An Essay toward Solving a Problem in the Doctrine of Chances* (Bayes, 1763), published posthumously. The denominator $\mathbb{P}(D)$ of this fraction, known as the *marginal likelihood*, is a normalization constant ensuring that the posterior is a probability distribution. Indeed, integrating both sides of Bayes' formula with respect to θ and using $\int \mathbb{P}(\theta|D)d\theta = 1$, we immediately obtain

$$\mathbb{P}(D) = \int \mathbb{P}(D|\theta)\mathbb{P}(\theta)d\theta.$$

The marginal likelihood plays an important role in model selection (see Section 1.4), but is otherwise not always essential. Ignoring this constant, the Bayesian paradigm boils down to the slogan *posterior \propto likelihood \times prior*.

The frequentist and Bayesian approaches are often presented as antagonistic and have indeed stirred much philosophical discussion in the past, (see, for example, Savage, 1972; De Finetti, 2017; Jaynes, 2003). A central contentious issue, apart from the interpretation of probabilities, is the necessity to specify a prior distribution in Bayesian statistics, which appears to introduce subjectivity into inference. However, it is also easy to argue that the prior is a very natural way to incorporate existing knowledge. We do not enter that debate here and refer to Diaconis and Skyrms (2017) for an excellent defense of the Bayesian cause. After all, one can take the modest view that the Bayesian model arises naturally by assuming more ingredients for the model, namely a prior distribution on the parameters. Whether or not this is appropriate is a decision for the statistician, who in the end must be wary of *all* model assumptions made, be they frequentist or Bayesian.

Example 1.15 (Allele frequency). Let us extend the allele frequency model under Hardy–Weinberg equilibrium given in equation (1.2) to n observed genotypes $\mathbf{g} = \{g_1, \dots, g_n\}$. Denoting by $\mathbf{n} = \{n_0, n_1, n_2\}$ the frequencies (counts) of the observed genotypes $g = 0, 1, 2$, the likelihood of the allele frequency f is

$$\mathcal{L}(f) = \mathbb{P}(\mathbf{g}|f) \propto f^{n_1+2n_2} (1-f)^{n_1+2n_0}. \quad (1.18)$$

We now have to choose a suitable prior distribution for f . Recall that a probability density belongs to the class of *beta distributions* Beta(α, β) if, for $0 \leq f \leq 1$, it is of the form

$$\mathbb{P}(f) = \frac{1}{B(\alpha, \beta)} \cdot f^{\alpha-1} (1-f)^{\beta-1} \quad (1.19)$$

for two *shape parameters* $\alpha, \beta > 0$, where $B(\alpha, \beta)$ is the *Beta function*. If we choose such a beta prior, it is obvious that the posterior distribution is a beta distribution as well:

$$\mathbb{P}(f|g) \sim \text{Beta}(\alpha + n_1 + 2n_2, \beta + n_1 + 2n_0). \quad (1.20)$$

This pragmatic decision on the prior class still leaves us much flexibility in the choice of the hyperparameters α, β . An example for $n = \{12, 11, 2\}$ simulated with $f^* = 1/3$ and with $\alpha = \beta = 1/2$ is discussed in Section 1.3.5 and shown in Figure 1.8. ■

1.3.1 Choice of Prior Distributions

If, as in the above example, the prior and the posterior lie in the same class of distributions, then this class is called *conjugate with respect to the likelihood*. Conjugate priors, if they exist, are often chosen because they lead to a well-known form of the posterior, which simplifies the calculations. Also note that for an independent sequence of observations it is sufficient to study conjugacy for one observation only. If the posterior after the first observation is in the same class as the prior, it will serve as the new prior of the second observation, which by definition stays in the conjugacy class, and so forth.

In other cases it may be useful to choose a prior that contains as little information about the parameter as possible. A first choice would, of course, be a locally uniform prior $\mathbb{P}(\theta) \propto 1$, and this is often used in practice. Under a uniform prior we have $\mathbb{P}(\theta|D) \propto \mathcal{L}(\theta)$ and so, for instance, the posterior mode equals the ML estimate. There are, however, problems associated with this prior choice. Firstly, if the parameter space is unbounded, the integral over the uniform prior will be equal to infinity and not equal to 1 as it should. Such priors are called *improper*. Despite this conceptual problem, improper priors are actually quite popular because they are sufficiently vague and, in a sense, lead back to the frequentist approach. They can, in fact, be used under the condition that at least the corresponding posterior distribution is proper. A second, more serious problem associated with uniform priors is that if one reparameterizes the parameter with some one-to-one transformation, then the corresponding prior of the transformed parameter will not, in general, be uniform anymore. The choice of parameter for a model is often arbitrary. For instance, in Example 1.15 above, is f or the odds $\theta = f/(1-f)$ the more intuitive parameter? If we choose a non-informative (uniform) prior for f , the corresponding prior for the odds θ will be informative, and vice versa. It turns out that there exists a particular choice of prior distribution that is invariant under reparameterizations. This is called *Jeffreys' prior*. In the case of a model with only one parameter θ and log-likelihood $\ell(\theta)$, Jeffreys' prior turns out to be

$$\mathbb{P}(\theta) \propto \sqrt{J(\theta)},$$

where $J(\theta)$ is the *expected Fisher information* defined by

$$J(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \log \ell(\theta)\right].$$

Example 1.16 (Allele frequency). Let us determine Jeffreys' prior for the Hardy–Weinberg model in Example 1.15. From equation (1.18) we obtain – up to a constant – the log-likelihood

$$\ell(f) = (n_1 + 2n_2) \log f + (n_1 + 2n_0) \log(1-f)$$

and thus

$$\frac{d^2}{df^2} \log \ell(f) = -\frac{n_1 + 2n_2}{f^2} - \frac{n_1 + 2n_0}{(1-f)^2}. \quad (1.21)$$

In order to calculate $J(\theta)$ we consider n_0, n_1, n_2 as random variables and use the expected values

$$\mathbb{E}[n_0] = n(1-f), \quad \mathbb{E}[n_1] = 2nf(1-f), \quad \mathbb{E}[n_2] = nf.$$

Now we can take the expectation of equation (1.21) and obtain

$$J(\theta) = -\mathbb{E}\left[\frac{d^2}{df^2} \log \ell(f)\right] = \frac{2n}{f(1-f)}.$$

Thus, Jeffreys' prior in this example is $\mathbb{P}(f) \propto f^{-1/2}(1-f)^{-1/2}$ which is exactly the beta prior with $\alpha = \beta = 1/2$ that we used for the numerical illustration in Example 1.15. ■

Jeffreys' prior is often hard to come by because the expected Fisher information is not always readily available. For an excellent overview of this topic we refer to Held and Sabanés Bové (2014, Ch. 6.3.3.).

1.3.2 Bayesian Point Estimates and Confidence Intervals

Often it is inconvenient to report the whole posterior and one would like to resort to point estimates, as in frequentist statistics. There are three popular Bayesian point estimates: the posterior mean, mode and median. One can give a decision-theoretic justification for these estimates via loss functions. To keep things simple, assume that the parameter θ is just a number, not a vector. We will estimate the parameter with a number a . A *loss function* $L(a, \theta)$ is a positive function quantifying the loss suffered when estimating θ ; the further away we are from θ , the larger the loss suffered. One then defines the corresponding estimator as the number a that minimizes the expected loss (called the *risk function*):

$$\hat{\theta} = \arg \min_a \mathbb{E}[L(a, \theta) | D] = \arg \min_a \int L(a, \theta) \mathbb{P}(\theta | D) d\theta.$$

Take, for instance, the quadratic loss $L(a, \theta) = (a - \theta)^2$. By linearity of the expectation we obtain the risk function

$$\mathbb{E}[L(a, \theta) | D] = \mathbb{E}[(a - \theta)^2 | D] = \mathbb{E}[\theta^2 | D] - 2a\mathbb{E}[\theta | D] + a^2.$$

This is a quadratic function in a which is minimal for the value

$$\hat{\theta} = \mathbb{E}[\theta | D] = \int \theta \mathbb{P}(\theta | D) d\theta.$$

We recognize this as the posterior mean estimate, often called *minimum mean squared error* (MMSE) estimate. Similarly, by choosing the absolute loss $L(a, \theta) = |a - \theta|$, one obtains the posterior median (see, for example, Held and Sabanés Bové, 2014, Thm. 6.3). Finally, the uncompromising 0–1 loss, which is equal to 0 if $a = \theta$ and equal to 1 if $a \neq \theta$, leads to the posterior mode estimate. This estimate is also called the maximum posterior or *maximum a posteriori* (MAP) estimate and corresponds simply to the parameter value with the highest posterior probability.

In addition to point estimates, as in frequentist statistics, one often wants a measure of confidence. The most popular such measure is the *highest posterior density* (HPD) region. For a given significance level α this is the set of most probable parameters that in total constitute $1 - \alpha$ of the posterior probability mass. More precisely, we first find a threshold value π such that for

the region $C_\alpha = \{\theta : \mathbb{P}(\theta|D) > \pi\}$ we get

$$\int_{C_\alpha} \mathbb{P}(\theta|D)d\theta = 1 - \alpha.$$

This region C_α , then, is the HPD region.

This short introduction to Bayesian statistics cannot, of course, do justice to this vast topic. There are many good book-length treatments of the field; we single out Robert (2007).

1.3.3 Markov Chain Monte Carlo

A common challenge in Bayesian inference is that the integral $\mathbb{P}(D) = \int \mathbb{P}(D|\theta)\mathbb{P}(\theta)d\theta$ cannot be solved analytically. Since $\mathbb{P}(D)$ is just a proportionality constant for inferring posterior distributions (see above), it may be inferred numerically using a variety of schemes. For instance, one might consider numerical integration using Riemann sums across a predefined set of parameter vectors θ_i . While feasible in very low dimensions, the number of necessary parameter vectors to accurately infer $\mathbb{P}(D)$ explodes very quickly in higher dimensions, a fact known as *the curse of dimensionality*.

A more robust approach that easily scales to higher dimensions is to numerically approximate the posterior distribution from a large number of samples drawn from it. The most widely used such method is Markov chain Monte Carlo (MCMC), which elegantly circumvents the necessity to evaluate $\mathbb{P}(D)$, as we will outline below. However, let us first recall a few important properties about Markov chains necessary to introduce the MCMC algorithm.

1.3.3.1 Markov Chains

Markov chains are named after the Russian probabilist Andrei A. Markov (Markov, 1906) and come in many varieties. To avoid technicalities, however, we restrict ourselves to the notion of *discrete-time* and *finite Markov chains* (DTMCs), and refer to the literature (e.g. Murphy, 2012; Allen, 2010) for a more complete overview, and to **Chapter 6** for an introduction to continuous-time Markov chains.

A DTMC is a sequence or process X_t , $t = 0, 1, 2, \dots$, of random variables taking values in a finite set of states $S = \{1, 2, \dots, K\}$. Moreover, this process has the property that its future behavior depends only on the present but not on its past history (*Markov property*); more precisely, that

$$\mathbb{P}(X_t = i_t \mid X_0 = i_0, \dots, X_{t-1} = i_{t-1}) = \mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}).$$

The $K \times K$ matrix \mathbf{P} formed by the elements

$$p_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$$

is called the *one-step transition matrix*. Here we assume that the Markov chain is *homogeneous*, that is, that the transition probabilities do not depend on time. Clearly, each row of \mathbf{P} adds up to one: $\sum_{j=1}^K p_{ij} = 1$. The probabilities $p_{ij}^{(t)}$ of transferring from state i into state j in exactly t steps satisfy the *Chapman–Kolmogorov equations*

$$p_{ij}^{(s+t)} = \sum_{k=1}^K p_{ik}^{(s)} p_{kj}^{(t)},$$

which means that the probability of getting from i to j in $s + t$ steps is just the probability of going from i to k in s steps and then from k to j in t steps. From this it follows that the t -step transition matrix is the t th power of the one-step transition matrix: $\mathbf{P}^{(t)} = \mathbf{P}^t$. If $\pi_j^{(t)} = \mathbb{P}(X_t = j)$

denotes the probability of being in state j at time t , we have $\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)}\mathbf{P}$, where the $\boldsymbol{\pi}^{(t)}$ are to be understood as row vectors with elements $\pi_1^{(t)}, \dots, \pi_K^{(t)}$. If we ever reach the stage where

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}, \quad (1.22)$$

we say we have reached the *stationary distribution*.

Example 1.17 (Wright–Fisher model). Let us use the classic Wright–Fisher model of genetic drift as an illustration of DTMCs. Consider a particular locus which has two alleles A and a . Denote by X_t the number of A s in a population of fixed size $2N$ at time t . The alleles of a generation at some time $t + 1$ are obtained by drawing $2N$ times with replacement from the previous generation at t . Since the new distribution of alleles depends only on the previous generation, X_t is a Markov chain with states $\mathcal{S} = \{0, 1, \dots, 2N\}$ and the binomial transition probabilities

$$p_{ij} = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}.$$

Once an allele is completely lost from a population, it never returns and so 0 and $2N$ are *absorbing states*: $p_{00} = p_{11} = 1$. It also follows that all distributions of the form $\boldsymbol{\pi} = (\pi_0, 0, \dots, 0, 1 - \pi_0)$ are stationary, that is, there is no unique stationary distribution for this chain. ■

We see from this example that a necessary condition for a unique stationary distribution is that we can go from any state to any other state; such chains are called *irreducible*. The Wright–Fisher model is an example of a reducible Markov chain. Now consider a Markov chain with only two states and where the probabilities of changing from one state to the other are $p_{12} = p_{21} = 0.1$ at each step. Obviously, the chain will spend 50% of its time in either state in the long run, regardless of the state in which the process starts. Therefore, $\boldsymbol{\pi} = (0.5, 0.5)$ will be the unique stationary distribution and also the *limiting distribution* in the sense that $\pi_j = \lim_{t \rightarrow \infty} P_{ij}^t$, independently of i . But now suppose that in our two-state Markov chain we have $p_{12} = p_{21} = 1$. If we start in state $X_0 = 1$, every odd time we will be in state 2 and every even time back in state 1. This chain clearly has no limiting distribution, but rather consists of a periodic loop. By definition, the period of a state i is the greatest common divisor of the set of times $t = 1, 2, \dots$ for which $P_{ii}^t > 0$. If all its states have period 1, the chain is called *aperiodic*. One can prove that *every irreducible aperiodic DTMC has a limiting distribution which is also its unique stationary distribution*. See Allen (2010) for a proof of this theorem.

Reversibility In order to fully understand the MCMC algorithm we need one more important notion of Markov chains: reversibility (Robert and Casella, 1999, Ch. 6.5.3). Let X_1, X_2, \dots, X_T be a homogeneous irreducible DTMC. Consider this chain backwards: X_T, X_{T-1}, \dots, X_0 . It may not be obvious, but this backward chain turns out also to be a Markov chain. Indeed,

$$\begin{aligned} \mathbb{P}(X_t = j | X_{t+1} = i, X_{t+2} = k, \dots) &= \frac{\mathbb{P}(X_t = j, X_{t+1} = i, X_{t+2} = k, \dots)}{\mathbb{P}(X_{t+1} = i, X_{t+2} = k, \dots)} \\ &= \frac{\mathbb{P}(X_{t+2} = k, \dots | X_{t+1} = i, X_t = j)}{\mathbb{P}(X_{t+2} = k, \dots | X_{t+1} = i)} \frac{\mathbb{P}(X_{t+1} = i, X_t = j)}{\mathbb{P}(X_{t+1} = i)} \\ &= \frac{\mathbb{P}(X_{t+2} = k, \dots | X_{t+1} = i)}{\mathbb{P}(X_{t+2} = k, \dots | X_{t+1} = i)} \mathbb{P}(X_t = j | X_{t+1} = i) \\ &= \mathbb{P}(X_t = j | X_{t+1} = i), \end{aligned}$$

where in the penultimate step we used the Markov property of the forward chain. We can compute the transition probabilities for the backward chain:

$$\begin{aligned}\mathbb{P}(X_t = j | X_{t+1} = i) &= \frac{\mathbb{P}(X_t = j, X_{t+1} = i)}{\mathbb{P}(X_{t+1} = i)} \\ &= \frac{\mathbb{P}(X_{t+1} = i | X_t = j)\mathbb{P}(X_t = j)}{\mathbb{P}(X_{t+1} = i)} = \frac{p_{ji}\mathbb{P}(X_t = j)}{\mathbb{P}(X_{t+1} = i)}.\end{aligned}$$

We see that these transition probabilities may depend on t , that is, the backward Markov chain need not be homogenous. But now assume that the forward chain has a stationary distribution π . Then the backward chain will be homogenous, too:

$$\mathbb{P}(X_t = j | X_{t+1} = i) = \frac{p_{ji}\pi_j}{\pi_i}. \quad (1.23)$$

The chain X_t is called *reversible* if the transition probabilities of the forward and backward chain are equal: $\mathbb{P}(X_t = j | X_{t+1} = i) = \mathbb{P}(X_{t+1} = j | X_t = i) = p_{ij}$. From this and equation (1.23) we get the *detailed balance equations*

$$\pi_i p_{ij} = \pi_j p_{ji}, \quad (1.24)$$

which state that the stationary flow from i to j equals the stationary flow from j to i . We actually do not need to know the stationary distribution in advance in order to check if a Markov chain is reversible. The following proposition even helps us find π . *If a homogeneous irreducible DTMC satisfies the detailed balance equations (1.24) for some probability distribution π then the chain is reversible and the distribution π is its unique stationary distribution.*

We can indeed easily check that the distribution must be stationary:

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j,$$

which in matrix notation reads $\pi \mathbf{P} = \pi$.

1.3.3.2 Metropolis–Hastings Algorithm

The general idea of any MCMC method is to construct an irreducible Markov chain that has as its limiting distribution and unique stationary distribution a particular distribution $\mathbb{P}(\mathbf{x})$ of interest. If run sufficiently long, such a Markov chain will converge to this stationary distribution and the states visited will constitute (correlated) samples from that distribution. An MCMC thus allows samples to be generated from distributions that are otherwise difficult to sample from.

The most commonly used MCMC variant is the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970):

1. Choose appropriate initial values \mathbf{x}_0 and set $t = 0$.
2. Propose a move $\mathbf{x}_t \rightarrow \mathbf{x}'_t$ according to some proposal kernel $q(\mathbf{x}, \mathbf{x}')$.
3. Accept move and set $\mathbf{x}_{t+1} = \mathbf{x}'_t$ with probability given by the Hastings ratio

$$h = \min \left(1, \frac{\mathbb{P}(\mathbf{x}') q(\mathbf{x}'_t, \mathbf{x}_t)}{\mathbb{P}(\mathbf{x}) q(\mathbf{x}_t, \mathbf{x}'_t)} \right),$$

- else reject move and set $\mathbf{x}_{t+1} = \mathbf{x}_t$.
4. Increment t and go back to step 2.

Under this algorithm, the probability of transitioning from state \mathbf{x} to state \mathbf{x}' is given by the product of the proposal and acceptance probabilities: in order to transition to state \mathbf{x}' , a move $\mathbf{x} \rightarrow \mathbf{x}'$ has to be proposed and then accepted. Hence

$$\mathbb{P}(\mathbf{x}'|\mathbf{x}) = q(\mathbf{x}, \mathbf{x}') \min\left(1, \frac{\mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x}_t)}{\mathbb{P}(\mathbf{x})q(\mathbf{x}_t, \mathbf{x}'_t)}\right). \quad (1.25)$$

It is straightforward to prove that a Markov chain with such transition probabilities has $\mathbb{P}(\mathbf{x})$ as its stationary distribution as $\mathbb{P}(\mathbf{x})$ fulfills the detailed balance

$$\mathbb{P}(\mathbf{x})\mathbb{P}(\mathbf{x}'|\mathbf{x}) = \mathbb{P}(\mathbf{x}')\mathbb{P}(\mathbf{x}|\mathbf{x}').$$

To see this, let us substitute $\mathbb{P}(\mathbf{x}'|\mathbf{x})$ and $\mathbb{P}(\mathbf{x}|\mathbf{x}')$ according to equation (1.25),

$$\mathbb{P}(\mathbf{x})q(\mathbf{x}, \mathbf{x}') \min\left(1, \frac{\mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x})}{\mathbb{P}(\mathbf{x})q(\mathbf{x}, \mathbf{x}')}\right) = \mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x}) \min\left(1, \frac{\mathbb{P}(\mathbf{x})q(\mathbf{x}, \mathbf{x}')}{\mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x})}\right),$$

from which we get

$$\min(\mathbb{P}(\mathbf{x})q(\mathbf{x}, \mathbf{x}'), \mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x})) = \min(\mathbb{P}(\mathbf{x}')q(\mathbf{x}', \mathbf{x}), \mathbb{P}(\mathbf{x})q(\mathbf{x}, \mathbf{x}')),$$

which is equal since $\min(a, b) = \min(b, a)$.

The only condition for this proof to hold is that the chosen proposal kernel is reversible, and hence that for each pair \mathbf{x} and \mathbf{x}' with non-zero $q(\mathbf{x}, \mathbf{x}')$, $q(\mathbf{x}', \mathbf{x})$ is also non-zero. Although not a requirement, symmetric proposal kernels with $q(\mathbf{x}, \mathbf{x}') = q(\mathbf{x}', \mathbf{x})$ are often chosen as they cancel out from the Hastings ratio.

Example 1.18 (Normal distribution). Let us develop an MCMC scheme to generate samples x_0, x_1, \dots, x_n from a normal distribution $x_i \sim \mathbb{P}(x|\mu, \sigma^2)$. Using the symmetric proposal kernel $q(x, x') \sim \mathcal{U}(x - \frac{d}{2}, x + \frac{d}{2})$ with width d , the Hastings ratio h becomes

$$\log h = \log \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x'-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \right) = \frac{1}{2\sigma^2} [(x - \mu)^2 - (x' - \mu)^2].$$

The results of an MCMC chain run with $d = 10$ and started at $x_0 = 0.0$ to sample from a normal distribution with $\mu = 0.0$ and $\sigma^2 = 1$ are shown in Figure 1.5(a). As seen in the other panels of Figure 1.5 and discussed below, the choice of d and x_0 can have a big impact on performance. ■

1.3.3.3 Convergence and Mixing

While the distribution of samples resulting from a Markov chain is guaranteed to match the stationary distribution $\mathbb{P}(\mathbf{x})$ in the limit, the number of iterations it will take to *converge* to that distribution within an acceptable error depends on multiple factors that require fine-tuning in practice. A particularly crucial aspect that often requires fine-tuning is the choice of the proposal distribution $q(\mathbf{x}, \mathbf{x}')$, which affects the way the MCMC chain explores the parameter space, a property commonly referred to as *mixing*. If the newly proposed values \mathbf{x}' are too far from \mathbf{x} , they are often hopelessly far from the posterior peak and will only very rarely be accepted. If the newly proposed values \mathbf{x}' are too close to \mathbf{x} , however, the chain will have a hard time exploring the entire posterior distribution. Both cases hamper the computational efficiency of the algorithm, and hence only very long chains will appropriately converge. Since an ideal choice of the

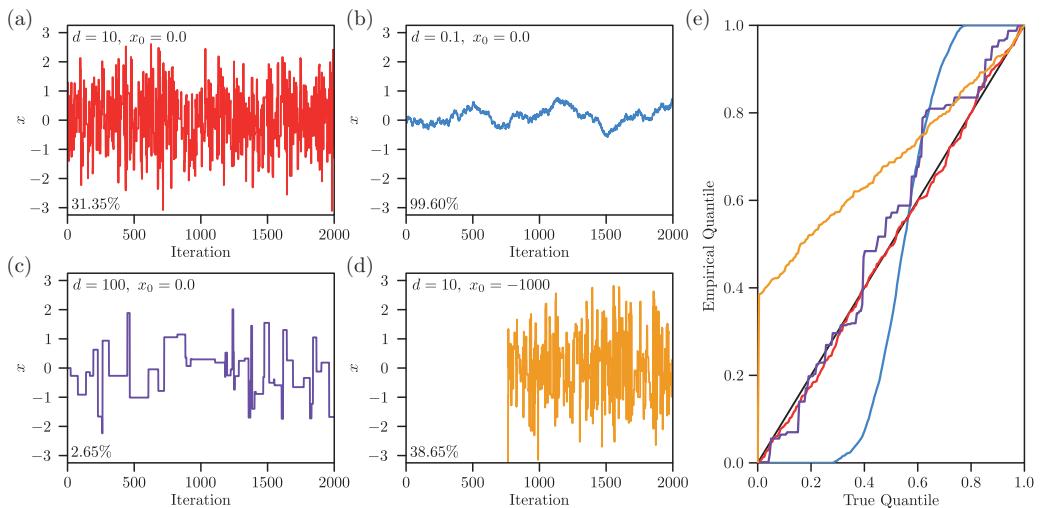


Figure 1.5 (a)–(d) Traces of MCMC chains with symmetric, uniform proposal kernels of different widths d and starting values x_0 to sample from a normal distribution with $\mu = 0.0$ and $\sigma^2 = 1$. Acceptance rates are given in the lower left corner. (e) Quantile–quantile plot comparing the empirical quantiles of the sampling distributions obtained with the different MCMC chains (colors as in (a)–(d)) against the true quantiles (black line) of the target distribution $\mathcal{N}(\mu, \sigma^2)$.

proposal distribution $q(\mathbf{x}, \mathbf{x}')$ requires knowledge on the distribution from which the MCMC samples, it is usually necessary to tinker with different proposal distributions and monitor the performance of the chain. As a rule of thumb, chains with an acceptance rate of $1/3$ are generally considered to mix well. See e.g. Robert and Casella (1999) for a more in-depth discussion.

Example 1.19 (Normal distribution). In Figure 1.5 we show the results of MCMC chains from Example 1.18 run with very low $d = 0.1$ and very high $d = 100$. As is visible from the traces, the former choice does not enable the chain to explore the full distribution within the finite number of iterations. The latter choice, in contrast, spends many iterations at the same location because most proposed values are far outside the region with high density and thus rarely accepted. As a consequence, the sampling distributions of both MCMC chains have not yet converged to the stationary distribution. ■

The choice of starting value \mathbf{x}_0 is equally crucial: if $\mathbb{P}(\mathbf{x}_0)$ is very small, say 10^{-x} , it will take at least 10^x iterations until the sampling distribution will correctly reflect that probability. Since MCMC is often used to characterize an unknown distribution, good (i.e. likely) starting values are generally unknown. A common strategy referred to as *burn-in* is to search for good starting values by running an initial MCMC chain, and to then to use the last value of that initial chain as starting value for the actual MCMC chain. This strategy, which is equivalent to just throwing out the first few iterations of the chain, is justified since the initial MCMC chain has the same limiting distribution.

Example 1.20 (Normal distribution). In Figure 1.5 we show the results of MCMC chains from Example 1.18 run with appropriate $d = 10$, but started at $x_0 = -1000$, and hence far out in the tail of the stationary distribution. As a result, the sampling distribution has a heavy lower tail, and it will take an enormous amount of additional iterations for the distribution to reach convergence. However, throwing out the first half of the chain as *burn-in* would result in acceptable convergence. ■

Importantly, even chains with agreeable acceptance rates may take a surprisingly long time to converge, for instance if the target distribution has multiple peaks, separated by a deep valley with very low density. It is thus paramount to properly assess the convergence of any MCMC chain, which is best done by running chains for a very large number of iterations. Alternatively, comparing multiple chains starting from different locations may also help assess convergence (see Robert and Casella, 1999, Ch. 12 for a detailed discussion on MCMC tuning).

1.3.3.4 Metropolis–Hastings Algorithm in Bayesian Inference

The Metropolis–Hastings algorithm is readily applied to generate samples from the posterior distribution $\mathbb{P}(\theta|D)$, in which case the Hastings ratio is calculated as

$$h = \min \left(1, \frac{\mathbb{P}(\theta'|D)q(\theta', \theta)}{\mathbb{P}(\theta|D)q(\theta, \theta')} \right) = \min \left(1, \frac{\mathbb{P}(D|\theta')\mathbb{P}(\theta')q(\theta', \theta)}{\mathbb{P}(D|\theta)\mathbb{P}(\theta)q(\theta, \theta')} \right), \quad (1.26)$$

which does not require the evaluation of $\mathbb{P}(D)$. The samples thus generated can then be used to approximate the posterior distribution, or to infer posterior point estimates and credible intervals.

Example 1.21 (Inbreeding). It is usually rather straightforward to develop an MCMC method using the Metropolis–Hastings algorithm, even for hierarchical models. To illustrate this, consider a model for learning about population-level inbreeding (or selfing rate) F from n diploid samples genotyped at L unlinked markers. Let us denote by f_l the unknown allele frequency at locus l and by g_{li} the genotype at locus l observed for individual i . Let us finally assume that the allele frequencies are beta distributed $f_l \sim \text{Beta}(\alpha, \alpha)$ with shape parameter α . See Figure 1.6(a) for a DAG of this model.

We are interested in the posterior distribution

$$\mathbb{P}(F, \alpha, f | g) = \frac{\mathbb{P}(g|F, \alpha, f)\mathbb{P}(F)\mathbb{P}(\alpha)\mathbb{P}(f)}{\mathbb{P}(g)},$$

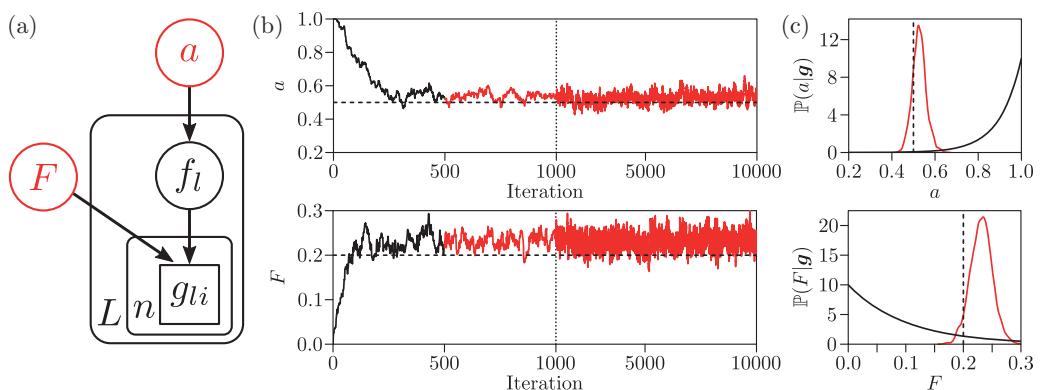


Figure 1.6 (a) DAG of the model discussed in Example 1.21 with the two hierarchical parameters F and α highlighted in red. (b) The trace of F and α of an MCMC with 10^5 iterations run on genotypes of 10 individuals at 500 loci simulated with $F = 0.2$ and $\alpha = 0.5$ (dashed lines). The first 500 iterations (black) were thrown out as burn-in. Note the change in scale after 1000 iterations. (c) Resulting marginal posterior distributions for F and α (red) against the prior distributions (black).

where $\mathbf{f} = \{f_1, \dots, f_L\}$, $\mathbf{g} = \{g_{11}, \dots, g_{L1}, \dots, g_{Ln}\}$. The likelihood function is

$$\mathcal{L}(F, a, \mathbf{f}) = \mathbb{P}(\mathbf{g}|F, a, \mathbf{f}) = \prod_{l=1}^L \mathbb{P}(f_l|a) \prod_{i=1}^n \mathbb{P}(g_{li}|F, f_l),$$

where under the classic model of inbreeding (Wright, 1921)

$$\mathbb{P}(g_{li}|F, f_l) = \begin{cases} (1-F)(1-f_l)^2 + F(1-f_l) & \text{if } g_{li} = 0, \\ (1-F)2f_l(1-f_l) & \text{if } g_{li} = 1, \\ (1-F)f_l^2 + Ff_l & \text{if } g_{li} = 2. \end{cases}$$

Let us define exponential prior distributions truncated to the interval $[0, 1]$ on the two hierarchical parameters F and a such that

$$\mathbb{P}(F) \sim \lambda_F e^{-F\lambda_F}, \quad \mathbb{P}(a) \sim \lambda_a e^{-(1-a)\lambda_a}.$$

As is commonly done in high-dimensional models, we will update each parameter in turn with the symmetric proposal kernel

$$q(x, x') \sim \mathcal{N}(x, \sigma_x^2), \quad x = F, a, f_l,$$

mirrored at 0 and 1 to respect the interval. The simplified Hastings ratios for updates $F \rightarrow F'$, $a \rightarrow a'$, and $f_l \rightarrow f'_l$ are

$$h_F = \frac{\mathbb{P}(F')}{\mathbb{P}(F)} \prod_{l=1}^L \prod_{i=1}^n \frac{\mathbb{P}(g_{li}|F', f_l)}{\mathbb{P}(g_{li}|F, f_l)}, \quad h_a = \frac{\mathbb{P}(a')}{\mathbb{P}(a)} \prod_{l=1}^L \frac{\mathbb{P}(f_l|a')}{\mathbb{P}(f_l|a)}, \quad h_{f_l} = \frac{\mathbb{P}(f_l|a')}{\mathbb{P}(f_l|a)} \prod_{i=1}^n \frac{\mathbb{P}(g_{li}|F, f'_l)}{\mathbb{P}(g_{li}|F, f_l)}.$$

Marginal posterior densities obtained by running this MCMC on simulated data (10 individuals and 500 loci) are shown in Figure 1.6(b,c). Despite the limited data, the prior has little influence and the marginal posterior distributions indicate low uncertainty. Note that the true value of $F = 0.2$ lies in the tail of the marginal posterior distribution on F . This does not indicate a problem with the inference, but rather highlights the true meaning of uncertainty as quantified by Bayesian approaches: $\mathbb{P}(D \leq 0.2|\mathbf{g}) = 0.057$, and hence the true value is to be expected in that tail in 5.7% of the cases.

To summarize the posterior distribution one might want to also calculate point estimates. While the MAP estimates must be obtained using kernel smoothing, the (marginal) posterior means or MMSE estimates are readily found to be

$$\hat{\theta}_{\text{MMSE}} \approx \frac{1}{I} \sum_{i=1}^I \theta_i, \quad i = 1, \dots, I,$$

where θ_i corresponds to the i th MCMC sample after the burn-in. In the case here, we obtain $\hat{a}_{\text{MMSE}} = 0.532$ and $\hat{F}_{\text{MMSE}} = 0.234$. ■

As shown in the above example, the Metropolis–Hastings algorithm extends naturally to hierarchical models of arbitrary complexity, as latent parameters can be updated and hence integrated out during the MCMC. However, they may also be integrated out analytically when calculating Hastings ratios of parameters at higher levels of the hierarchy. Finally, and also as shown on the above example, marginal distributions are obtained directly from the MCMC samples. However, we advise also studying posterior distributions of higher order to identify potential correlations in the parameters that are easily missed when focusing on marginal distributions alone (see below for an example).

1.3.4 Empirical Bayes for Latent Variable Problems

Although numerical schemes such as MCMC naturally allow for the inference of hierarchical models, they may not be computationally efficient. This is particularly true if the main interest lies in the marginal posterior distributions of variables at lower levels of the hierarchy that could be calculated analytically without the need for stochastic schemes if the parameters at higher levels were known. A popular strategy for such cases is to resort to *empirical Bayes* methods (Robbins, 1964). In a first step these methods infer point estimates for the parameters at higher levels by integrating out those at lower levels, and then infer posterior distributions for lower-level parameters while setting those at higher level to their point estimates.

Consider a latent variable model similar to equation (1.12) with a vector $\mathbf{x} = \{x_1, \dots, x_n\}$ of n observed data, a vector $\mathbf{z} = \{z_1, \dots, z_n\}$ of n latent data and hierarchical parameters θ such that all x_i are independent of all other $x_j, i \neq j$, if conditioned on z_i (usually denoted by $x_i \perp x_j | z_i, i \neq j$), and similarly $z_i \perp z_j | \theta, i \neq j$. The likelihood of such a model is

$$\mathcal{L}(\mathbf{z}, \theta) = \prod_{i=1}^n \mathbb{P}(x_i | z_i) \mathbb{P}(z_i | \theta).$$

Note that the inference of the multidimensional posterior distribution

$$\mathbb{P}(\mathbf{z}, \theta | \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} | \mathbf{z}) \mathbb{P}(\mathbf{z} | \theta) \mathbb{P}(\theta)}{\mathbb{P}(\mathbf{x})}$$

usually requires numerical integration because of the intractability of $\mathbb{P}(\mathbf{x})$. Often there is interest in the inference of the latent variables \mathbf{z} , and in particular the marginal posterior distributions

$$\mathbb{P}(z_i | \mathbf{x}) = \iint \mathbb{P}(\mathbf{z}, \theta | \mathbf{x}) d\theta dz_{-i},$$

where \mathbf{z}_{-i} denotes the vector \mathbf{z} without its i th element. Importantly, this differs from $\mathbb{P}(z_i | x_i)$ in that the hierarchical model also accounts for information about the population \mathbf{z} contained in \mathbf{x}_{-i} and characterized by θ .

The empirical Bayes method now consists of first estimating a value $\hat{\theta}$ from the data, and then estimating $\mathbb{P}(z_i | \mathbf{x})$ according to

$$\mathbb{P}(z_i | \mathbf{x}) \approx \mathbb{P}(z_i | x_i, \hat{\theta}) = \frac{\mathbb{P}(x_i | z_i) \mathbb{P}(z_i | \hat{\theta})}{\int \mathbb{P}(x_i | z) \mathbb{P}(z | \hat{\theta}) dz}, \quad (1.27)$$

which can often be done analytically. Here, $\mathbb{P}(z_i | \hat{\theta})$ is interpreted as the prior distribution on z_i , parameterized by $\hat{\theta}$ learned from the data. This may seem to contradict the standard Bayesian philosophy under which prior distributions reflect the *a priori* belief before any data were observed (Efron, 2012). It is justified in hierarchical models like these, however, if there is so much information on the hierarchical parameters that the marginal posterior on θ ,

$$\mathbb{P}(\theta | \mathbf{x}) = \int \mathbb{P}(\mathbf{z}, \theta | \mathbf{x}) dz,$$

very narrowly peaks at a single value. Indeed, as the sample size $n \rightarrow \infty$, the posterior on θ will shrink to a point mass on the true parameter θ^* , and hence $\mathbb{P}(z_i | \mathbf{x}) = \mathbb{P}(z_i | x_i, \theta^*)$. For finite data, using equation (1.27) with an initial estimate $\hat{\theta} \approx \theta^*$ can thus be considered a computationally efficient approximation to a standard Bayesian treatment of the full model.

Example 1.22 (Genotype calling). As an example, consider a common strategy for genotype calling from next-generation sequencing data of a population sample. Let us denote by $\mathbb{P}(d_i|g_i)$ the likelihood of observing the sequencing data of individual $i = 1, \dots, n$ given the unknown genotype g_i of that individual. Under the assumption of Hardy–Weinberg equilibrium in the population, we can build the hierarchical model

$$\mathcal{L}(f, g) = \prod_{i=1}^n \mathbb{P}(d_i|g_i)\mathbb{P}(g_i|f),$$

where f is the allele frequency in the population, $g = \{g_1, \dots, g_n\}$ and $\mathbb{P}(g_i|f)$ is given by equation (1.2).

Note that the evaluation of the multidimensional posterior distribution $\mathbb{P}(f, g|\mathbf{d})$, $\mathbf{d} = \{d_1, \dots, d_n\}$, requires a numerical scheme such as MCMC. However, if f were known, the posterior distribution on each g_i would be readily and analytically calculated as

$$\mathbb{P}(g_i|d_i, f) = \frac{\mathbb{P}(d_i|g_i)\mathbb{P}(g_i|f)}{\sum_{h=0}^2 \mathbb{P}(d_i|h)\mathbb{P}(h|f)}, \quad (1.28)$$

where the Hardy–Weinberg equilibrium serves as a natural prior on individual genotypes.

One might want to benefit from this computational advantage by using an empirical Bayes scheme, in which a point estimate of the allele frequency \hat{f} is obtained in a first step by treating all g_i as latent variables and integrating them out, for instance via an EM algorithm similar to Example 1.10. Then, in a second step, the posterior distributions on all g_i are calculated using equation (1.28) with $f = \hat{f}$.

To illustrate the accuracy of this strategy, we compared the marginal posterior distributions $\mathbb{P}(g_i|d_i, \hat{f})$ to the corresponding distributions $\mathbb{P}(g_i|\mathbf{d})$ of the full hierarchical model, which we obtained using 10^5 MCMC samples. We then quantified this difference using the L_1 distance

$$L_1 = \frac{1}{3n} \sum_{i=1}^n \sum_{g=0}^2 |\mathbb{P}(g_i|d_i, \hat{f}) - \mathbb{P}(g_i|\mathbf{d})|, \quad (1.29)$$

reported in Figure 1.7 for different sample sizes and sequencing depths. In the case of informative data (depth of $10\times$), we found the L_1 distances to be rather small (often less than 1%), even for low sample sizes. However, in the case of limited data (depth of $2\times$), the median L_1 was often rather large (up to 10% or more), even for moderate sample sizes of $n = 100$. In these cases, it is thus important to account for the uncertainty associated with f , rather than relying on a single and noisy estimate \hat{f} . Of course, this changes if larger samples are used, in which case the posterior on f gets narrower, and thus justifying the empirical Bayes approach. ■

1.3.5 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a class of simulation-based techniques to conduct Bayesian inference under models with intractable likelihoods. Consider some data \mathcal{D} generated under an arbitrary model $\mathbb{P}(\mathcal{D}|\theta)$ with parameters θ and prior $\mathbb{P}(\theta)$. In addition, let $s_{\text{obs}} = T(\mathcal{D})$ be a vector of observed summary statistics calculated on the full data. The most basic ABC algorithm, also termed *rejection algorithm* or *ABC-REJ*, proceeds as follows (Tavaré et al., 1997; Weiss and von Haeseler, 1998):

1. Sample a parameter vector θ' from the prior distribution $\mathbb{P}(\theta)$.
2. Simulate data under the model $\mathcal{D}' \sim \mathbb{P}(\mathcal{D}'|\theta')$ and calculate summary statistics $s' = T(\mathcal{D}')$.
3. Accept θ' if $\|s_{\text{obs}} - s'\| < \delta$.
4. Go to step 1.

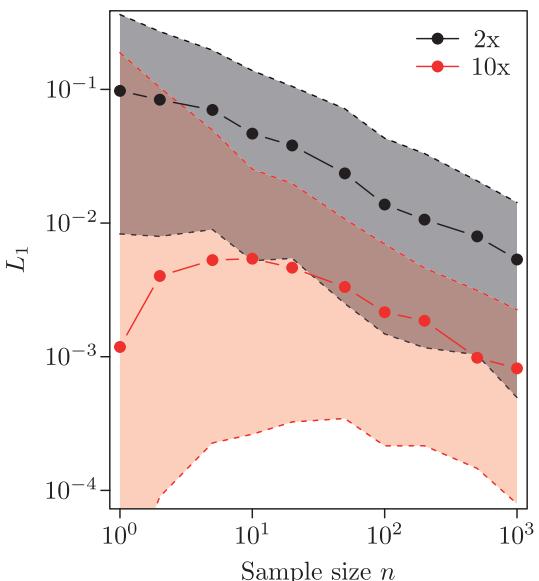


Figure 1.7 Comparison of marginal posterior estimates obtained with the full model (using MCMC) and with an empirical Bayes scheme for different sample sizes n , quantified by the average difference in posterior probabilities L_1 (see text). Connected dots and shaded areas give the median L_1 and 90% quantile across 200 replicate simulations with sequencing depth of 2x (black) and 10x (red).

This algorithm has the agreeable property of being applicable to any model that can be simulated, even if its likelihood function is intractable or of arbitrary complexity. In contrast to the Monte Carlo solution presented in Section 1.2.5.2, it is not limited to the case of intractable integrals (or sums). The price to pay for this flexibility are two approximations common to all ABC algorithms: (1) The data is summarized by some summary statistics $s = T(\mathcal{D})$. (2) Proposed parameter values θ' are accepted if the simulated summary statistics $s' = T(\mathcal{D}')$ are sufficiently close to the observed summary statistics s_{obs} , according to some distance metric $\|\cdot\|$ and a threshold δ . The resulting algorithm does generate samples from $\mathbb{P}(\theta||s_{\text{obs}} - s' < \delta)$, which matches the true posterior distribution $\mathbb{P}(\theta|\mathcal{D})$ under the sublime conditions that the summary statistics $s = T(\mathcal{D})$ are sufficient for the data \mathcal{D} and $\delta = 0$. If the summary statistics are near-sufficient and δ is small, the sampling distribution should constitute a reasonable approximation to the full posterior.

Example 1.23 (Allele frequency). Let us revisit the problem of inferring the allele frequency f from n observed genotypes $\mathbf{g} = \{g_1, \dots, g_n\}$, for which we derived the analytical posterior distribution in Example 1.15 when using a beta prior $f \sim \text{Beta}(\alpha, \beta)$. Let us further use the genotype counts $\mathbf{n} = \{n_0, n_1, n_2\}$ of the genotypes $g = 0, 1, 2$ as sufficient summary statistics for the data (see Example 1.15). ABC samples can now be generated very easily under this model with the ABC-REJ algorithm:

1. Sample f' from the prior distribution $\text{Beta}(\alpha, \beta)$.
2. Simulate n genotypes given f' and summarize them with the counts \mathbf{n}' .
3. Accept f' if the Euclidean norm $d = \|\mathbf{n}_{\text{obs}} - \mathbf{n}'\| < \delta$.
4. Go to step 1.

Example posterior estimates for $\mathbf{n} = \{12, 11, 2\}$ from Example 1.15 simulated with $f^* = 1/3$ are shown in Figure 1.8(a). ■

As shown in the above example (Figure 1.8(a)), ABC approximations are fairly accurate for low δ , but become easily distorted if δ is large. Actually, larger δ values weaken the contribution

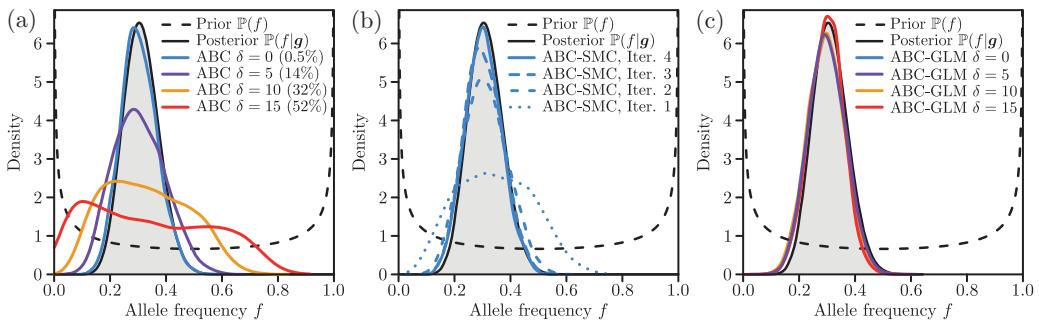


Figure 1.8 Comparison of ABC estimates (color) against the true posterior distribution given by Expression (1.20) (black, solid with gray fill) of the allele frequency f from $n = 25$ genotypes simulated with $f^* = 1/3$ resulting in $n = \{12, 11, 2\}$ and with prior distribution $\mathbb{P}(f) = \text{Beta}(\alpha, \beta)$ with $\alpha = \beta = 0.5$ (black, dashed). (a) ABC-REJ posterior estimates from $5 \cdot 10^3$ simulations obtained with different thresholds δ . Percentages in parentheses give the acceptance rates of the ABC algorithms. (b) ABC posterior estimates from the first four iterations of a sequential Monte Carlo algorithm with $N = 2 \cdot 10^4$ and $\epsilon = 0.25$. (c) Posterior estimates obtained with ABC-GLM from the retained simulations under (a).

of the likelihood to the posterior, such that $\mathbb{P}(\theta | \|s - s'\| < \delta) \rightarrow \mathbb{P}(\theta)$ as $\delta \rightarrow \infty$. Since it is usually rather difficult to know a good choice of δ in practice, most applications use a slightly modified rejection algorithm in that a large number of n simulations are conducted first, and the threshold δ is then defined according to some tolerance ϵ such that $\mathbb{P}(\|s - s'\| < \delta | \theta)$ is ϵ for $\theta \sim \mathbb{P}(\theta)$.

1.3.5.1 Improved ABC Sampling Techniques

ABC-MCMC Given the obvious trade-off between computational effort and accuracy in inference inherent in ABC-REJ, multiple algorithms to improve sampling efficiency have been proposed. The first (Marjoram et al., 2003; Wegmann et al., 2009) was an ABC version of an MCMC with Metropolis–Hastings updates named *ABC-MCMC*:

1. Choose an appropriate initial parameter vector θ_0 . Set $t = 0$.
2. Propose a move $\theta_t \rightarrow \theta'_t$ according to some proposal kernel $q(\theta, \theta')$.
3. Simulate data D' with parameters θ'_t and calculate summary statistics $s' = T(D')$.
4. If $d = \|s_{\text{obs}} - s'\| < \delta$, accept move and set $\theta_{t+1} = \theta'_t$ with probability

$$h_{abc} = \min \left(1, \frac{\mathbb{P}(\theta'_t) q(\theta'_t, \theta_t)}{\mathbb{P}(\theta_t) q(\theta_t, \theta'_t)} \right),$$

- else reject move and set $\theta_{t+1} = \theta_t$.
5. Increment t and go back to step 2.

Note that in contrast to the standard algorithm (Section 1.3.3.2), the acceptance step is split into two parts: summary statistics s' resulting from a simulation conducted with proposed parameter vectors θ' need to be sufficiently close to the observed summary statistics s_{obs} , and are then accepted with probability h .

In contrast to the standard MCMC algorithm, and similarly to ABC-REJ, acceptance rates are therefore driven by the absolute likelihood $\mathbb{P}(\|s_{\text{obs}} - s'\| < \delta | \theta)$, rather than the ratio of the likelihood at the new and old locations (compare step 4 to equation (1.26) in Section 1.3.3.2). As a result, large thresholds δ often have to be used to ensure the chain is adequately exploring

the parameter space, particularly in high-dimensional models. For a promising way to circumvent this issue we refer to a recently proposed variant ABC-MCMC algorithm that updates one parameter at the time and uses summary statistics specific to that parameter (Kousathanas et al., 2016).

ABC-SMC Another algorithm with improved sampling efficiency is an ABC variant of particle filters termed *sequential Monte Carlo* (Sisson et al., 2007; Beaumont et al., 2009), leading to the *ABC-SMC* algorithm. The basic idea of particle filters is to have a population of parameter vectors (named particles) $\Theta_t = \{\theta^{(1)}, \dots, \theta^{(N)}\}$, from which a new sample Θ_{t+1} is generated by resampling the most relevant parameter vectors from Θ_t with variation. Here, relevance is quantified in terms of how well they match the target distribution (the posterior) captured by some weights $w_t = \{w_t^{(1)}, \dots, w_t^{(N)}\}$, and variation is added via a transition kernel q . While an initial sample Θ_0 is usually taken from the prior distribution, successive samples match the target distribution (the posterior) more and more closely.

As the posterior density cannot be calculated analytically in an ABC setting, a variation of the SMC with rejection step with decreasing tolerance δ_t is used (Sisson et al., 2007; Beaumont et al., 2009). However, it is usually difficult to make a good choice of δ_t *a priori*, and we thus present here a simple adaptive version (Wegmann et al., 2010) in which an acceptance rate ϵ is defined and δ_t is chosen accordingly, similar to what is often done in ABC-RE:

1. Sample a population of parameter vectors Θ_0 from the prior distribution $\theta_0^{(i)} \sim \mathbb{P}(\theta)$, $i = 1, \dots, N$, and set all weights $w_0^{(i)} = \frac{1}{N}$. Set $t = 0$.
2. Generate $j = 1, \dots, N/\epsilon$ parameter vectors from Θ_t by repeating the following steps:
 - (a) Sample a parameter vector $\theta_t^{(i)}$ according to the weights w_t .
 - (b) Propose a new parameter vector θ^j according to some proposal kernel $q(\theta_t^{(i)}, \theta^j)$.
 - (c) Simulate data D^j with θ^j , calculate summary statistics $s^j = T(D^j)$, and calculate distance $d^j = \|s_{\text{obs}} - s^j\|$.
3. Accept the N of the N/ϵ parameter vectors that have the smallest distance as a new population Θ_{t+1} and calculate their corresponding weights as

$$w_{t+1}^{(i)} \propto \frac{\mathbb{P}(\theta_{t+1}^{(i)})}{\sum_{k=1}^N w_t^{(k)} q(\theta_t^{(k)}, \theta_{t+1}^{(i)})}.$$

4. Increment t and go back to step 2.

It is easy to intuitively understand why this algorithm works, as the weights of each sampled parameter vector appropriately reflect the likelihood, the prior density, as well as the choice of proposal kernel: all simulations with $d_j > \delta_t$ have zero weight, while all others have a weight given by their prior density, penalized for how frequently they are proposed (see Beaumont et al., 2009, for a proof). More advanced adaptive schemes also exist (see, for example, Moreno-Mayar et al., 2018).

Example 1.24 (Allele frequency). In Figure 1.8(b) we show successive ABC posteriors for the model and data given in Example 1.23 obtained using an iterative SMC algorithm with $N = 5 \cdot 10^3$ and with $\epsilon = 0.25$. Hence, we generated $2 \cdot 10^4$ parameter vectors in each iteration, of which the 25% with smallest distance constituted the next population. Simulations were performed as in Example 1.23. ■

The above example illustrates a considerable improvement in efficiency over ABC-REJ: after four iterations (10^5 simulations in total), accuracy was comparable to those obtained with ABC-REJ at a tolerance $\delta = 0$, which resulted in an acceptance rate of 0.5%, and hence required 10^6 simulations to generate $5 \cdot 10^3$ samples. Similar improvements were also reported for other toy examples (e.g. Sisson et al., 2007) as well as for applications of ABC-MCMC for demographic inference (e.g. Wegmann et al., 2009).

However, a major drawback of these methods is that the simulation process is specific to s_{obs} , and has to be repeated if the data changes. This is particularly problematic when validating these methods, as the algorithm has to be rerun for each pseudo-observed data set. Yet a proper validation must be part of any ABC analysis if the summary statistics are not sufficient (see below).

1.3.5.2 Post-Sampling Adjustments

A complementary approach to reducing the simulation burden is to accept with large tolerances δ , but to correct for the too permissive sampling of parameter vectors by exploiting the local relationship between summary statistics and parameters. This relationship is easily learned from the retained simulations. The first such method was introduced by Beaumont et al. (2002) in their seminal paper coining the term ‘approximate Bayesian computation’, and these types of methods are now known as *post-sampling adjustments*.

ABC-REG Consider some ABC samples $\theta_1, \dots, \theta_n$ with corresponding summary statistics s_1, \dots, s_n sampled from the ABC posterior $\theta_i \sim \mathbb{P}(\theta | \|s_{\text{obs}} - s_i\| < \delta)$ with some threshold $\delta > 0$. The basic idea of the method by Beaumont et al. (2002) is now to fit a linear regression

$$\theta_i = \theta_0 + \theta_i^T \beta + \epsilon_i$$

and then to project the parameter vectors to s as

$$\tilde{\theta}_i = \theta_i - (s_i - s_{\text{obs}})^T \hat{\beta},$$

where $\hat{\beta}$ is the least-squares estimator of β . If the relationship between θ and s was truly linear, $\tilde{\theta}_i \sim \mathbb{P}(\theta | s_i = s_{\text{obs}})$. Linearity is unlikely to hold over the entire prior range $\mathbb{P}(\theta)$. But locally around the observed s_{obs} linearity is likely a very good approximation, in which case the effect of choosing a large δ can be mitigated without extra simulations (Beaumont et al., 2002). The linearity condition may be relaxed by fitting more elaborate models (Blum and François, 2009), but this usually requires more simulations.

ABC-GLM While applied successfully in many applications, ABC-REG suffers from two shortcomings. First, the projection $\theta_i \rightarrow \tilde{\theta}_i$ does not account for the prior distribution if $\mathbb{P}(\theta_i) \neq \mathbb{P}(\tilde{\theta}_i)$. Second, it does not easily lend itself to classic Bayesian techniques such as model averaging or model selection via marginal densities. While the former shortcoming could easily be fixed by introducing weights $w_i = \mathbb{P}(\theta_i)/\mathbb{P}(\tilde{\theta}_i)$ when estimating posterior densities, we will focus here on an alternative post-sampling adjustment *ABC-GLM* (Leuenberger and Wegmann, 2010) that fits a local likelihood model, rather than projecting posterior samples.

Consider again ABC samples of m -dimensional parameter vectors $\theta_1, \dots, \theta_n$ with corresponding summary statistics s_1, \dots, s_n sampled from the ABC posterior $\theta_i \sim \mathbb{P}(\theta | \|s_{\text{obs}} - s_i\| < \delta)$ with some tolerance $\delta > 0$. The likelihood of this truncated model $\mathcal{M}_\delta(s)$ is given by

$$\mathbb{P}_\delta(s|\theta) = \text{Ind}(s \in \mathcal{B}_\delta) \cdot \mathbb{P}(s|\theta) \cdot \left(\int_{\mathcal{B}_\delta} \mathbb{P}(s|\theta) ds \right)^{-1}, \quad (1.30)$$

where \mathcal{B}_δ is the δ -ball in the space of summary statistics around s_{obs} . $\text{Ind}(s \in \mathcal{B}_\delta)$ indicates if s is inside that ball and the integral gives the normalizing constant. The truncation similarly affects the prior distribution, which becomes

$$\mathbb{P}_\delta(\theta) = \frac{\mathbb{P}(\theta) \int_{\mathcal{B}_\delta} \mathbb{P}(s|\theta) ds}{\int \mathbb{P}(\theta) \int_{\mathcal{B}_\delta} \mathbb{P}(s|\theta) ds d\theta}. \quad (1.31)$$

Combining equations (1.30) and (1.31), we get (Leuenberger and Wegmann, 2010)

$$\mathbb{P}(\theta | s_{\text{obs}}) = \frac{\mathbb{P}(s_{\text{obs}}|\theta) \mathbb{P}(\theta)}{\int \mathbb{P}(s_{\text{obs}}|\theta) \mathbb{P}(\theta) d\theta} = \frac{\mathbb{P}_\delta(s_{\text{obs}}|\theta) \mathbb{P}_\delta(\theta)}{\int \mathbb{P}_\delta(s_{\text{obs}}|\theta) \mathbb{P}_\delta(\theta) d\theta}.$$

Thus, the posterior distribution of the full model $\mathbb{P}(s|\theta)$ for $s = s_{\text{obs}}$ given the prior $\mathbb{P}(\theta)$ is exactly equal to the posterior distribution of the truncated model $\mathbb{P}_\delta(s|\theta)$ given the truncated prior $\mathbb{P}_\delta(\theta)$. A reasonable approximation of the posterior $\mathbb{P}(\theta | s_{\text{obs}})$ is thus obtained by learning the truncated prior and truncated likelihood from the retained simulations.

The smoothed distribution of the retained parameter vectors constitutes an empirical estimate of the truncated prior $\mathbb{P}_\delta(\theta)$. Leuenberger and Wegmann (2010) proposed achieving this by placing a sharp Gaussian peak over each parameter value θ_i as

$$\mathbb{P}_\delta(\theta) = \frac{1}{n} \sum_{i=1}^n \phi(\theta | \theta_i, \Sigma_\theta),$$

where $\phi(\theta | \theta_i, \Sigma_\theta)$ is the multivariate normal density at θ with mean θ_i and covariance matrix $\Sigma_\theta = \text{diag}(\sigma_1, \dots, \sigma_m)$, determining the sharpness of the peaks and hence the amount of smoothing.

In order to estimate the truncated likelihood $\mathbb{P}_\delta(s|\theta)$, an educated guess of a parametric statistical model must be made. For simplicity and given the success of the linear version of ABC-REG, Leuenberger and Wegmann (2010) proposed using the general linear model (hence the name ABC-GLM)

$$s_i = s_0 + \mathbf{C}\theta_i + \epsilon_i,$$

where s_0 an intercept, \mathbf{C} is a matrix of constants, and ϵ a random vector with a multivariate normal distribution of zero mean and covariance matrix Σ_s . Such a GLM has the advantage that it also accounts for the strong correlation normally present between the components of the summary statistics, which can be learned robustly with ordinary least squares. Finally, the proposed GLM is guaranteed to converge to the true posterior distribution in the limit $\delta \rightarrow 0$.

Example 1.25 (Allele frequency). Let us revisit the allele frequency model described in Examples 1.23 and 1.24 by applying ABC-GLM to the retained simulations obtained with ABC-REJ for different tolerances (Figure 1.8(a)). As can be seen from the resulting posterior estimates in Figure 1.8(c), the ABC-GLM approach mitigates the effect of large tolerances almost completely. ■

1.3.5.3 A Note on Insufficient Summary Statistics

A particularly crucial aspect in any ABC analysis is the choice of summary statistics $s = T(D)$ for two often conflicting reasons. First, they should be sufficient or nearly sufficient for the inferred parameters in order for the ABC approximations to hold. Second, they should be low-dimensional for the comparison of distances to be meaningful. Multiple methods have been proposed to extract a low-dimensional set of summary statistics (e.g. Wegmann et al., 2009; Fearnhead and Prangle, 2012; Aeschbacher et al., 2012), yet none of these guarantees sufficiency. Since users resort to ABC methods for intractable likelihoods, it is often even impossible to formally test for sufficiency. A proper validation using pseudo-observed data sets with known parameter values is thus strongly advised (Wegmann et al., 2009; Leuenberger and Wegmann, 2010; Wegmann et al., 2010).

1.4 Model Selection

'All models are wrong but some are useful' is a famous quip attributed to George Box. Indeed, in many situations the choice of a specific model is not obvious and one may have several candidate models explaining the data D . Consequently, many different criteria have been proposed to compare model fit and choose the best model. Here we just present, as illustrations, the most commonly used methods in frequentist and Bayesian statistics: the *likelihood ratio statistic*, model posterior probabilities and *Bayes factors*. Apart from these, there exist other model selection techniques, such as cross-validation, *Akaike's information criterion* (AIC) and the *Bayesian information criterion* (BIC), to give both frequentist and Bayesian examples. For a thorough discussion, we refer particularly to the very readable introduction to this vast topic in Held and Sabanés Bové (2014, Ch. 7).

1.4.1 Likelihood Ratio Statistic

Suppose we have to choose between two models: a restricted model \mathcal{M}_1 , which is nested in the more complex model \mathcal{M}_2 . 'Nested' here means that \mathcal{M}_1 is similar to a version of \mathcal{M}_2 in which some parameters are restricted to particular values. Let $\mathcal{L}(\theta)$ be the likelihood of the more complex model \mathcal{M}_2 and $\hat{\theta}_2$ the ML estimates given data D . The ML estimates under \mathcal{M}_1 (i.e. under the parameter restriction) are $\hat{\theta}_1$. The two models can be compared using the *likelihood ratio statistic*

$$W = 2 \log \frac{\mathcal{L}(\hat{\theta}_2)}{\mathcal{L}(\hat{\theta}_1)} = 2(\ell(\hat{\theta}_2) - \ell(\hat{\theta}_1)), \quad (1.32)$$

where $\ell(\hat{\theta}_1)$ and $\ell(\hat{\theta}_2)$ are the log-likelihoods at the corresponding ML estimates.

Clearly, $\ell(\hat{\theta}_2) \geq \ell(\hat{\theta}_1)$ as any parameter combination under the restricted model \mathcal{M}_1 is also valid under \mathcal{M}_2 . Hence, W will always be non-negative, even if the data were drawn from the restricted model. The question is therefore whether W is sufficiently larger than zero to prefer the more complex \mathcal{M}_2 over the simpler model \mathcal{M}_1 . In a frequentist setting, we therefore interpret W as a test statistic and seek to calculate $p = \mathbb{P}(W \geq w | \mathcal{M}_1)$ to obtain W equal to or larger than an observed value w if model \mathcal{M}_1 was correct. The simpler model \mathcal{M}_1 is then rejected if this p -value is below a chosen significance level α (often 5%). This test, usually referred to as the *likelihood ratio test*, exploits the fact that W is asymptotically χ_k^2 distributed, if \mathcal{M}_1 was correct, with the number of degrees of freedom k equal to the number of restrictions (see, for example, Keener, 2011, Ch. 17 for the rather subtle proof).

Example 1.26 (Genotype frequencies). Let us illustrate this approach by testing whether some observed genotype frequencies $\mathbf{n} = \{n_0, n_1, n_2\}$ follow Hardy–Weinberg proportions. As the more complex alternative \mathcal{M}_2 , let us consider the multinomial genotype frequency model from Example 1.5, which has two free parameters π_0 and π_2 . From equation (1.6) we obtain the log-likelihood

$$\ell_2(\pi_0, \pi_2) = \log n! - \sum_{g=0}^2 \log n_g! + \sum_{g=0}^2 n_g \log \pi_g$$

with $\pi_1 = 1 - \pi_0 - \pi_2$. The unrestricted ML estimate for the two parameters was calculated in Example 1.5:

$$\hat{\theta}_2 = \left(\frac{n_0}{n}, \frac{n_2}{n} \right).$$

As restricted model \mathcal{M}_1 we use the genotype frequency model under Hardy–Weinberg from Example 1.2 with just one parameter f . This model is nested in \mathcal{M}_2 with the restriction that $\pi_0 = f^2$ and $\pi_2 = (1-f)^2$. From equation (1.2) we get the likelihood

$$\mathcal{L}_1(f) = \frac{n!}{n_0! n_1! n_2!} 2^{n_1} f^{n_1+2n_2} (1-f)^{n_1+2n_0}, \quad (1.33)$$

and the corresponding log-likelihood

$$\ell_1(f) = \log n! - \sum_{g=0}^2 \log n_g! + n_1 \log 2 + (n_1 + 2n_2) \log f + (n_1 + 2n_0) \log(1-f).$$

From this we easily get the ML estimate $\hat{f} = (n_1 + 2n_2)/2n$, and thus the restricted ML estimate

$$\hat{\theta}_1 = ((1-\hat{f})^2, \hat{f}^2) = \left(\frac{(2n_0 + n_1)^2}{4n^2}, \frac{(n_1 + 2n_2)^2}{4n^2} \right). \quad (1.34)$$

Now we can determine the value of the likelihood ratio statistic W as defined in equation (1.32). Reusing the numerical values $\mathbf{n} = \{12, 11, 2\}$ from Example 1.15 that were generated under Hardy–Weinberg with $f^* = 1/3$, we obtain $w = 0.27$. Since \mathcal{M}_2 has two parameters and \mathcal{M}_1 one parameter, W is approximately χ^2 distributed with one degree of freedom if \mathcal{M}_1 is correct: $W \sim \chi_1^2$. We get $p = \mathbb{P}(W \geq w | \mathcal{M}_1) \approx 1 - F_1(w) = 0.60$, where F_1 is the cumulative distribution of χ_1^2 , and thus we cannot reject the simpler Hardy–Weinberg model \mathcal{M}_1 at the $\alpha = 5\%$ level. ■

1.4.2 Bayesian Model Choice

1.4.2.1 Model Posterior Probabilities

Suppose we have to choose between K models $\mathcal{M}_1, \dots, \mathcal{M}_K$, possibly with different parameter sets $\theta_1, \dots, \theta_K$, to which we assign prior probabilities $\mathbb{P}(\mathcal{M}_1), \dots, \mathbb{P}(\mathcal{M}_K)$ summing to 1. In Bayesian manner, these models come with their respective prior distributions $\mathbb{P}(\theta_k | \mathcal{M}_k)$, $k = 1, \dots, K$. According to Bayes' theorem, the posterior probability of model i is given by

$$\mathbb{P}(\mathcal{M}_i | D) = \frac{\mathbb{P}(D | \mathcal{M}_i) \mathbb{P}(\mathcal{M}_i)}{\sum_{k=1}^K \mathbb{P}(D | \mathcal{M}_k) \mathbb{P}(\mathcal{M}_k)},$$

and similarly for any of the alternative models. Here,

$$\mathbb{P}(D | \mathcal{M}_k) = \int \mathbb{P}(D | \theta_k, \mathcal{M}_k) \mathbb{P}(\theta_k | \mathcal{M}_k) d\theta_k$$

is the marginal likelihood of model \mathcal{M}_k , obtained by integrating out the parameters θ_k .

1.4.2.2 Bayes Factors

While the posterior probabilities naturally reflect the uncertainty associated with model choice, a more commonly reported metric to choose among two models $\mathcal{M}_1, \mathcal{M}_2$ is the so-called *Bayes factor* BF_{12} , defined as the ratio of the marginal likelihoods

$$\text{BF}_{12} := \frac{\mathbb{P}(D|\mathcal{M}_1)}{\mathbb{P}(D|\mathcal{M}_2)}.$$

The Bayes factor is directly related to the *posterior odds*

$$\frac{\mathbb{P}(\mathcal{M}_1|D)}{\mathbb{P}(\mathcal{M}_2|D)} = \text{BF}_{12} \cdot \frac{\mathbb{P}(\mathcal{M}_1)}{\mathbb{P}(\mathcal{M}_2)},$$

which are the product of the BF_{12} and the prior odds. Hence, reporting the Bayes factor implies that the posterior odds can easily be recalculated for any choice of prior odds. Nonetheless, the Bayes factor is often directly used for model choice. Obviously, a $\text{BF}_{12} > 1$ is evidence in favor of the first model, but this evidence is usually only considered as strong if $\text{BF}_{12} > 20$ (Kass and Raftery, 1995).

Note that, in contrast to the likelihood ratio approach introduced above, Bayesian posterior probabilities and Bayes factors are not restricted to nested models. Indeed, there is no restriction on the choice of parameter sets θ_k , or on the number of alternative models considered.

Example 1.27 (Genotype frequencies). Let us compare the same models as in Example 1.26, but this time with aid of the Bayes factor. For our first model \mathcal{M}_1 we chose the beta prior from Example 1.15. We can calculate the marginal likelihood of \mathcal{M}_1 explicitly (Held and Sabanés Bové, 2014, p. 234). First, observe that from Bayes' formula,

$$\mathbb{P}(\mathbf{g}|\mathcal{M}_1) = \frac{\mathbb{P}(\mathbf{g}|f)\mathbb{P}(f)}{\mathbb{P}(f|\mathbf{g})}. \quad (1.35)$$

Here, $\mathbb{P}(\mathbf{g}|f)$ and $\mathbb{P}(f)$ are given by equations (1.33) and (1.19), respectively. From Expression (1.20) we further get

$$\mathbb{P}(f|\mathbf{g}) = \frac{f^{\alpha+n_1+2n_2-1}(1-f)^{\beta+n_1+2n_0-1}}{B(\alpha+n_1+2n_2, \beta+n_1+2n_0)}.$$

Inserting these terms into equation (1.35), we obtain for \mathcal{M}_1 the marginal likelihood

$$\mathbb{P}(\mathbf{g}|\mathcal{M}_1) = \frac{n!}{n_0!n_1!n_2!} 2^{n_1} \frac{B(\alpha+n_1+2n_2, \beta+n_1+2n_0)}{B(\alpha, \beta)}. \quad (1.36)$$

For our second model \mathcal{M}_2 , the multinomial model from Example 1.5, the conjugate prior is the *Dirichlet distribution* given by

$$\mathbb{P}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{g=0}^2 \pi_k^{\alpha_g-1},$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ are the prior parameters and where we used the *multivariate beta function*

$$B(\boldsymbol{\alpha}) = \frac{\prod_{g=0}^2 \Gamma(\alpha_g)}{\Gamma(\sum_{g=0}^2 \alpha_g)}.$$

The posterior will then also be Dirichlet distributed and one can show (see Murphy, 2012, Ch. 3.4) that the marginal likelihood is given by

$$\mathbb{P}(g|\mathcal{M}_2) = \frac{n!}{n_0!n_1!n_2!} \frac{\text{B}(\alpha + n)}{\text{B}(\alpha)}. \quad (1.37)$$

From equation (1.36) and (1.37) we get the posterior probabilities and the Bayes factor for our model comparison. Applied to the numerical values $n = \{12, 11, 2\}$ from Example 1.15 and using Jeffreys' priors with parameters $\alpha = \beta = 1/2$ for \mathcal{M}_1 and $\alpha_0 = \alpha_1 = \alpha_2 = 1/2$ for \mathcal{M}_2 (see Example 1.16) and equal prior probabilities $\mathbb{P}(\mathcal{M}_1) = \mathbb{P}(\mathcal{M}_2) = 1/2$, we get the posterior probabilities $\mathbb{P}(\mathcal{M}_1|g) = 0.29$ and $\mathbb{P}(\mathcal{M}_2|g) = 0.71$, and the Bayes factor $\text{BF}_{12} = 0.41$. Hence, and in line with the frequentist results from Example 1.26, there is no reason to prefer the more complex model \mathcal{M}_2 . ■

1.5 Hidden Markov Models

Many models in statistical genetics assume an underlying Markov model for which the states are not observed directly, but can only be inferred indirectly via some observations. Such models are called *hidden Markov models* (HMMs) and are frequently used to infer processes such as genetic drift and selection from time series data (e.g. Ferrer-Admetlla et al., 2016), or more commonly specific features along chromosomes, as we will see in the examples below. While HMMs were developed by Baum and Petrie (1966); Baum et al. (1970), we refer to Murphy (2012, Ch. 17.3) or Press (2007, Ch. 16.3.1) for excellent overviews of the theory. The goal of this section is to present the algorithms most commonly used in statistical genetics to conduct inference under HMMs.

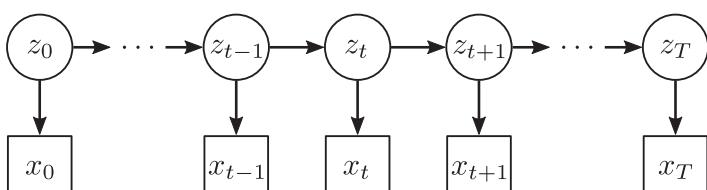
Formally, an HMM consists of a finite Markov chain with hidden (unobserved) states z_t running from time $t = 0$ to $t = T$. Moreover, we are given an observation model with emission probabilities $\mathbb{P}(x_t|z_t)$. While both the hidden state space and the time may be continuous, we will focus here on the standard type consisting of a DTMC (see Section 1.3.3.1) with discrete hidden states $z_t = 1, \dots, K$ (see the DAG in Figure 1.9). We denote by $x_{0:t}$ the vector of all observations up to time t :

$$x_{0:t} = (x_0, x_1, \dots, x_t).$$

If $\mathbb{P}(z_0)$ denotes the initial distribution of the states then the joint distribution of the HMM is

$$\mathbb{P}(z_{0:T}, x_{0:T}) = \mathbb{P}(z_0) \prod_{t=1}^T \mathbb{P}(z_t|z_{t-1}) \prod_{t=0}^T \mathbb{P}(x_t|z_t), \quad (1.38)$$

where $\mathbb{P}(z_t|z_{t-1})$ are the transition probabilities of the DTMC.



Example 1.28 (Chromosome painting). We will illustrate the use of HMMs in genetics through the problem of inferring local ancestry along the genome in an admixed individual from genotyping data. This inference is also called ‘chromosome painting’ as the aim is to ‘paint’ each segment along a chromosome according to a set of ancestry labels, such as ‘European’ and ‘African’ for African-American individuals. HMMs are a natural choice for modeling ancestry along a chromosome for two reasons. First, and because recombination occurs only at few locations per meiosis, ancestry segments generally span many loci and hence knowledge on the ancestry of one locus is very informative about ancestry at the next locus. Second, ancestry cannot usually be directly observed. Indeed, genotypes may rarely be assigned unambiguously to a specific ancestry because variation is usually shared among populations, particularly in humans. But since allele frequencies often differ among populations, the observed genotypes of an admixed individual do contain information about the underlying ancestry.

While most commonly used models use a reference panel for each ancestry label to exploit information contained in the haplotype structure (e.g. Wegmann et al., 2011; Price et al., 2009), we will adopt here a much more simplistic model for illustrative purposes. Specifically, let us denote by $g_l = 0, 1, 2$ the observed genotypes at loci $l = 0, \dots, L$ of an individual that derives its ancestry from two distinct sources. The goal is now to use an HMM to infer the hidden ancestries z_l for each locus, where $z_l = 0, 1, 2$ denotes the number of alleles with ancestry from the second source at locus l .

Three probability distributions need to be specified:

1. **Transition probabilities** $\mathbb{P}(z_l|z_{l-1})$. These will be given by the per locus recombination rates $\rho_l = \{\rho_{l1}, \rho_{l2}\}$ with which a chromosome of ancestry 2 recombines with a chromosome of ancestry 1, and vice versa. For the purpose of illustration we shall make here the artificial assumption of equal recombination rates among all successive loci: $\rho_l = \rho = \{\rho_1, \rho_2\}$. For a diploid individual we thus have the full transition matrix \mathbf{P} with elements $p_{ij} = \mathbb{P}(z_l = j|z_{l-1} = i, \rho)$ (see Section 1.3.3.1)

$$\mathbf{P} = \begin{pmatrix} (1 - \rho_2)^2 & 2\rho_2(1 - \rho_2) & \rho_2^2 \\ \rho_1(1 - \rho_2) & (1 - \rho_1)(1 - \rho_2) + \rho_1\rho_2 & (1 - \rho_1)\rho_2 \\ \rho_1^2 & 2\rho_1(1 - \rho_1) & (1 - \rho_1)^2 \end{pmatrix}. \quad (1.39)$$

2. **Initial state distribution** $\mathbb{P}(z_0)$. We will assume $\mathbb{P}(z_0)$ follows the stationary distribution of the DTMC given by \mathbf{P} . Under the chosen parameterization, the stationary distribution of ancestry along a single chromosome is given by $\rho_1/(\rho_1 + \rho_2)$ and $\rho_2/(\rho_1 + \rho_2)$ for ancestry 1 and 2, respectively. For a diploid individual we thus have

$$\mathbb{P}(z_0|\rho) = \begin{cases} \frac{\rho_1^2}{(\rho_1 + \rho_2)^2} & \text{if } z_0 = 0, \\ \frac{2\rho_1\rho_2}{(\rho_1 + \rho_2)^2} & \text{if } z_0 = 1, \\ \frac{\rho_2^2}{(\rho_1 + \rho_2)^2} & \text{if } z_0 = 2. \end{cases}$$

3. **Emission probabilities** $\mathbb{P}(g_l|z_l)$, These will be a function of the allele frequencies $f_l = \{f_{l1}, f_{l2}\}$ at loci l in populations 1 and 2, respectively. We thus have the emission matrix

$$\mathbf{E} = \begin{pmatrix} (1-f_{l1})^2 & 2f_{l1}(1-f_{l1}) & f_{l1}^2 \\ (1-f_{l1})(1-f_{l2}) & f_{l1}(1-f_{l2}) + (1-f_{l1})f_{l2} & f_{l1}f_{l2} \\ (1-f_{l2})^2 & 2f_{l2}(1-f_{l2}) & f_{l2}^2 \end{pmatrix}$$

with elements $e_{ij} = \mathbb{P}(g_l = j | z_l = i, \mathbf{f})$. ■

1.5.1 Bayesian Inference of Hidden States Using Forward-Backward Algorithm

A general aim in all HMMs is to infer the hidden states z_t , given either the evidence $\mathbf{x}_{0:t}$ up to the present moment (*filtering*, online) or the total evidence $\mathbf{x}_{0:T}$ (*smoothing*, offline). This can be done elegantly in a Bayesian setting where the interest lies in calculating the posterior probabilities $\mathbb{P}(z_t | \mathbf{x}_{0:T})$ of the hidden state at all time points t given the full data $\mathbf{x}_{0:T}$. The forward-backward algorithm offers an efficient way to calculate these posterior probabilities (and other important quantities of HMMs) and is motivated by the observation that conditioning the chain on a particular state z_t at time t separates past and future in the sense that the joint smoothing probability splits as

$$\mathbb{P}(z_t, \mathbf{x}_{0:T}) = \mathbb{P}(z_t, \mathbf{x}_{0:t})\mathbb{P}(\mathbf{x}_{t+1:T} | z_t) := \alpha_t(z_t)\beta_t(z_t). \quad (1.40)$$

Here, $\alpha_t(z_t) := \mathbb{P}(z_t, \mathbf{x}_{0:t})$ is the probability of observing the past data up to time t and of being in state z_t , and $\beta_t(z_t) := \mathbb{P}(\mathbf{x}_{t+1:T} | z_t)$ is the probability of all future data conditions on being in state z_t at time t . Both these probabilities can be calculated efficiently with recursions:

- 1. Forward recursion.** At first it seems that in order to calculate $\alpha_t(z_t)$ we have to sum over all possible paths to get to state z_t at time t , an impossible task in practice. But a moment's reflection makes clear that z_t is conditionally independent of everything but z_{t-1} and \mathbf{x}_t is conditionally independent of everything but z_t . Thus we can determine the α s with the recursion

$$\alpha_t(z_t) = \mathbb{P}(\mathbf{x}_t | z_t) \sum_{z_{t-1}=1}^K P(z_t | z_{t-1})\alpha_{t-1}(z_{t-1}).$$

The recursion starts with $\alpha_0(z_0) = \mathbb{P}(\mathbf{x}_0 | z_0)\mathbb{P}(z_0)$. The number of multiplications necessary to determine all α s is of order $K^2 \cdot T$. The α s can get very small. For this reason it is practical to normalize at each time step and to work with the filtered forward posteriors

$$\mathbb{P}(z_t | \mathbf{x}_{0:t}) = \frac{\mathbb{P}(z_t, \mathbf{x}_{0:t})}{\sum_{z_t} \mathbb{P}(z_t, \mathbf{x}_{0:t})} = \frac{\alpha_t(z_t)}{\sum_{z_t} \alpha_t(z_t)}.$$

- 2. Backward recursion.** The β s satisfy a backward recursion

$$\beta(z_t) = \sum_{z_{t+1}=1}^K \mathbb{P}(z_{t+1} | z_t)\mathbb{P}(\mathbf{x}_{t+1} | z_{t+1})\beta(z_{t+1}),$$

starting at $\beta(z_T) = 1$. As above, to avoid numerical difficulties it is recommended to normalize the β s in each backward pass.

From equation (1.40) we see that we only have to normalize the joint smoothing probability in order to obtain the smoothed posterior

$$\gamma_t(z_t) := \mathbb{P}(z_t | \mathbf{x}_{0:T}) = \frac{\alpha_t(z_t)\beta_t(z_t)}{\sum_{z_t} \alpha_t(z_t)\beta_t(z_t)}. \quad (1.41)$$

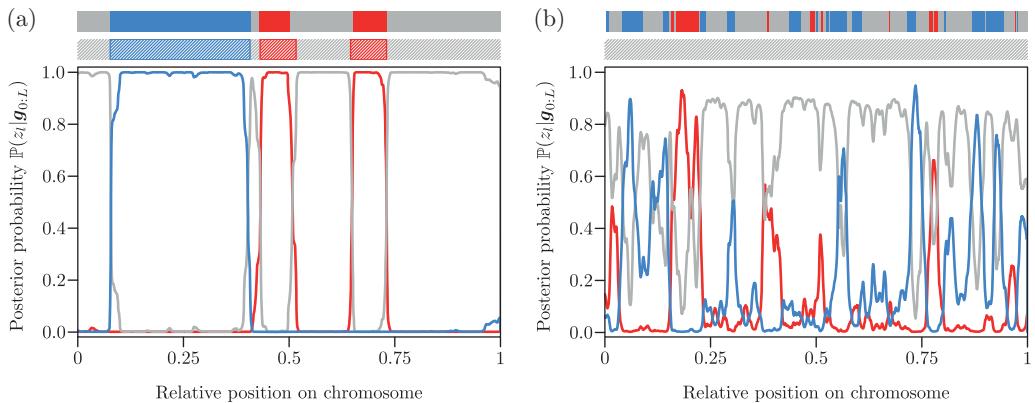


Figure 1.10 Estimates of the hidden ancestry at each locus $l = 0, \dots, L = 10^4$ from genotypes of an admixed individual simulated according to the model described in Example 1.28. The true ancestry makeup is shown as solid segments on top, followed by the most likely path $\hat{z}_{0:L}$ as inferred with the Viterbi algorithm (Example 1.30) is below. Finally, the posterior probabilities $\mathbb{P}(z_l | g_{0:L})$ inferred as detailed in Example 1.29 are given. In all cases, the colors red, gray and blue indicate $z_l = 0, 1$ and 2, respectively. (a) Data was simulated with $\rho_1 = \rho_2 = 2.5 \cdot 10^{-4}$ and the allele frequencies were set to $f_{11} = 0.4$ and $f_{12} = 0.6$ for all loci. (b) As (a) but with $\rho_1 = \rho_2 = 2.5 \cdot 10^{-3}$, $f_{11} = 0.45$ and $f_{12} = 0.55$ for all loci.

Example 1.29 (Chromosome painting). In Figure 1.10 we show posterior estimates $\mathbb{P}(z_l | g_{0:L})$ calculated using the forward-backward algorithm with the ρ and f used to simulate genotypes $g_{0:L}$ of a diploid individual. If ancestry segments are relatively large and allele frequencies sufficiently different between populations, the model allows for a near perfect inference of the true underlying ancestry (Figure 1.10(a) with $\rho_1 = \rho_2 = 2.5 \cdot 10^{-4}$, $f_{11} = 0.4$ and $f_{12} = 0.6$). However, the inference is much less powerful if segments are shorter and / or allele frequencies less informative (Figure 1.10(a) with $\rho_1 = \rho_2 = 2.5 \cdot 10^{-3}$, $f_{11} = 0.45$ and $f_{12} = 0.55$). ■

1.5.2 Finding the Most Likely Hidden Path (Viterbi Algorithm)

Sometimes it is interesting to infer the most likely path through the hidden states. Specifically, let us infer the path $\hat{z}_{0:T}$ that maximizes the joint distribution $\mathbb{P}(z_{0:T}, x_{0:T})$ given in equation (1.38). Note that the most likely path does not necessarily correspond to the sequence of mostly likely state estimates as $\mathbb{P}(z_t | x_{0:T})$ using the forward-backward algorithm. However, Viterbi (1967) introduced an equally elegant recursion, commonly called the *Viterbi algorithm*, to find the most likely path.

The key insight is that the most likely path to state z_t must consist of the most likely path through one of the states z_{t-1} . If we denote by $\delta_t(z_t)$ the probability of the most likely path to state z_t given the observations $x_{0:t}$, we therefore have

$$\delta_t(z_t) := \max_{z_{t-1}} \delta_{t-1}(z_{t-1}) \mathbb{P}(z_t | z_{t-1}) \mathbb{P}(x_t | z_t).$$

Initializing this recursion with $\delta_0(z_0) = \mathbb{P}(z_0) \mathbb{P}(x_0 | z_0)$ and computing until T , we obtain the final state of the most likely path

$$\hat{z}_T = \arg \max_{z_T} \delta_T(z_T).$$

We can now trace back the most likely path as follows. Knowing the state \hat{z}_t at time t of the most likely path, we can identify \hat{z}_{t-1} as

$$\hat{z}_{t-1} = \arg \max_{z_{t-1}} \delta_{t-1}(z_{t-1}) \mathbb{P}(\hat{z}_t | z_{t-1}) \mathbb{P}(x_t | \hat{z}_t).$$

To speed up this backward tracing, we may want to store the previous state for the most likely path leading to z_t for each state z_t during the δ recursion as

$$d_t(z_t) := \arg \max_{z_{t-1}} \delta_{t-1}(z_{t-1}) \mathbb{P}(z_t | z_{t-1}) \mathbb{P}(x_t | z_t).$$

The backward tracing then becomes $\hat{z}_{t-1} = d_t(\hat{z}_t)$. Similar to the forward-backward case, the δ s may get very small. Again we advise normalizing the δ s at each iteration. Alternatively, and in contrast to the forward-backward case, one can also conduct the entire recursion in logs.

Example 1.30 (Chromosome painting). In Figure 1.10 we show the most likely ancestry path for the model and data used in Examples 1.28 and 1.29. Similar to the Bayesian estimates of hidden states shown above (Example 1.29), the inference is rather accurate in the case of long ancestry segments (Figure 1.10(a)). If ancestry segments are short, however, the most likely path fails to detect any segments of homozygous ancestry (Figure 1.10(b)). The comparison to the Bayesian inference of the same data nicely illustrates the loss of information if a single point estimate $\hat{z}_{0:T}$ is considered, rather than an evaluation of the full uncertainty associated with the inference of hidden states. ■

1.5.3 MLE Inference of Hierarchical Parameters (Baum–Welch Algorithm)

So far we assumed transition and emission probabilities as given. In practice, however, they almost always have to be inferred from the observations. Since we do not know through which hidden states the trajectories went, we can apply the EM algorithm in order to estimate the parameters of the HMM, treating the hidden states as latent variables. This algorithm is usually referred to as the *Baum–Welch algorithm* after Leonard E. Baum and Lloyd R. Welch.

Let us first discuss the most general form of the Baum–Welch algorithm in which emission and transition probabilities are estimated individually for each time point t . This is obviously only possible if we possess multiple observations $x_{0:T}^n$, $n = 1, \dots, N$, of N independent runs of our Markov chain (e.g. genotypes of multiple admixed individuals).

Let us denote by θ the set of hierarchical parameters: the initial distribution $\mathbb{P}(z_0)$, the transition probabilities $\mathbb{P}(z_t | z_{t-1})$ between the states, and the emission probabilities $\mathbb{P}(x_t | z_t)$ of the observations. As introduced in Section 1.2.4, ML estimates of these parameters can then be obtained by iteratively maximizing the Q -function $Q(\theta | \theta') = \mathbb{E}[\ell_c(\theta) | \theta', x_{0:T}]$, where θ' denotes the vector of current parameter estimates and the complete-data log-likelihood is given by

$$\ell_c(\theta) = \log \mathbb{P}(z_0) + \sum_{t=1}^T \log \mathbb{P}(z_t | z_{t-1}) + \sum_{t=0}^T \log \mathbb{P}(x_t | z_t).$$

We get

$$\begin{aligned} Q(\theta|\theta') &= \sum_{n=1}^N \left[\mathbb{E} [\log \mathbb{P}(z_0) | \mathbf{x}_{0:T}^n, \theta'] + \sum_{t=1}^T \mathbb{E} [\log \mathbb{P}(z_t | z_{t-1}) | \mathbf{x}_{0:T}^n, \theta'] + \dots \right. \\ &\quad \left. + \sum_{t=0}^T \mathbb{E} [\log \mathbb{P}(x_t^n | z_t) | \theta'] \right] \\ &= \sum_{n=1}^N \left[\sum_{z_0} \gamma_0^n(z_0) \log \mathbb{P}(z_0) + \sum_{t=1}^T \sum_{z_t} \sum_{z_{t-1}} \xi_t^n(z_t, z_{t-1}) \log \mathbb{P}(z_t | z_{t-1}) + \dots \right. \\ &\quad \left. + \sum_{t=0}^T \sum_{z_t} \gamma_t^n(z_t) \log \mathbb{P}(x_t^n | z_t) \right], \end{aligned}$$

where the $\gamma_t^n(z_t)$ and $\xi_t^n(z_t, z_{t-1})$ denote the expectation weights $\mathbb{P}(z_t | x_{0:T}^n)$ and $\mathbb{P}(z_t, z_{t-1} | x_{0:T}^n)$, respectively. Importantly, these weights can be calculated efficiently with the standard forward-backward algorithm introduced above. Indeed, the $\gamma_t^n(z_t)$ are given by equation (1.41) and all

$$\xi_t(z_t, z_{t-1}) := \mathbb{P}(z_t, z_{t-1} | \mathbf{x}_{0:T}) \propto \alpha_{t-1}(z_{t-1}) \mathbb{P}(z_t | z_{t-1}) \beta_t(z_t) \mathbb{P}(x_t | z_t),$$

where one has to normalize such that the sum over z_t is 1.

The EM algorithm proceeds as follows. In the E-step, a forward-backward pass of the HMM is run using the current parameter estimates θ' and the relevant α s and β s are determined, from which we calculate the γ s, and ξ s. In the M-step, the new parameter estimates θ are then identified as those that maximize the function $Q(\theta|\theta')$. The update for the initial distribution is

$$\mathbb{P}(z_0) = \frac{1}{N} \sum_{n=1}^N \gamma_0^n(z_0),$$

which is just the expected number of the N independent runs that started in state z_0 . Similarly, we obtain

$$\mathbb{P}(z_t | z_{t-1}) = \frac{\sum_{n=1}^N \xi_t^n(z_t, z_{t-1})}{\sum_{n=1}^N \sum_{z_t} \xi_t^n(z_t, z_{t-1})}$$

for the new transition probabilities, corresponding to the expected number of transitions $z_{t-1} \rightarrow z_t$ among all independent runs. Finally, we get for the new emission probabilities

$$\mathbb{P}(x_t | z_t) = \frac{1}{N} \sum_{n=1}^N \gamma_t^n(z_t) \text{Ind}(x_t^n = x_t),$$

which are the expected number independent runs with observed state x_t that were in state z_t .

Unless the number of independent observations N is huge, the HMM is easily overparameterized. If there are K possible hidden states, the number of free transition probabilities to be inferred at each time point is $K(K - 1)$, and hence the number of observations N should be at least this large. For that reason, one often puts additional constraints on the different parameters. One might, for instance, assume that the model is homogeneous, that is, that the transition probabilities and the emission probabilities are independent of time, in which case the corresponding update equations above have to be averaged over t . Another strategy is to constrain the emission and transition probabilities with hierarchical parameters, as we did in Example 1.28. The M-step updates to these parameters, however, have to be derived anew for

each model. Note that both of these strategies may render the inference possible even from a single observed run.

Example 1.31 (Chromosome painting). Let us revisit the model given in Example 1.28 and derive a Baum–Welch algorithm to infer the hierarchical parameters ρ from the observations g_0, \dots, g_L of an admixed individual. The MLE of ρ is found iteratively as follows. In the E-step, the forward-backward algorithm is run to calculate all $\xi_l(z_l, z_{l-1})$ using the current estimates ρ' as shown above. Then, in the M-step, new estimates ρ are found by maximizing $Q(\rho|\rho')$, the relevant part of which is

$$Q_\rho(\rho|\rho') = \sum_{z_0} \gamma_0(z_0) \log \mathbb{P}(z_0|\rho) + \sum_{l=1}^L \sum_{z_{l-1}} \sum_{z_l} \xi_l(z_l, z_{l-1}) \log \mathbb{P}(z_l|z_{l-1}, \rho).$$

If L is large, we can safely ignore the terms corresponding to locus 0 since they will be numerically swamped by the other loci. From matrix \mathbf{P} given in equation (1.39) we read off

$$Q_\rho(\rho|\rho') = A_1 \rho_1^2 + A_2 \rho_2^2 + B \rho_1 \rho_2 + C_1 \rho_1 + C_2 \rho_2 + D,$$

where

$$\begin{aligned} A_1 &:= \sum_{l=1}^L [\xi_l(0, 2) - 2\xi_l(1, 2) + \xi_l(2, 2)], \\ A_2 &:= \sum_{l=1}^L [\xi_l(0, 0) - 2\xi_l(1, 0) + \xi_l(2, 0)], \\ B &:= \sum_{l=1}^L [-\xi_l(0, 1) + 2\xi_l(1, 1) - \xi_l(2, 1)], \\ C_1 &:= \sum_{l=1}^L [\xi_l(0, 1) - \xi_l(1, 1) + 2\xi_l(1, 2) - 2\xi_l(2, 2)], \\ C_2 &:= \sum_{l=1}^L [-2\xi_l(0, 0) + 2\xi_l(1, 0) - \xi_l(1, 1) + \xi_l(2, 1)], \end{aligned}$$

and the constant D is irrelevant for maximizing $Q_\rho(\rho|\rho')$. Taking the derivatives in respect to ρ_1 and ρ_2 , we get

$$\begin{aligned} \frac{\partial}{\partial \rho_1} Q_\rho(\rho|\rho') &= 2A_1 \rho_1 + B \rho_2 + C_1, \\ \frac{\partial}{\partial \rho_2} Q_\rho(\rho|\rho') &= 2A_2 \rho_2 + B \rho_1 + C_2. \end{aligned}$$

Setting these to zero, we obtain a linear system that is easily solved for ρ_1 and ρ_2 . ■

Acknowledgements

This chapter has evolved from notes and examples initially compiled for a graduate course on model-based inference in bioinformatics at the University of Fribourg, Switzerland. However, turning these notes into a coherent chapter would not have been possible without the extremely helpful and constructive feedback of Ida Moltke, Vivian Link and two anonymous reviewers.

References

- Aeschbacher, S., Beaumont, M.A. and Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics* **192**, 1027–1047.
- Allen, L.J.S. (2010). *An Introduction to Stochastic Processes with Applications to Biology*. CRC Press, Boca Raton, FL.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge.
- Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* **37**(6), 1554–1563.
- Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**(1), 164–171.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**, 370–418.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.
- Beaumont, M.A., Cornuet, J.-M., Marin, J.-M. and Robert, C.P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96**(4), 983–990.
- Blum, M.G.B. and François, O. (2009). Non-linear regression models for approximate Bayesian computation. *Statistics and Computing* **20**(1), 63–73.
- Broyden, C.G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics* **6**(1), 76–90.
- Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd edition. Duxbury, Pacific Grove, CA.
- Davison, A.C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.
- De Finetti, B. (2017). *Theory of Probability: A Critical Introductory Treatment*. John Wiley & Sons, Chichester.
- Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38.
- Diaconis, P. and Skyrms, B. (2017). *Ten Great Ideas about Chance*. Princeton University Press, Princeton, NJ.
- Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge.
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10):e1003905.
- Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society, Series B* **74**(3), 419–474.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**(6), 368–376.
- Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**, 521–565.
- Ferrer-Admetlla, A., Leuenberger, C., Jensen, J.D. and Wegmann, D. (2016). An approximate Markov model for the Wright-Fisher diffusion. *Genetics* **203**(2), 831–846.
- Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.
- Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal* **13**(3), 317–322.

- Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **31**(4), 1208–1211.
- Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation* **24**(109), 23–26.
- Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.
- Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science* **28**(706), 49–50.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109.
- Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference: Likelihood and Bayes*. Springer, Berlin.
- Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., Kirsanow, K., Cassidy, L.M., Martiniano, R., Strobel, M., Scheu, A., Kotsakis, K., Halstead, P., Triantaphyllou, S., Kyparissi-Apostolika, N., Urem-Kotsou, D.D., Ziota, C., Adaktylou, F., Gopalan, S., Bobo, D.M.D., Winkelbach, L., Blöcher, J., Unterländer, M., Leuenberger, C., Çilingiroğlu, Ç., Horejs, B., Gerritsen, F., Shennan, S.J., Bradley, D.G., Currat, M., Veeramah, K.R., Wegmann, D., Thomas, M.G., Papageorgopoulou, C. and Burger, J. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences of the United States of America* **113**(25), 6886–6891.
- Jaynes, E.T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.
- Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**(430), 773–795.
- Keener, R.W. (2011). *Theoretical Statistics: Topics for a Core Course*. Springer, New York.
- Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43.
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.
- Kousathanas, A., Leuenberger, C., Helfer, J., Quinodoz, M., Foll, M. and Wegmann, D. (2016). Likelihood-free inference in high-dimensional models. *Genetics* **203**(2), 893–904.
- Kousathanas, A., Leuenberger, C., Link, V., Sell, C., Burger, J. and Wegmann, D. (2017). Inferring heterozygosity from ancient and low coverage genomes. *Genetics* **205**(1), 317–332.
- Lange, K. (2004). *Optimization*. Springer, New York.
- Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer, New York.
- Lehmann, E.L. and Casella, G. (2006). *Theory of Point Estimation*. Springer, New York.
- Lehmann, E.L. and Romano, J.P. (2006). *Testing Statistical Hypotheses*. Springer, New York.
- Leuenberger, C. and Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics* **184**(1), 243–252.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21), 2987–2993.
- Lindsay, B.G. (1988). Composite likelihood methods. In N.U. Prabhu (ed.), *Statistical Inference from Stochastic Processes*, volume 80 of *Contemporary Mathematics*. American Mathematical Society, Providence, RI, pp. 221–239.
- Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A. and Wegmann, D. (2017). ATLAS: Analysis tools for low-depth and ancient samples. Preprint, bioRxiv 105346.

- Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**(26), 15324–15328.
- Markov, A.A. (1906). Extension of the law of large numbers to dependent quantities. *Izvestiya Fiziko-Matematicheskikh Obschestva Kazan University (2nd Ser.)* **15**, 135–156.
- McLachlan, G. and Krishnan, T. (2007). *The EM Algorithm and Extensions*. John Wiley & Sons, Hoboken, NJ.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**(6), 1087–1092.
- Moreno-Mayar, J.V., Potter, B.A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J.A., Albrechtsen, A., Malaspina, A.-S., Sikora, M., Reuther, J.D., Irish, J.D., Malhi, R.S., Orlando, L., Song, Y.S., Nielsen, R., Meltzer, D.J. and Willerslev, E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**(7687), 203–207.
- Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.
- Nocedal, J. and Wright, S.J. (1999). *Numerical Optimization*. Springer, New York.
- Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, Oxford.
- Press, W.H. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York.
- Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* **5**(6), e1000519.
- Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics* **35**(1), 1–20.
- Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, New York.
- Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York.
- Savage, L.J. (1972). *The Foundations of Statistics*. Dover, New York.
- Shanno, D.F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* **24**(111), 647–656.
- Sisson, S.A., Fan, Y. and Tanaka, M.M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **104**(6), 1760–1765.
- Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA.
- Tavaré, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**(2), 505–518.
- Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics* **35**(2), 335–353.
- Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42.
- Venkatesh, S.S. (2013). *The Theory of Probability: Explorations and Applications*. Cambridge University Press, Cambridge.
- Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* **13**(2), 260–269.
- Wegmann, D., Leuenberger, C. and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**(4), 1207–1218.

- Wegmann, D., Leuenberger, C., Neuenschwander, S. and Excoffier, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116.
- Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., Becker, L.C., Ruczinski, I., Beaty, T.H., Kardia, S.L.R., Meyers, D.A., Barnes, K.C., Becker, D.M., Freimer, N.B. and Novembre, J. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics* **43**(9), 847–853.
- Weinberg, W. (1908). Ueber den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für Vaterländische Naturkunde in Württemberg* **64**, 369–382.
- Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics* **149**(3), 1539–1546.
- Wright, S. (1921). Systems of mating I. the biometric relations between parent and offspring. *Genetics* **6**(2), 111–123.

2

Linkage Disequilibrium, Recombination and Haplotype Structure

Gil McVean and Jerome Kelleher

University of Oxford, UK

Abstract

Every chromosome carries a unique sequence of DNA, or haplotype. However, certain combinations of variants are shared between individuals. The extent of this sharing, or haplotype structure, is referred to as linkage disequilibrium, and its distribution in natural populations can be informative about diverse processes, from the rate of recombination to the influence of adaptive evolution. This chapter aims to provide a foundation for understanding haplotype structure and linkage disequilibrium, discussing how it can be measured, how it relates to the underlying genealogical process and how it can be informative about underlying molecular, historical and evolutionary processes. These ideas play an important role in modern statistical and computational genomics, influencing diverse applications from human migration to genome compression.

2.1 What Is Linkage Disequilibrium?

Every human genome carries a unique combination of DNA letters that act and interact with the environment and each other to create a unique individual. However, while every individual carries roughly a hundred new mutations (Scally and Durbin, 2012), most of the genetic diversity results from the shuffling of existing variants (or alleles) through the meiotic processes of chromosomal segregation and recombination. Consequently, while every genome may be unique, certain combinations of variants are shared: sometimes by just a few individuals, sometimes by a large fraction of the population. The term 'linkage disequilibrium' (LD) is broadly used to refer to the non-random sharing (or lack thereof) of combinations of variants. It would, perhaps, be better to talk about 'haplotype structure' or 'allelic association' (both terms we will also use). Nevertheless, though neither requiring linkage (physical association on a chromosome) nor being particularly a disequilibrium (e.g. one can discuss the equilibrium level of LD), the term LD has stuck.

To be clear about what we mean by LD, consider the three example data sets in Figure 2.1. In each case the marginal allele frequencies at each polymorphic site are approximately equal; what is different between the data sets is the degree of structuring of the variation or linkage disequilibrium. Specifically, Figure 2.1(a) shows a very high level of structuring, Figure 2.1(c) shows a very low level of structuring and Figure 2.1(b) shows something in between. Such differences in how genetic variation is structured point to differences in the underlying biological processes



Figure 2.1 Haplotype patterns with different levels of linkage disequilibrium. Each panel shows a sample of 100 chromosomes drawn from a population with (a) no, (b) low and (c) high recombination. Each row is a chromosome and the rarer allele at each site has been shaded black. Data sets all have approximately the same set of marginal allele frequencies, so the only difference is in terms of the degree of structuring, or linkage disequilibrium.

experienced by the populations from which the samples are drawn. Here, the primary difference is in the recombination rate; the data have been simulated with zero, some and lots of recombination, respectively. However, other molecular and historical processes, including mutation, natural selection, geographical isolation and changes in population size, will also influence the structuring of genetic variation in populations. It is the goal of population genetics to make inference about such processes from observations of genetic variation in contemporary populations. Naturally, we wish to use the relevant information contained in patterns of LD. The aim of this chapter is to explore how we can understand patterns of LD observed in empirical data.

It is also worth giving a more formal definition of LD. Consider a sample of chromosomes where polymorphism has been observed at a series of three loci, x , y , and z . For simplicity we will assume that each locus has only two alleles (A/a , B/b and C/c , respectively). The obvious description of the sample is in terms of the number of times we observe each haplotype, n_{ABC} , n_{abc} , etc., or alternatively their sample proportions, f_{ABC} , f_{abc} , etc. However, we can equivalently describe the data in terms of the marginal allele frequencies at each locus, f_A , f_a , etc., and a series of terms that reflect the extent to which combinations of alleles are found more or less frequently than expected, assuming independence between the alleles at each locus. For example:

$$f_{ABC} = f_A f_B f_C + f_A D_{BC} + f_B D_{AC} + f_C D_{AB} + D_{ABC}, \quad (2.1)$$

where

$$\begin{aligned} D_{AB} &= f_{AB} - f_A f_B, \\ D_{AC} &= f_{AC} - f_A f_C, \\ D_{BC} &= f_{BC} - f_B f_C, \\ D_{ABC} &= f_{ABC} - f_A D_{BC} - f_B D_{AC} - f_C D_{AB} - f_A f_B f_C. \end{aligned} \quad (2.2)$$

The D terms, which we will refer to as LD coefficients, therefore measure the difference between the observed frequency of pairs or triples of alleles and that expected from the marginal allele frequencies and other D terms of lower order (e.g. the D terms for triples contain the D terms for pairs). Similar expressions apply to any data set of any complexity (in terms of numbers of loci and numbers of alleles at each locus). However, the number of terms clearly explodes as the number of loci increases. Nevertheless, the point should be clear: patterns of genetic variation can be described in terms of the marginal allele frequencies and a series of terms relating to the degree of association between pairs, triples, etc. of alleles. It is these terms that are broadly referred to as LD. Informally, when a strong association exists between the alleles at two or more loci, they are often said to 'be in' or 'show' high or strong LD. Conversely, when no association exists and the loci are independent, they are sometimes said to be in 'linkage equilibrium'.

Before progressing, there is a rather subtle (but ultimately profound) point to make. In the previous paragraph we talked about how to describe genetic variation within a sample of

chromosomes. Historically, population genetics has focused more on describing genetic variation within *populations*: idealised entities consisting of (effectively) infinite numbers of individuals whose genetic composition can be described in terms of allele frequencies and coefficients of LD, just as in the sample. While the notion of a population is very helpful (and will be used extensively in this chapter) a focus on the sample has three benefits. First, the sample is all that we have, although we can of course use the sample to make inferences about populations. Second, in reality there is no such thing as a population, just a series of individuals with their own and overlapping histories. Thirdly, thinking about the history of the sample, specifically the genealogical history, provides a coherent way of linking what we observe in patterns of genetic variation to underlying biological and historical processes (Hudson, 1990). For these reasons, this chapter will focus heavily on the interpretation of LD within a sample.

This chapter is divided into three parts. In the first we explore how to summarise LD in empirical data. In the second we consider how simple probabilistic models of genealogical history can be used to explore the effects of various molecular and historical processes on patterns of LD. Finally, we discuss how these ideas influence approaches in modern statistical and computational genomics.

2.2 Measuring Linkage Disequilibrium

As stated above, our ultimate aim is to make inference about underlying biological and historical processes from patterns of genetic variation, including the structuring of variants, or LD. To motivate the problem, consider the three data sets shown in Figure 2.2. Each panel shows the inferred haplotypes for single nucleotide polymorphism (SNP) data from the same 100 kb

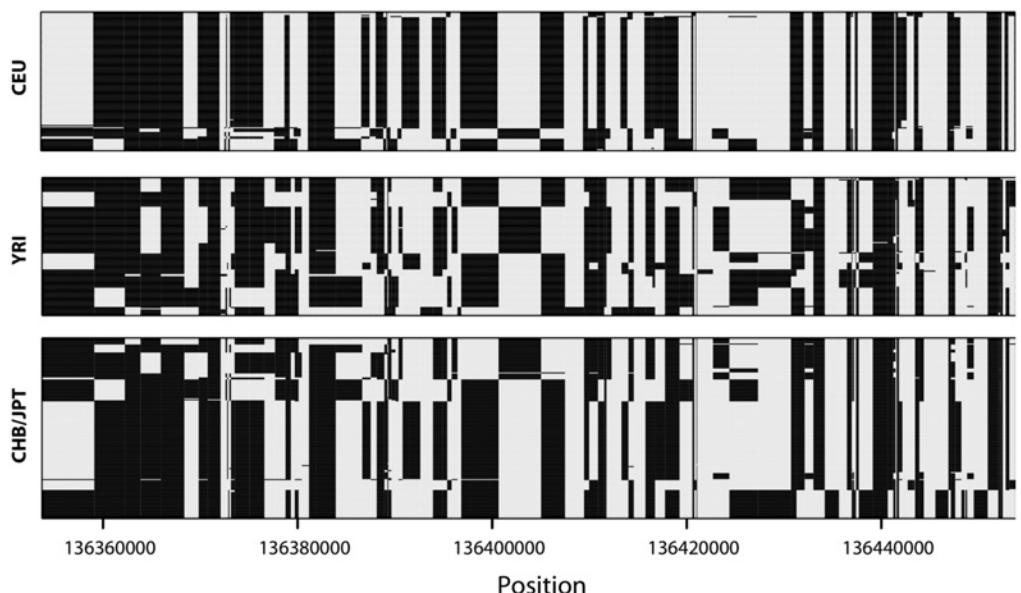


Figure 2.2 Haplotype structure in a 100 kb region surrounding the *Lactase* gene on human chromosome 2 for the four HapMap population samples (International HapMap Consortium, 2005). CEU, Individuals of European origin from Utah; YRI, Yoruba from Nigeria; CHB/JPT, Han Chinese from Beijing and Japanese from the Tokyo region (60 unrelated individuals in each of CEU and YRI, 90 in CHB/JPT). The CEU panel is dominated by a single haplotype that extends over the entire region. Much higher haplotype diversity is found in the other populations (see also Table 2.1).

surrounding the *Lactase* gene on human chromosome 2, but from three different samples of individuals: individuals of European ancestry living in Utah (referred to as CEU), individuals of Yoruba origin from Nigeria (referred to as YRI) and a combination of Han Chinese from Beijing and Japanese people living in Tokyo (referred to collectively as CHB/JPT); the data is from the International HapMap Project (International HapMap Consortium, 2005). This gene is important because a mutation in the promoter, found at high frequencies in European populations, results in the ability to digest lactose in milk persisting into adulthood; in other populations this ability ceases between 5 and 10 years of age (Swallow, 2003). The high frequency of this mutation in Europe is thought to have been the result of strong selection for lactose persistence associated with the innovation of dairy farming. This hypothesis is strongly supported by the genetic variation data, which shows the European population dominated by a single haplotype in marked comparison to the other populations (Bersaglieri *et al.*, 2004). Presumably the dominant haplotype is that on which the advantageous mutation arose and it has been swept up to high frequency through hitch-hiking (Maynard Smith and Haigh, 1974) with the selected mutation (see **Section 14.3.1.3** for more details on hitch-hiking).

However, suppose we knew nothing about the gene and its function. Faced with such haplotype data, how might we begin to assemble a coherent picture of the underlying processes? By way of an aside it should be noted that throughout we will typically assume that the haplotype ‘phase’ (see **Chapter 3**) is known, through a combination of genotyping in pedigrees and/or statistical analysis (Marchini *et al.*, 2006; Browning and Browning, 2011). Our first approach might be to present the data graphically, as in Figure 2.2. This shows the marked differences between populations and is clearly the most complete representation of the data (*it is the data*). Nevertheless, it would also be useful to have some low-dimensional summaries of the data that could be used to compare populations, or perhaps this region to somewhere else in the genome.

The aim of this section is to introduce a range of low-dimensional summaries of LD that could be applied to such data. In reality, we would also use low-dimensional summaries of the data that are functions of the SNP allele frequencies rather than their structuring – statistics such as Tajima’s D (Tajima, 1989), Fay and Wu’s H (Fay and Wu, 2000), Fu and Li’s D (Fu and Li, 1993) and F_{ST} (see **Section 14.4.3**). Indeed, it does not make much sense to separate out the analysis of LD from the analysis of allele frequencies. However, the diversity of measures of LD is sufficient to make devoting an entire section to them worthwhile. One thing that must be stressed, however, is that no single summary of LD is ‘best’ in the sense that it captures all information about underlying processes (e.g. no single-number summary is *sufficient* in the statistical sense that it captures all information about some parameter of a chosen model; see **Chapter 1**). Different summaries are more or less useful for identifying the effects of different underlying processes. In the analysis of empirical data it is therefore important to stress the use of multiple summaries, each of which may give some more insight.

2.2.1 Single-Number Summaries of LD

A natural starting point in the analysis of LD would be to ask whether there are any single-number statistics of the data that are informative about LD, in the same way, for example, that the average pairwise diversity or Tajima’s D are used in the analysis of allele frequency data. If the region is relatively short, one useful summary is simply the number of distinct haplotypes observed (the greater the LD, the fewer haplotypes). Similarly, haplotype homozygosity (the probability that two haplotypes picked without replacement from the sample are identical) indicates how skewed the haplotype frequencies are. There is, however, a problem with such summaries. As the length of the region surveyed increases (or more SNPs are typed), we would

eventually expect to reach a point where every haplotype is unique (and haplotype homozygosity is zero). We might therefore want a summary whose value is not so arbitrarily determined by the length of the region analysed.

There are many possibilities for such statistics. One approach is to attempt to break the data up into a series of blocks, each of which represents a region with (arbitrarily) low numbers of haplotypes and high haplotype homozygosity (Anderson and Novembre, 2003; Gabriel *et al.*, 2002; Wang *et al.*, 2002; Zhang *et al.*, 2002). The number of such blocks is therefore a measure of the degree of structure in the data. However, any choice about how to break the data up is arbitrary, and the concept of such ‘haplotype blocks’ has been little used in recent years. A related idea, motivated by the design of association studies to map the genetic basis of phenotypic variation, is to identify ‘tag’ SNPs that capture (in ways that are discussed below) variation within a region (Carlson *et al.*, 2004; de Bakker *et al.*, 2005; Johnson *et al.*, 2001). The number of tag SNPs required for a region is therefore a measure of how structured the variation is (e.g. if only two distinct haplotypes were observed only one tag SNP would be required). However, because there is no single most useful measure of how well variation is ‘captured’, numbers of tag SNPs are hard to compare between studies.

Another approach to summarising LD within a region is to estimate the influence of recombination. For example, nonparametric techniques (Hudson and Kaplan, 1985; Myers and Griffiths, 2003; Song *et al.*, 2005) can be used to estimate the minimum number of recombination events that have occurred in the history of the sampled chromosomes. Similarly, parametric, coalescent-based techniques (discussed later) can be used to estimate the ‘population recombination rate’, $4N_e c$, where c is the genetic distance across the region and N_e is the effective population size (Stumpf and McVean, 2003; Auton and McVean, 2012). Broadly speaking, a high recombination rate (with a large number of detectable recombination events) will tend to result in little genetic structuring (like Figure 2.1(c)), while low rates will tend to result in data sets like that in Figure 2.1(a). However, the association between recombination rate and LD is far from perfect, particularly if the region of interest has experienced adaptive evolution, the demographic history is not well approximated by a randomly mating population of constant size (e.g. there have been dramatic changes in population size or geographical subdivision), or there is considerable gene conversion.

By way of example, single-number summaries of LD for the *Lactase* gene region of Figure 2.2 are presented in Table 2.1. The strong structuring of the CEU sample is shown clearly in the reduced number of distinct haplotypes, the increased homozygosity and the smaller number of detectable recombination events relative to the other populations. The similarity in

Table 2.1 Single-number summaries of LD for the *Lactase* gene

	YRI	CEU	CHB/JPT
Number of chromosomes	120	120	180
Number of distinct haplotypes	34	18	35
Haplotype homozygosity	0.05	0.53	0.15
Recombination events*	23	10	23
Estimated $4N_e c/\text{kb}^\dagger$	0.12	0.10	0.07

*Lower bound on the minimum number of recombination events estimated by the method of Myers and Griffiths (2003).

†Estimated using the method of McVean *et al.* (2002) assuming a constant crossover rate and $\theta = 0.001$ per site ($\theta = 4N_e u$ is the population mutation rate).

haplotype numbers and detectable recombination events between YRI and CHB/JPT is complicated by the larger sample size of the latter; sub-samples of 120 chromosomes from CHB/JPT typically show values intermediate between YRI and CEU. Interestingly, the parametric estimates of the population recombination rates show the rate in CEU to be slightly higher than CHB/JPT, a pattern generally seen across the genome (Myers *et al.*, 2006; International HapMap Consortium, 2005). It is important to note that the model-based estimates of the population recombination rate assume neutrality (and a very simple demographic history). The action of natural selection is just one force that can result in different summaries of LD giving apparently conflicting indications as to the relative amount of LD.

2.2.2 The Spatial Distribution of LD

Any single-number summary of LD will fail to capture heterogeneity in the observed structuring of variation along a chromosome. Some regions may have greater structuring than others or some combinations of variants may show more or less structuring than others. For example, crossing over during meiosis will tend to lead to systematically lower levels of LD for variants at distantly separated loci compared to closely situated ones. Alternatively, gene conversion or mutational hotspots might create variants that are much more randomly distributed than their neighbours.

There are two approaches to summarising the spatial distribution of LD. One possibility is to make inferences about the spatial nature of the underlying biological or evolutionary processes (crossing over, gene conversion, mutation, natural selection, etc.). For example, LD in humans is strongly influenced by the concentration of meiotic crossing-over events into short regions called recombination hotspots. Consequently, inferences about the underlying recombination landscape reflect, at least in part, how LD changes along a chromosome (Jeffreys *et al.*, 2001, 2005; McVean *et al.*, 2005; International HapMap Consortium, 2005). The alternative is to make summaries of LD for subsets of the data (e.g. pairs of sites) and show, usually graphically, spatial patterns in these summaries. The following discussion focuses on two-locus summaries of LD as these are the most widely used summaries of LD for genetic variation data.

Consider a pair of loci, at which exactly two different alleles are observed in the population; these being A/a at the first locus and B/b at the second. These are most naturally thought of as single nucleotide polymorphisms, but they might also be insertion-deletion polymorphisms or restriction fragment length polymorphisms. For the moment assume that the haplotype phase of the alleles is known. As described above, the standard coefficient of LD between the alleles at the two loci is defined as

$$\begin{aligned} D_{AB} &= f_{AB} - f_A f_B \\ &= f_{AB}f_{ab} - f_{Ab}f_{aB}, \end{aligned} \tag{2.3}$$

where f_{AB} is the frequency of haplotypes carrying the A and B alleles and f_A is the marginal allele frequency of allele A . D_{AB} therefore measures the difference between the frequency of the AB haplotype and that expected if the haplotype frequencies were simply given by the product of the marginal allele frequencies. Any deviation from this expectation results in a non-zero value for D_{AB} , with a positive value indicating that the AB haplotype is found more often than expected assuming independence, and a negative value indicating that it is found less often than expected. Although (2.3) focuses on the AB haplotype, the coefficient of LD for any other haplotype is given by the simple relationship $D_{AB} = -D_{ab} = -D_{Ab} = D_{aB}$.

As described above, the coefficient is computed from the sample haplotype frequencies. However, we might also be interested in asking how the sample coefficient relates to that of

the population (if we believe that one exists). If we let D_{AB} be the population coefficient, the sample coefficient \hat{D}_{AB} has the properties (Hill, 1974)

$$\begin{aligned}\hat{D}_{AB} &= \hat{f}_{AB} - \hat{f}_A \hat{f}_B, \\ \mathbb{E}[\hat{D}_{AB}] &= \frac{n-1}{n} D_{AB}, \\ \text{Var}(\hat{D}_{AB}) &= \frac{1}{n} (f_A f_a f_B f_b + (f_A - f_a)(f_B - f_b) D_{AB} - D_{AB}^2).\end{aligned}\quad (2.4)$$

Here n is the sample size and \hat{f}_{AB} means the obvious estimate of f_{AB} (the population frequency) from the sample, n_{AB}/n , where n_{AB} is the number of AB haplotypes in the sample. The most important point about (2.4) is that the variance in the estimate is strongly influenced by the allele frequencies at the two loci. Furthermore, the range of values \hat{D}_{AB} can take is strongly influenced by the allele frequencies. If we arbitrarily define the A and B alleles to be the rarer alleles at each locus and enforce (without loss of generality) $\hat{f}_B \leq \hat{f}_A$, it follows that

$$-\hat{f}_A \hat{f}_B \leq \hat{D}_{AB} \leq \hat{f}_a \hat{f}_B. \quad (2.5)$$

The strong dependency on allele frequency for the standard coefficient of LD is an undesirable property because it makes comparison between pairs of alleles with different allele frequencies difficult. Consequently, several other measures of LD have been proposed that (at least in some ways) are less sensitive to marginal allele frequencies (Hedrick, 1987).

The most useful of these is the r^2 measure (Hill and Robertson, 1968). Consider assigning an allelic value, X_A , which is 1 if the allele at the first locus is A and 0 if the allele is a . Also assign an allelic value, X_B , with equivalent properties at the second locus. The quantity measured by (2.3) can then be interpreted as the covariance in allelic value between the loci. A natural way to transform the covariance is to measure the Pearson correlation coefficient,

$$r_{AB} = \frac{\text{Cov}(X_A, X_B)}{\sqrt{\text{Var}(X_A)\text{Var}(X_B)}} = \frac{D_{AB}}{\sqrt{f_A f_a f_B f_b}}. \quad (2.6)$$

In fact, for several reasons (not least because (2.6) has an arbitrary sign depending on how the allelic values are assigned), it is actually more useful to consider the square of the correlation coefficient,

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}. \quad (2.7)$$

The r^2 measure has many useful properties. First, as indicated by the lack of subscripts for D and r in (2.7), it has the same value however the alleles at the two loci are labelled. Second, there are simple relationships between the r^2 statistic and two features of interest, the power of association studies (Chapman *et al.*, 2003; Pritchard and Przeworski, 2001) and properties of the underlying genealogical history (McVean, 2002). Third, there is a direct relationship between the sample estimate of the r^2 coefficient, obtained by replacing population values by the sample values in (2.7), and the power to detect significant association, that is, to reject the null hypothesis $H_0 : D = 0$. An obvious test to consider is the contingency-table test where, under the null, the test statistic

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2.8)$$

is approximately χ^2 distributed with 1 df as the sample size tends to infinity. Here O_{ij} and E_{ij} are, respectively, the observed and expected counts of the ij haplotype, where the expectation

is calculated assuming independence between the loci. The relationship between (2.8) and the r^2 measure of LD is

$$X^2 = n\hat{r}^2. \quad (2.9)$$

The null hypothesis of no association can thus be rejected at a specified level, α , if $n\hat{r}^2$ is greater than the appropriate critical value of the test statistic.

Another test that might be considered is the likelihood ratio test, where the test statistic

$$\Lambda = 2 \log \left(\frac{\mathcal{L}(D = \hat{D})}{\mathcal{L}(D = 0)} \right) \quad (2.10)$$

is also approximately χ^2 distributed with 1 df under the null hypothesis. Here $\mathcal{L}(D)$ indicates the likelihood of the LD coefficient calculated using the multinomial distribution and the specified value of D . It can be shown that

$$\begin{aligned} \Lambda &= 2n \sum_{ij} (\hat{f}_i \hat{f}_j + \hat{D}_{ij}) \log \left(1 + \frac{\hat{D}_{ij}}{\hat{f}_i \hat{f}_j} \right) \\ &= n\hat{r}^2 + o(D^3). \end{aligned} \quad (2.11)$$

Consequently the test statistics (and hence the power) of the contingency-table and likelihood ratio tests are approximately equal and a function only of the sample size and observed r^2 measure of LD. Note that for small sample sizes the χ^2 approximation is unlikely to hold. In these circumstances it is possible to use standard permutation procedures or exact tests to estimate the significance of the observed correlation.

Although r^2 has many useful properties, it is far from the only measure in use. For example, the $|D'|$ measure (Lewontin, 1964) is defined as the absolute value of the ratio of the observed D to the most extreme value it could take given the observed allele frequencies:

$$|D'| = \begin{cases} \frac{-\hat{D}_{AB}}{\min(\hat{f}_A \hat{f}_B, \hat{f}_a \hat{f}_b)} & \hat{D}_{AB} < 0 \\ \frac{\hat{D}_{AB}}{\min(\hat{f}_A \hat{f}_B, \hat{f}_a \hat{f}_b)} & \hat{D}_{AB} > 0. \end{cases} \quad (2.12)$$

The main use of $|D'|$ is that it measures (in a nonparametric way) the evidence for recombination between the loci. A feature of (2.12) is that $|D'|$ can only ever be less than 1 if all four possible haplotypes are observed in the sample. If the mutation rate is low, such that repeat or back mutation is unlikely, then if all four possible haplotypes are observed it can be inferred that at least one recombination event must have occurred in the history of the sample (Hudson and Kaplan, 1985). Conversely, if anything less than the four combinations are observed the data are compatible with a history in which no recombination has occurred. Furthermore, the greater the recombination that has occurred, the more likely two loci are to be in linkage equilibrium (i.e., independent). So a value of $|D'| = 1$ can be interpreted as evidence for no recombination, while a value near 0 can be interpreted as evidence for considerable recombination. There is, however, a problem with such an interpretation for rare alleles. Even if all four combinations are present in the population, it may be unlikely to see all four in a finite sample if at least one haplotype is at low frequency (Devlin and Risch, 1995; Hedrick, 1987; Lewontin, 1988). For this reason, the interpretation of a $|D'|$ of 1 is highly dependent on the sample allele frequencies (as shown in Figure 2.3) and constructing confidence intervals for $|D'|$ is highly recommended (Gabriel *et al.*, 2002). Furthermore, if the primary interest of a study is to learn about

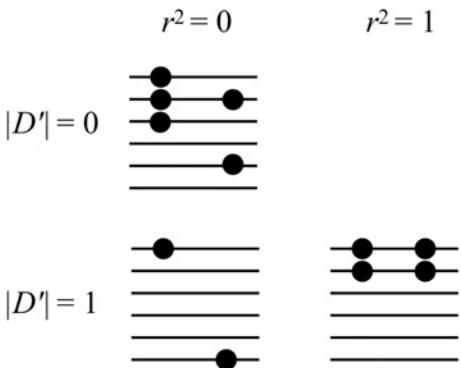


Figure 2.3 The relationship between sample configuration and the r^2 and $|D'|$ measures of two-locus LD. Each panel shows a configuration that corresponds to either high or low values of the two measures. Each bar is a chromosome and each circle is an allele. For the diagonal plots the measures agree. However, for the bottom-left panel, r^2 is near 0 while $|D'|$ is 1, demonstrating how the two measures focus on different aspects of the sample configuration.

recombination, it makes considerably more sense to use nonparametric or parametric approaches to learning about recombination directly.

As discussed above, an informative approach to summarising the spatial structure of LD is to compute two-locus statistics for all pairs of polymorphic loci and to represent these values graphically. Figure 2.4 shows three example data sets and their corresponding LD plots that demonstrate how the spatial distribution of LD can vary. In Figure 2.4(a) there is a tendency for closely sited alleles to show strong to moderate LD, while more distant ones show much weaker LD. In Figure 2.4(b) there are two strong blocks of LD separated by a point at which LD breaks down almost completely. In Figure 2.4(c) there is no apparent spatial structure to LD: alleles can be in strong association whether they are near or far away from each other. These differences in LD patterns reflect different underlying processes: a region with a moderate and constant recombination rate, a region with a strong recombination hotspot and low background recombination rate and a series of unlinked loci sampled from two highly differentiated populations,

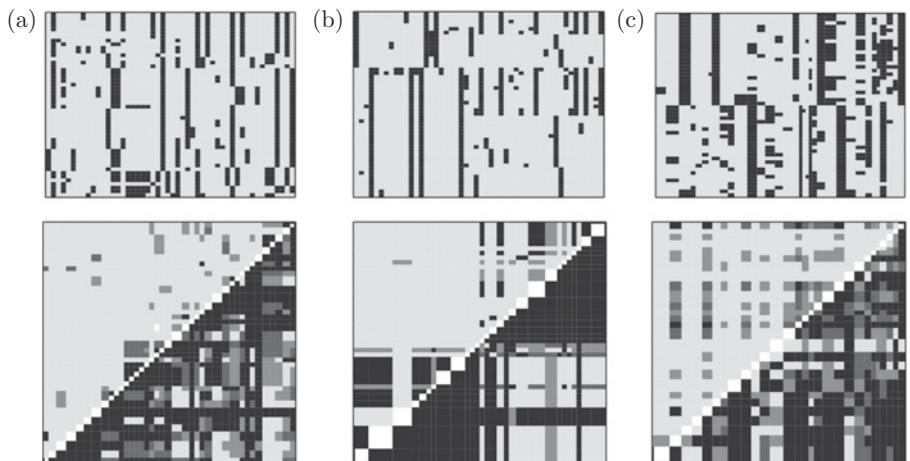


Figure 2.4 The spatial structure of LD. In each panel the upper plot shows the haplotypes and the lower plot shows a matrix of pairwise r^2 (upper left) and $|D'|$ values (bottom right) shaded by magnitude from black (values near 1) to light grey (values near 0). (A) In a region of constant crossover rate sites close to each other show strong LD, which gradually decreases the further sites are apart ($n = 50$, $\theta = 10$, $C = 30$). (B) In a region with a central recombination hotspot, LD appears block-like, with regions of very high LD separated by points at which the association breaks down ($n = 50$, $\theta = 10$, $C = 50$ concentrated on a single central hotspot). (C) Where LD is generated by the mixture of individuals from two differentiated populations there is no spatial structure to LD; sites near or far can be in strong LD ($n = 25$ in each population, 80 loci, data simulated under a beta-binomial model of differentiation (Balding and Nichols, 1995) with $F_{ST} = 0.9$).

respectively. Although the pictures are somewhat noisy it is clear that an understanding of the spatial distribution of LD can greatly help in the interpretation of underlying processes.

2.2.3 Various Extensions of Two-Locus LD Measures

The previous situation considered how to measure LD in the setting where loci are bi-allelic and the allelic phase is known. But in many situations neither may be true. If the sampled individuals are diploid and the haplotype phase of alleles is unknown it is possible to estimate the haplotype frequencies (Weir, 1996), using, for example, maximum likelihood inference (see **Chapter 1** for more details on maximum likelihood inference). For two bi-allelic loci, lack of phase information adds remarkably little uncertainty to estimates of LD (Hill, 1974), because the only two-locus genotype where phase cannot be accurately inferred is when the individual is heterozygous at both loci. For multi-allelic loci (such as microsatellites), haplotype estimation can also be achieved by maximum likelihood, for example by the expectation maximisation (EM) algorithm (see **Chapter 1** for more details on EM algorithms).

The key problem for multi-allelic systems is how to summarise LD. One approach, motivated by the relationship between the r^2 measure and the χ^2 test in the bi-allelic case, is to use the statistic (Hill, 1975)

$$Q = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^l \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^k \sum_{j=1}^l \frac{\hat{D}_{ij}^2}{\hat{f}_i \hat{f}_j}. \quad (2.13)$$

Here, as before, O_{ij} and E_{ij} are the observed and expected haplotype counts when there are k and l alleles respectively at the two loci. Again, under the null hypothesis $H_0 : D_{ij} = 0$, for all i and j , Qn will be approximately χ^2 distributed, though with $(k-1)(l-1)$ degrees of freedom. Of course, as the number of alleles increases, so that the expected haplotype counts tend to decrease, it becomes more important to use permutation methods, rather than the χ^2 approximation, to assess significance. There are also multi-allele versions of $|D'|$ (Hedrick, 1987). Again, it is natural to present the result of such analyses graphically.

2.3 Modelling Linkage Disequilibrium and Genealogical History

It should be clear by now that many different forces can influence the distribution of LD. These include molecular processes such as mutation and recombination, historical processes such as natural selection and population history and various aspects of experimental design (the choice of loci to study, the ascertainment process, the way in which the sampled individuals were assembled, the numbers of individuals sampled, etc.). If we are to have any hope of making useful inferences about the underlying processes from empirical data, we need to have a coherent framework to assess the way in which such forces can affect the patterns of variation we observe. A natural approach is to use simple probabilistic models to explore the distribution of patterns of LD we might observe under different scenarios. The aim of this section is to introduce such models and illustrate how they can be used to provide insights from empirical data. A key feature will be the idea that genealogical models, specifically the coalescent with recombination, provide a flexible and intuitive approach to modelling genetic variation. We will also show how features of the underlying genealogical history relate to properties of LD.

2.3.1 A Historical Perspective

Before introducing the coalescent perspective it is useful to provide a brief historical sketch of mathematical treatments of LD. These approaches have given considerable insight into the

nature of LD and, in contrast to most of the Monte Carlo based coalescent work, do provide simple analytical results about quantities of interest. The description we give below is not chronological. All of these models assume a population of constant size and random mating (see **Chapter 4** for more details on such population models).

2.3.1.1 The Relationship between LD and Two-Locus IBD

Intuitively, the level of LD observed between two loci must be related to the extent to which they share a common ancestry. Indeed, this is really the point of coalescent modelling. Completely linked loci share exactly the same ancestry and typically show high levels of LD, while unlinked loci have independent ancestries and typically show low LD. One way of quantifying the degree of shared ancestry between two loci is two-locus identity by descent (IBD). Single-locus IBD measures the probability that two chromosomes sampled at random from a population share a common ancestor before some defined point in the past (note that all chromosomes ultimately share a common ancestor, so the equilibrium value of IBD is 1). The two-locus version simply extends the notion to measure the probability that two chromosomes sampled at random share a single common ancestor at both loci before some defined point in the past (and that there has been no recombination on either pathway from the sample chromosomes to the common ancestor). Note that IBD does not refer to identity in state (i.e. whether two chromosomes carry the same alleles), rather it refers to relatedness. Over time IBD increases through genetic drift and decreases through recombination.

Sved (1971) derived an expression for the change in two-locus IBD over time as a function of the diploid population size, N , and the genetic map distance between the two loci, $c \ll 1$. Define Q as the probability of two randomly chosen individuals being IBD at both loci, conditional on their being IBD at one locus. The recurrence relation

$$Q_{t+1} = \frac{1}{2N}(1 - c)^2 + \left(1 - \frac{1}{2N}\right)Q_t(1 - c)^2 \quad (2.14)$$

then describes the dynamics of Q over time, leading to the equilibrium value

$$\tilde{Q} = \frac{1}{1 + C}, \quad (2.15)$$

where $C = 4Nc$. Sved assumes that $\mathbb{E}[r^2] \approx Q$ and so obtains

$$r^2 \approx \tilde{Q} = \frac{1}{1 + C}. \quad (2.16)$$

In short, the argument suggests that the expected value of the r^2 is near 1 for very small recombination rates and approaches $1/C$ for $C \gg 1$.

Although this approximation is a useful heuristic (Chakravarti *et al.*, 1984) and widely quoted (Jobling *et al.*, 2004), it is limited in application (Weir and Hill, 1986) because two chromosomes may share a common ancestor yet be different in allelic state due to a more recent mutation. Also, implicit within Sved's argument is the assumption that allele frequencies do not change over time. For these reasons, (2.16) will only be a good approximation when the time-scale over which two chromosomes at two loci may share a common ancestor before recombining is very short, which is only true for large values of C .

2.3.1.2 Matrix Methods and Diffusion Approximations

There are two alternative, and rather more rigorous, approaches to obtaining results about the expected value of LD statistics under simple population models. One approach is to use matrix recursions to describe the change in moments of LD statistics over time (Hill, 1975, 1977; Hill and Robertson, 1966, 1968). The other is to use a diffusion approximation, replacing the discrete nature of genes in populations by a continuous space of allele frequencies

(Ohta and Kimura, 1969a,b, 1971). Although these methods appear somewhat different at first, they are actually closely related, and can be used both to examine the dynamics of change in LD over time and to obtain expressions for quantities of interest at equilibrium. For example, although it is not possible to calculate the expected value of r^2 at equilibrium it is possible to calculate a related quantity,

$$\sigma_d^2 = \frac{\mathbb{E}[D_{AB}^2]}{\mathbb{E}[f_A(1-f_A)f_B(1-f_B)]}, \quad (2.17)$$

for a pair (or a set of equivalent pairs) of bi-allelic loci, where f_A and f_B are the allele frequencies at two loci. (The σ_d^2 quantity here takes the expectations of the numerator and denominator of (2.7); in particular, it does not denote variance.) Under the infinite-sites model (Karlin and McGregor, 1967; Kimura, 1969), the diffusion approximation leads to the solution (Ohta and Kimura, 1971)

$$\sigma_d^2 = \frac{10 + C}{22 + 13C + C^2}. \quad (2.18)$$

The same result can be obtained from models of bi-allelic loci with a low and symmetric rate of mutation between alleles (Hill, 1975; Ohta and Kimura, 1969a). Like the simple expression of Sved, this result predicts that, for large C , the expected value of r^2 is approximately $1/C$. The main difference between the predictions of (2.16) and (2.18) is for small C where (2.18) predicts a value considerably less than 1 (a value of $5/11$ for $C = 0$). Figure 2.5 compares estimates from Monte Carlo coalescent simulation. Neither approximation provides a particularly accurate prediction for the expected value of r^2 , unless rare variants (loci where the rare variant is less than 10% in frequency) are excluded. Nevertheless, (2.18) does predict the general shape of the decrease in average r^2 with increasing C .

2.3.2 Coalescent Modelling

The single most striking feature about Figure 2.5 is just how noisy LD is; the mean value of r^2 between loci at a given genetic distance captures very little of the complexity of the full distribution. This has two implications. First, it is hard to obtain an intuitive understanding of LD by thinking about ‘expected’ values of LD statistics. Second, the analysis of empirical data by comparing observed LD statistics to their ‘expected’ values is likely to be only weakly informative.

In order to capture the full complexity of LD patterns it is necessary to use stochastic modelling techniques to simulate the types of patterns one might observe under difference

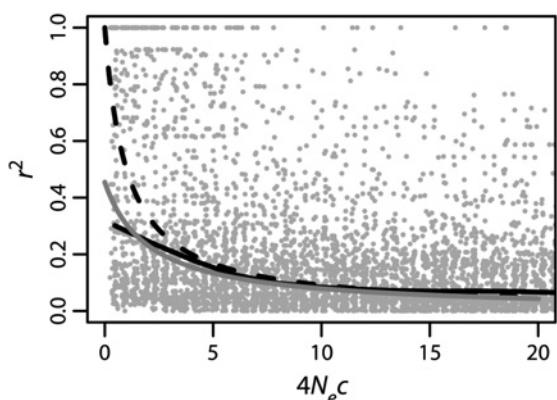


Figure 2.5 Analytical approximations for the expected r^2 between pairs of alleles as a function of the population genetic distance, $4N_e c$, between sites. Sved’s approximation (2.16): dashed black line. Ohta and Kimura’s approximation (2.18): solid grey line. Also shown (grey dots) are the values of r^2 between pairs of alleles at the corresponding genetic distance obtained from a single coalescent simulation with 50 chromosomes, $\theta = 100$, $C = 100$ and the sliding average (solid black line) for all pairs of sites where the minor allele is at frequency of at least 10%. Although the Ohta and Kimura approximation performs quite well at predicting the mean r^2 , its predictive power for any pair of sites is extremely poor due to high variance.

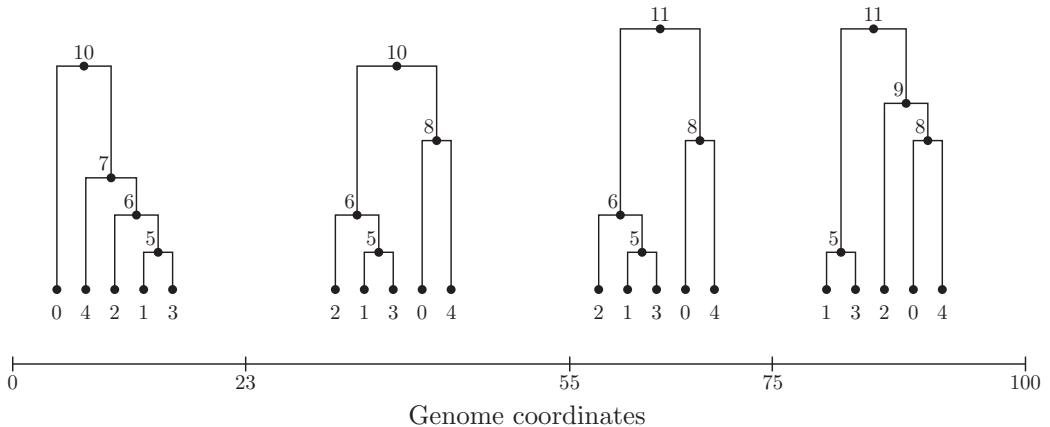


Figure 2.6 A sequence of trees produced by the coalescent with recombination for five samples and 100 loci. Each tree in the sequence differs from the previous by a subtree-prune-and-regraft operation (Wiuf and Hein, 1999; Song, 2006). For example, the tree covering the interval 23–55 is obtained from the first tree by pruning the subtree rooted at 4 and inserting it into the branch joining 0 to 10. As a result, adjacent trees differ from each other, but are highly correlated. For example, the subtree rooted at 6 is identical in all trees from 0 to 75. See also Figure 2.8 for an illustration of a subtree-prune-and-regraft operation.

scenarios. The advent of coalescent modelling (Hudson, 1983b; Kingman, 1982; Tajima, 1983), and specifically the coalescent with recombination (Hudson, 1983a), has led to a revolution in the way genetic variation data is approached. Coalescent models focus on properties of the sample by considering the genealogical history that relates a set of chromosomes to each other (see **Chapter 5**). For recombining data, a single tree is replaced by a sequence of correlated trees (Hudson, 1983a; Kelleher *et al.*, 2016). Figure 2.6 shows the sequence of trees resulting from a simulation of the coalescent with recombination. The trees are highly correlated, and share much of their structure. Mutations that lead to variation within the sample occur along the branches of these trees. Consequently, the structure of LD reflects the correlation structure of the underlying genealogical history. Put another way, the coalescent perspective states that the best way of understanding genetic variation is to think about the structure of the underlying genealogy. Different evolutionary forces have different effects on the shape and correlation structure of these underlying genealogies. Although this view is strongly driven by a neutral perspective (the mutations we observe have not themselves influenced genealogical history), the effects of certain types of natural selection, such as selective sweeps or balancing selection, on patterns of linked neutral variation can also be considered from a coalescent or genealogical viewpoint.

In addition to the various theoretical insights that coalescent theory has made possible, a genealogical approach has greatly enabled the analysis of genetic variation through the ability to simulate data under various evolutionary models. In the rest of this subsection we will describe how a genealogical framework can be used to understand the distribution of LD.

2.3.2.1 LD Patterns in the Absence of Recombination

It may sound strange, but we can actually learn a lot about LD by studying its behaviour in regions where there is no recombination (Slatkin, 1994). Consider the three data sets and their corresponding genealogies in Figure 2.7, corresponding respectively to simulations with a constant population size, a growing population and a population that has experienced a very strong recent bottleneck. Apart from the differences in the numbers of polymorphic sites, there are also strong differences in their structuring. First note that any two mutations that occur on

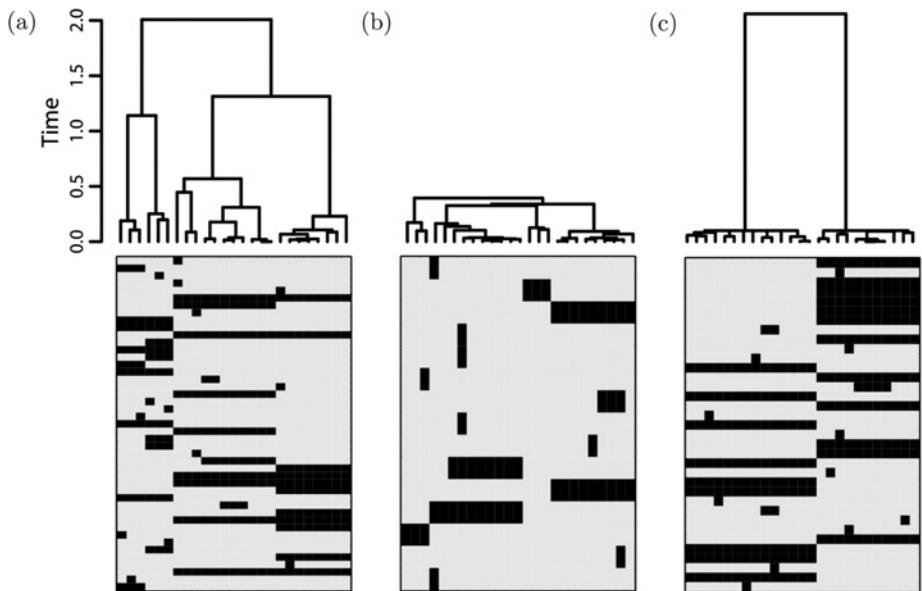


Figure 2.7 Linkage disequilibrium in non-recombining regions. The data in each of the regions are shown below the underlying genealogy as a column for each chromosome. Despite each example having no recombination, there can be low, moderate or high LD, depending on the structure of the underlying genealogy. Data simulated with (a) a constant population size, (b) a strongly growing population or (c) a population that has experienced a severe recent bottleneck.

the same branch in the tree will occur in exactly the same members of the sample, hence be in ‘perfect’ LD (i.e. $r^2 = 1$). Next note that mutations on different branches will be only very weakly correlated ($r^2 \ll 1$) if the branches are in different parts of the tree and particularly if one or both are near the tips. It follows that the extent of LD reflects the extent to which the tree shape is dominated by a few long branches, as in the case of the bottlenecked population (on which lots of mutations occur and which are therefore in perfect association), or not, as in the case of the growing population (mutations occur on branches in different parts of the tree and therefore typically show weak association). It follows that (at least for the r^2 measure) LD is strongest for the bottlenecked population and weakest for the growing one. The constant-size population shows a mixture of highly correlated and weakly correlated alleles, reflecting the distribution of mutations all across the tree. Of course, because of the inherent stochasticity in the coalescent process it would be unwise to make inferences about population history from a single non-recombining locus that showed the patterns in either Figure 2.7(b) or Figure 2.7(c).

Digressing slightly, it is interesting to note how classical population genetics theory concerning the distribution of genetic variation under the infinite-alleles model can be related to the structure of LD in infinite-sites models without recombination. For example, the number of distinct haplotypes in the sample is equivalent to the number of distinct alleles, k . The classic result then gives an expectation for these quantities for the case of constant-sized neutral populations (Ewens, 1972),

$$\mathbb{E}[k] = \sum_{i=0}^{n-1} \frac{\theta}{i + \theta}. \quad (2.19)$$

Here $\theta = 4N_e u L$ is the population mutation rate over the region of interest, where u is the per site, per generation mutation rate and L is the number of sites. Similarly, the Ewens sampling

formula (Ewens, 1972) describes the distribution of the numbers of each distinct haplotype conditional on the total number of observed haplotypes,

$$\mathbb{P}[n_1, n_2, \dots, n_k | k, n] \propto \frac{1}{n_1 n_2 \dots n_k}. \quad (2.20)$$

This result inspired the first statistical test for the hypothesis that the region of interest is evolving neutrally: the Ewens–Watterson homozygosity test (Watterson, 1977, 1978), which compares the observed (haplotype) homozygosity to the distribution expected from the above formula. Indeed, it was thinking about the effect of recombination on this test that first led to the development of the coalescent model with recombination (Hudson, 1983a; Strobeck and Morgan, 1978). Recombination tends to increase the number of observed alleles (haplotypes) and reduces the skew in allele (haplotype) frequency, resulting in a systematic decrease in (haplotype) homozygosity.

2.3.2.2 LD in Recombining Regions

When recombination occurs within a region, different positions will have different, though correlated, trees (Figure 2.6). This raises a series of questions. How does recombination change tree structure? What is the relationship between correlation in tree structure and LD? How should we measure the correlation in trees? To answer these questions it is first helpful to consider the two complementary ways in which we can view the construction of marginal genealogies. The first approach is to work backwards in time, modelling the ancestry of our sample through the effects of common ancestor and recombination events (Hudson, 1983a, 1990; Kelleher *et al.*, 2016). Common ancestor events merge the ancestral material of a pair of lineages, and may give rise to coalescences in the marginal trees. Recombination events split the ancestral material for a given lineage, creating two lineages that can then evolve independently (potentially creating different marginal trees on either side of the split). Equivalently, we can view this backwards-in-time process as a graph (the ancestral recombination graph), in which we have nodes for common ancestor and recombination events (Griffiths, 1991; Griffiths and Marjoram, 1997). This backwards-in-time process is easy to describe and simulate, but deriving analytical results can be challenging due to the complex distribution of blocks of ancestral material among ancestors.

The second approach to modelling genealogies under the coalescent with recombination is to think about how trees ‘evolve’ along a sequence through recombination (Wiuf and Hein, 1999). While this spatial process is considerably more complex than the backward-in-time approach outlined above, it is amenable to an approximation known as the sequentially Markov coalescent (SMC) (McVean and Cardin, 2005; Marjoram and Wall, 2006). The relative simplicity of the SMC (and slightly refined SMC') has enabled significant progress to be made on a variety of different inference problems (Li and Durbin, 2011; Harris and Nielsen, 2013; Sheehan *et al.*, 2013; Schiffels and Durbin, 2014; Carmi *et al.*, 2014; Rasmussen *et al.*, 2014; Zheng *et al.*, 2014; Terhorst *et al.*, 2017). The SMC has been shown to be a good approximation to the coalescent with recombination in the pairwise case (Wilton *et al.*, 2015) and for sample sizes up to 4 (Hobolth and Jensen, 2014), and in simulations produces patterns of LD indistinguishable from the full coalescent (McVean and Cardin, 2005; Marjoram and Wall, 2006). In the following, we give a sense of how genealogical history changes through recombination by considering how to simulate a sequence of trees along a unit region over which there is a constant recombination rate of C under this approximation.

1. Simulate a coalescent tree (i.e. no recombination) at the far left-hand edge of the region. The total tree length (in time scaled by $2N_e$ generations) is T_L .

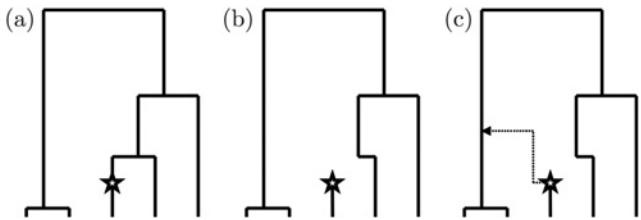


Figure 2.8 Recombination as a point process along the sequence (Wiuf and Hein, 1999). (a) At any point along the sequence there is an underlying genealogical tree. The distance to the next recombination event is exponentially distributed with rate dependent on the total branch length of the current tree and the recombination rate. Having chosen the physical position of the recombination event, the position on the tree, indicated by the star, is chosen uniformly along the branch lengths. (b) In an approximation to the coalescent process, called the SMC (McVean and Cardin, 2005), the branch immediately above the recombination point is erased, leaving a floating lineage. (c) This lineage then coalesces back into the remaining tree and the process is repeated. In the coalescent, no branch can be erased leading to non-Markovian behaviour.

2. The distance along the region to the next recombination event is exponentially distributed with rate $T_L C/2$. If the next point lies within the unit interval continue, otherwise stop.
3. Choose a point to recombine uniformly along the branches of the tree. Erase the remainder of the branch immediately above the chosen point.
4. Allow the recombined lineage to coalesce back to the remaining lineages at a rate proportional to the number of non-recombined lineages present (note this has to be updated if there are coalescent events among these). Note also that the recombined lineage could coalesce beyond the most recent common ancestor (MRCA) of the current tree.
5. Now assign the total length of the new tree to T_L . Return to step 2.

These steps are illustrated in Figure 2.8. Several features should be clear from this approximation. First, trees change over the region by small steps. Of course, lots of steps (resulting from large recombination rates) result in lots of changes, so the tree at the right-hand side of the sequence looks very different from that at the left-hand side of the sequence. Second, because of the structure of coalescent trees, many of the recombination events occur deep in the tree when there are relatively few lineages. These often have remarkably little effect on the distribution of genetic variation, and recombination events during the phase when only two lineages are present are essentially undetectable. Third, if a pair or group of sequences shares a very recent common ancestor, this part of the tree will persist over considerable genetic distances. Consequently, it is often possible to identify pairs or small groups of sequences that are identical over extremely long regions. To illustrate the effect, consider the statistic T_S , the sum of the lengths of the branches shared by the trees at the start and end of a sequence. Under the SMC the expectation is

$$\mathbb{E}[T_S] = 2 \sum_{i=1}^{n-1} \frac{1}{i + C}. \quad (2.21)$$

For $n > C$ the proportion of the total expected tree length that is shared is approximately $1 - \log(C)/\log(n)$ (note that in the coalescent the shared time is likely to be slightly higher). Consequently, for large sample sizes a considerable proportion of the tree can be shared even when the total recombination rate is very high. For example, with 1000 sequences, over 15% of the total expected time is shared by points separated by $C = 400$, which in humans of European origin corresponds to about 1 cM. Of course, if the tree shape is strongly dominated by large recent clades, as for example happens if there has been a recent and strong but partial selective sweep, such parts of the tree can persist over much greater genetic distances.

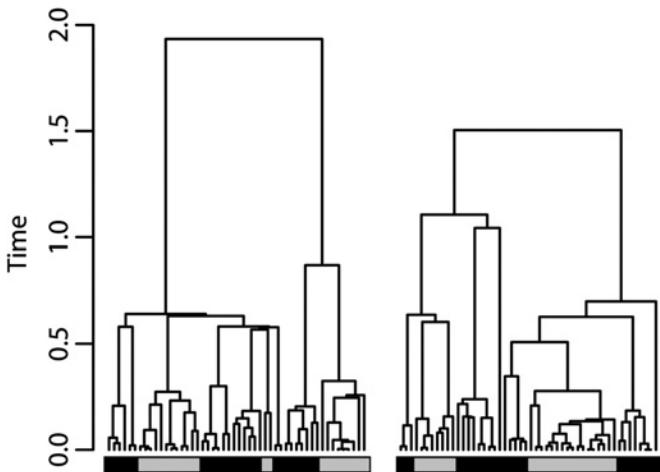


Figure 2.9 Correlations in genealogical history at unlinked loci. Individuals from two populations, indicated by the grey and black bars below, have been sampled at two unlinked loci. Although the trees are independent, there is nevertheless genealogical correlation because a pair of individuals sampled from within a population is likely to have a more recent common ancestor at both loci than a pair of individuals where one from each population has been sampled. Data simulated with 25 individuals sampled from each of two populations of equal size that diverged exactly N_e generations ago from an ancestral population of the same size.

2.3.2.3 LD in Populations with Geographical Subdivision and/or Admixture

Although the notion of a tree changing along the sequence provides useful insights into the nature of LD, there is one aspect that, at least at first glance, it fails to describe: LD between alleles at unlinked loci in structured populations. As demonstrated in Figure 2.4, when the sample contains individuals from one or more populations (and/or individuals who are admixed) LD can exist between even unlinked loci because of variation in allele frequencies between populations. How can we understand this phenomenon within a genealogical context?

The answer is simple. LD is determined by the correlation in genealogical history along a chromosome. Recombination acts to break down such correlations, but if there are biases in terms of which lineages can coalesce, some correlation will persist indefinitely. Figure 2.9 illustrates this idea by considering independent coalescent trees sampled from a pair of populations that diverged some time ago. Despite their independence, both trees show a strong clustering of individuals from the same populations. In short, while directly sharing genealogical trees results in genealogical correlation and hence LD, genealogical correlations can also arise indirectly by forces shaping the nature of coalescence.

2.3.3 Relating Genealogical History to LD

It should be clear by now that LD is a reflection of correlation in the genealogical history of samples. Informally, if the coalescence time for a pair of sequences sampled at a given locus is informative about the coalescence time for the same pair of sequences at a second locus (relative to the sample as a whole), we expect variation at the two loci to exhibit significant LD. But exactly what is the relationship? Is it possible to be more quantitative about which aspects of genealogical correlation relate to which measures of LD?

A partial answer to this question comes from studying the quantity σ_d^2 ; see (2.17). It is well known that the expectation of D^2 between alleles at a pair of loci, x and y , can be written in terms of two-locus identity coefficients (Hudson, 1985; Strobeck and Morgan, 1978)

$$D_{xy}^2 = F(C_{ij}^x, C_{ij}^y) - 2F(C_{ij}^x, C_{ik}^y) + F(C_{ij}^x, C_{kl}^y). \quad (2.22)$$

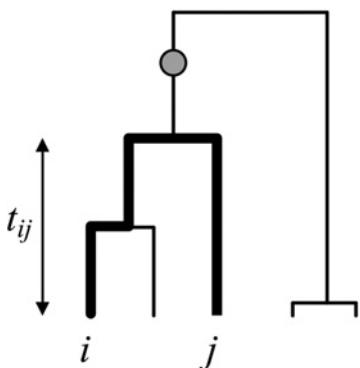


Figure 2.10 Identity in state and genealogical history. Two chromosomes, i and j , will be identical in state if no mutation occurs on those branches of the tree that lead to their common ancestor, at time t_{ij} . It follows that two-locus identity in state reflects whether or not mutations at the two trees fall on the branches to the pair's common ancestor. In the limit of a low mutation rate, but conditioning on segregation at the two loci in a sample of size n , identity in state can be written as a function of the correlations in coalescence time at the two loci.

The three terms relate to the probabilities of identity at the two loci for pairs of chromosomes sampled in different ways. Consider sampling four chromosomes and labelling them i, j, k and l . The first identity coefficient compares chromosomes i and j at both the x and y loci. The second compares chromosomes i and j at the x locus, but i and k at the y locus. The third compares chromosomes i and j at the x locus and k and l at the y locus. The expectation of these identity coefficients over evolutionary replicates therefore determines the average strength of LD.

The link to genealogical history is straightforward: each identity coefficient relates to the probability that a mutation occurs on the branches of the tree that separates the pairs of chromosomes at each locus (see Figure 2.10). If we condition on a single mutation occurring at each of the two loci within a sample of size n and let the mutation rate tend towards zero, we get the result (McVean, 2002)

$$\begin{aligned} \mathbb{E}[F(C_{ij}^x, C_{ij}^y)] &= \lim_{\theta \rightarrow 0} \frac{\mathbb{E}[(T_x - 2t_{ij}^x)(T_y - 2t_{ij}^y)] e^{-\theta(T_x + T_y)/2}}{\mathbb{E}[T_x T_y e^{-\theta(T_x + T_y)/2}]} \\ &= \frac{\mathbb{E}[(T_x - 2t_{ij}^x)(T_y - 2t_{ij}^y)]}{\mathbb{E}[T_x T_y]}, \end{aligned} \quad (2.23)$$

where T_x is the sum of the branch lengths in the tree at locus x and t_{ij}^x is the coalescent time for chromosomes i and j at locus x . Similar results can be found for the other identity coefficients for the denominator in (2.17). Combining these equations, we get the following result:

$$\sigma_d^2 = \frac{\rho(t_{ij}^x, t_{ij}^y) - 2\rho(t_{ij}^x, t_{ik}^y) - \rho(t_{ij}^x, t_{kl}^y)}{\mathbb{E}[t]^2 / \text{Var}(t) + \rho(t_{ij}^x, t_{kl}^y)}, \quad (2.24)$$

where $\rho(t_{ij}^x, t_{kl}^y)$ indicates the Pearson correlation coefficient between the coalescent time for chromosomes i and j at locus x , and the coalescent time between chromosomes k and l at locus y and the first term in the denominator is the ratio of the square of the expected coalescence time for a pair of sequences to its variance.

Equation (2.24) has two important features. First, the required correlations in time to the common ancestor can be obtained from coalescent theory (Griffiths, 1981; Pluzhnikov and Donnelly, 1996), replicating the result of (2.18). More importantly, (2.24) gives a way of understanding the behaviour of LD, or perhaps more appropriately r^2 , under more complex population genetic models. For example, the increase in LD that accompanies population bottlenecks can largely be understood through its effect on the ratio of the mean coalescence time to its standard deviation. Bottlenecks increase the variance in coalescence time considerably, leading

to a reduction in the denominator of (2.24) and an increase in LD (McVean, 2002). It is also possible to describe the behaviour of LD under models of population structure (McVean, 2002; Wakeley and Lessard, 2003) and even selective sweeps (McVean, 2007) using the same approach.

2.4 Data Analysis

Coalescent modelling provides a framework within which to understand patterns of LD (see also **Chapter 5**). But what we are usually interested in is making inferences about underlying processes from the patterns of genetic variation observed in an experiment. The aim of this section is to discuss some key statistical applications of the concepts introduced earlier, and show how these relate to understanding patterns of genetic variation.

2.4.1 Estimating Recombination Rates

To illustrate various approaches to parameter estimation in the context of understanding LD we will consider the problem of how to estimate the recombination rate, or rather the population recombination rate, over a region. All of the methods described consider only a neutral population of constant population size. The point of covering the various estimators is to give an indication of the range of possible approaches.

2.4.1.1 Moment Methods

The first population genetic method for estimating the population recombination rate (Hudson, 1987; Wakeley, 1997) used a method of moments approach. Using the results mentioned above that relate LD coefficients to two-locus sampling identities, Hudson derived an expression for the expected sample variance of the number of nucleotide differences between pairs of sequences under the infinite-sites model. If the number of pairwise differences between sequences i and j is π_{ij} then

$$\begin{aligned}\tilde{\pi} &= \frac{1}{n^2} \sum_{i,j} \pi_{ij}, \\ S_\pi^2 &= \frac{1}{n^2} \sum_{i,j} \pi_{ij}^2 - \tilde{\pi}^2, \\ \mathbb{E}[\tilde{\pi}] &= (1 - 1/n)\theta, \\ \mathbb{E}[S_\pi^2] &= f(\theta, C, n),\end{aligned}\tag{2.25}$$

where θ is the scaled mutation rate and $f(\theta, C, n)$ is a known function of the two parameters and the sample size (Hudson, 1987). To obtain a point estimate of C , θ is replaced with a point estimate from the sample (also obtained by the method of moments) and the equation is solved (if possible) for C . Similar approaches could be constructed for other single-number summaries of the data (e.g. those in Table 2.1), although Monte Carlo methods would have to be used to obtain estimates.

The great strength of this estimator is its simplicity. Unfortunately, the estimator also has very poor properties (bias, high variance, lack of statistical consistency, undefined values). This is partly because of the inherent stochasticity of the coalescent process, so for any given value of C and θ , there is a huge amount of variation in the observed patterns of variation (hence any estimator of C is expected to have considerable variance). But it is partly also because it only

uses a small fraction of the total information about recombination present in the data. It is not known how well moment estimators from other sample properties might perform.

2.4.1.2 Likelihood Methods

A natural quantity to compute is the likelihood of the parameter (proportional to the probability, or probability density, of observing the data given the specified parameter value). Ideally, we would like to calculate the probability of observing the data given the coalescent model and specified values of the population parameters, perhaps choosing the values that maximise the likelihood as the estimates. While this problem has received considerable attention (Fearnhead and Donnelly, 2001; Kuhner *et al.*, 2000; Nielsen, 2000), currently such approaches are only computationally feasible for small to moderate sized data sets (Fearnhead *et al.*, 2004). A common feature of all these methods is the use of Monte Carlo techniques (particularly Markov chain Monte Carlo and importance sampling; see **Chapter 1**). For example, in importance sampling, a proposal function, Q , is used to generate coalescent histories, H (a series of coalescent, mutation and recombination events), compatible with the data for given values of θ and C . The likelihood of the data can be estimated from

$$\mathcal{L}(\theta, C; x) = \mathbb{E}_Q \left[\frac{\mathbb{P}[H|\theta, C]}{Q(H|\theta, C, x)} \mathbb{P}[x|H], \right] \quad (2.26)$$

where $\mathbb{P}[H|\theta, C]$ is the coalescent probability of the history (note that this is not a function of the data), $\mathbb{P}[x|H]$ is the probability of the data given the history (which is always exactly 1 here), and $Q(H|\theta, C, x)$ is the proposal probability of the history (note that this is a function of the data). The main problem is that unless Q is close to the optimal proposal scheme

$$Q_{\text{OPT}} = \mathbb{P}[H|\theta, C, x], \quad (2.27)$$

the large variance in likelihood estimates across simulations makes obtaining accurate estimates of the likelihood nearly impossible.

For larger data sets a more practical alternative is to calculate the likelihood of some informative summary (or summaries) of the data using Monte Carlo techniques (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Padhukasahasram *et al.*, 2006; Wall, 2000). Wall's estimator, based on the number of distinct haplotypes and a nonparametric estimate of the minimum number of detectable recombination events, while conditioning on the number of segregating sites, is good in terms of having low bias and variance comparable to the best alternatives. A great strength of these methods is that they can provide useful summaries of uncertainty in estimates, for example through estimating the Bayesian posterior distribution of the parameter. Because some information is thrown away, the resulting uncertainty will be greater than the true uncertainty. However, there is no guarantee that the likelihood surface computed from a summary statistic will reflect the likelihood surface computed from the full data (indeed, they may disagree completely). Furthermore, these methods (like the moment estimators) may not be robust to deviations from the assumed model. For example, the number of distinct haplotypes observed will be sensitive to whether sequence data have been collected by SNP genotyping or complete sequencing.

2.4.1.3 Approximating the Likelihood

Although it may not be practical to calculate the coalescent likelihood of the entire data, it is possible to approximate the likelihood function through multiplying the likelihoods for subsets of the data (Fearnhead *et al.*, 2004; Hey and Wakeley, 1997; Hudson, 2001; McVean *et al.*, 2002, 2004; Wall, 2004), the resulting quantity being referred to as a composite likelihood (see **Chapter 1** for further details on composite likelihood methods). In effect, the idea is to treat

subsets of the data as if they were independent of each other, whereas in reality these subsets (pairs, triples or non-overlapping sets) are clearly not. Nevertheless, finding the value of the recombination parameter that maximises this function appears to provide estimates that are at least as accurate as any other approach.

Perhaps the greatest strength of the composite likelihood approaches lies in their flexibility, particularly in the use of estimating variable recombination rates and identifying recombination hotspots (Fearnhead and Donnelly, 2002; Fearnhead and Smith, 2005; McVean *et al.*, 2004; Myers *et al.*, 2005). For example, consider Hudson's composite likelihood approach, which considers pairs of sites. For each pair of sites it is possible to pre-compute the coalescent likelihood for a specified value of θ per site and a range of recombination rates between the sites, for example, using the importance sampling approach (McVean *et al.*, 2002). The likelihood of a genetic map $g = \{g_1, g_2, \dots, g_m\}$, where each g_i is the map position of the i th site in a set of m ordered sites, is approximated by

$$\mathcal{L}_C(\theta, g; x) = \prod_{\substack{ij \\ j>i}} \mathcal{L}(\theta, g_j - g_i; x_{ij}). \quad (2.28)$$

Here \mathcal{L}_C denotes the composite likelihood, \mathcal{L} denotes the coalescent likelihood for the genetic distance between loci i and j , and x_{ij} is the data at those two sites. In practice, the value of θ is estimated previously using a moment method. Because of the pre-computation, searching over the space of g is computationally feasible. McVean *et al.* (2004) used a Monte Carlo technique called reversible jump Markov chain Monte Carlo (Green, 1995) to explore the space of possible genetic maps. While these methods seem to be robust to many deviations from the assumed model (McVean *et al.*, 2004), their greatest weakness is the difficulty in assessing uncertainty. Specifically, the composite likelihood surface is typically more sharply peaked compared to the true likelihood surface (even though it uses less information), resulting in considerable underestimation of uncertainty.

2.4.1.4 Approximating the Coalescent

An alternative approach to approximating the coalescent likelihood function is to devise an alternative model for the data, motivated by an understanding of the coalescent, but under which it is straightforward to calculate the likelihood (Crawford *et al.*, 2004; Li and Stephens, 2003). The product of approximate conditionals (PAC) scheme uses the following decomposition to motivate an approximate model. Consider the data $x = \{x_1, x_2, \dots, x_n\}$, where x_i is the i th haplotype in a sample of size n . The likelihood function can be written as the product of a series of conditional distributions (the dependence on θ has been dropped for simplicity)

$$\mathcal{L}(C; x) = \mathbb{P}[x_1|C] \times \mathbb{P}[x_2|x_1, C] \times \cdots \times \mathbb{P}[x_n|x_1, x_2, \dots, x_{n-1}, C]. \quad (2.29)$$

Of course, knowing the conditional probabilities is equivalent to knowing the coalescent likelihood. However, the PAC scheme approximates the conditional probabilities by considering the k th haplotype as an imperfect mosaic of the previous $k - 1$. Specifically, the model employed has the structure of a hidden Markov model in which the underlying state at a given nucleotide position refers to which of the $k - 1$ other chromosomes the k th is derived from. The transition probabilities are a function of the recombination rate and the emission probabilities are a function of the mutation rate. This model captures many of the key features of genetic variation, such as the relatedness between chromosomes, how this changes through recombination, and how, as the sample size increases, additional chromosomes tend to look more and more like existing ones, but it is entirely non-genealogical.

This approach was actually derived from the work mentioned above that uses importance sampling to calculate the full coalescent likelihood of the data (Fearnhead and Donnelly, 2001; Stephens and Donnelly, 2000). Because of the Markovian structure of the coalescent, each history that is compatible with the data can be broken down into a series of events that can be sampled sequentially. The optimal proposal density chooses an event (coalescence, mutation, and recombination), e , according to its probability given the data

$$Q_{\text{OPT}}(e|\theta, C, x) = \mathbb{P}[e|\theta, C] \frac{\mathbb{P}[x + e|\theta, C]}{\mathbb{P}[x|\theta, C]}, \quad (2.30)$$

where $\mathbb{P}[e|\theta, C]$ is the coalescent probability of the event and $x + e$ is the original data, x , modified by event e . The key insight is to note that $x + e$ and x are very similar. For example, if e is a coalescent event between the i th and j th sequences (which are also identical) it follows that

$$\begin{aligned} \frac{\mathbb{P}[x + e|\theta, C]}{\mathbb{P}[x|\theta, C]} &= \frac{\mathbb{P}[x_1, \dots, x_i, \dots, x_k|\theta, C]}{\mathbb{P}[x_1, \dots, x_i, x_j, \dots, x_k|\theta, C]} \\ &= \frac{1}{\mathbb{P}[x_j|x_1, \dots, x_i, \dots, x_k, \theta, C]}. \end{aligned} \quad (2.31)$$

This is exactly the same conditional probability which is approximated in the PAC scheme. For other types of events, similar expressions involving the conditional distributions can be found (Fearnhead and Donnelly, 2001).

The strengths of the PAC scheme are efficiency and computational tractability. It is also very flexible and can be extended to include features such as geographical structuring (Wasser *et al.*, 2004). Nevertheless, because the model is not the coalescent, parameter estimates are typically biased in ways that are unpredictable (and the estimated uncertainty in estimates may not necessarily reflect the ‘true’ uncertainty). For example, these methods typically infer recombination when there is none and do not necessarily ‘spot’ signals of recombination such as pairs of sites for which all four possible combinations of haplotypes are observed in the sample (Hudson and Kaplan, 1985). The approximation also introduces an order-dependency into the likelihood function (the true conditionals would, of course, give rise to the same likelihood no matter how the sequences were considered). This can be partly overcome by averaging inference over multiple orderings. (See **Section 2.4.2.1**, and **Chapters 3** and **8** for applications of the model.)

2.4.2 Methods Exploiting Haplotype Structure

The previous sections have introduced concepts underlying LD and discussed the central role of recombination. Driven by the rapid growth in availability of large-scale whole genome sequencing and SNP typing data, there is also great interest in being able to use LD and haplotype structure for a wide range of problems in statistical, population and computational genomics. In this subsection we discuss some of these applications.

2.4.2.1 Detecting and Using Recent Common Ancestry

Recent common ancestry between individuals typically results in the sharing of long stretches of haplotype identity. Many methods in statistical and population genetics exploit this observation for the purpose of detecting relatives in data sets (Purcell *et al.*, 2007; Gusev *et al.*, 2009; Albrechtse *et al.*, 2009; Browning and Browning, 2013); haplotype phasing and imputation (see **Chapter 3**) with or without reference panels (Scheet and Stephens, 2006; Browning and Browning, 2007; Marchini *et al.*, 2007; Kong *et al.*, 2008; Howie *et al.*, 2009; Li *et al.*, 2010; Williams *et al.*, 2012; Delaneau *et al.*, 2012, 2013; Loh *et al.*, 2016; Davies *et al.*,

2016b; O'Connell *et al.*, 2016); IBD-based disease mapping (Albrechtsen *et al.*, 2009; Moltke *et al.*, 2011; see also **Chapter 20**); and inferring ancestry along a chromosome – so-called 'chromosome-painting' methods such as HapMix (Price *et al.*, 2009), fineStructure, ChromoPainter (Lawson *et al.*, 2012) and GLOBETROTTER (Hellenthal *et al.*, 2014). Moreover, because there is a simple relationship between the expected length of haplotype sharing and the number of meioses separating two chromosomes, recombination 'clocks' have become an important component of efforts to date demographic events, such as time since admixture (Hellenthal *et al.*, 2008; Lawson *et al.*, 2012; Lawson and Falush, 2012) and changes in population size (Ralph and Coop, 2013). When applied at a population scale, such methods have been able to reveal great detail about the impact of population migrations, geographical isolation and major events on human populations (Leslie *et al.*, 2015; Arauna *et al.*, 2016; Busby *et al.*, 2015, 2016; Brucato *et al.*, 2017; Byrne *et al.*, 2017; Gilbert *et al.*, 2017; Takeuchi *et al.*, 2017). Further information on these uses of haplotype structure can be found in **Chapter 8**.

These approaches aim to identify individuals that share recent common ancestors, often exploiting the hidden Markov model structure of the Li and Stephens algorithm for computational efficiency, either choosing the most likely (Viterbi) path to represent relatedness or sampling from the posterior distribution. However, it is worth noting that the extent of haplotype sharing is a noisy indicator of genealogical relatedness. Due to the stochastic nature of recombination, it is quite possible that at a given position on a chromosome, the person that you share the greatest extent of haplotype identity with is not your nearest genealogical relative. For example, under the neutral coalescent, simulations show that the haplotype with whom you share the longest identity is not among your genealogical nearest neighbours approximately 14% of the time (independent of sample size) and in a further 14% of cases it is not uniquely among your genealogical nearest neighbours (Xifara, 2014). In short, while extended haplotype identity is typically indicative of shared recent common ancestry, the converse is not necessarily true. One consequence is that even before the complications of genotyping error, recurrent mutation, back mutation and gene conversion, there are fundamental limits on problems such as rare variant imputation or reconstruction of pedigrees that require very accurate inference of genealogical relationships (see also Harris, 2011).

2.4.2.2 Identifying Incomplete Selective Sweeps

The notion that shared haplotype identity is a proxy for age of the MRCA has also been exploited in the search for genetic variants that have undergone partial selective sweeps. Such loci are likely to show a disparity between their age as inferred by allele frequency (high frequency implies ancient origin) and by haplotype sharing (extended haplotype sharing implies recent origin). The extended haplotype homozygosity (EHH) test (Sabeti *et al.*, 2002) was the first approach to detecting positive selection that made use of this observation. Subsequent improvements, such as the iHS test (Voight *et al.*, 2006), correct for variation in the recombination rate across the genome (regions of low recombination will tend to exhibit extended haplotype sharing). The approach has also been refined to detect sweeps specific to a particular population (Sabeti *et al.*, 2007) and to distinguish hard and soft selective sweeps (Garud *et al.*, 2015). (See **Section 14.4.5** for more details on haplotype methods for detecting selection.) To date, relatively few cases of strong partial sweeps have been identified in humans, with some notable exceptions around particular genes. For example, *LCT*, where variation affects lactose digestion and correlates with the geographic and temporal origins of pastoralism (Tishkoff *et al.*, 2007); *SLC24A5* and other genes that influence skin pigmentation (Lamason *et al.*, 2005); and *EDAR*, which influences hair type and sweat gland density production (Kamberov *et al.*, 2013). All of these variants also show strong geographic differentiation,

which further supports the hypothesis of adaptation following exposure to new environmental selection pressures.

2.4.2.3 Data Compression

The amount of genetic data available to researchers has increased tremendously in recent years. For example, population-scale sequencing projects (Genome of the Netherlands Consortium *et al.*, 2014; UK10K Consortium *et al.*, 2015; 1000 Genomes Project Consortium, 2015; Gudbjartsson *et al.*, 2015; Sudlow *et al.*, 2015; Bycroft *et al.*, 2017) are producing genetic data for hundreds of thousands of humans, and this rate of data collection is expected to continue (Stephens *et al.*, 2015). The dramatic increase in data volume has presented serious problems for existing tool chains, both in terms of the size of data files that must be processed and the running time of algorithms. Specialised compression methods for various types of genetic data have been investigated for many years (Giancarlo *et al.*, 2009). We are interested here in compressing and analysing variation data, that is, the observed differences between the samples and some reference.

The simplest approach to compressing such data is to organise it in a ‘variant-centric’ fashion, such that all observed genotypes for a given variant are stored in a contiguous block. Because most variants are rare, this will result in long runs of identical values that compress well using standard methods. This is the approach taken by the Variant Call Format (Danecek *et al.*, 2011), and its more efficient binary encoding, BCF. The compression levels achieved by this straightforward approach are good, and quite competitive with more sophisticated methods. The disadvantage is that many queries require full decompression of the data, which can be prohibitively time-consuming. The SpeedGene data format and software library (Qiao *et al.*, 2012) chooses from one of three encoding methods for each SNP determined by the allele frequency; Sambo *et al.* (2014) extend this idea by identifying blocks of SNPs in LD, and adding two additional potential encodings utilising this information.

An alternative to this variant-centric approach is to store genotypes in a ‘sample-centric’ fashion. Here, all genotypes for a particular sample are stored consecutively. While this breaks the simple repetition structure of data stored in variant-centric form, other methods can be employed to find compressible structure. For example, TGC (Deorowicz *et al.*, 2013) compresses sample genotypes using a custom Lempel–Ziv style approach on runs of identity among samples. This explicit use of LD structure results in excellent compression performance (e.g. 32 MB for all SNPs in 1000 Genomes chromosome 1); unfortunately, querying a data set requires full decompression, making the format unsuitable for online analysis. The GQT toolkit (Layer *et al.*, 2016) also uses the sample-centric organisation of genotype data, but takes a different approach to compression. Variants are sorted by allele frequency (resulting in longer runs of identical values within samples) and then compressed using an efficient bit-vector representation (Wu *et al.*, 2002). The resulting file-sizes are similar to compressed BCF, but many queries can be performed directly on the compressed representation, resulting in a considerable increase in speed.

The positional Burrows–Wheeler transform (PBWT) (Durbin, 2014) is another sample-centric method. Building on the success of Burrows–Wheeler transform applications in genomics (Langmead *et al.*, 2009; Li and Durbin, 2009; Li *et al.*, 2009), the PBWT provides very good compression performance and efficient algorithms for finding haplotype matches. The algorithm builds a representation of the data based on sorted haplotype prefixes in linear time, and the LD structure of the data ensures that the sorting orders between adjacent sites change very little on average. The method has been successfully applied to phasing (Loh *et al.*, 2016), detection of IBD segments (Naseri *et al.*, 2017), improving the performance of the Li and

Stephens model (Lunter, 2016), and a general query engine for genotype data (Li, 2016). Recent extensions include privacy preserving search (Shimizu *et al.*, 2016) and generalisation to the setting of graph genomes (Novak *et al.*, 2017a).

2.5 Prospects

The past decade has seen major advances in our understanding of the biological processes influencing recombination. The identification of the recombination hotspot positioning gene, *Prdm9* (Myers *et al.*, 2010; Parvanov *et al.*, 2010; Baudat *et al.*, 2010), led to many surprises about how recombination can evolve rapidly (Myers *et al.*, 2008; Paigen and Petkov, 2010; Kong *et al.*, 2010; Auton *et al.*, 2012; Baudat *et al.*, 2013), drive genome evolution (Coop and Myers, 2007; Duret and Galtier, 2009) and even contribute to speciation (Flachs *et al.*, 2012; Kono *et al.*, 2014; Davies *et al.*, 2016a). Related processes, such as gene conversion (Jeffreys and May, 2004; Chen *et al.*, 2007) and non-allelic homologous recombination (Gu *et al.*, 2008; Sasaki *et al.*, 2010) have also been shown to have a major impact on genome diversity and evolution. Such complications mean that the simple models of recombination described here are very much an approximation to what occurs in biology, though still provide a useful framework for understanding the structure of genetic diversity.

The next decade will see a remarkable explosion of genomic data, resulting in data sets consisting of millions of humans. The scale of such data provides huge opportunities to learn about human history and the link between genetic variation and human biology and disease, but also brings major challenges in relation to data handling and our ability to fit ever more complex models. There is increasing interest in developing methods that take advantage of the underlying tree structure of genetic data. One such data structure, (known as a ‘succinct tree sequence’) has been shown to perform well for simulated data (Kelleher *et al.*, 2016, 2018); for example, a simulation of 500,000 human-like chromosomes (200 Mb, with $N_e = 10^4$ and per-base mutation and recombination rates of 10^{-8} per generation), gives a 157 MiB file, while the corresponding VCF file (giving only the genotype information) would require about 1 TiB (Kelleher *et al.*, 2018). Moreover, the encoded genealogical structures naturally lead to efficient algorithms for computing many statistics of interest. However, whether efficient approaches for inferring such structures from real data are possible remains an open problem.

Finally, as our understanding of genomic diversity across multiple species grows, so too does our appreciation of the plasticity of genomes and the variation in gene content and order that can occur. Even within humans, regions such as the MHC, KIR and olfactory gene clusters exhibit vast diversity in structure and content (Horton *et al.*, 2008; Jiang *et al.*, 2012; Trask *et al.*, 1998), while in bacteria, any one isolate is likely to harbour a small minority of the pangenome (Tettelin *et al.*, 2005; Rasko *et al.*, 2008; Tettelin *et al.*, 2008; Touchon *et al.*, 2009). Such observations have motivated the search for solutions to representing genomic variation through generative graph structures that represent genome diversity (Dilthey *et al.*, 2015; Novak *et al.*, 2017b; Garrison *et al.*, 2017). These represent a new interface between population genetics and genomics, though they are very much in their infancy.

Acknowledgements

We would like to thank Ida Moltke and Anthony Wilder Wohns for helpful comments. This work was supported by the Wellcome Trust [100956/Z/13/Z].

References

- 1000 Genomes Project Consortium. A global reference for human genetic variation (2015). *Nature* **526**(7571), 68–74.
- Albrechtsen, A., Korneliussen, T.S., Moltke, I., van Overeem Hansen, T., Nielsen, F.C. and Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic epidemiology* **33**(3), 266–274.
- Anderson, E.C. and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics* **73**(2), 336–354.
- Arauna, L.R., Mendoza-Revilla, J., Mas-Sandoval, A., Izaabel, H., Bekada, A., Benhamamouch, S., Fadhloui-Zid, K., Zalloua, P., Hellenthal, G. and Comas, D. (2016). Recent historical migrations have shaped the gene pool of Arabs and Berbers in North Africa. *Molecular Biology and Evolution* **34**(2), 318–329.
- Auton, A. and McVean, G. (2012). Estimating recombination rates from genetic variation in humans. In M. Anisimova (ed.), *Evolutionary Genomics, Volume 2: Statistical and Computational Methods*. Humana Press, New York, pp. 217–237.
- Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., et al. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**(6078), 193–198.
- Balding, D.J. and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In B.S., Weir (ed.), *Human Identification: The Use of DNA Markers*. Springer Netherlands, Dordrecht, pp. 3–12.
- Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G. and De Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**(5967), 836–840.
- Baudat, F., Imai, Y. and De Massy, B. (2013). Meiotic recombination in mammals: Localization and regulation. *Nature Reviews Genetics* **14**(11), 794–806.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.
- Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**(6), 1111–1120.
- Browning, B.L. and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**(2), 459–471.
- Browning, S.R. and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**(5), 1084–1097.
- Browning, S.R. and Browning, B.L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics* **12**(10), 703–714.
- Brucato, N., Kusuma, P., Beaujard, P., Sudoyo, H., Cox, M.P. and Ricaut, F.-X. (2017). Genomic admixture tracks pulses of economic activity over 2,000 years in the Indian Ocean trading network. *Scientific Reports*, 7.
- Busby, G.B., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., Rudan, I., Zemunik, T., Hayward, C., Toncheva, D., Karachanak-Yankova, S., et al. (2015). The role of recent admixture in forming the contemporary West Eurasian genomic landscape. *Current Biology* **25**(19), 2518–2526.

- Busby, G.B., Band, G., Le, Q.S., Jallow, M., Bougama, E., Mangano, V.D., Amenga-Etego, L.N., Enimil, A., Apinjoh, T., Ndila, C.M., et al. (2016). Admixture into and within sub-Saharan Africa. *eLife*, 5, e15266.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint, bioRxiv 166298.
- Byrne, R.P., Martiniano, R., Cassidy, L.M., Carrigan, M., Hellenthal, G., Hardiman, O., Bradley, D.G. and McLaughlin, R.L. (2017). Insular Celtic population structure and genomic footprints of migration. Preprint, bioRxiv 230797.
- Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* 74(1), 106–120.
- Carmi, S., Wilton, P.R., Wakeley, J. and Pe'er, I. (2014). A renewal theory approach to IBD sharing. *Theoretical Population Biology* 97, 35–48.
- Chakravarti, A., Buetow, K.H., Antonarakis, S., Waber, P., Boehm, C. and Kazazian, H. (1984). Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, 36(6), 1239.
- Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity*, 56(1–3), 18–31.
- Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C. and Patrinos, G.P. (2007). Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics* 8(10), 762–775.
- Coop, G. and Myers, S.R. (2007). Live hot, die young: Transmission distortion in recombination hotspots. *PLoS Genetics*, 3(3), e35.
- Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A. and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* 36(7), 700–706.
- Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27(15), 2156–2158.
- Davies, B., Hatton, E., Altemose, N., Hussin, J.G., Pratto, F., Zhang, G., Hinch, A.G., Moralli, D., Biggs, D., Diaz, R., et al (2016a). Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* 530(7589), 171–176.
- Davies, R.W., Flint, J., Myers, S. and Mott, R. (2016b). Rapid genotype imputation from sequence without reference panels. *Nature Genetics* 48(8), 965–969.
- de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics* 37(11), 1217–1223.
- Delaneau, O., Marchini, J. and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods* 9(2), 179–181.
- Delaneau, O., Zagury, J.-F. and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10(1), 5–6.
- Deorowicz, S., Danek, A. and Grabowski, S. (2013). Genome compression: A novel approach for large collections. *Bioinformatics* 29(20), 2572–2578.
- Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29(2), 311–322.
- Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R. and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics* 47(6), 682–688.

- Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**(9), 1266–1272.
- Duret, L. and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* **10**, 285–311.
- EWENS, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**(1), 87–112.
- Fay, J.C. and Wu, C.-I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**(3), 1405–1413.
- Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**(3), 1299–1318.
- Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society, Series B* **64**(4), 657–680.
- Fearnhead, P. and Smith, N.G. (2005). A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *American Journal of Human Genetics* **77**(5), 781–794.
- Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. and Donnelly, P. (2004). Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics* **167**(4), 2067–2081.
- Flachs, P., Mihola, O., Šimeček, P., Gregorová, S., Schimenti, J.C., Matsui, Y., Baudat, F., de Massy, B., Piálek, J., Forejt, J. and Trachtulec, Z. (2012). Interallelic and intergenic incompatibilities of the Prdm9 (Hst1) gene in mouse hybrid sterility. *PLoS Genetics*, **8**(11), e1003044.
- Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**(3), 693–709.
- Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. *Science* **296**(5576), 2225–2229.
- Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Lin, M.F., Paten, B. and Durbin, R. (2017). Sequence variation aware genome references and read mapping with the variation graph toolkit. Preprint, bioRxiv 234856.
- Garud, N.R., Messer, P.W., Buzbas, E.O. and Petrov, D.A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, **11**(2), e1005004.
- Genome of the Netherlands Consortium, et al. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**(8), 818–825.
- Giancarlo, R., Scaturro, D. and Utro, F. (2009). Textual data compression in computational biology: A synopsis. *Bioinformatics* **25**(13), 1575–1586.
- Gilbert, E., O'Reilly, S., Merrigan, M., McGettigan, D., Molloy, A.M., Brody, L.C., Bodmer, W., Hutnik, K., Ennis, S., Lawson, D.J., et al. (2017). The Irish DNA atlas: Revealing fine-scale population structure and history within Ireland. *Scientific Reports*, **7**(1), 17199.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4), 711–732.
- Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology* **19**(2), 169–186.
- Griffiths, R.C. (1991). The two-locus ancestral graph. In *Selected Proceedings of the Sheffield Symposium on Applied Probability*, volume 18, pp. 100–117.
- Griffiths, R.C. and Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly and S., Tavaré (eds.), *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, volume 87. Springer-Verlag, Berlin, pp. 257–270.

- Gu, W., Zhang, F. and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics*, **1**(1), 4.
- Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **47**(5), 435–444.
- Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M. and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19**(2), 318–326.
- Harris, K. (2011). The relationship of identity by state to identity by descent and imputation accuracy in population sequencing data. Master's thesis, University of Cambridge.
- Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, **9**(6), e1003521.
- Hedrick, P.W. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics* **117**(2), 331–341.
- Hellenthal, G., Auton, A. and Falush, D. (2008). Inferring human colonization history using a copying model. *PLoS Genetics*, **4**(5), e1000078.
- Hellenthal, G., Busby, G.B., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014). A genetic atlas of human admixture history. *Science* **343**(6172), 747–751.
- Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics* **145**(3), 833–846.
- Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**(6), 226–231.
- Hill, W.G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**(2), 229–239.
- Hill, W.G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**(2), 117–126.
- Hill, W.G. (1977). Correlation of gene frequencies between neutral linked genes in finite populations. *Theoretical Population Biology* **11**(2), 239–248.
- Hill, W.G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research* **8**(3), 269–294.
- Hobolth, A. and Jensen, J.L. (2014). Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology* **98**, 48–58.
- Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R.J., Almeida, J., Forbes, S., Gilbert, J.G., Halls, K., Harrow, J.L., et al. (2008). Variation analysis and gene annotation of eight MHC haplotypes: The MHC haplotype project. *Immunogenetics* **60**(1), 1–18.
- Howie, B.N., Donnelly, P. and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**(6), e1000529.
- Hudson, R.R. (1983a). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**(2), 183–201.
- Hudson, R.R. (1983b). Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**(1), 203–217.
- Hudson, R.R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**(3), 611–631.
- Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetics Research* **50**(3), 245–250.
- Hudson, R.R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1–44.

- Hudson, R.R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**(4), 1805–1817.
- Hudson, R.R. and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**(1), 147–164.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**(7063), 1299–1320.
- Jeffreys, A.J. and May, C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics* **36**(2), 151–156.
- Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29**(2), 217–222.
- Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S. and Donnelly, P. (2005). Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics* **37**(6), 601–606.
- Jiang, W., Johnson, C., Jayaraman, J., Simecek, N., Noble, J., Moffatt, M.F., Cookson, W.O., Trowsdale, J. and Traherne, J.A. (2012). Copy number variation leads to considerable diversity for B but not A haplotypes of the human KIR genes encoding NK cell receptors. *Genome Research* **22**(10), 1845–1854.
- Jobling, M., Hurles, M. and Tyler-Smith, C. (2004). *Human Evolutionary Genetics: Origins, Peoples & Disease*. Garland Science, New York.
- Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**(2), 233–237.
- Kamberov, Y.G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., et al. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**(4), 691–702.
- Karlin, S. and McGregor, J. (1967). The number of mutant forms maintained in a population. In L., LeCam and J., Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability*, volume 4. University of California Press, Berkeley, pp. 415–438.
- Kelleher, J., Etheridge, A.M. and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, **12**(5), e1004842.
- Kelleher, J., Thornton, K., Ashander, J. and Ralph, P. (2018). Efficient pedigree recording for fast population genetics simulation. Preprint, bioRxiv 248500.
- Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4), 893.
- Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**(3), 235–248.
- Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**(9), 1068–1075.
- Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdóttir, A., Walters, G.B., Jonasdóttir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**(7319), 1099–1103.
- Kono, H., Tamura, M., Osada, N., Suzuki, H., Abe, K., Moriwaki, K., Ohta, K. and Shiroishi, T. (2014). Prdm9 polymorphism unveils mouse evolutionary tracks. *DNA Research* **21**(3), 315–326.
- Kuhner, M.K., Yamato, J. and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**(3), 1393–1401.
- Lamason, R.L., Mohideen, M.-A.P., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynev, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. (2005). SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310**(5755), 1782–1786.
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**(3), R25.

- Lawson, D.J. and Falush, D. (2012). Population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13.
- Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), e1002453.
- Layer, R.M., Kindlon, N., Karczewski, K.J., Exome Aggregation Consortium and Quinlan, A.R. (2016). Efficient genotype compression and analysis of large genetic-variation data sets. *Nature Methods* 13(1), 63–65.
- Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Rorvik, E.C., Cunliffe, B., Lawson, D.J., et al. (2015). The fine-scale genetic structure of the British population. *Nature* 519(7543), 309–314.
- Lewontin, R. (1964). The interaction of selection and linkage I. General considerations; heterotic models. *Genetics*, 49(1), 49.
- Lewontin, R. (1988). On measures of gametic disequilibrium. *Genetics* 120(3), 849–852.
- Li, H. (2016). BGT: Efficient and flexible genotype query across many samples. *Bioinformatics* 32(4), 590–592.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14), 1754–1760.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165(4), 2213–2233.
- Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25(15), 1966–1967.
- Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* 34(8), 816–834.
- Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics* 48(11), 1443–1448.
- Lunter, G. (2016). Fast haplotype matching in very large cohorts using the Li and Stephens model. Preprint, bioRxiv 048280.
- Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* 78(3), 437–450.
- Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* 39(7), 906–913.
- Marjoram, P. and Wall, J.D. (2006). Fast ‘coalescent’ simulation. *BMC Genetics*, 7(1), 16.
- Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100(26), 15324–15328.
- Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* 23(1), 23–35.
- McVean, G. (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* 175(3), 1395–1406.
- McVean, G., Awadalla, P. and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160(3), 1231–1241.
- McVean, G., Spencer, C.C. and Chaix, R. (2005). Perspectives on human genetic variation from the HapMap Project. *PLoS Genetics*, 1(4), e54.
- McVean, G.A.T. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* 162(2), 987–991.

- McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B* **360**(1459), 1387–1393.
- McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**(5670), 581–584.
- Moltke, I., Albrechtsen, A., Hansen, T.V., Nielsen, F.C. and Nielsen, R. (2011). A method for detecting ibd regions simultaneously in multiple individuals – with applications to disease genetics. *Genome Research* **21**(7), 1168–1180.
- Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746), 321–324.
- Myers, S., Spencer, C., Auton, A., Bottolo, L., Freeman, C., Donnelly, P. and McVean, G. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochemical Society Transactions* **34**, 526–530.
- Myers, S., Freeman, C., Auton, A., Donnelly, P. and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics* **40**(9), 1124–1129.
- Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G. and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**(5967), 876–879.
- Myers, S.R. and Griffiths, R.C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**(1), 375–394.
- Naseri, A., Liu, X., Zhang, S. and Zhi, D. (2017). Ultra-fast identity by descent detection in biobank-scale cohorts using positional Burrows-Wheeler transform. Preprint, bioRxiv 103325.
- Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**(2), 931–942.
- Novak, A.M., Garrison, E. and Paten, B. (2017a). A graph extension of the positional Burrows-Wheeler transform and its applications. *Algorithms for Molecular Biology*, **12**(1), 18.
- Novak, A.M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M.A.S., Guthrie, S., Kahles, A., Keenan, S., Kelleher, J., Kural, D., Li, H., Lin, M.F., Miga, K., Ouyang, N., Rakoccevic, G., Smuga-Otto, M., Zaraneck, A.W., Durbin, R., McVean, G., Haussler, D. and Paten, B. (2017b). Genome graphs. Preprint, bioRxiv 101378.
- O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O. and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature Genetics* **48**(7), 817–820.
- Ohta, T. and Kimura, M. (1969a). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*, **63**(1), 229.
- Ohta, T. and Kimura, M. (1969b). Linkage disequilibrium due to random genetic drift. *Genetics Research* **13**(1), 47–55.
- Ohta, T. and Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, **68**(4), 571.
- Padhukasahasram, B., Wall, J.D., Marjoram, P. and Nordborg, M. (2006). Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics* **174**(3), 1517–1528.
- Paigen, K. and Petkov, P. (2010). Mammalian recombination hot spots: Properties, control and evolution. *Nature Reviews Genetics* **11**(3), 221–233.
- Parvanov, E.D., Petkov, P.M. and Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science* **327**(5967), 835.
- Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**(3), 1247–1262.

- Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, **5**(6), e1000519.
- Pritchard, J.K. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* **69**(1), 1–14.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3), 559–575.
- Qiao, D., Yip, W.-K. and Lange, C. (2012). Handling the data management needs of high-throughput sequencing data: Speedgene, a compression algorithm for the efficient storage of genetic data. *BMC Bioinformatics*, **13**(1), 100.
- Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biology*, **11**(5), e1001555.
- Rasko, D.A., Rosovitz, M., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., et al. (2008). The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology* **190**(20), 6881–6893.
- Rasmussen, M.D., Hubisz, M.J., Gronau, I. and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, **10**(5), e1004342.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**(6909), 832–837.
- Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**(7164), 913.
- Sambo, F., Di Camillo, B., Toffolo, G. and Cobelli, C. (2014). Compression and fast retrieval of SNP data. *Bioinformatics* **30**(21), 3078–3085.
- Sasaki, M., Lange, J. and Keeney, S. (2010). Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology* **11**(3), 182–195.
- Scally, A. and Durbin, R. (2012). Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics* **13**(10), 745–753.
- Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**(4), 629–644.
- Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**, 919–925.
- Sheehan, S., Harris, K. and Song, Y.S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194**(3), 647–662.
- Shimizu, K., Nuida, K. and Rätsch, G. (2016). Efficient privacy-preserving string search and an application in genomics. *Bioinformatics* **32**(11), 1652–1661.
- Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**(1), 331–336.
- Song, Y.S. (2006). Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Annals of Combinatorics* **10**(1), 147–163.
- Song, Y.S., Wu, Y. and Gusfield, D. (2005). Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution. *Bioinformatics*, **21**(suppl_1), i413–i422.
- Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society, Series B* **62**(4), 605–635.

- Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S. and Robinson, G.E. (2015). Big data: Astronomical or genomic? *PLoS Biology*, **13**(7), e1002195.
- Strobeck, C. and Morgan, K. (1978). The effect of intragenic recombination on the number of alleles in a finite population. *Genetics* **88**(4), 829–844.
- Stumpf, M.P. and McVean, G.A.T. (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics* **4**(12), 959–968.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, **12**(3), e1001779.
- Sved, J. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**(2), 125–141.
- Swallow, D.M. (2003). Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics* **37**(1), 197–219.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2), 437–460.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3), 585–595.
- Takeuchi, F., Katsuya, T., Kimura, R., Nabika, T., Isomura, M., Ohkubo, T., Tabara, Y., Yamamoto, K., Yokota, M., Liu, X., et al. (2017). The fine-scale genetic structure and evolution of the Japanese population. *PloS One*, **12**(11), e0185487.
- Terhorst, J., Kamm, J.A. and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**(2), 303–309.
- Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Anguoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome’. *Proceedings of the National Academy of Sciences of the United States of America* **102**(39), 13950–13955.
- Tettelin, H., Riley, D., Cattuto, C. and Medini, D. (2008). Comparative genomics: The bacterial pan-genome. *Current Opinion in Microbiology* **11**(5), 472–477.
- Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**(1), 31–40.
- Touchon, M., Hoede, C., Tenaillyon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics*, **5**(1), e1000344.
- Trask, B.J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen, O.T., Eichler, E., Van Den Engh, G., Rouquier, S., Shizuya, H., et al. (1998). Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Human Molecular Genetics* **7**(13), 2007–2020.
- UK10K Consortium, et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* **526**(7571), 82–90.
- Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, **4**(3), e72.
- Wakeley, J. (1997). Using the variance of pairwise differences to estimate the recombination rate. *Genetics Research* **69**(1), 45–48.
- Wakeley, J. and Lessard, S. (2003). Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics* **164**(3), 1043–1053.

- Wall, J.D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution* **17**(1), 156–163.
- Wall, J.D. (2004). Estimating recombination rates using three-site likelihoods. *Genetics* **167**(3), 1461–1473.
- Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *American Journal of Human Genetics* **71**(5), 1227–1234.
- Wasser, S.K., Shedlock, A.M., Comstock, K., Ostrander, E.A., Mutayoba, B. and Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America* **101**(41), 14847–14852.
- Watterson, G. (1977). Heterosis or neutrality? *Genetics* **85**(4), 789–814.
- Watterson, G. (1978). The homozygosity test of neutrality. *Genetics* **88**(2), 405–417.
- Weir, B. and Hill, W. (1986). Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, **38**(5), 776.
- Weir, B.S. (1996). *Genetic Data Analysis: Methods for Discrete Population Genetic Data*, 2nd edition. Sinauer Associates, Sunderland, MA.
- Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H. and Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**(2), 238–251.
- Wilton, P.R., Carmi, S. and Hobolth, A. (2015). The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics* **200**(1), 343–355.
- Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology* **55**(3), 248–259.
- Wu, K., Otoo, E.J. and Shoshani, A. (2002). Compressing bitmap indexes for faster search operations. In *Scientific and Statistical Database Management, 2002. Proceedings. 14th International Conference on*, pp. 99–108, Los Alamitos, CA. IEEE Computer Society.
- Xifara, D.-K. (2014). *The detection, structure and uses of extended haplotype identity in population genetic data*. PhD thesis, University of Oxford.
- Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences* **99**(11), 7335–7339.
- Zheng, C., Kuhner, M.K. and Thompson, E.A. (2014). Bayesian inference of local trees along chromosomes by the sequential Markov coalescent. *Journal of molecular evolution* **78**(5), 279–292.

3

Haplotype Estimation and Genotype Imputation

Jonathan Marchini

Regeneron Genetics Center, Tarrytown, NY, USA

Abstract

Humans inherit one copy of each chromosome from each parent. These chromosome copies consist of sequences of alleles and are referred to as haplotypes. Modern SNP microarrays only assay genotypes at a subset of all polymorphic sites and do not directly measure haplotypes. This chapter provides an overview of statistical models and computational approaches for inferring haplotypes from genotypes and for imputing (predicting) unmeasured genotypes based on dense reference sets of haplotypes. Haplotypes and imputed genotypes form the basis of many downstream analyses in the study of human disease and population genetics.

3.1 Haplotype Estimation

A haplotype is defined as a combination of alleles at a set of loci on a single chromosome, that have all been inherited from a specific parent. In some applications a haplotype might span a whole chromosome, while in others interest may lie in a haplotype spanning a shorter stretch of sequence such as a gene. Genotyping technologies produce genotype data that does not include explicit haplotype information. At each locus the unordered pair of alleles is measured, with no information as to the parent of origin of each allele. Haplotype estimation, also known as phasing or haplotyping, describes the statistical problem of recovering the haplotypes that underlie the genotypes at a set of genotyped markers. Markers are most often single nucleotide polymorphisms (SNPs), but can also be short insertions/deletions (indels) or structural variants.

Since haplotypes are the fundamental units of inheritance, they form the preferred data type for human genetic studies. Haplotypes are used in a variety of applications such as studies of human genetic variation (International HapMap Consortium, 2005; 1000 Genomes Project Consortium *et al.*, 2015), detection of selection (Sabeti *et al.*, 2007), estimation of recombination rates (Myers *et al.*, 2005), ancestry estimation (Hellenthal *et al.*, 2014) and disease association studies (Wellcome Trust Case Control Consortium, 2007). The accuracy of haplotypes can have a downstream effect on these studies, so phasing has become a well-studied problem in statistical genetics, with many different approaches being proposed.

Phasing can be challenging as the number of possible solutions increases exponentially with the length of the sequence of genotypes being phased. Consider genotype data collected on N

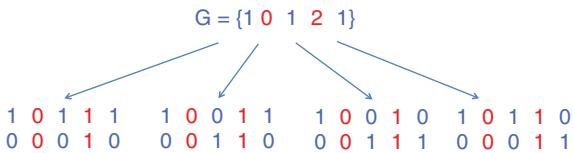


Figure 3.1 A vector of five genotypes (top) that consists of three heterozygous sites (blue) and two homozygous sites (red). The 2^2 possible haplotype pairs (diplotypes) are shown below.

individuals at L SNPs denoted by $G_i = (G_{i1}, \dots, G_{iL})$, $i = 1, \dots, N$, with $G_{il} \in \{0, 1, 2\}$ denoting the count of a specific allele, for example, the non-reference allele, for individual i at site l . The number of different possible pairs of haplotypes (or diplotypes) that would be consistent with the i th individual's genotype vector, G_i , is 2^{C-1} , where $C = \sum_{l=1}^L I(G_{il} = 1)$ is the number of heterozygous genotypes in G_i and $I(x)$ denotes the indicator function on the true/false condition x . See Figure 3.1 for a simple example, where 1 and 0 at a site on a haplotype denote whether the allele present at this site is the non-reference allele or not. In many applications genotypes will span whole chromosomes, leading to a vast possible solution space for each individual.

3.1.1 A Simple Haplotype Frequency Model

One of the earliest approaches to haplotype estimation involves posing the problem as a missing-data problem (Excoffier and Slatkin, 1995). In this approach, if $G = (G_1, \dots, G_N)$ denotes the genotype data on N individuals, then the true haplotypes of each individual, denoted by $h_i = \{h_{i1}, h_{i2}\}$, $i = 1, \dots, N$, where $h_{ij} = (h_{ij1}, \dots, h_{ijL})$, $h_{ijl} \in \{0, 1\}$, are considered missing data. Assuming no genotyping errors, there will be a combined set of D unique haplotypes that could underlie these genotypes. The notation $h_a \oplus h_b = G_l$ can be used to denote that two haplotypes h_a and h_b are consistent with G_l .

A generative model is then used in which each of the D haplotypes is assumed to have sample frequency given by the elements of the D -dimensional vector $\phi = \{\phi_1, \dots, \phi_D\}$, with $\sum_{d=1}^D \phi_d = 1$. Individuals are assumed independent and the likelihood of each individual's genotype vector is then modeled by averaging over all the possible combinations of consistent haplotypes. The likelihood is

$$L(\phi) = \prod_{i=1}^N P(G_i|\phi) = \prod_{i=1}^N \sum_{a,b:h_a \oplus h_b = G_i} \phi_a \phi_b. \quad (3.1)$$

An EM algorithm is used to find a maximum likelihood estimate of ϕ . This involves alternating between two algorithmic steps, an E-step and an M-step. At the t th iteration of the algorithm the D -dimensional haplotype frequency vector $\phi^{(t)}$ is updated and is guaranteed to increase the likelihood from the previous iteration due to the properties of the EM algorithm (see Chapter 1).

The E-step involves the *complete-data log-likelihood* which assumes that the missing haplotypes are known:

$$\log P(G, H|\theta) = \sum_{i=1}^N \sum_{d=1}^D (I(h_{i1} = s_d) + I(h_{i2} = s_d)) \log \phi_d, \quad (3.2)$$

where $\{s_1, \dots, s_D\}$ are the D haplotypes that are consistent with the data. The expectation of this quantity is then taken over the conditional distribution of the missing data, $H = \{h_1, \dots, h_N\}$, given the observed data G and current estimate $\phi^{(t)}$. For the i th individual, if $h_a \oplus h_b = G_i$ then this conditional distribution is given by

$$P(h_a, h_b|G_i, \phi^{(t)}) \propto P(G_i|h_a, h_b, \phi^{(t)}) = [2 - I(h_a = h_b)]\phi_a \phi_b. \quad (3.3)$$

The M-step then involves maximizing the expected complete-data log-likelihood to obtain a new estimate of ϕ , which occurs when

$$\phi_d^{(t+1)} = \frac{1}{2N} \sum_{i=1}^N \sum_{h_a \oplus h_b = G_i} P(h_a, h_b | \phi^{(t)}) [I(h_a = s_d) + I(h_b = s_d)]. \quad (3.4)$$

The algorithm terminates when the likelihood stops increasing significantly, and the method is usually run multiple times from different start points to better estimate the maximum likelihood solution. The conditional distributions in equation (3.3) can be used to infer most likely haplotypes, or carry through uncertainty into downstream analysis such as haplotype association tests (Morris, 2005). As the length of sequence grows, D increases exponentially. Since the model involves parameterizing the whole possible solution space, the method becomes intractable for long sequences. It is unlikely that all D haplotypes will exist in G , so much effort is spent estimating parameters likely to be 0.

3.1.2 Hidden Markov Models for Phasing

A key flaw with the simple haplotype frequency model in Section 3.1.1 is that it includes no dependence between the elements of ϕ . In other words, the model assumes that haplotypes are conditionally independent given ϕ . A study of population genetics provides the perspective that the joint distribution of a set of haplotypes can be described by a coalescent model (Kingman, 1982, 2000). This model has the property that if a sample contains many copies of a given haplotype h it is likely that the other haplotypes in the sample will be similar haplotypes to h in terms of mutation and recombination.

3.1.2.1 PHASE and the Li and Stephens model

PHASE v1 was the first method to suggest a model incorporating this idea (Stephens *et al.*, 2001). The authors suggested a Markov chain Monte Carlo (MCMC) algorithm to sample from the joint set of haplotypes H (for more on MCMC algorithms, see Chapter 1). The algorithm scans through individuals in a random order, and at each stage updates the current estimates of the i th individual's haplotypes h_i by sampling from a distribution of haplotypes conditional upon the remaining set of other haplotypes H_{-i} , denoted by $P(h_i | H_{-i})$. The authors suggested decomposing this as

$$P(h_i | H_{-i}) = P(h_{i1} | H_{-i})P(h_{i2} | H_{-i}, h_{i1}). \quad (3.5)$$

PHASE v1 used a coalescent approximation that assumed haplotypes differed only via mutation, and this was later extended (Stephens and Scheet, 2005) to allow for recombination using a hidden Markov model (HMM) formulation based on the popular and influential Li and Stephens (2003) model. For a detailed description of HMMs, see Chapter 1.

Both the terms in the right-hand side of equation (3.5) have the form $P(f|F)$, where $f = (f_1, \dots, f_L)$ is a haplotype at a set of L bi-allelic sites with $f_l \in \{0, 1\}$ and F consists of M fully observed haplotypes at the same L sites such that $F_{ml} \in \{0, 1\}$. The Li and Stephens model is a generative model that specifies that f is modeled as an *imperfect mosaic* of a set of other haplotypes in F . The model can be written as

$$P(f|F, \theta, \rho) = \sum_Z P(f|Z, F, \theta)P(Z|F, \rho), \quad (3.6)$$

where $Z \in \{Z_1, \dots, Z_L\}$ and $Z_l \in \{1, \dots, M\}$ is a sequence of unobserved copying states for the L sites, indicating which of the M haplotypes f is an imperfect copy of at each of the L sites. The term $P(Z|F, \rho)$ models the *transition probabilities* of the HMM and models how the copying state vector Z changes along the sequence. The parameter ρ controls the rate of change of

switching between copying states. The term $P(f|Z, F, \theta)$ models the *emission probabilities* and allows each observed haplotype to differ from the haplotypes specified by the copying vector Z .

The parameters ρ and θ have explicit links to the fine-scale recombination rate along the sequence and mutation rate, respectively. The transition probabilities are modeled as

$$P(Z|F, \rho) = P(Z_1) \prod_{l=1}^{L-1} P(Z_{l+1}|Z_l),$$

where

$$P(Z_1) = \frac{1}{M}, \quad P(Z_{l+1} = a|Z_l = b) = \begin{cases} e^{\frac{-\rho_l d_l}{M}} + \left(1 - e^{\frac{-\rho_l d_l}{M}}\right) \frac{1}{M} & \text{if } a = b, \\ \left(1 - e^{\frac{-\rho_l d_l}{M}}\right) \frac{1}{M} & \text{if } a \neq b. \end{cases} \quad (3.7)$$

Here d_l is the physical distance between sites l and $l + 1$, $\rho_l = 4N_e c_l$, where N_e is the effective diploid population size, and c_l is the average rate of crossover per unit physical distance per meiosis between sites l and $l + 1$. PHASE estimates the fine-scale recombination rates.

This part of the model captures the important feature of real data sets that as the number of conditioning haplotypes M increases the transition rate to a different state decreases. In other words, a given haplotype is more likely to share long stretches of sequence with another haplotype in F , as M increases.

The emission probabilities are modeled as

$$P(f_l = a|Z_l = k, F, \theta) = \begin{cases} M/(M + \theta) + (1/2)\theta/(M + \theta) & F_{kl} = a, \\ (1/2)\theta/(M + \theta) & F_{kl} \neq a. \end{cases} \quad (3.8)$$

Li and Stephens suggested using the Watterson estimator $\theta = (\sum_{m=1}^{M-1} (1/m))^{-1}$, which is based on the assumption of an expected number of mutation events per site of 1.

The PHASE software (Stephens and Donnelly, 2003) used a partition-ligation scheme to fit the model. First the data set is divided up into blocks of sites and the MCMC sampler run on each block. To do this, the whole distribution $P(f|F)$ was enumerated at each step. Blocks were then combined and the method rerun, but with low-probability solutions from each block excluded. This strategy for phasing is comparatively very slow compared to current methods, but at the time this method was fast enough that it could be used to phase the HapMap Project data set which consisted of hundreds of samples at ~ 2 million SNP sites (Marchini *et al.*, 2006). In the following, a few of the key current methods will be described.

3.1.2.2 IMPUTE and MaCH

MaCH (Li *et al.*, 2010) and IMPUTE v2 (Howie *et al.*, 2009) developed similar phasing methods based on the same MCMC sampling idea of PHASE, but avoid the use of partition ligation and complete enumeration of the conditional distributions $P(f|F)$. Instead, these methods use a diploid version of the Li and Stephens model with a pair of unobserved copying vectors. The HMM forward-backward calculations in this algorithm have $O(N^2 L)$ time complexity, where N is the number of individuals being phased.

MaCH proposed to reduce the time complexity when N is large by choosing a random subset of haplotypes to condition on at each update step. IMPUTE v2 also uses a subset of the haplotype at each stage but attempts to choose a ‘best’ set of haplotypes. When updating the haplotypes of the i th individual the set of K haplotypes that are close to the current haplotype estimates in terms of Hamming distance are chosen as the conditioning set. IMPUTE v2 uses pre-calculated fine-scale recombination rates that have been estimated in humans (Myers

et al., 2005). The resulting HMM calculations in this algorithm have $O(K^2L)$ time complexity with $K < N$.

3.1.2.3 fastPHASE

fastPHASE (Scheet and Stephens, 2006) proposed an HMM that specifies a set of B unobserved states or clusters designed to represent common haplotypes. The b th cluster is assigned a weight α_{bl} that denotes the fraction of haplotypes it contains at site l , with $\sum_{b=1}^B \alpha_{bl} = 1$. Each cluster also has an associated frequency λ_{bl} of allele 1 at each site. Each individual's genotype data is then modeled as an HMM on this state space with transitions between states controlled by a further set of parameters, r , at each site,

$$P(G_i|\alpha, \lambda, r) = \sum_Z P(G_i|Z, \lambda)P(Z|\alpha, r). \quad (3.9)$$

Here $P(G_i|Z, \lambda)$ models the emission probabilities, that is, how likely the observed genotypes are given the underlying states, and $P(Z_i|\alpha, r)$ models the transition probabilities, that is, the patterns of switching between states. An EM algorithm is used to fit the model and estimate α , λ and r . Pairs of diplotypes are then sampled from $P(h_i|G_i, \alpha, \lambda, r)$. The authors recommend using $B = 20$ clusters and averaging results over 10 random starts.

3.1.2.4 Beagle

Beagle (Browning, 2006; Browning and Browning, 2007, 2009) uses a bifurcating tree structure (or localized haplotype-cluster model) to represent the haplotype frequencies of consecutive SNPs. Given an initial set of haplotype estimates, a bifurcating tree is constructed from left to right across the set of haplotypes, with tree edges weighted by the number of haplotypes that pass along it. The tree is then pruned to produce a more parsimonious characterization of the data set. At each level of the tree, pairs of nodes are compared in terms of their downstream haplotype frequencies by summing the squared differences of their downstream partial haplotype frequencies; if this number exceeds a threshold, then the nodes are not similar enough to combine. Possibly the best way to understand the model is by looking at the small example given in Browning (2006, Figure 2 and Table 1). At each iteration of the algorithm the tree structure is built using the current phased haplotypes of all the individuals. Then each individual's haplotypes are updated by sampling from the induced diploid HMM, conditional upon the individual's genotypes.

The complexity of the tree will vary dependent upon the local linkage disequilibrium (LD) structure of the data and has the attractive property that it can adapt to the local haplotype diversity. The model can be thought of as a local haplotype-clustering model, similar to fastPHASE, but with a variable number of clusters across a region.

A key problem with using a bifurcating tree model is that haplotypes of a given individual are estimated *conditional* upon that model at each iteration, and are not directly compared to each other, as is done in the IMPUTE v2 approach. This means that long stretches of shared sequence between two individuals are not always captured by the model, and this degrades performance, especially as data sets increase in size. For this reason, Beagle v5 (Browning *et al.*, 2018) has recently abandoned the bifurcating tree model in favour the IMPUTE v2 model, but with a novel, fast method of constructing a custom conditioning set of haplotypes.

3.1.2.5 SHAPEIT

The main advance of the SHAPEIT approach was to reduce the complexity of HMM calculations in the models used by MaCH and IMPUTE v2 from quadratic, $O(K^2L)$, in the number of conditioning states K , to $O(KL)$. This was achieved by using a graphical model, denoted S_{G_i} , to represent all the possible haplotypes underlying a given individual's genotypes G_i . The nodes of the graphical model are defined by consecutive chunks of sequence with J heterozygous sites

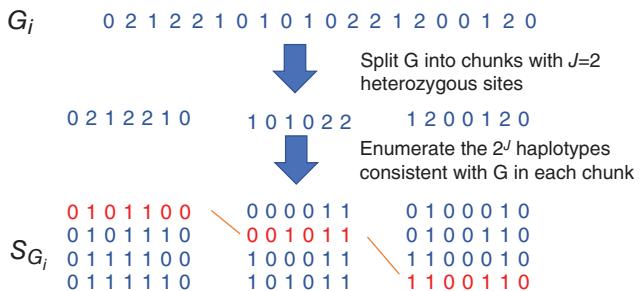


Figure 3.2 In the SHAPEIT v2 model each individual's genotype vector (top) is split into chunks with J heterozygous sites (middle). Within each chunk all 2^J consistent haplotypes are listed and represented as nodes in a graph S_{G_i} . A consistent haplotype spanning the whole length of G_i can be represented as a path through S_{G_i} . An example path is shown in red.

(see Figure 3.2). The key idea is then that the distribution of paths through S_{G_i} conditional upon a set of known haplotypes can be modeled as a Markov chain. The transition probabilities of this Markov chain can be learned using a forward-backward algorithm for HMMs with a complexity $O(KL)$ (for details on this algorithm, see **Chapter 1**). Pairs of haplotypes consistent with G_i can then be quickly sampled. A further key part of this approach involves gradually pruning away states in S_{G_i} that seem unlikely given the data, and then merging chunks together. This approach has the property that the solution space is gradually reduced as the method proceeds.

SHAPEIT v1 (Delaneau *et al.*, 2012) and SHAPEIT v2 (Delaneau *et al.*, 2013b) differ in how they represent the set of conditioning haplotypes. SHAPEIT v1 used a collapsed state space reduction approach in which the conditioning haplotypes are split into blocks of sites such that within each block there will be just a small number of unique haplotypes. SHAPEIT v2 chooses a local subset of conditioning haplotypes (as in IMPUTE v2) for each individual, at each iteration of the algorithm, resulting in much better performance than SHAPEIT v1 (Delaneau *et al.*, 2013b).

Choosing the local subset of haplotypes involves a $O(4N^2)$ Hamming distance calculation step, and this can become a problem in large data sets with more than 20,000 samples. For this reason, and also motivated by the problem of phasing the genotype data from the UK Biobank project which has 500,000 samples, SHAPEIT v3 was developed (O'Connell *et al.*, 2016). Large sample sizes will lead to increased local similarity between groups of haplotypes due to the higher probability of more recent shared ancestry. This idea is exploited by using a recursive clustering algorithm to partition the haplotypes into clusters of similar haplotypes of specified size $Q \ll 2N$. Distances are computed only between haplotypes within a cluster. This reduces the complexity of this step to $O(Q^2)$ such that the complexity of the whole algorithm is dominated by the $O(N \log N)$ scaling of the clustering routine. The algorithm also stops updating a haplotype if it detects a perfect match with another haplotype across a long stretch of sequence.

SHAPEIT v4 (Delaneau *et al.*, 2018) makes a further advance on the IMPUTE v2 method of choosing a local subset of haplotypes that makes use of the positional Burrows–Wheeler Transform (pBWT) that is described in more detail in Section 3.2.3.4. The method stores the current haplotypes at each iteration in a pBWT data structure, which facilitates locally matching haplotypes to be identified in a given window. This approach has the attractive property that the number of conditioning haplotypes varies from window to window, and locally adapts to the haplotype structure in the data set. This, together with a new, fast initialization method (also based on the pBWT) and re-factoring the SHAPEIT code base, resulted in a highly accurate and computationally efficient method that outperforms SHAPEIT v3, EAGLE v2 and Beagle v5 on large data sets such as the UK Biobank.

3.1.2.6 HAPI-UR

HAPI-UR (Haplotype Inference for Unrelated Samples) uses a haploid HMM in which markers are created from multiple flanking SNPs (Williams *et al.*, 2012), in a similar way to the collapsed state space reduction approach used by the SHAPEIT v1 model, although there are differences in how these markers are defined. HAPI-UR constructs individual-specific diploid HMMs that enumerate just those diploid states that are consistent with the individual's genotypes, which is also very similar to the method employed by the SHAPEIT models. A key difference is that SHAPEIT allows for mutation in the emission part of its model. The method uses MCMC to iteratively update the haploid HMM and sample new haplotype estimates from the individual-specific diploid HMMs.

3.1.3 Phasing in Related Samples

A special case of the phasing problem arises when the samples to be phased consist of close relatives. The simplest case occurs when genotype data has been collected on mother–father–child trios. In this case Mendel's rules dictate strong constraints on the solution space of haplotypes. If the genotypes of the mother, father, and child are fully observed at an SNP, then the parental origin of the two alleles carried by the child can be completely determined in all cases except when all three individuals are heterozygous. Combining this idea across sites allows sequences of alleles all with the same parental origin to be inferred, and produces haplotype estimates in the child and both parents. Sites very far apart on a chromosome can have their genotype successfully phased, and trio phasing typically results in highly accurate estimated haplotypes. A similar scenario arises with parent–child duos, although less phase information can be inferred in this case. Beagle and SHAPEIT v2 both have specific options to phase trios and parent–child duos.

Two caveats to this approach are that missing genotype data may also cause ambiguous phase, and that estimated haplotypes in parents are inferred as those transmitted and untransmitted to the child, which will be recombined versions of the true parental haplotypes.

Approaches from the literature on linkage analysis also exist for handling more complex family structures or pedigrees (Lange *et al.*, 2013; Sobel and Lange, 1996; Abecasis *et al.*, 2002; Gudbjartsson *et al.*, 2000). However, such methods face several limitations; approaches based on the Lander–Green algorithm have computational and space complexity that scale exponentially with sample size; they can be sensitive to genotyping error and they can only phase sites where at least one member of the pedigree is not heterozygous. The last point is particularly crucial, as it means the haplotypes will not be 'complete' and cannot be easily used in applications such as imputation, which require complete haplotypes at a full set of sites.

The SHAPEIT v2 model does remarkably well at haplotype estimation on very closely related individuals ignoring any known pedigree information (O'Connell *et al.*, 2014). However, the resulting haplotypes will not generally be completely consistent with any known pedigree information. The duoHMM approach (O'Connell *et al.*, 2014) proposes a model-based approach to update an initial set of haplotype estimates based on the available family/pedigree information. An HMM is used to model the unobserved 'gene flow' between parent and child duos. Given two haplotypes in each of a parent and child, the gene flow is defined as the pattern of inheritance between these four haplotypes along the sequence, that is, which pair of haplotypes is shared between parent and child at each site. Along a chromosome the true gene flow will remain constant for long stretches due to the low rate of recombination in any given meiosis. Errors in phasing in either parent or child will occur more frequently than real recombination events and can cause switches in gene flow. By inferring the switching rates of the HMM it is possible to infer the underlying gene flow and obtain updated haplotype estimates and probabilities of recombination at each locus. The number of recombinations detected by this approach was

shown to be in good agreement with expectations from genetic map distances and much lower than those produced by MERLIN (Abecasis *et al.*, 2002). The duoHMM approach is incorporated into the SHAPEIT v2 software.

Kong *et al.* (2008) introduced an idea known as long-range phasing (LRP) that leverages detected segments of shared sequence between individuals. The method searches for pairs of individuals who share long stretches of sequence. These are used to select a set of ‘surrogate parents’ for each individual in a given window. Trio phasing rules are then applied to infer phase. This approach was applied successfully in a rule-based way to data from an Icelandic population. This is a relatively small and isolated population, a large proportion of whom have been genotyped. Implementations of this approach include a model-based version called Systematic Long Range Phasing (SLRP) (Palin *et al.*, 2011) and the ALPHAPHASE software developed for livestock populations (Hickey *et al.*, 2011). However, both approaches suffer from the problem that phase can only be inferred for genomic regions where identity by descent (IBD) sharing is detected.

The EAGLE v1 method (Loh *et al.*, 2016b) implemented an LRP approach which makes initial phase calls based on > 4 cM tracts of IBD sharing between pairs of individuals, identified by looking for long stretches of identical homozygous genotypes. This is then followed up with two approximate HMM decoding iterations to refine phase.

3.1.4 Phasing Using Sequencing Data

High-throughput sequencing is becoming widely used in all aspects of human disease genetics and population genetics. Estimating haplotypes from sequencing data is challenging due to the higher density and generally lower frequency of polymorphic sites than on typical SNP microarrays. Low-frequency sites can be harder to phase, and the increased SNP density increases the necessary computation. Furthermore, when sequencing coverage is low, genotypes are only partially observed, with each genotype having some level of uncertainty depending on the number of reads that cover the site. It has become the norm to represent this uncertainty using a *genotype likelihood* (GL) at each site in each individual. These GLs take the form $P(\text{Reads}|G_{il})$, where $G_{il} \in \{0, 1, 2\}$. Beagle, IMPUTE v2, and a version of the MaCH known as Thunder are all able to handle GLs using their respective haplotype models.

These approaches all produce haplotype estimates for each sample, and as a consequence can produce genotypes at each site. For this reason this process is sometimes referred to as *genotype calling from sequencing*, rather than phasing. This is distinct from the *genotype calling from microarrays*, for which a different set of methods are required (Rabbee and Speed, 2006; Teo *et al.*, 2007).

The SNPtools approach (Wang *et al.*, 2013) is a novel approach for phasing sequence data and genotype calling. Like other approaches, the method employs an MCMC algorithm that iteratively updates an individual’s haplotypes using a Li and Stephens HMM, but only conditional upon a very small subset of all other individuals’ haplotypes of size 4, which is fast to compute. The four haplotypes are referred to as ‘parental’ haplotypes and can be thought of as unknown parameters of the model. A separate Metropolis–Hastings (MH) update step is used to sample from the space of all possible haplotype subsets of size 4, and it is this step, which is repeated many times per iteration for each individual, that takes up the majority of the computation time. A variant of this approach was used to call genotypes for the Haplotype Reference Consortium (HRC) project which consisted of mostly low-coverage sequence data on ~32,500 samples (McCarthy *et al.*, 2016). In this study an initial set of haplotypes on the samples was used to constrain the search space of the MH update step for the conditioning haplotypes for each individual.

The SHAPEIT v2 model is not the ideal model for handling GL data as it involves enumerating all possible haplotypes consistent with a set of genotypes. If genotypes are uncertain, as they are when considering GLs, the space of possible haplotypes becomes very large, and this causes problems of convergence. However, the method can handle a reasonable number of uncertain genotypes. For phasing the 1000 Genomes Project data (1000 Genomes Project Consortium *et al.*, 2015) a hybrid approach was developed which used an initial run of Beagle to determine a set of confident genotype calls. These were then fixed and the remaining genotype calls were made using SHAPEIT v2. This approach also leveraged the fact that many of the 1000 Genomes samples had also been genotyped on SNP microarrays together with many of their close first-degree relatives. Accurate haplotypes estimated at these microarray sites were used to define a set of sites at which phase was fixed. As a result, this constrains and improves the genotype calling at the remaining sites (Delaneau *et al.*, 2014).

When sequencing coverage is low ($2\times$ or lower) there can be considerable uncertainty about the underlying genotypes, which increases the difficulty of the phasing and genotyping problem. The Beagle model has been the most widely used method in this setting and was used on the CONVERGE study that collected $1.7\times$ sequencing data from 11,670 Han Chinese women (CONVERGE Consortium, 2015). It has also been proposed that for fixed study funds, low-coverage sequencing augmented with phasing and imputation may be a more powerful approach for genome-wide association studies (GWASs) than using a genotyping chip (Pasaniuc *et al.*, 2012).

The STITCH (Sequencing To Imputation Through Constructing Haplotypes) method was developed for genotype calling from very low sequencing data. The model used is based on the fastPHASE model, with a set of K unknown ancestral haplotypes (or clusters), but with a modified set of emission probabilities that model read level data. The model scales $O(K^2L)$ and was shown to work well on $0.15\times$ sequencing reads on outbred laboratory mice. For application to human data, where a larger value of K is needed, the authors proposed a *pseudo-haploid* model that attaches unobserved labels to each read denoting their parental origin. The resulting EM-based maximum likelihood method has complexity $O(KL)$. This method was shown to outperform Beagle when applied to the CONVERGE data, and has been applied at scale to low-coverage sequence data from non-invasive prenatal testing (Liu *et al.*, 2018).

Sequencing reads are short stretches of sequence and can be thought of as mini-haplotypes and so can contain phase information that can potentially aid performance. SHAPEIT v2 has been extended (Delaneau *et al.*, 2013a) to use phase informative reads but is primarily designed for processing high-coverage sequence data or data sets that have already had genotypes called. In this setting, SHAPEIT v2 will build two separate Markov models on the space of consistent haplotypes for each individual at each iteration: one based on all the other haplotypes in the sample, and one based on the read information. These are then combined and an update of each individual's haplotypes occurs through sampling from the Markov chain. This approach is especially beneficial at the rarest SNPs and can phase a substantial fraction of singleton SNPs. This methodology has been retained in the SHAPEIT v4 software (Delaneau *et al.*, 2018), but now relies on the WhatsHap (Patterson *et al.*, 2015) method as a pre-processing method to extract the phase-informative read information by grouping heterozygous genotypes into phase sets when they are overlapped by the same sequencing reads.

3.1.5 Phasing from a Reference Panel

Over time the availability of phased haplotype reference panels in humans has gradually increased both in the number of samples and number of SNPs, indels and structural variants (see Table 3.1). These haplotypes can be used to help phase new unphased samples. A special

Table 3.1 History of haplotype reference panels used for phasing and genotype imputation

Reference	Year	Number of haplotypes	Number of populations	Number of variants	SNPs	Indels + SVs
HapMap 2	2006	420	3	2,139,483	✓	
HapMap 3	2009	2,368	11	1,440,616	✓	
1000 Genomes Pilot	2010	358	3	14,894,361	✓	
1000 Genomes 2010	2010	1,258	14	22,242,654	✓	
1000 Genomes Interim	2011	2,186	14	38,558,931	✓	
1000 Genomes Phase 1	2012	2,186	14	38,219,282	✓	✓
1000 Genomes Phase 3	2014	5,008	25	~88,000,000	✓	✓
UK10K	2014	7,562	1	26,032,603	✓	✓
HRC	2016	64,976	30	39,235,157	✓	

case of this approach occurs in applications of personalized genetic medicine where a single individual has been sequenced at high coverage and needs to be phased. SHAPEIT v2 and Beagle both include functionality of phasing against a reference panel.

The EAGLE2 method (Loh *et al.*, 2016a) condenses the reference panel into a compact data structure, called a HapHedge, that can be thought of as a set of tree structures that model the local haplotype structure in a lossless fashion. The method selectively explores the space of diplotypes using a branching-and-pruning beam search that only expends computation on the most likely phase paths.

The recently introduced SHAPEITR method is specifically focused on phasing high-coverage sequenced samples and phasing them using large reference panels such as the HRC (Sharp *et al.*, 2016). This approach leverages the rarest variants in the reference panel to help inform an HMM about which reference haplotypes to choose as possible copying states. If two individuals share a rare variant then this increases the chance that these individuals share a long stretch of sequence around that site. Using rare variants in this way was shown to improve the accuracy of the default SHAPEIT v2 approach that iteratively chooses copying states using variants of all frequencies. Accuracy can be further enhanced through the use of phase informative reads. This approach is publicly available via the Oxford Phasing Server (<https://phasingserver.stats.ox.ac.uk/>).

Simulating haplotype data, conditional on a haplotype reference panel, can be achieved using the HAPGEN program and can be useful when testing new methods (Su *et al.*, 2011). This software can also simulate haplotype data in case–control samples by specifying a small number of causal loci together with the relative risks of the disease alleles.

3.1.6 Measuring Phasing Performance

The performance of phasing methods is usually measured using samples with known phase derived from family information. For example, mother–father–child trio genotypes can be used to infer the parental origin of most sites across a chromosome. By including just the parents, or the child, in sets of samples to be phased, the inferred haplotypes in these individuals can be compared to true haplotypes, at sites inferred from the full trio. An alternative approach combines male chromosome X haplotypes together to form unphased genotypes that can then be phased and compared to the true haplotypes.

Errors in phasing occur as ‘switches’ in phase between consecutive heterozygous sites. If T is the total number of heterozygous sites in a sample across a stretch of sequence of length W and

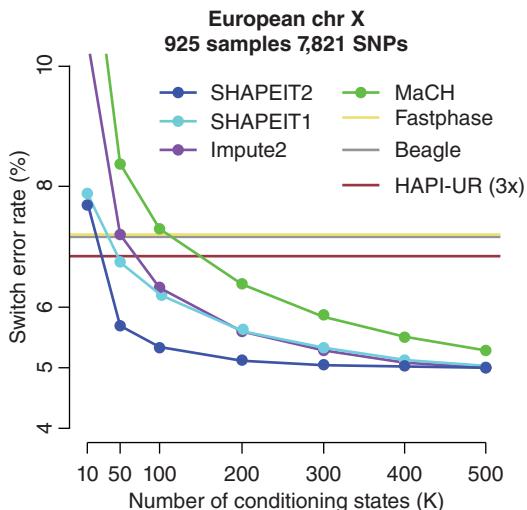


Figure 3.3 Accuracy comparison of different phasing methods when applied to a test data set derived from phase-known male X chromosome haplotypes. The y-axis shows the switch error of the different methods. The x-axis shows the number of conditioning states used by some of the methods, which can be used to control accuracy. Figure reproduced from the SHAPEIT v2 paper (Delaneau *et al.*, 2013b).

S is the number of switches needed to correct the inferred haplotypes so that they agree with the true haplotypes then the switch error is defined as $e = S/T$ and the mean distance between switches as $d = W/S$. Alternatively, the proportion of phase correct segments of some length, say 10 Mb, can be reported. The factors that affect accuracy will be similar to those that affect genotype imputation performance (discussed in Section 3.2.6). These include the density and allele frequency of the sites being phased, the ancestry, relatedness and number of the samples being phased, and whether a reference panel is used to help with phasing. It is important to take this into account when comparing error rates across experiments and data sets.

Several papers have compared the performance of different phasing methods (Sheet and Stephens, 2006; Delaneau *et al.*, 2012, 2013b; O'Connell *et al.*, 2014; Loh *et al.*, 2016a,b; Browning and Browning, 2011). Figure 3.3 reproduces a figure from the SHAPEIT v2 paper (Delaneau *et al.*, 2013b) that compares many of the methods described in this chapter using a chromosome X test data set. The accuracy of many methods can be controlled by a user-defined parameter which is the number of conditioning states (x -axis), but this will also impact running time.

Herzig *et al.* (2018) compared the performance of ALPHAPHASE, SLRP, SHAPEIT v2, SHAPEIT v3, Beagle, EAGLE1 and EAGLE2 in the setting of phasing data from isolated populations, where haplotype sharing between samples is expected to be relatively high. The conclusion was that SHAPEIT v2, SHAPEIT v3 and EAGLE2 produced the most accurate results in these tests.

On the very large cohorts such as the UK Biobank data set switch error rates can become very low. SHAPEIT v3, EAGLE1 and EAGLE2 have reported switch error rates as low as 0.3%, which corresponds to just a handful of switch errors across a chromosome in many cases. This highlights the ability of these methods to detect and leverage the long stretches of shared sequence between individuals as sample size gets very large.

3.2 Genotype Imputation

In a typical GWAS in humans a large number of markers (usually hundreds of thousands of SNPs) are assayed across a genome in thousands of individuals. The markers on commercial genotyping arrays are usually designed so that a large fraction of common genetic variation is captured, but there will always remain SNPs that have not been directly genotyped. These

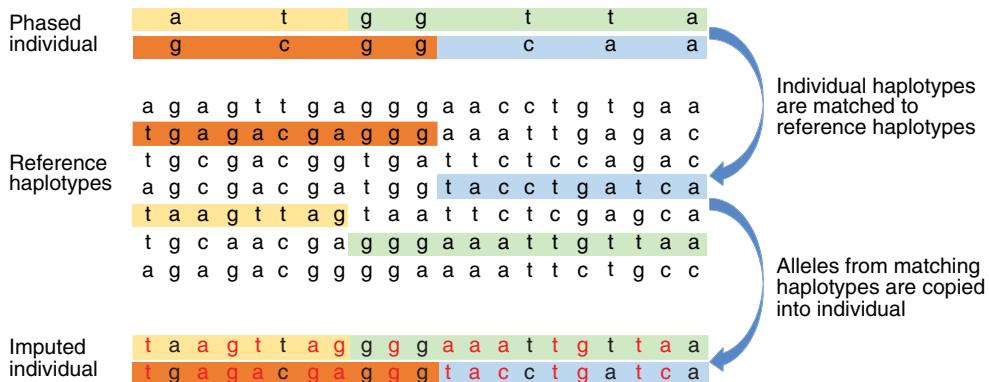


Figure 3.4 Basic idea of genotype imputation. The phased haplotypes of a given individual (top) are compared to a set of reference haplotypes at a denser set of sites (middle). All imputation methods work by modeling matches between the reference haplotypes and individual haplotypes (shown as colored stretches of sequence). The alleles on matching reference haplotypes are then copied into the phased individual's haplotypes to produce imputed alleles (bottom). Matches are not normally as unique as those in this simple example, and methods average over many possible matches and produce a probability distribution for the unobserved alleles/genotypes.

ungenotyped markers will mostly have low minor allele frequencies, and may not be in strong LD with any single marker assayed by the array (see **Chapter 2** for a detailed description of linkage disequilibrium). It would be desirable to be able to test these ungenotyped markers for association with any given phenotype in the study individuals. For this reason methods have been developed that attempt to predict, or impute, the ungenotyped sites. These methods take advantage of additional data sets of genetic variation that typically contain a much larger and more complete set of markers. Such data sets are typically phased and referred to as *haplotype reference panels*. Imputation methods take advantage of sharing of haplotypes of relatively short stretches of sequence between study individuals and haplotype reference panels. Figure 3.4 illustrates the basic idea behind the genotype imputation process.

3.2.1 Uses of Imputation in GWASs

There are three main reasons to carry out genotype imputation within the context of a GWAS. Firstly genotype imputation will usually greatly increase the number of SNPs that can be tested for association. For example, the UK Biobank data set started with 800,000 variants and imputation increased this to over 96 million variants (Bycroft *et al.*, 2018). This adds greatly to the resolution in a given region and helps infer (or 'fine-map') the most likely causal variant (Benner *et al.*, 2016). Secondly, imputation can increase the power to detect a causal variant, and it can often be observed that in regions harboring true causal associations the imputed SNPs have more significant signals of association than the directly genotyped SNPs. These first two points are illustrated in Figure 3.5 which shows association test results from a study of brain imaging phenotypes in the UK Biobank (Elliott *et al.*, 2018). Finally, imputation can facilitate meta-analysis of studies. If two or more studies have been genotyped using different SNP microarrays, and so consist of different sets of SNPs, then after imputation using a given reference panel both studies will consist of genotypes at an almost identical set of SNPs. This allows researchers to combine results, via meta-analysis approaches, to boost sample sizes and uncover more associations (Zeggini *et al.*, 2008; for more details, see **Chapter 22**). This has led

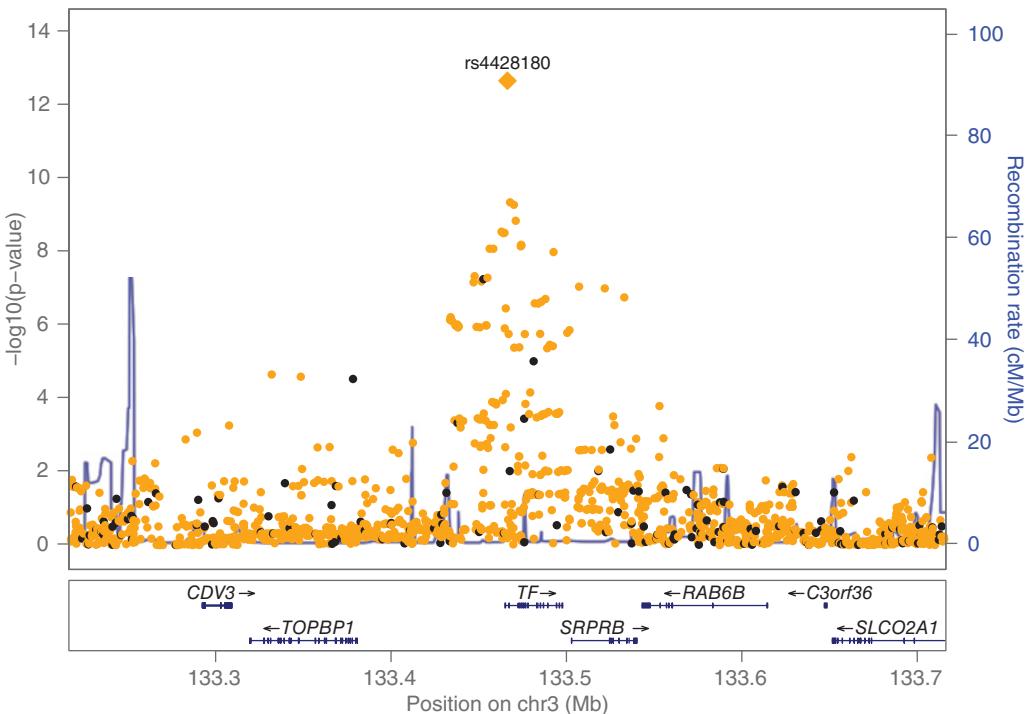


Figure 3.5 Genetic association for right Pallidum SWI T2* MRI measurements in UK Biobank participants. Each plotted point represents the association test results for imputed SNPs (orange) and genotyped SNPs (black). The x-axis is the physical position of the SNPs. The y-axis is the $-\log_{10} p$ -value for the association test. The plot highlights how imputed SNPs provide more significant associations than the genotyped SNPs, with one imputed SNP (rs4428180) clearly identified.

to a culture change in human disease genetics where it is now normal to share and combine association summary statistics across studies.

3.2.2 Haploid Imputation

Imputation usually starts by first phasing the individuals to be imputed, followed by a haploid imputation step in which each haplotype is imputed separately, as illustrated in Figure 3.4. The first step of this process is known as ‘pre-phasing’, and is normally carried out using one of the methods described in Section 3.1.

Several of the existing methods for haploid imputation use an HMM based on the Li and Stephens model described in Section 3.1.2.1. In this approach alleles are imputed into the haplotypes of a given individual independently of each other. As above we use the notation $h_{ij} = (h_{ij1}, \dots, h_{ijL})$ to denote the j th haplotype of the i th diploid individual at a set of L bi-allelic sites $i \in \{1, \dots, N\}$, $j \in \{1, 2\}$. The key difference here is that we will assume that only some of the alleles are observed, $h_{ijl} \in \{0, 1\}$, and the majority will be missing. Here we use H to denote the haplotype reference panel that consists of M fully observed haplotypes at the same L sites such that $H_{ml} \in \{0, 1\}$. The model can be written as

$$P(h_{ij}|H, \theta, \rho) = \sum_Z P(h_{ij}|Z, H, \theta)P(Z|H, \rho), \quad (3.10)$$

where $Z \in \{Z_1, \dots, Z_L\}$ and $Z_l \in \{1, \dots, M\}$ is a sequence of unobserved copying states for the L sites.

The idea of the Li and Stephens imperfect mosaic model captures features we expect to see in real data. Haplotypes of a given individual will tend to be the same as, or similar to, those in the reference panel over some stretch of sequence. The length of stretches of sharing between a study haplotype and the reference panel will depend on several factors such as the size of the reference panel, the allele frequency and number of genotyped sites, the match in ancestry between study and reference samples, the quality of the genotypes, and the phase accuracy of the data sets.

Imputation involves inferring the probability distribution of each of the unobserved alleles. If we use Ω and Ψ to denote the set of indices of those ungenotyped and genotyped sites respectively, then for each $l \in \Omega$ the marginal posterior distribution of the copying state can be written as

$$P(Z_l | h_{ij}, H, \theta, \rho) \propto P(\{h_{ij1}, \dots, h_{ijl}\}, Z_l | H, \theta, \rho) P(\{h_{ij(l+1)}, \dots, h_{ijL}\} | Z_l, H, \theta, \rho), \quad (3.11)$$

where the two terms on the right-hand side are known as the forward and backward probabilities and can be efficiently calculated recursively. The most efficient implementations calculate and store these quantities (vectors of length M) at just the $L - |\Omega|$ genotyped sites, with distributions at all sites in Ω being calculable by from these quantities. Due to the symmetry in the Li and Stephens transition probabilities, the algorithm scales $O(ML)$.

The marginal posterior distribution of the unobserved alleles at all the sites in Ω is then calculated as

$$P(h_{ijl} = d | h_{ij}, H, \rho, \theta) = \sum_{Z_l=1}^M P(h_{ijl} = d | Z_l, \theta) P(Z_l | h_{ij}, H, \theta, \rho), \quad \text{where } d \in \{0, 1\}. \quad (3.12)$$

There are some assumptions implicit in the above formulation that may not be true in practice. Genotype imputation brings together two data sets: the haplotype reference panel and the phased set of genotypes in the study samples. It is crucial that these two data sets align in two important ways. Firstly, the positions of a given SNP in the two data sets must be relative to the same build of the human genome. Secondly, the coding of the alleles at a SNP must agree. Most haplotype reference panels code alleles using the forward (or +) strand of the human genome reference sequence, so it is important the allele coding of the study samples is the same. Most of the widely used software packages included functionality to check and/or correct for so-called strand problems.

3.2.3 Imputation Methods

3.2.3.1 IMPUTE

IMPUTE v1 implemented a diploid version of the model in equations (3.10) –(3.12), where the study sample was unphased. This model involves *two* unknown copying vectors, one for each unknown haplotype, and in effect averages over phase when imputing alleles. This avoids the need to phase the study samples but has a disadvantage: the HMM algorithms' complexity is $O(M^2L)$, which becomes increasingly problematic with increasing reference panel size.

IMPUTE v2 (Howie *et al.*, 2009, 2011) implemented a more flexible range of imputation options than IMPUTE v1. Study samples are phased jointly and iteratively using a Li and Stephens based model just at the set of sites in common between the study and reference samples. At each stage of the phasing only a subset of K_1 'best matching' haplotypes are used to update the phase of each individual, which reduces the complexity of the HMM in the phasing

to $O(K_1^2 L)$. At each iteration a haploid imputation step is used to impute the unobserved alleles in each study sample, this time using just a subset of K_2 haplotypes.

IMPUTE v2 also implements the haploid imputation model in equations (3.10) –(3.12), based on the pre-phasing approach (Howie *et al.*, 2012). It can also impute from a pair of reference panels, usually where the first panel is almost a strict subset of the sites in the second panel. In addition, IMPUTE v2 allows two haplotype panels to be merged together to a common set of sites. This approach was used to combine the UK10K project reference panel with the 1000 Genomes Phase 3 panel (Huang *et al.*, 2015).

IMPUTE v4 was developed with the motivation to make the haploid imputation step as fast as possible for imputing the ~500,000 individuals genotyped for the UK Biobank study (Bycroft *et al.*, 2018). It uses compressed data structures for storing the haplotypes and a novel strategy for interpolating copying probabilities between genotyped sites. Suppose l is one of the sites in the ungenotyped set Ω and u and v are the indices of left and right flanking sites that have been genotyped. The marginal copying probabilities at site l are then linearly interpolated from those at sites u and v :

$$P(Z_l|h_{ij}, H, \theta, \rho) = \pi P(Z_u|h_{ij}, H, \theta, \rho) + (1 - \pi)P(Z_v|h_{ij}, H, \theta, \rho), \quad (3.13)$$

where π is the relative genetic position of site l between sites u and v . An advantage of this approach is that since

$$\sum_{m=1}^M P(Z_l = m|h_{ij}, H, \theta, \rho) = 1, \quad (3.14)$$

this term need only be calculated for those states $Z_l = m$ such that H_{ml} carries the rare allele at that site, which can be done efficiently by storing allele indices.

3.2.3.2 MaCH/minimac

The MaCH (Markov Chain Haplotyping) method (Li *et al.*, 2010) is also based on a Li and Stephens HMM. This approach iteratively updated the phase of each study sample, as well as learning crossover and mutation parameters ρ and θ on a per site basis. A hybrid approach in which the parameters are estimated using just a subset of individuals was also proposed.

This approach evolved into the minimac method with the advent of the pre-phasing approach to imputation (Howie *et al.*, 2012). minimac3 uses ideas of state space reduction (Delaneau *et al.*, 2012; Paul and Song, 2012) to reduce computational cost (Das *et al.*, 2016). In this approach the reference panel of haplotypes is split into blocks of sites (with a single overlapping site). Within each block there will be a relatively small number of unique haplotypes. The HMM forward-backward algorithm is able to iterate over only these unique haplotypes without loss of accuracy. In effect the HMM works with the haplotype reference panel represented as a compressed data structure. A data storage file format called M3VCF was introduced to allow haplotype data compression and faster data input into minimac3.

3.2.3.3 Beagle

The original versions of imputation functionality in Beagle collapse the haplotype reference panel into a compact graph structure described in Section 3.1.2.4. An individual's haplotypes are then imputed within the method conditional upon this graph, but no direct matching to the original reference panel haplotypes takes place and so information regarding long stretches of shared sequence between individuals is lost. This likely explains the lower accuracy of this approach when compared to others and was likely the reason that Beagle v4.1 adopted a very similar HMM to the Li and Stephens model (Browning and Browning, 2016). The method

groups SNP markers together within 0.005 cM windows to form aggregated markers with more than two possible alleles (or mini haplotypes). The transition rates of the HMM use the mean position of SNPs with each aggregated marker. For a marker that aggregates k SNPs, the model will emit exactly the same sequence of k alleles as the copied state with probability $\max(1 - k\epsilon, 0.5)$, and all other possible sequences with equal probability, where ϵ is a user-defined constant error rate. If each aggregated marker contains a single SNP then this model reduces to one very similar to equations (3.10)–(3.12). The method uses a linear interpolation scheme for state probabilities, and takes advantage of redundancies in calculations when reference haplotypes are identical between any two aggregated markers. In addition, a binary reference panel data format, called bref, was proposed that speeds up the initial data input stage of the process. Building on these ideas, Beagle v5 (Browning *et al.*, 2018) has adopted the IMPUTE v2 idea of conditioning only on a subset of the haplotypes in the reference panel, and uses a novel, fast method of constructing a custom conditioning set of haplotypes.

3.2.3.4 Positional Burrows–Wheeler Transform

The Sanger Imputation Server (<https://imputation.sanger.ac.uk/>) implements an unpublished method that carries out imputation from a haplotype reference panel stored in a positional Burrows–Wheeler transform data structure (Durbin, 2014). Assuming that the haplotype reference panel H consists of M haplotypes at L sites, then the data structure is based on recursively sorting the haplotypes in the reference panel H according to their prefix at site l , and storing the ordered alleles at site l as a vector (Y_{1l}, \dots, Y_{Ml}) . Since site l will often be in strong LD with sites preceding it, the prefix ordering will tend to induce long stretches of identical alleles in (Y_{1l}, \dots, Y_{Ml}) which are highly compressible using run length encoding. In addition, if we let (A_{1l}, \dots, A_{Ml}) be a vector that stores the haplotype indices of the haplotypes in the prefix sorted order at site l , then the relationship between H , Y , and A is $Y_{il} = H_{A_{il}l}$. The vectors (A_{1l}, \dots, A_{Ml}) will not readily compress but can be stored at just a subset of sites. A simple algorithm allows calculation of the index vector at a site $l + 1$ from (Y_{1l}, \dots, Y_{Ml}) and (A_{1l}, \dots, A_{Ml}) . This data structure also allows fast searching of locally matching haplotypes. The imputation algorithm (unpublished) finds locally matching haplotypes to the study individual's haplotypes and uses these in a weighted sum to impute alleles.

This approach has the property that once the haplotype data is placed into the pBWT data structure the matching operations are independent of the number of haplotypes. So while the most basic implementation of the haploid imputation model in equations (3.10)–(3.12) has complexity $O(ML)$, the pBWT imputation step has complexity $O(L)$.

The pBWT data structure has also been used at the core of SHAPEIT v4, which is one of the most accurate phasing methods (Delaneau *et al.*, 2018) and the most recent version (v5) of the IMPUTE software (unpublished).

3.2.3.5 SNP Tagging Approaches

Large genetic variation projects such as the HapMap Project were the first to uncover the true extent of LD between SNPs in the human genome. Based on these observations a set of simple and quick approaches based on the idea of SNP tagging were proposed for imputation. For each SNP to be imputed, a small set of E flanking genotyped SNPs from the reference panel were chosen to produce good predictive performance of the SNP. The genotyped data at these E flanking SNPs was phased together with the reference panel at the $E + 1$ SNPs. The missing genotypes at the SNP of interest are imputed as part of the phasing process. While quick and simple to implement, these approaches proved not to be as accurate as HMM-based methods that are able to use more of the flanking data on either side of each SNP to be imputed

(Howie *et al.*, 2009). SNP tagging imputation is implemented in software packages PLINK (Purcell *et al.*, 2007) and TUNA (Nicolae, 2006).

3.2.3.6 Imputation Servers

The majority of haplotype reference panels, such as those produced from the HapMap Project and 1000 Genomes Project, have been made publicly available. This has allowed researchers to run imputation using their own local computational resources. The Haplotype Reference Consortium reference panel could not be made public due to not all contributing data sets having consent for public release. Hence web-based services for imputation and phasing using this reference panel have been developed. These free services allow users to securely upload their own data sets, which are imputed using centralized computational resources and then returned to users. Servers are available at <https://imputation.sanger.ac.uk/> and <https://imputationserver.sph.umich.edu>.

3.2.4 Testing Imputed Genotypes for Association

As already described, the main use of imputed genotypes is in GWASs. Typically, genotypes are imputed ignoring any phenotype information on the individuals. The genotypes can then be tested for association using any phenotype that has been measured. It has been suggested that imputation could be improved using phenotype information in a joint model (Lin *et al.*, 2008), but would require reimputation every time a new phenotype is tested.

After haploid imputation, it is usual to ignore the phase information and consider the distribution of the imputed genotype $G_{il} = h_{i1l} + h_{i2l} \in \{0, 1, 2\}$, which is given as

$$p_{ilg} = P(G_{il} = g | h_i, H, \rho, \theta) = \sum_{h_{i1l} + h_{i2l} = g} P(h_{i1l} | h_{i1}, H, \rho, \theta) + P(h_{i2l} | h_{i2}, H, \rho, \theta). \quad (3.15)$$

The simplest way to use these imputed probabilistic genotypes in association testing is to take the most likely genotype, or just those genotypes with $p_{ilg} > a$ for some threshold probability a , but this can lead to false positives and loss of power.

If $\Phi = (\Phi_1, \dots, \Phi_N)$ denotes the vector of phenotypes on N imputed study samples then a model-based approach to testing for association can proceed via averaging over the genotype uncertainty at each SNP. The likelihood for such a model can be written as

$$P(\Phi | G_{il}, H, \beta) = \prod_{i=1}^N \sum_{g=0}^2 P(\Phi_i | G_{il} = g, \beta) p_{ilg}. \quad (3.16)$$

Here the term $P(\Phi_i | G_{il} = g, \beta)$ is the part of the model that links genotype to phenotype, for example using a linear or logistic model, with β denoting parameters of that model. The software package SNPTEST implements score tests, Wald tests and maximum likelihood ratio tests and a set of Bayesian tests based on this likelihood (Marchini and Howie, 2010; for more details about linear and logistic model based association tests and GWAS in general, see Chapter 21).

Another common approach has been to infer an expected allele count, or *dosage*, for each imputed genotype at the l th site in the i th individual, defined as $e_{il} = \sum_{g=0}^2 p_{ilg}g$. This continuous genotype, $e_{il} \in [0, 2]$, is then used in the association model, usually with no need to change the test. This method does not quite use all the information. For example, a dosage of 1 can be produced by completely certain, $p_{il} = (0, 1, 0)$, or uncertain distributions, $p_{il} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$. In practice this seems to not have a large effect (Liu *et al.*, 2013), and this method has the advantage

of simplicity. Software packages that can use imputed dosages include PLINK (Purcell *et al.*, 2007), SNPTEST v2 (Marchini and Howie, 2010), BGENIE (Bycroft *et al.*, 2018), and BOLT-LMM (Loh *et al.*, 2015).

One pitfall in using imputed data in GWASs can be encountered if cases and controls for a binary phenotype have been genotyped using two different SNP microarrays, and then imputed. A poorly imputed SNP in either cases or controls will tend to have its allele frequency pulled towards the frequency that is in the reference panel, and this can lead to false positive associations.

3.2.5 Summary Statistic Imputation

In many disease genetics settings a key quantity of interest is the allele frequency of each SNP. Wen and Stephens (2010) considered the scenario where just the allele frequencies at genotyped SNPs are available and proposed a model to impute the allele frequencies at ungenotyped SNPs. A special case of this problem considers each individual separately with genotypes used in place of the allele frequency at each SNP. The authors proposed to use a Gaussian distribution to model the distribution of a haplotype h_{ij} conditional upon a reference set of haplotypes H ,

$$h_{ij} \sim N_L(\hat{\mu}, \hat{\Sigma}). \quad (3.17)$$

The estimates of $\hat{\mu}$ and $\hat{\Sigma}$ are derived by calculating the expectation and variance of the Li and Stephens model distribution $P(h_{ij}|H, \theta, \rho)$ in equation (3.10), and reduce to shrunken estimates of the mean and covariance of the SNPs in the reference panel. Unobserved genotypes can be imputed using their conditional Gaussian distribution given the observed genotypes. These can be computed analytically and take the form of a linear weighted contribution from each SNP. This approach is implemented in software called BLIMP (Best Linear Imputation).

A variant of this approach, called MVNCALL (Menelaou and Marchini, 2013), has been developed to infer genotypes at a single locus from a set of genotype likelihoods (see Section 3.1.4) conditional upon a flanking set of phased SNP sites. This approach was used in the 1000 Genomes Project to call genotypes and phase indels, multi-allelic SNPs, complex and structural variants, and short tandem repeats onto the main SNP haplotypes.

This approach has been further developed to directly impute the GWAS summary statistics, or Z-scores, at ungenotyped SNPs from the Z-scores at genotyped SNPs (Lee *et al.*, 2013; Pasaniuc *et al.*, 2014). While these approaches are likely to work well at imputing common variants, it is expected that accuracy will fall off at very low frequencies.

3.2.6 Factors Affecting Accuracy

It has become standard to measure imputation accuracy using experiments with test data sets derived from high-coverage sequencing that provide very dense and accurate sets of genotypes. A typical experiment will involve choosing the genotypes at sites on one of several possible commercial SNP microarrays, hiding the genotypes at all the remaining sites, and imputing them from a reference panel. Imputed genotypes are then stratified into B bins according to the allele frequency of their corresponding SNP in the reference panel, so for example all SNPs with an allele frequency of $\sim 1\%$ are placed in the same bin. In the b th bin, the squared correlation (r^2) between the imputed dosages and the true genotypes is calculated and the r^2 values for all the B bins are then plotted against the allele frequency of the SNPs in each of the bins. Usually allele frequency is plotted on a log scale to accentuate the lower allele frequency range, which is increasingly of interest in the hunt for disease susceptibility genes.

Squared correlation is considered to be a good measure because it has a direct link to the power of an association test at a marker genotype in LD with a causal locus (Chapman *et al.*, 2003). The power of the test is a function of the expected chi-squared test statistic, which has the form

$$E(\chi^2) \simeq Nr^2\beta^2p(1-p), \quad (3.18)$$

where p and β are the allele frequency and additive effect size of the causal locus and r^2 is the squared correlation of the marker genotype and the true causal locus. In the rest of this subsection, some of the factors that affect the performance of imputation (measured using r^2) will be discussed.

3.2.6.1 Reference Panel Size and SNP Allele Frequency

Imputation will work well when there are long stretches of shared haplotypes between the study individuals and the individuals in the reference panel. Imputation would be almost perfect if the reference panel contained many close relatives of the study individuals. In many populations around the world it will take some time until whole genome sequencing (WGS) is so common that huge reference panels are available that make this possible. However, it has clearly been observed that increasing reference panel size increases imputation accuracy.

Figure 3.6 is reproduced from the HRC paper (McCarthy *et al.*, 2016) and compares the performance of reference panels from the 1000 Genomes Project, UK10K Project, and the HRC when imputing genotypes in individuals of European ancestry. Firstly, this shows that in general imputation becomes more challenging as the allele frequency decreases. It also shows that increasing reference panel size has most effect at low allele frequencies and that there are diminishing returns of increasing reference panel size.

Reference panels have gradually evolved through increases in the number and ancestral diversity of the samples (see Table 3.1). The first panels were produced by the HapMap Project (International HapMap Consortium, 2005; International HapMap Consortium *et al.*, 2007) which was focused on detecting and genotyping SNPs with minor allele frequencies of 5% or greater. The 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2010, 2012, 2015) produced a number of different panels using WGS to detect and genotype genetic variants (including SNPs, indels, and structural variants) with minor allele frequencies down to frequencies of 1% and greater in a more global set of populations than HapMap. As the cost of sequencing has decreased WGS has been applied to larger sample sizes in single populations such as the UK10K project (Huang *et al.*, 2015; UK10K Consortium *et al.*, 2015). The HRC brought together 20 different data sets of WGS data to build a reference panel of 65,000 SNP haplotypes (McCarthy *et al.*, 2016).

Reference panels will continue to grow in size. Current efforts under way include the 100,000 Genomes Project in the UK which will use high-coverage WGS on ~75,000 germline samples and ~25,000 cancer cell genomes (<https://www.genomicsengland.co.uk/the-100000-genomes-project/>). The TopMed study (<https://www.ncbiwg.org/>) is collecting WGS on over 100,000 samples from a diverse set of disease samples. It also seems likely that the ~500,000 samples from the UK Biobank study will be sequenced in the not too distant future (<http://www.ukbiobank.ac.uk/>).

3.2.6.2 Ancestry

Imputation performance will depend also upon the ancestry of the individual being imputed in combination with the ancestral composition of the samples in the reference panel. Imputation experiments using the cosmopolitan 1000 Genomes Project reference panel (1000 Genomes Project Consortium *et al.*, 2015) suggest that at common variants imputation performance is

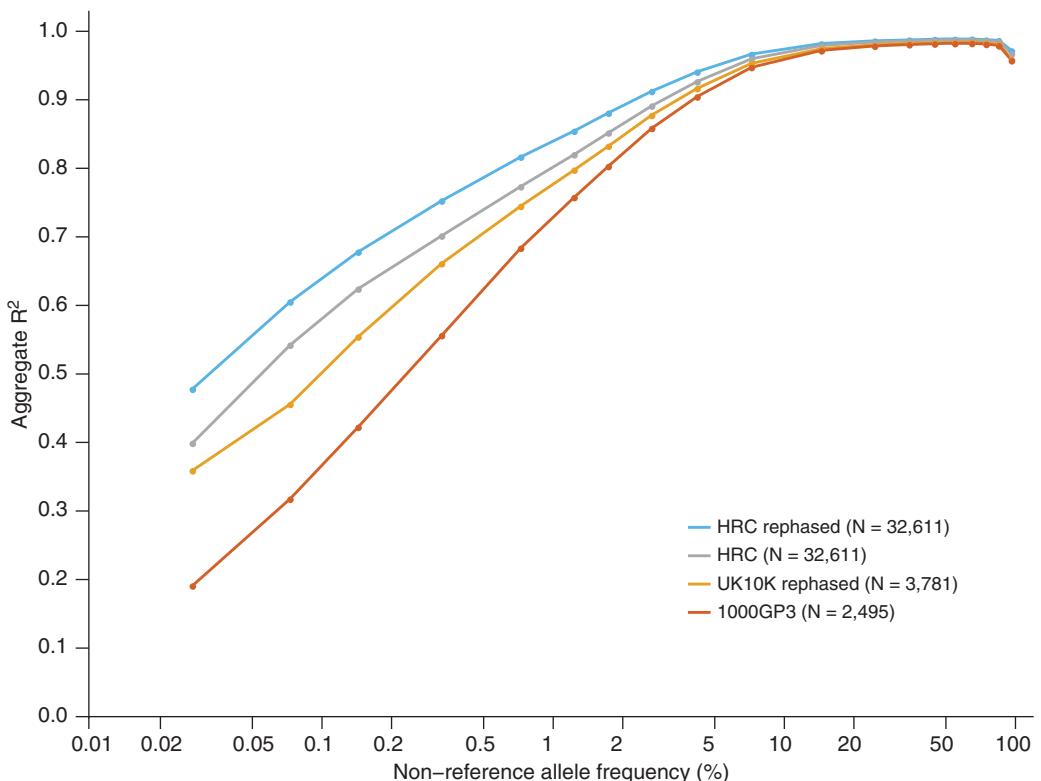


Figure 3.6 Performance of imputation using different reference panels. The x-axis shows the non-reference allele frequency of the SNP being imputed on a log scale. The y-axis shows imputation accuracy measured by aggregate R^2 value (on the plot denoted R^2) when imputing SNP genotypes into ten CEU samples, that is, samples from individuals of European origin from Utah. These results are based on using genotypes from sites on the Illumina Omni1M SNP array as pseudo-GWAS data. The HRC rephased panel was produced by rephasing the original HRC genotypes using SHAPEIT v3 to produce better-quality reference haplotypes. Reproduced from McCarthy *et al.* (2016).

higher in populations with high levels of LD such as European samples (cyan line in Figure 3.7), but at low allele frequencies it appears that imputation is most accurate in African ancestry populations (yellow and brown lines in Figure 3.7). Figure 3.7 is reproduced from the 1000 Genomes Phase 3 paper (1000 Genomes Project Consortium *et al.*, 2015). The hypothesis here is that greater genetic diversity results in a larger number of haplotypes and improves the chances that a rare variant is tagged by a characteristic haplotype.

3.2.6.3 Genotyping Microarray

The choice of genotyping microarray is also a factor in imputation quality. Whole genome genotyping arrays are mostly produced by two commercial companies (Illumina and Affymetrix), which have produced a large number of different arrays over the years. Most arrays contain a set of SNPs chosen to capture as much genetic variation as possible, and in some cases to maximize imputation performance (Bycroft *et al.*, 2018). SNPs and markers of direct interest, for example SNPs in the HLA region on chromosome 6, in exome regions or known to be important to specific diseases are also included as markers on some arrays. The imputation performance of 10 different arrays when imputing European samples is shown in Figure 3.8 using the HRC reference panel and highlights that there can be clear differences between arrays.

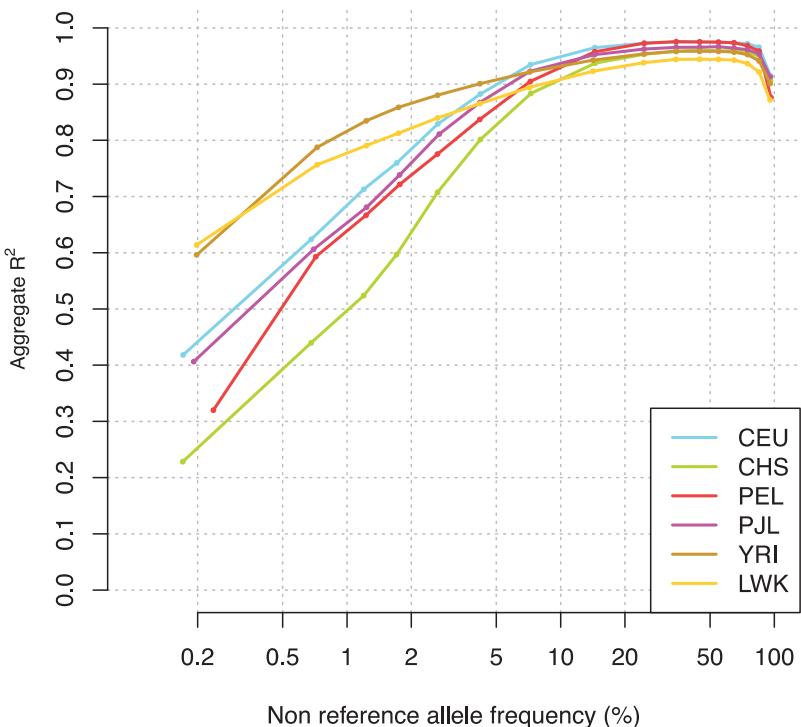


Figure 3.7 Imputation accuracy in individuals from six different populations using the 1000 Genomes Phase 3 reference panel. Populations are CEU = European, CHS = Han Chinese South, PEL = Peruvian in Lima, Peru, YRI = Yoruba in Ibadan, Nigeria, LWK = Luhya in Webuye, Kenya, PJL = Punjabi in Lahore, Pakistan. Imputation r^2 (y-axis, on the plot denoted R^2) is plotted against log allele frequency (x-axis). Reproduced from 1000 Genomes Project Consortium *et al.* (2015).

3.2.6.4 Imputation Method

Various studies have been carried out to compare the accuracy and performance of different imputation methods (Howie *et al.*, 2009, 2012; Das *et al.*, 2016; Browning and Browning, 2016). A comparison of many of the most widely used methods that are based on haploid imputation (IMPUTE v2, IMPUTE v4, minimac3, Beagle, pBWT) showed that minimac3, IMPUTE v4, and IMPUTE v2 all produce very similar results, and better accuracy than Beagle and pBWT (Herzig *et al.*, 2018).

3.2.7 Quality Control for Imputed Data

Genotype probabilities produced by the methods that use HMMs tend to be well calibrated in the sense that genotypes imputed with probability q are correct on average $100q\%$ of the time (Marchini *et al.*, 2007). While the quality of imputation is very good in most cases, it is generally a good idea to apply a filter to remove (or at least flag) SNPs that might not be imputed well.

One way to measure the quality of imputed data at an SNP site l is to consider using the imputed data to estimate the allele frequency θ_l at the SNP site. The empirical and expected variances of that estimate (assuming Hardy–Weinberg equilibrium) are then compared. If p_{ilg} denotes the imputation probability that the l th site of the i th individual is g (from equation (3.15)), $e_{il} = \sum_{g=0}^2 p_{ilg}g$ denotes the dosage for the i th individual at the l th site, $\hat{\theta}_l = \sum_{i=1}^N e_{il}/2N$ is the estimate of θ_l , and $\lambda = \sum_{i=1}^N e_{il}^2/N$, then a quality measure can be defined

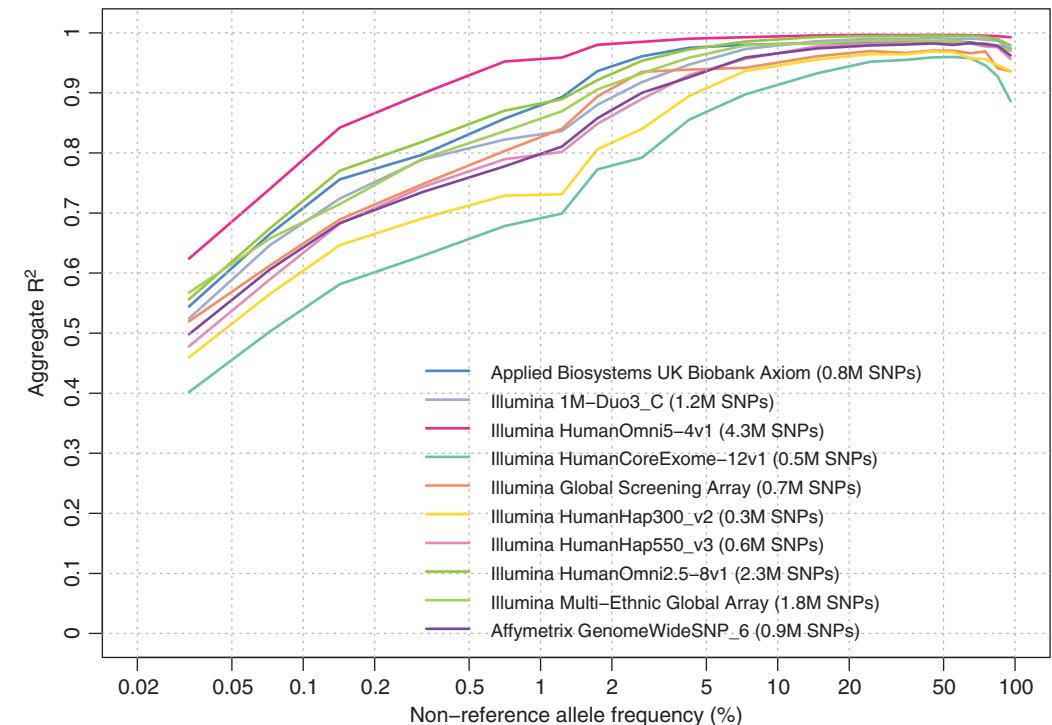


Figure 3.8 Comparison of imputation performance of several commercially available genotyping arrays. The x-axis of each plot shows non-reference allele frequency on a log scale, which accentuates low allele frequencies. The y-axis shows imputation performance in terms of r^2 (on the plot denoted R^2).

(Li *et al.*, 2010) as

$$\hat{r}^2 = \begin{cases} \frac{\lambda - (2\hat{\theta}_l)^2}{2\hat{\theta}_l(1 - \hat{\theta}_l)} & \text{if } 0 < \hat{\theta}_l < 1, \\ 1 & \text{if } \hat{\theta}_l = 0, \hat{\theta}_l = 1. \end{cases} \quad (3.19)$$

A very similar quality measure can be derived using a related concept of measuring the statistical information about the allele frequency estimate with a missing-data likelihood framework (Marchini and Howie, 2010):

$$I_A = \begin{cases} 1 - \frac{\omega}{2\hat{\theta}_l(1 - \hat{\theta}_l)} & \text{if } 0 < \hat{\theta}_l < 1, \\ 1 & \text{if } \hat{\theta}_l = 0, \hat{\theta}_l = 1, \end{cases} \quad (3.20)$$

where $f_{il} = \sum_{g=0}^2 p_{ilg} g^2$ and $\omega = (\sum_{i=1}^N (f_{il} - e_{il}^2)) / N$.

The two quality measures, routinely referred to as *information measures*, have been shown to be highly correlated (Marchini and Howie, 2010). The interpretation is that an information measure of α on a sample of N individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample of size αN . Most GWAS studies have tended to filter out SNPs with an information measure threshold between 0.3 and 0.5. However, in larger association studies, such as those possible when working with the UK Biobank data set (Bycroft *et al.*, 2018), it may be possible to use a lower

threshold. For example, SNPs with an information measure of 0.1 in a sample size of 500,000 samples are equivalent to perfect data in a sample size of 50,000 samples.

3.3 Future Directions

Large cohorts of high-coverage whole genome sequencing data consisting of millions of individuals will soon be collected. These will be invaluable as imputation reference panels, and it seems likely the imputation servers will continue to be an important way in which data sets can be phased and imputed. These large reference panels pose challenges in terms of the sheer number of variants they will likely contain, so data compression and computational efficiency will continue to be important. The pBWT based method is a good start in this direction. It may make sense to compress very rare variants separately from more common variants.

References

- 1000 Genomes Project Consortium *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.
- 1000 Genomes Project Consortium *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.
- 1000 Genomes Project Consortium *et al.* (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.
- Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.
- Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S. and Pirinen M. (2016). FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501.
- Browning, B.L. and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223.
- Browning, B.L. and Browning, S.R. (2016). Genotype imputation with millions of reference samples. *American Journal of Human Genetics*, **98**, 116–126.
- Browning, S.R. (2006). Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics* **78**, 903–913.
- Browning, S.R. and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**, 1084–1097.
- Browning, S.R. and Browning, B.L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics* **12**, 703–714.
- Browning, B.L., Zhou, Y., Browning, S.R. (2018). A one-penny imputed genome from next generation reference panels. *American Journal of Human Genetics* **103**, 338–348.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209.
- Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Human Heredity* **56**, 18–31.

- CONVERGE Consortium (2015). Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591.
- Das, S., Forer, L., Schönherr, S., et al. (2016). Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284–1287.
- Delaneau, O., Marchini, J. and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181.
- Delaneau, O., Howie, B., Cox, A.J., Zagury, J.-F. and Marchini, J. (2013a). Haplotype estimation using sequencing reads. *American Journal of Human Genetics* **93**, 687–696.
- Delaneau, O., Zagury, J.-F. and Marchini, J. (2013b). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5–6.
- Delaneau, O., Marchini, J. and 1000 Genomes Project Consortium (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5**, 3934.
- Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J. & Dermitzakis, E. (2018). Integrative haplotype estimation with sub-linear complexity. Preprint, bioRxiv 493403.
- Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272.
- Elliott, L.T., Sharp, K., Alfaro-Almagro, F., Shi, S., Douaud, G., Miller, K., Marchini, J. and Smith, S. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216.
- Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.
- Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* **25**, 12–13.
- Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014). A genetic atlas of human admixture history. *Science* **343**, 747–751.
- Herzig, A.F., Nutille, T., Babron, M.C., Ciullo, M., Bellenguez, C. and Leutenegger, A.L. (2018). Strategies for phasing and imputation in a population isolate. *Genetic Epidemiology* **42**(2), 201–213.
- Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N. and van der Werf, J.H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics, Selection, Evolution* **43**, 12.
- Howie, B., Donnelly, P. and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* **5**, e1000529.
- Howie, B., Marchini, J. and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470.
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955–959.
- Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Mallerba, G., Trabetti, E., Zheng, H.-F., UK10K Consortium, Gambaro, G., Richards, J.B., Durbin, R., Timpson, N.J., Marchini, J. and Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications* **6**, 8111.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.
- International HapMap Consortium et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.
- Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**, 235–248.
- Kingman, J.F.C. (2000). Origins of the coalescent: 1974–1982. *Genetics* **156**, 1461–1463 (2000).

- Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdóttir, U., Gudbjartsson, D.F., Stefansson, H. and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics* **40**, 1068–1075.
- Lange, K., Papp, J.C., Sinsheimer, J.S., Sripracha, R., Zhou, H. and Sobel, E.M. (2013). Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics* **29**, 1568–1570.
- Lee, D., Bigdeli, T.B., Riley, B.P., Fanous, A.H. and Bacanu, S.-A. (2013). DIST: Direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29**, 2925–2927.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.
- Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816–834.
- Lin, D.Y., Hu, Y. and Huang, B.E. (2008). Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics* **82**, 444–452.
- Liu, K., Luedtke, A. and Tintle, N. (2013). Optimal methods for using posterior probabilities in association testing. *Human Heredity* **75**, 2–11.
- Liu, S., Huang, E., Chen, F., et al. (2018). Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359.e14.
- Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N. and Price, A.L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290.
- Loh, P.-R., Danecek, P., Palamara, P.F., et al. (2016a). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448.
- Loh, P.-R., Palamara, P.F. and Price, A.L. (2016b). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* **48**, 811–816.
- Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511.
- Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., Donnelly, P. and International HapMap Consortium (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics* **78**, 437–450.
- Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.
- McCarthy, S., Das, S., Kretzschmar, W., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283.
- Menelaou, A. and Marchini, J. (2013). Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91.
- Morris, A.P. (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genetic Epidemiology* **29**, 91–107.
- Myers, S., Bottolo, L., Freeman, C., McVean, G. and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324.
- Nicolae, D.L. (2006). Testing Untyped Alleles (TUNA) – applications to genome-wide association studies. *Genetic Epidemiology* **30**, 718–727.
- O'Connell, J., Gurdasani, D., Delaneau, O., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* **10**, e1004234.

- O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.-F., Delaneau, O. and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nature Genetics* **48**, 817–820.
- Palin, K., Campbell, H., Wright, A.F., Wilson, J.F. and Durbin, R. (2011). Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic Epidemiology* **35**, 853–860.
- Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., Sullivan, P.F., Bergen, S., Moran, J.L., Hultman, C.M., Lichtenstein, P., Magnusson, P., Purcell, S.M., Haas, D.W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P.I., Reich, D. and Price, A.L. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics* **44**, 631–635.
- Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D.P., Patterson, N. and Price, A.L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914.
- Patterson, M. et al. (2015). WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology* **22**, 498–509.
- Paul, J.S. and Song, Y.S. (2012). Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics* **28**, 2008–2015.
- Purcell, S. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.
- Rabbee, N. and Speed, T.P. (2006). A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12.
- Sabeti, P.C., Varilly, P., Fry, B., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918.
- Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629–644.
- Sharp, K., Kretzschmar, W., Delaneau, O. and Marchini, J. (2016). Phasing for medical sequencing using rare variants and large haplotype reference panels. *Bioinformatics* **32**, 1974–1980.
- Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.
- Stephens, M. and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics* **73**, 1162–1169.
- Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**, 449–462.
- Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.
- Su, Z., Marchini, J. and Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305.
- Teo, Y.Y., Inouye, M., Small, K.S., Gwilliam, R., Deloukas, P., Kwiatkowski, D.P. and Clark, T.G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.
- UK10K Consortium et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* **526**(7571), 82–90.
- Wang, Y., Lu, J., Yu, J., Gibbs, R.A. and Yu, F. (2013). An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Research* **23**, 833–842.

- Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
- Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Annals of Applied Statistics* **4**, 1158–1182.
- Williams, A.L., Patterson, N., Glessner, J., Hakonarson, H. and Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics* **91**, 238–251.
- Zeggini, E., Scott, L.J., Saxena, R., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* **40**, 638–645.

4

Mathematical Models in Population Genetics

Nick Barton¹ and Alison Etheridge²

¹Institute of Science and Technology Austria, Klosterneuburg, Austria

²University of Oxford, UK

Abstract

We review the history of population genetics, starting with its origins a century ago from the synthesis between Mendel and Darwin's ideas, through to the recent development of sophisticated schemes of inference from sequence data, based on the coalescent. We explain the close relation between the coalescent and a diffusion process, which we illustrate by their application to understand spatial structure. We summarise the powerful methods available for analysis of multiple loci, when linkage equilibrium can be assumed, and then discuss approaches to the more challenging case, where associations between alleles require that we follow genotype, rather than allele, frequencies. Though we can hardly cover the whole of population genetics, we give an overview of the current state of the subject, and future challenges to it.

4.1 Introduction

The importance of mathematical models in population genetics can be traced back to the modern evolutionary synthesis, in which Darwin's theory of natural selection was finally reconciled with Mendelian inheritance. In Darwin's argument, evolution is a continuous process; complex well-adapted organisms evolve through selection acting on large numbers of slight variants in traits. Mendel, by contrast, focuses on discontinuous changes in traits determined by a single gene. Fisher's resolution of the apparent mismatch (Fisher, 1918) showed that biometric measurements could be explained by multiple Mendelian factors, each of small effect, together with random, non-genetic, influences. In the process, he developed some of the most fundamental tools in modern statistics.

Around the same time, Haldane used mathematical models to analyse the effect on a population of differences in survival or reproduction due to one or two Mendelian factors (Haldane, 1932), and Wright quantified the way in which the random process of reproduction in a finite population leads to changes in allele frequencies through his theory of random genetic drift (Wright, 1931). He suggested that selection pushes a population towards local maxima in fitness, but genetic drift can drive it away from those maxima, allowing it to explore the whole 'adaptive landscape'.

The twentieth century saw extraordinary developments in stochastic modelling, and as new mathematical tools became available, they were rapidly adopted as tools for understanding the interactions between different forces of evolution. Stochastic differential equations were

embraced as models of allele frequencies at a single locus (Feller, 1951), and spatially distributed populations were captured through coupled systems of equations in Kimura's stepping stone model (Kimura, 1952); Malécot (1948) and Wright (1943) modelled populations evolving in spatial continua as Poisson random fields; and Bulmer (1980) developed the infinitesimal model, implicit in Fisher (1918), to describe the within and between family distribution of an additive trait that depends on a large number of unlinked loci.

At the time of the modern evolutionary synthesis, genetic variability could not be observed directly, but by the mid -1960s, scientists began to study genetic variation at the level of DNA and RNA. It soon became apparent that there was far more genetic variation within species than had been expected, prompting Kimura (1968) and King and Jukes (1969) to question the importance of natural selection as the driving force of evolution. Instead, they proposed that most variation was selectively neutral. This further drove the need for mathematical models, and corresponding statistical tests, to provide the tools with which to test the 'null model' of neutrality, against alternative models with selection.

As sequencing became cheaper, it became routine for geneticists to use patterns of genetic variation to infer the genealogical trees relating individuals in a sample from a population. In order to compare to experiments, mathematical models for the way that populations evolve forwards in time started to come hand in hand with consistent (backwards in time) models for genealogies. Kingman (1982) introduced his celebrated coalescent, describing genealogies under the neutral Wright–Fisher model, and it was rapidly extended to incorporate spatial structure and multiple loci (Hudson, 1990). Adding selection proved more challenging; as one traces the genealogy of a sample backwards in time, one is faced with a model that is not a Markov process, but, at least for some forms of selection, Neuhauser and Krone (1997) introduced the ancestral selection graph, an extension of the coalescent from which the genealogy of a sample can be extracted.

There are very significant mathematical and computational challenges associated with coalescent models that have stimulated a great deal of statistical and computational innovation. Meanwhile, forwards in time models of theoretical population genetics provide an invaluable tool for understanding how different forces of evolution will interact for different parameter regimes and time-scales.

In view of this plethora of mathematical tools, we can do no more than skim the surface. We begin with single-locus models of panmictic populations, before introducing spatial structure. We shall then turn to multi-locus models. While in principle this extension is straightforward, in practice the number of variables required to describe the population, and fitness of individuals within that population, becomes prohibitively large. We describe some of the approximations that are used to help understand the complexities of real populations.

4.2 Single-Locus Models

We begin in Section 4.2.1 with models of panmictic (random mating) populations of fixed (finite) size. Even for a single locus, we can write down almost arbitrarily complicated models for the interaction of selection, mutation, and random genetic drift. These models are typically analytically intractable, but nonetheless, in the process of seeking approximations, one often gains insight into the evolutionary process by identifying the parameter combinations that lead to nontrivial limits.

Of particular importance is the limit as the population size tends to infinity, which we investigate in Section 4.2.2. Our starting point will be the Wright–Fisher model. Even for the bi-allelic neutral Wright–Fisher model with mutation, Feller (1951) identified three distinct regimes: (i) if mutation is much weaker than random genetic drift, then the limiting diffusion is one

of pure random drift; (ii) if mutation and random drift are comparable, then the diffusion also incorporates a deterministic term reflecting mutation; (iii) if mutation is much stronger than genetic drift, then the limit is deterministic, but one can identify small Gaussian fluctuations about that limit.

Spatial structure, to which we turn our attention in Section 4.2.3, offers new challenges. We introduce a generalisation of the Wright–Fisher model due to Wright and Malécot that aims to follow allele frequencies for a population evolving in a spatial continuum. We illustrate Felsenstein’s ‘pain in the torus’ through simulation, before turning to Kimura’s stepping stone model from which we derive the Wright–Malécot formula for the way in which correlations in allele frequencies decay with spatial separation. Before turning to multiple loci, we close this section with a very brief description of the resolution of the pain in the torus provided by the spatial Lambda–Fleming–Viot model for a population in a spatial continuum.

4.2.1 Random Drift and the Kingman Coalescent

4.2.1.1 The Neutral Wright–Fisher Model

We begin with the simplest imaginable model of inheritance: take a panmictic, haploid population of size N , evolving in discrete generations. We consider a single locus at which there are two alleles that we shall denote by P and Q . We write $p(t)$ for the proportion of P -alleles. During reproduction, each individual produces a very large (effectively infinite) number of gametes (of the same type as the parent) which go into a pool from which N individuals are sampled to form the next generation. Under this model, given that $p(t) = p$, the number of type P alleles in generation $t + 1$ has a $\text{Binom}(N, p)$ distribution. In particular, writing $\Delta p(t) = p(t + 1) - p(t)$ for the change in allele frequencies over a single generation, if $p(t) = p$, we have

$$\mathbb{E}[\Delta p(t)] = 0, \quad \text{var}(\Delta p(t)) = \frac{1}{N}p(1-p). \quad (4.1)$$

That the expected change in allele frequencies is zero reflects neutrality; that the variance is order $1/N$ tells us that we can expect to see substantial changes in allele frequencies due to random fluctuations over time-scales of the order of N generations.

4.2.1.2 The Kingman Coalescent

By recording only $p(t)$ in the Wright–Fisher model, we lose almost all information about the way in which individuals in the population are related to one another. However, we can also think of the Wright–Fisher model as defining a model of inheritance without reference to type: each individual in generation $t + 1$ samples their parent uniformly at random (with replacement) from the individuals in generation t . (We recover the model of allele frequencies by dictating that offspring inherit the type of their parent.) From this perspective, it is straightforward to describe the genealogical trees relating individuals in a random sample from the population. Consider first a sample of size 2. The only information in the genealogical tree is the number of generations since their most recent common ancestor. The chance that they ‘chose’ the same parent in the previous generation is just $1/N$. If they did not, then the chance that they have a common grandparent is again $1/N$; and so on. The number of generations that we must trace back to reach a common ancestor is a geometric random variable with mean N . Just as we saw for allele frequencies, we only see non-trivial shared ancestry over time periods of order N generations and so we shall choose ‘ N generations’ to be our basic unit of time. In these units, provided that N is large, the geometrically distributed time back to the most recent common ancestor of a sample of size 2 can be approximated by an exponentially distributed random variable with mean 1.

For larger samples, of size $k \geq 3$ say, note that the chance that three or more individuals choose a common parent is order $(1/N^2)$ and so if N is large, the chance that we see such an

event before the ancestral lineages have all merged through the pairwise mergers is negligible. Similarly, we do not expect to see any *simultaneous* mergers of pairs of lineages ancestral to the sample.

Putting this together, measuring time in units of N generations, for large N , we can approximate the genealogical trees relating individuals in a sample through the Kingman coalescent: starting from a sample of size k , we trace back an exponential time with mean $1/(k)$ (in rescaled time) at which time a pair of ancestral lineages (picked uniformly at random from all possible pairs) merges into a single lineage. (We have used the fact that the minimum of $\binom{k}{2}$ independent exponential random variables with parameter 1 is an exponential random variable with parameter $\binom{k}{2}$.) We then trace back an independent exponential time with mean $1/(k-1)$ until the next merger, which is again equally likely to involve any of the $\binom{k-1}{2}$ pairs of ancestral lineages available; and so on.

The Kingman coalescent describes the genealogy of a random sample from the population. The distribution of types in the sample is determined by assigning types to ancestral lineages at some time t in the past and tracing their descent through the genealogy to the present. Crucially, we do *not* get to specify the types in the sample. If we condition on knowing the types in the sample, this constrains the genealogy and it is no longer simply a Kingman coalescent. The implications of this are important; the coalescent model defines the genealogy of a sample given only the sample size and not the allelic types. It is not an inference model; once we observe the alleles, the model no longer applies.

4.2.1.3 Adding Mutation

So far, we have not included mutation. In practice, if we wish to reconstruct the genealogical trees from data, we use differences between the DNA sequences of individuals in the sample to infer relatedness. If we are to see such differences, then we need to see mutations of order 1 on a ‘branch’ of the genealogical tree: any fewer and we will not see differences between individuals; any more and there will be so many differences that we will not be able to distinguish from a sample of unrelated individuals. Patterns observed in data reflect the evolution of the population over time-scales dictated by the neutral mutation rate.

Under the neutral model, mutations are superposed onto the Kingman coalescent as a Poisson process. Thus, if a branch has length L , say, then the number of mutations that occur along the ancestral lineage is Poisson with parameter $\theta L/2$. The factor of 2 is chosen so that if two individuals had a common ancestor at time T in the past, they will differ by a Poisson number of mutations with parameter θT (since the tree relating them has two branches each of length T). If u is the probability of a mutation per individual per generation, then $\theta = 2Nu$ (because of the scaling of time).

Many models of mutation have been proposed. For example, Watterson (1975) introduced the so-called *infinitely many sites model*, in which each individual is represented by a string of infinitely many completely linked sites. The mutation rate per site is assumed to be very small so that the total number of mutations per individual per generation is finite and we can ignore back mutations. In the infinitely many sites model we can infer the entire mutational history of an individual. By contrast, in the *infinitely many alleles model* introduced by Kimura and Crow (1964), when a mutation occurs, it leads to an entirely new haplotype, never before seen in the population. In the infinitely many sites model, observed haplotypes can have different degrees of similarity, but this is no longer true for the infinitely many alleles model.

4.2.1.4 The Cannings Model

In the neutral Wright–Fisher model, if the individuals in generation t are labelled $\{1, 2, \dots, N\}$, then writing v_i for the number of offspring of the i th individual, the vector (v_1, \dots, v_N) has a multinomial distribution with equal probabilities. The mean number of offspring of each

individual is 1, as it must be in a model with fixed population size in which all individuals are equally fit. However, under the Wright–Fisher model the variance in family size of an individual is forced to be $1 - 1/N$, which is unduly restrictive. Under the Cannings model (Cannings, 1974), this condition is relaxed by allowing (v_1, \dots, v_N) to be any *exchangeable random vector*. This just means that the probability of the vector (v_1, \dots, v_N) is the same as the probability of $(v_{\pi(1)}, \dots, v_{\pi(N)})$ for any permutation $(\pi(1), \dots, \pi(N))$ of the labels of the individuals in the parental generation. It is this exchangeability that preserves the neutrality.

As we saw in the Wright–Fisher model, a natural time-scale for the process (and the corresponding Kingman coalescent) is determined by the probability that two individuals sampled from the population have a common parent in the previous generation. For the Cannings model, this is

$$c_N = \frac{\mathbb{E}[v_1(v_1 - 1)]}{N - 1}.$$

(For the neutral Wright–Fisher model, $c_N = 1/N$.) Measuring time in units of $1/c_N$ generations, the genealogy of a sample will once again converge to a Kingman coalescent provided that

$$\frac{\mathbb{E}[v_1(v_1 - 1)(v_1 - 2)]}{(N - 1)(N - 2)c_N} \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad (4.2)$$

(Möhle, 2000). The quantity on the left is $1/c_N$ times the probability that three individuals sampled from the current population have a common parent. Condition (4.2) guarantees that for large enough populations, with high probability, all lineages ancestral to a finite random sample will have coalesced through pairwise mergers before we see an event in which at least three lineages merge.

4.2.1.5 Limitations

There are (at least) two obvious ways in which the arguments that led us from neutral models of Wright–Fisher/Cannings type to the Kingman coalescent can fail. The first is that the quantity in (4.2) may not tend to zero. In this case we may see simultaneous and/or multiple mergers (by which we mean merging of at least three lineages in a single event) on time-scales of $1/c_N$ generations (Möhle and Sagitov, 2001). The more general coalescents that arise in this way are known as Λ and Ξ coalescents. They may be appropriate for modelling highly fecund populations with sweepstakes reproduction (Eldon and Wakeley, 2006). Their study is beyond our scope here; instead we refer to Berestycki (2009) for a review. The second way in which our arguments can fail is that they assume that the sample size is small compared to the (effective) population size. We have estimated the chance of a merger of three lineages, but for a sample of size k , there are of the order of k^3 different triples that might merge, so we really need that k^3 times the quantity on the left of (4.2) is negligible compared to $k^2 c_N$, the rate of pairwise mergers. Worse, if our population is evolving in discrete time, then the chance that we see two pairwise mergers in the same generation is of the order of $k^4 c_N^2$, which for the neutral Wright–Fisher model is no longer negligible once sample size is of the order of the square root of population size. As sample sizes grow, it becomes less and less accurate to ignore simultaneous and multiple mergers.

If we wish to compare the predictions of the Kingman coalescent to genealogical trees inferred from data, then we must convert back into units of single generations, so that coalescence in a sample of k ancestral lineages will take place at rate $\frac{1}{N} \binom{k}{2}$. Of course no real population evolves according to precisely the Wright–Fisher or Cannings models. However, the Kingman coalescent has proved to be a useful model for an astonishing array of organisms provided that individuals are sampled at sufficiently large spatial separation, and that we replace the census population size N by an effective population size N_e . This is particularly remarkable when we consider that for natural populations the effective population size can be many orders

of magnitude different from the census population; for example, for the human population it is of the order of 10,000 rather than 7 billion (Charlesworth, 2009).

4.2.1.6 Overlapping Generations: the Moran Model

In the models above, the population evolves in discrete non-overlapping generations. It is often mathematically simpler to consider models of overlapping generations. The most popular model of this type is the Moran model (Moran, 1958). In the Moran model for a neutral haploid population of size N , reproduction events are separated by independent exponential random variables. At such an event, a pair of individuals is sampled uniformly at random from the population, one dies and the other produces two offspring. The offspring inherit the type of the parent.

There is no accepted convention for the rate at which events take place in the Moran model. If we choose the mean time between events to be $1/(N)$, even for finite N , the genealogy of a random sample from the population is *exactly* given by the Kingman coalescent. This tells us that although, in contrast to the Wright–Fisher model, under the Moran model any individual can only have either zero or two offspring, for large populations, the predictions of the two models are the same (provided sample size is not too large).

4.2.2 Diffusion Approximations

In approximating the genealogies under the neutral Wright–Fisher model by the Kingman coalescent, we are taking a limit. Forwards in time, this corresponds to approximating the allele frequencies by a one-dimensional *diffusion process*. A diffusion $\{X(t)\}_{t \geq 0}$ evolves continuously in time and is characterised in terms of two quantities, usually called the drift and diffusion coefficients, that describe the mean and variance of the change in X_t over an infinitesimally small time period. Thus, if we write $\Delta_h X(t) = X(t+h) - X(t)$ for the change in X over the time interval $(t, t+h)$, the drift coefficient is

$$a(t, x) = \lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[\Delta_h X(t) | X(t) = x], \quad (4.3)$$

and the diffusion coefficient is

$$b(t, x) = \lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[(\Delta_h X(t))^2 | X(t) = x].$$

Roughly speaking, for very small h ,

$$X(t+h) \approx X(t) + h a(t, X(t)) + \sqrt{h b(t, X(t))} \xi(t)$$

for a standard normal random variable $\xi(t)$ (with $\xi(t')$ independent of $\xi(t)$ if $t \neq t'$). We write this as a *stochastic differential equation*,

$$dX_t = a(t, X_t) dt + \sqrt{b(t, X_t)} dB_t, \quad (4.4)$$

where $\{B_t\}_{t \geq 0}$ is Brownian motion.

Diffusion processes often arise as limits of discrete time and/or space Markov processes that move through a sequence of frequent small jumps. (The prototype is Brownian motion, which can be seen as a scaling limit of simple random walk.) Suppose that we have a sequence of such processes $\{\tilde{X}^{(h)}(nh)\}_{n \in \mathbb{N}}$ and write $\Delta \tilde{X}^{(h)}(t) = \tilde{X}^{(h)}(t+h) - \tilde{X}^{(h)}(t)$. Then if

$$a(t, x) = \lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[\Delta_h \tilde{X}^{(h)}(t) | \tilde{X}^{(h)}(t) = x] \quad (4.5)$$

and

$$b(t, x) = \lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[(\Delta_h \tilde{X}^{(h)}(t))^2 | \tilde{X}^{(h)}(t) = x],$$

both exist, and

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{E}[(\Delta_h \tilde{X}^{(h)}(t))^4 | \tilde{X}^{(h)}(t) = x] = 0, \quad (4.6)$$

then as h tends to zero, we can approximate $\tilde{X}^{(h)}$ by the diffusion process with drift and diffusion coefficients a and b , respectively. The fourth-moment condition in (4.6) can be relaxed, but is often rather easy to check. For a more thorough treatment we refer to Karlin and Taylor (1981) or Durrett (1996).

But let us see how it works for the Wright–Fisher model. We are interested in approximating the way that allele frequencies (denoted by p , in keeping with our previous notation) evolve with time (measured in units of N generations). We already calculated the mean and variance of the increment over a single generation (4.1), and the fourth centred moment of a $\text{Binom}(N, p)$ is $Np(1-p)(3p(1-p)(N-2)+1)$, which is order N^2 , so that the fourth moment of the change in allele frequencies over a single generation is order $1/N^2$. Setting $h = 1/N$, we see that for large N , measuring time in units of N generations, the allele frequencies can be approximated by the one-dimensional diffusion process with zero drift (reflecting neutrality) and diffusion coefficient $p(1-p)$. Written as a stochastic differential equation, this becomes

$$dp_t = \sqrt{p_t(1-p_t)} dB_t,$$

where B_t is a Brownian motion. The solution to this equation is often called the Wright–Fisher diffusion. It is an unfortunate accident of history that the term that mathematicians call ‘diffusion’ is modelling what geneticists call ‘genetic drift’; to a mathematician, ‘drift’ is the systematic ‘deterministic’ part of a diffusion model, determined by (4.3).

This forwards in time model for allele frequencies comes hand in hand with the Kingman coalescent as a model for the genealogical trees relating individuals in a random sample from the population. Changing back to ‘real’ time units in our Kingman coalescent, by multiplying the coalescence rate by $1/N_e$, corresponds to the same time-change in the Wright–Fisher diffusion, which then becomes

$$dp_t = \sqrt{\frac{1}{N_e} p_t(1-p_t)} dB_t.$$

4.2.2.1 Adding Selection and Mutation

The neutral Wright–Fisher model described above can be viewed as a simple prototype for a wealth of models that incorporate more realistic biological assumptions. We illustrate by considering a population of N diploid individuals, with (possibly frequency-dependent) selection and mutation between allelic types P and Q . We use p to denote the proportion of type P chromosomes and $q = 1 - p$ for the proportion of type Q . Note that the total number of chromosomes is now $2N$ and this number will play the role played by population size in our haploid model.

Each diploid individual has type PP , PQ , or QQ , and we write P_{11} , P_{12} and P_{22} for the corresponding proportions of each type. Then $p = P_{11} + \frac{1}{2}P_{12}$ and $q = P_{22} + \frac{1}{2}P_{12}$. During the reproductive process, each individual produces a large (effectively infinite) number of germ cells (cells of the same genotype) that split into gametes (cells containing just one chromosome from each pair). A total of $2N$ gametes are sampled from this infinite pool, and they fuse at random to form the next generation of N diploid individuals. We assume that there is selection in favour of certain genotypes. Further, there is mutation from type P to Q and vice versa.

Suppose that immediately before the reproductive step, the proportion of type P is p . For simplicity we assume multiplicative selection. That is, relative fitnesses of $PP : PQ : QQ$ can be expressed in the form $\eta_1^2 : \eta_1\eta_2 : \eta_2^2$. After selection, the proportion of type P alleles will be

$$p^* = \frac{\eta_1^2 p^2 + \eta_1 \eta_2 p q}{\eta_1^2 p^2 + 2\eta_1 \eta_2 p q + \eta_2^2 q^2} = \frac{\eta_1 p}{\eta_1 p + \eta_2 q}.$$

This means that we can model selection as acting on haploids. Since we only care about relative fitness, without loss of generality we set $\eta_2 = 1$ and then after selection the proportion of type P will be $p^* = \frac{p(1+s_0)}{1+s_0 p}$ where $s_0 = \eta_1 + 1$. In the case of directional selection, s_0 is just a constant, but by taking s_0 to be frequency dependent (i.e. a function of p), we can approximate more complicated selection acting on the diploid population. For example, a form of balancing selection can be modelled by assuming that $s_0(p) = \tilde{s}_0(p_0 - p)$ for some $0 < p_0 < 1$ and constant \tilde{s}_0 . If the population size is sufficiently large, this is close to a model of heterozygous advantage with relative diploid fitnesses $PP : PQ : QQ$ of $1 - \tilde{s}_0 q_0 : 1 : 1 - \tilde{s}_0 p_0$, where $q_0 = 1 - p_0$.

We now account for mutation between P and Q . Suppose that in each generation a mutation from P to Q has probability μ_1 and from Q to P has probability μ_2 . After the mutation step, the proportion of type P is $p^{**} = (1 - \mu_1)p^* + \mu_2(1 - p^*)$. Finally, $2N$ gametes are chosen at random to form the next generation. The resulting number of type P chromosomes in the population will then be binomially distributed with $2N$ trials and success probability p^{**} .

If selection is not too strong, and mutation is not too common, then measuring time in units of $2N$ generations, we can approximate the dynamics of allele frequencies by a one-dimensional diffusion. To establish what form that takes, define $\Delta p(t)$ to be the change in the frequency of P -alleles between generations t and $t + 1$. Since the population is no longer neutral, this quantity will no longer have zero expectation. Instead

$$\begin{aligned}\mathbb{E}[\Delta p(t)|p(t) = p] &= p^{**} - p = \frac{(1 - \mu_1)(1 + s_0)p + \mu_2(1 - p)}{1 + s_0 p} - p \\ &= \frac{s_0 p(1 - p) - \mu_1 p + \mu_2(1 - p) - \mu_1 s_0 p}{1 + s_0 p}.\end{aligned}$$

We take $h = 1/2N$ in (4.5) and, in order to obtain a non-trivial limit, we suppose that $2Ns_0 \rightarrow s$, $2N\mu_1 \rightarrow u$ and $2N\mu_2 \rightarrow v$ as $N \rightarrow \infty$. In that case

$$\begin{aligned}\lim_{N \rightarrow \infty} 2N\mathbb{E}[\Delta p(t)|p(t) = p] &= sp(1 - p) - up + vp(1 - p), \\ 2N\mathbb{E}[(\Delta p(t))^2|p(t) = p] &= p(1 - p) + \mathcal{O}\left(\frac{1}{N}\right)\end{aligned}$$

and, just as in the neutral case,

$$\mathbb{E}[(\Delta p(t))^4|p(t) = p] = \mathcal{O}\left(\frac{1}{N^2}\right).$$

We conclude that for sufficiently large N , measuring time in units of $2N$ generations, the process of allele frequencies can be approximated by

$$dp_t = (sp_t(1 - p_t) - up_t + vp(1 - p_t))dt + \sqrt{p_t(1 - p_t)}dB_t. \quad (4.7)$$

For frequency dependent selection, we simply replace $s = 2Ns_0$ by $s(p) = 2Ns_0(p)$. Scaling time by $2N$ to recover an equation in ‘real’ time units, this becomes

$$dp_t = (s_0 p_t(1 - p_t) - \mu_1 p_t + \mu_2(1 - p_t))dt + \sqrt{\frac{1}{2N}p_t(1 - p_t)}dB_t. \quad (4.8)$$

4.2.2.2 The Kolmogorov Equations

One of the most powerful tools in studying the distribution of a solution of a stochastic differential equation is the corresponding infinitesimal generator. This allows us to characterise the distribution in terms of a differential equation. Suppose that X_t is the one-dimensional diffusion that solves the stochastic differential equation (4.4). For any set B , and times $t < T$, let

$$p_B(t, x; T, y) = \mathbb{P}[X_T \in B | X_t = x] = \int_B p(t, x; T, y) dy$$

be the probability that X_T is in B given that $X_t = x$. Then the *transition density* $p(t, x; T, y)$ satisfies the *Kolmogorov backward equation*:

$$\begin{aligned} \frac{\partial p}{\partial t}(t, x; T, y) + \mathcal{A}p(t, x; T, y) &= 0 \\ p(t, x; T, y) &\rightarrow \delta_y(x) \quad \text{as } t \rightarrow T, \end{aligned}$$

where $\delta_y(x)$ is the Dirac delta function at y , and

$$\mathcal{A}f(t, x) = a(t, x) \frac{\partial f}{\partial x} + \frac{1}{2} b(t, x) \frac{\partial^2 f}{\partial x^2}$$

is the infinitesimal generator of the process $\{X_t\}_{t \geq 0}$ (note the factor of 1/2 in front of b). The name ‘backward equation’ comes from the fact that it operates on the ‘backwards in time’ variables (t, x) . In the time-homogeneous case, we often write

$$\mathcal{A}f(x) = \frac{d}{dt} \mathbb{E}[f(X_t) | X_0 = x] \Big|_{t=0}.$$

The form of the generator is then immediate from Taylor’s theorem, since

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[f(X_t) | X_0 = x] \Big|_{t=0} &= \lim_{h \downarrow 0} \frac{1}{h} \left\{ \mathbb{E}[\Delta_h X | X_0 = x] f'(x) \right. \\ &\quad \left. + \frac{1}{2} \mathbb{E}[(\Delta_h X)^2 | X_0 = x] f''(x) + \mathcal{O}(\mathbb{E}[(\Delta_h X)^3 | X_0 = x]) \right\}. \end{aligned}$$

We can also write down an equation acting on the ‘forwards variables’ (T, y) . This is known both as the *Kolmogorov forward equation* and the *Fokker–Planck equation*. It takes the form

$$\frac{\partial p}{\partial T}(t, x; T, y) = \mathcal{A}^* p(t, x; T, y),$$

where

$$\mathcal{A}^* f(T, y) = -\frac{\partial}{\partial y}(a(T, y)f(T, y)) + \frac{1}{2} \frac{\partial^2}{\partial y^2}(b(T, Y)f(T, y)).$$

Things are especially convenient if a and b are independent of time. This is true, for example, for Brownian motion, for which $a = 0$ and $b = 1$, so $\mathcal{A}f = \frac{1}{2} \frac{\partial^2 f}{\partial x^2}$, and the Wright–Fisher diffusion (for which we shall use $p \in [0, 1]$ for the variable representing the allele frequency), for which $a = 0$ and $b = p(1 - p)$, so $\mathcal{A} = \frac{1}{2}p(1 - p) \frac{\partial^2 f}{\partial p^2}$. A nice feature of time-homogeneous one-dimensional diffusions is that many quantities can be calculated explicitly. This is because (except at some singular points which for models of allele frequencies are generally $p = 0$ and $p = 1$), they can be transformed into a Brownian motion through first a change of space variable and then a time-change. This is known as the theory of speed and scale and is laid out with a particular emphasis on genetic models in Etheridge (2011).

Let us give one application of the forward equation for the diffusion in (4.7). Suppose that we are interested in the equilibrium distribution of allele frequencies. The conditional probability

density should converge as $T \rightarrow \infty$ to a limit that is independent of the initial allele frequency. Then taking limits in the Kolmogorov forwards equation, we find that the density $\varphi(p)$ should satisfy

$$\frac{1}{2} \frac{d^2}{dp^2} [p(1-p)\varphi(p)] - \frac{1}{2} \frac{d}{dp} [(sp(1-p) - u p + v p(1-p))\varphi(p)] = 0.$$

Integrating, this yields

$$\varphi(p) = Cp^{2v-1}(1-p)^{2u-1}e^{2sp}, \quad 0 \leq p \leq 1,$$

where the constant C is chosen so that $\int_0^1 \varphi(p) dp = 1$. Kimura (1956) was the first to use diffusion theory to establish this formula.

Suppose that we write $\bar{W}(p)$ for the mean fitness in the population if the frequency of P -alleles is p . Then $\bar{W}(p) = \eta_1^2 p^2 + 2\eta_1\eta_2 p(1-p) + \eta_2^2(1-p)^2 = (\eta_1 p + \eta_2 q)^2 = (1 + s_0 p)^2 \approx e^{2s_0 p}$. (In a haploid population this would instead be $e^{s_0 p}$.) Rewriting φ in terms of the unscaled coefficients of our model, we then find for our diploid model

$$\varphi(p) \approx Cp^{4N_e\mu_2-1}(1-p)^{4N_e\mu_1-1}\bar{W}(p)^{2N_e}. \quad (4.9)$$

Wright (1949) derived this expression from the consideration that the mean and variance of the stationary distribution must be unchanged in successive generations. If s_0 is replaced by a function of p , then the same expression for the stationary distribution remains valid on replacing $e^{s_0 p}$ by $\exp(\int_0^p s_0(x)dx)$.

4.2.2.3 Multiple Alleles

So far we have restricted ourselves to the case in which there are just two possible alleles, P and Q , at the locus under consideration. This allowed us to characterise allele frequencies in terms of a one-dimensional diffusion. The same ideas can be extended to more than two types, resulting in a multi-dimensional diffusion. We illustrate this in the simplest case of a neutral haploid population.

Suppose that there are K allelic types, A_1, \dots, A_K . Once again we start from a Wright–Fisher model. The population configuration at any time can be described by a vector $\underline{X} = (X_1, X_2, \dots, X_K)$ where X_i is the number of genes of allelic type A_i and we assume that $X_1 + \dots + X_K = N$. In the neutral Wright–Fisher model without mutation,

$$\begin{aligned} \mathbb{P}[\underline{X}(t+1) = (Y_1, \dots, Y_K) | \underline{X}(t) = (X_1, \dots, X_K)] \\ = \frac{N!}{Y_1! Y_2! \dots Y_K!} \psi_1^{Y_1} \psi_2^{Y_2} \dots \psi_K^{Y_K}, \end{aligned}$$

where $\psi_i = \frac{X_i}{N}$ and $\sum_{i=1}^K Y_i = N$ (the probability is zero if this last condition is not satisfied).

Write $\underline{p}(t) = (p_1(t), \dots, p_K(t))$ where $p_i(t) = X_i(t)/N$ and consider the increment $\Delta p_i = p_i(t+1) - p_i(t)$. By ‘pooling’ all the alleles A_j for $j \neq i$ into a single class (‘not A_i ’), we recover the Wright–Fisher model for two alleles for which we already checked that

$$\mathbb{E}[\Delta p_i] = 0, \quad \text{var}(\Delta p_i) = \frac{1}{N} p_i(1-p_i)$$

and

$$\mathbb{E}[(\Delta p_i)^k] = \mathcal{O}\left(\frac{1}{N^2}\right) \quad \forall k \geq 3.$$

To complete the picture we need the *covariances*, that is, we must calculate, for $i \neq j$,

$$\begin{aligned}\mathbb{E}[\Delta p_i \Delta p_j] &= \frac{1}{N^2} \mathbb{E}[(X_i(t+1) - X_i(t))(X_j(t+1) - X_j(t)) | \underline{p}(t)] \\ &= \frac{1}{N^2} \mathbb{E}[X_i(t+1)X_j(t+1) | \underline{p}(t)] - p_i(t)p_j(t).\end{aligned}\quad (4.10)$$

Now

$$\begin{aligned}\mathbb{E}_t[X_i(t+1)X_j(t+1)] \\ = \frac{1}{2} \left\{ \mathbb{E}_t[(X_i(t+1) + X_j(t+1))^2] - \mathbb{E}_t[X_i(t+1)^2] - \mathbb{E}_t[X_j(t+1)^2] \right\}\end{aligned}$$

and, again ‘pooling’ genes of type A_i and A_j , we just need to recall that for a $\text{Binom}(N, p)$ distribution,

$$\mathbb{E}[X^2] = Np(1-p) + N^2p^2.$$

This gives

$$\begin{aligned}\mathbb{E}_t[X_i(t+1)X_j(t+1)] &= \frac{1}{2} \left\{ N(p_i + p_j)(1 - (p_i + p_j)) + N^2(p_i + p_j)^2 \right. \\ &\quad \left. - Np_i(1 - p_i) - N^2p_i^2 - Np_j(1 - p_j) - N^2p_j^2 \right\} = -Np_ip_j + N^2p_ip_j\end{aligned}$$

and so

$$\mathbb{E}[\Delta p_i \Delta p_j] = -\frac{1}{N}p_ip_j.$$

If we consider functions $f(p_1, \dots, p_{K-1})$ (note that $p_K = 1 - \sum_{j=1}^{K-1} p_j$), rescale time so that the intergeneration time is $1/N$, and let $N \rightarrow \infty$, then we can use Taylor’s theorem to identify the infinitesimal generator of the limiting model. This gives

$$\mathcal{L}f := \frac{d}{dt} \mathbb{E}[f(p_1, \dots, p_{K-1})] \Big|_{t=0} = \frac{1}{2} \sum_{i,j \in \{1, \dots, K-1\}} p_i(\delta_{ij} - p_j) \frac{\partial^2 f}{\partial p_i \partial p_j},$$

where $\delta_{ij} = 1$ if $i = j$, 0 otherwise.

We can also add mutation. Suppose that for $i \neq j$, μ_{ij} is the probability that an offspring of a type A_i parent is of type A_j (assumed to be independent for different offspring). Then writing $\beta_{ij} = \lim_{N \rightarrow \infty} N\mu_{ij}$, the generator of the K -allele Wright–Fisher diffusion with mutation is

$$\frac{1}{2} \sum_{i,j \in \{1, \dots, K-1\}} p_i(\delta_{ij} - p_j) \frac{\partial^2 f}{\partial p_i \partial p_j} + \sum_{i=1}^{K-1} \left(-p_i \sum_{j=1, j \neq i}^K \beta_{ij} + \sum_{j=1, j \neq i}^K p_j \beta_{ji} \right) \frac{\partial f}{\partial p_i}.$$

4.2.2.4 Gaussian Fluctuations and Drift Load

Equation (4.7) provided a way to approximate p when mutation, selection and random drift are all comparable. If mutation and selection are much weaker than drift, $2N_e s_0 \ll 1$, $2N_e \mu_i \ll 1$, then we recover a model of pure drift. The third regime to have been extensively studied is that in which selection and mutation are weak, but still dominate drift. In the notation of our diffusion approximation for the Wright–Fisher model, $s_0 = \varepsilon_N s$ and $\mu_1 = \varepsilon_N u$, $\mu_2 = \varepsilon_N v$, with $\varepsilon_N \rightarrow 0$ but $N\varepsilon_N \rightarrow \infty$ as $N \rightarrow \infty$.

In this setting, one can model allele frequencies on time-scales of $1/\varepsilon_N$ generations by the deterministic equation

$$\frac{d\tilde{p}}{dt} = s(\tilde{p})\tilde{p}(1 - \tilde{p}) - u\tilde{p} + v(1 - \tilde{p}).$$

Under some mild conditions, for very large N_e the fluctuations about the deterministic limit are a Gaussian process and can also be characterised. Rather than give details of that theory, let us consider a special example.

Suppose that we are modelling overdominance with relative fitnesses $PP : PQ : QQ$ of $1 - s_0 / 2 : 1 : 1 - s_0 / 2$. We can approximate allele frequencies using the analogue of equation (4.8) for density-dependent selection. We will also suppose that $\mu_i \ll s$, and so the equation becomes

$$dp = s_0 p(1-p) \left(\frac{1}{2} - p \right) dt + \sqrt{\frac{1}{2N_e} p(1-p)} dB_t,$$

where now we are supposing that $N_e s \gg 1$, so that although eventually this diffusion will be absorbed at $p = 0$ or $p = 1$, we expect it to spend a long time close to $p = 1/2$ before this happens. The deterministic limiting equation (obtained with $N_e = \infty$) has an equilibrium at $\bar{p} = 1/2$. Notice that at this equilibrium, the mean fitness in the population is

$$\left(1 - \frac{s_0}{2}\right)\bar{p}^2 + 2\bar{p}(1-\bar{p}) + \left(1 - \frac{s_0}{2}\right)(1-\bar{p})^2 = 1 - \frac{s_0}{4},$$

whereas the maximum fitness in the population is 1. The quantity $s_0/4$ is the *segregation load*.

Now consider the stochastic equation close to equilibrium and for very large N_e . Writing $z = \bar{p} - p = \frac{1}{2} - p$, we approximate by the equation linearised about \bar{p} which yields

$$dz = -\frac{s_0}{4} z dt + \sqrt{\frac{1}{4} \frac{1}{2N_e}} dB_t.$$

The solution to this equation is an Ornstein–Uhlenbeck process which has a Gaussian stationary distribution with mean 0 and variance $1/(4N_e s_0)$. Although our argument is heuristic, one can pass directly from an individual-based model to this model for fluctuations about a deterministic limiting equation for allele frequencies (Norman, 1975; Nagylaki, 1990).

If we include the fluctuations in the population, the mean fitness is

$$1 - \frac{s_0}{4} - s_0(p - \bar{p})^2$$

when the frequency of P -alleles is p which, at stationarity, has mean

$$1 - \frac{s_0}{4} - s_0 \frac{1}{4N_e s_0} = 1 - \frac{s_0}{4} - \frac{1}{4N_e}. \quad (4.11)$$

The reduction in mean fitness of $1/(4N_e)$ is called the *drift load*. Notice that it is independent of the strength of selection. Increasing s_0 reduces the size of the fluctuations around the deterministic equilibrium, but this is exactly cancelled by the increased loss of fitness associated with a deviation of a given size. Robertson (1970) obtained this result from Wright's formula. If there are K alleles, then the drift load is $(K - 1)/(4N_e)$.

4.2.3 Spatially Structured Populations

4.2.3.1 Continuous Space: the Wright–Malécot Formula and the Pain in the Torus

Most populations are distributed in space, often a continuous two-dimensional environment. Wright (1943) and Malécot (1948) both suggested extending the Wright–Fisher model to a spatial context. Suppose that the population is distributed across two-dimensional Euclidean space \mathbb{R}^2 . Recall that in the Wright–Fisher model, the number of offspring of an individual has a $\text{Binom}(N, 1/N)$ distribution which, for large N , is approximately a Poisson random variable with mean 1. In the spatial context, it seems natural to take the number of offspring of each individual to be $\text{Poiss}(1)$. Offspring will disperse around the position of the parent, and Wright and Malécot both assume that, at maturity, they are distributed according to (independent) symmetric Gaussian random variables centred on the position of the parent. They suppose

that individuals in the parental population are distributed across space according to a Poisson random field with constant intensity ρ . This means that the number of individuals in a region A of area $|A|$ is Poisson with intensity $\rho|A|$, the numbers of individuals in disjoint regions are independent, and if we know that there are k points in a region A , then the location of each of those points is an independent uniform random variable over A . The parameter ρ reflects local population density. Wright and Malécot assumed that with this initial condition, the offspring would also be distributed according to a Poisson random field with intensity ρ . In mathematical language, they assumed that the distribution of the Poisson random field was stationary for their model. Under this assumption, and with an infinitely many alleles mutation model, one can then write down recursions for the probability of identity in state of two individuals sampled at separation x from the population; that is, if the probability of a mutation between parent and child is μ , independently for each child in each generation, then they write down an expression for the probability that there have been no mutations along the ancestral lineages of the two individuals since their most recent common ancestor. We recall their argument in the context of the Kimura stepping stone model below.

Felsenstein (1975) was the first to notice that there is an inconsistency in the Wright–Malécot assumptions. In fact, the Poisson random field is not stationary for the Wright–Malécot model. In one and two spatial dimensions a population evolving according to the Wright–Malécot model, which in mathematical language is a critical branching random walk, will die out locally, but on the road to extinction it will develop ‘clumps’ of arbitrarily large density and extent.¹ One can try to overcome this by working instead, for example, on a torus (a circle in one dimension and a square with opposite sides identified in two dimensions). To see what then happens, consider the simulation in Figure 4.1 (kindly supplied by Jerome Kelleher). At time zero, 1000 individuals are thrown down uniformly at random in the torus, resulting in the top left-hand frame. Starting from this initial condition, the population evolves according to the Wright–Malécot model. For this realisation, the frames show the state of the population after 10, 100 and 1000 generations. Since individuals reproduce independently of one another, the total population size is a critical Galton–Watson branching process, and so the population will eventually die out. The probability that an individual ancestor leaves descendants after t generations declines like $1/t$. Even if we allow the mean number of offspring of individuals to differ from 1, unless we introduce some dependence between individuals’ reproductive success, the total population will always either go to zero or grow without bound. One might hope that, conditioning on the total population size being constant, say, we would arrive at a sensible model, but as we see from our simulation, in which the population was very close to 1000 for the first 100 generations, this does not overcome the problem of clumping. Felsenstein dubbed the problem ‘the pain in the torus’.

4.2.3.2 Kimura’s Stepping Stone Model

The lesson that we learn from Felsenstein’s work is that in order to overcome clumping we will have to introduce some *local* regulation of population size. An extreme version of this is to constrain the population to live in demes of constant size, situated at the vertices of a graph. This leads us to the stepping stone model of Kimura (1952).

We take the demes to be at the vertices of \mathbb{Z}^2 . There are $2N$ genes in each deme. We suppose that the population evolves in discrete generations. In each generation, first offspring are generated by Wright–Fisher sampling within each deme. Next, a proportion $g_1(x - y)$ of the offspring in deme x migrate to deme y .

¹ In fact there is an error in Felsenstein’s paper: he concluded that this clumping took place in all dimensions. This was corrected by Sudbury (1977).

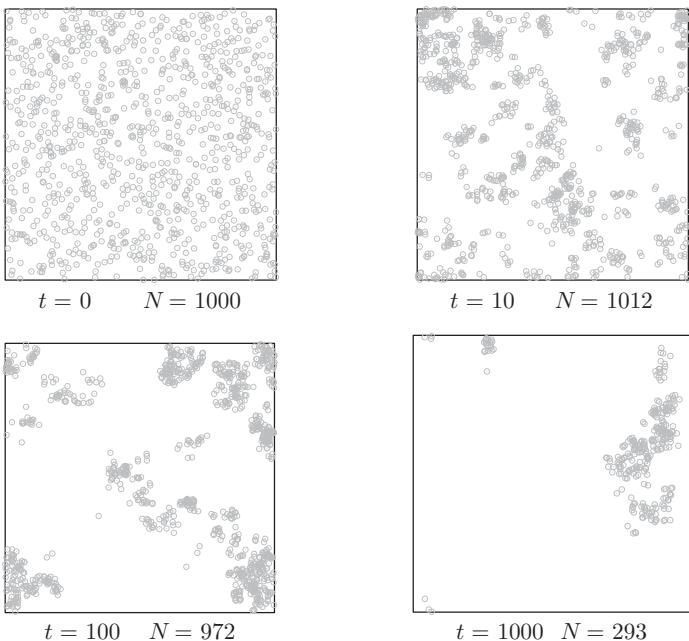


Figure 4.1 The pain in the torus: a population evolving according to the Wright–Malécot model on a torus. In each generation, each individual (independently) produces Poiss(1) offspring, distributed (independently) in a Gaussian distribution around the location of the parent. Although the population size in this example remains fairly stable for time-scales of hundreds of generations, the population develops ‘clumps’. Top row, starting configuration of individuals and configuration after 10 generations; bottom row, configuration after 100 and 1000 generations.

We derive the Wright–Malécot formula in this setting; that is, we find an expression for the probability that two genes sampled at separation x (a vector in \mathbb{Z}^2) are identical in state. If the probability of a mutation per individual per generation is $1 - e^{-\mu}$, then this is equivalent to finding $\mathbb{E}_x[e^{-2\mu T}]$ where T is the time back to the most recent common ancestor, and the subscript x refers to the sampling separation. We shall write $\phi(z, x) = \mathbb{E}_x[z^T]$. Our derivation parallels the approach of Wright (1943).

Let $\psi_t(x)$ be the probability that two genes sampled at separation x had their most recent common ancestor exactly t generations in the past. For $t \geq 1$, we decompose this quantity according to the separation of the immediate ancestors of the two genes. For $t = 1$, $\psi_1(x)$ is $1/2N$ times the probability that two genes at separation x arose as migrants from the same deme, that is,

$$\psi_1(x) = \frac{1}{2N} G_1(x),$$

where $G_1(x)$ is the convolution of two copies of g_1 (corresponding to modelling the *separation* of two lineages). If, on the other hand, they have distinct parents, at separation y , then the chance that their most recent common ancestor was t generations in the past is $\psi_{t-1}(y)$. For $t > 1$, we arrive at the recursion

$$\begin{aligned} \psi_t(x) &= \sum_y \left\{ G_1(x-y)\psi_{t-1}(y) - \frac{\mathbf{1}_{\{y=0\}}}{2N} G_1(x-y)\psi_{t-1}(0) \right\} \\ &= \frac{1}{2N} \left(G_t(x) - \sum_{\tau=1}^{t-1} G_{t-\tau}(x)\psi_\tau(0) \right), \end{aligned}$$

where G_t is the t -fold convolution of G_1 . Multiplying by z^t and summing over t yields

$$\phi(z, x) = \frac{\tilde{G}(z, x)}{2N} (1 - \phi(z, 0)),$$

where \tilde{G} denotes the discrete Laplace transform of G ,

$$\tilde{G}(z, x) = \sum_{t=1}^{\infty} G_t(x) z^t.$$

Setting $x = 0$ to find an expression for $\phi(z, 0)$ and substituting gives

$$\phi(z, x) = \frac{\tilde{G}(z, x)}{2N + \tilde{G}(z, 0)}. \quad (4.12)$$

This takes a particularly simple form if g_1 is a discretised Gaussian kernel which we can then approximate by a strictly Gaussian dispersal kernel. On an infinite range,

$$\frac{1}{2N} G_t(x) = \frac{1}{2\mathcal{N}t} \exp\left(-\frac{|x|^2}{4\sigma^2 t}\right),$$

where $\mathcal{N} = 4N\pi\sigma^2$ is the *neighbourhood size*. (This corresponds to dispersal of individual lineages at rate $\sigma^2/2$, the extra factor of 2 arising because G_1 governs the separation between two lineages.) With this continuous approximation to G_t ,

$$\frac{1}{2N} \tilde{G}(z, 0) = \frac{1}{2\mathcal{N}} \sum_{t=1}^{\infty} \frac{z^t}{t} = \frac{1}{\mathcal{N}} \log\left(\frac{1}{\sqrt{1-z}}\right)$$

and

$$\frac{1}{2N} \tilde{G}(z, x) = \frac{1}{\mathcal{N}} \sum_{t=1}^{\infty} \frac{z^t}{2t} \exp\left(-\frac{|x|^2}{4\sigma^2 t}\right). \quad (4.13)$$

Substituting $z = e^{-2\mu}$, this gives an exact expression for the probability of identity. In fact it is usual to approximate it. Provided that $|x|\sqrt{1-z}/\sigma$ is not too small, $|x|/\sigma > 2$, say, and $z > 0.5$, the quantity in (4.13) is approximately

$$\frac{1}{\mathcal{N}} K_0\left(\frac{|x|}{\sigma} \sqrt{1-z}\right),$$

where K_0 is the modified Bessel function of the second kind of degree 0. To circumvent the divergence of the Bessel function as $x \rightarrow 0$, it is often convenient to proceed as in Barton *et al.* (2002) and declare there to be a local scale κ over which the generating function is approximately constant and equal to $\tilde{\phi}(z, 0)$. Writing equation (4.13) as

$$\phi(z, x) = \frac{1 - \tilde{\phi}(z, 0)}{\mathcal{N}} K_0\left(\frac{|x|}{\sigma} \sqrt{1-z}\right),$$

equating $\phi(z, \kappa)$ to $\tilde{\phi}(z, 0)$ and rearranging (using the fact that $K_0(y) \approx -\log y$ as $y \downarrow 0$), we obtain

$$\phi(z, x) \approx \frac{K_0\left(\frac{|x|}{\sigma} \sqrt{1-z}\right)}{\mathcal{N} - \log\left(\frac{\kappa}{\sigma} \sqrt{1-z}\right)}. \quad (4.14)$$

Setting $z = e^{-2\mu}$, we recover the Wright–Malécot formula:

$$\phi(e^{-2\mu}, x) = \mathbb{E}_x[e^{-2\mu T}] \approx \frac{K_0(x/\ell_\mu)}{\mathcal{N} + \log(\ell_\mu/\kappa)}, \quad \text{for } |x| > \kappa, \quad (4.15)$$

where $\ell_\mu = \sigma/\sqrt{2\mu}$ and

$$\phi(e^{-2\mu}, 0) = \frac{\log(\ell_\mu/\kappa)}{\mathcal{N} + \log(\ell_\mu/\kappa)}.$$

Kimura and Weiss (1964) obtained this formula for the stepping stone model. Essentially the same proof applies to any dispersal distribution; for example, Durrett (2008, Theorem 5.7) finds an expression in the case of nearest-neighbour random walk.

4.2.3.3 Back to Continuous Space

The stepping stone model is one rather extreme form of local population regulation. It seems somewhat unnatural to force the population to live in a discrete space. In Berestycki *et al.* (2009), an alternative approach to local population regulation is considered. Reproduction ‘events’ are determined by a Poisson point process which prescribes the time t when the event occurs, a region of space A in which the population will be changed by the event, and an ‘impact’ u which determines, for each individual within the region, the probability that it will die during the event. Each individual living within the region has an equal chance of being chosen as the parent of the event (and thus reproductive success is lower in crowded regions). Individuals living within the region, independently, die with probability u and a $\text{Poiss}(\rho u |A|)$ number of offspring, of the same genetic type as the parent, are thrown down uniformly at random across A . (Here $|A|$ denotes the area of A and ρ is population density.)

If ρ is sufficiently large, then the process has a non-trivial stationary distribution which is stochastically bounded by a Poisson random field of intensity ρ (and so, in particular, does not have clumps). In other words, by basing reproduction on events rather than on individuals we can overcome the pain in the torus. One can write down expressions for the genealogical trees relating individuals in a sample from the population and they converge rapidly as ρ increases. It turns out to be convenient to pass to the limit as $\rho \rightarrow \infty$. The resulting process, known as the *spatial Lambda–Fleming–Viot process* was introduced in Etheridge (2008) and rigorously constructed in Barton *et al.* (2010).

There are many variants of this model, which provides a flexible framework into which one readily incorporates, for example, selection, recombination, large-scale extinction–recolonisation events and so on. One can also write down a discrete generation version of the model, in which a random tessellation of the plane determines disjoint regions in which reproduction events take place. The approach is reviewed in Barton *et al.* (2013b).

4.3 Multiple Loci

In principle, the population genetics of multiple loci is straightforward: one simply tracks the proportions of all possible genotypes, as they change under the various evolutionary processes. However, the number of variables needed to describe the state of the population, and the number of parameters needed to describe how fitness depends on genotype, both grow geometrically with the number of genes. This makes a direct extension of single-locus methods to large numbers of loci infeasible, even numerically. Fortunately, there are ways to represent and to approximate the evolution of multiple loci that can help us understand the complexities of real populations.

If mating is random, and recombination is fast relative to other processes, then genotype frequencies are just the product of the frequencies of the component alleles; we say that the population is at *linkage equilibrium*. We discuss this situation in Section 4.3.1. The state of the

population is described by the allele frequencies, and the selection on any allele depends only on its marginal effect on fitness, which equals the gradient of mean fitness with respect to allele frequency. Thus, mean fitness acts as a potential function, which leads to some simple results, even when many loci interact to determine fitness (i.e. when there is epistasis).

Analysis is much harder when the rate of recombination is comparable with that of other processes, so that allele frequencies at different loci are not independent. In order to determine the evolution of the population, it no longer suffices to follow the allele frequencies: the state space is now far larger. Usually, it is not sensible to work directly with genotype frequencies, since this gives little intuition, and since even in a large population, only a small fraction of genotypes may be present. Instead populations are best represented by the associations among sets of alleles, termed *linkage disequilibria*. Although general results are elusive, useful approximations can be made by assuming that recombination is fast enough that these associations are weak (quasi-linkage equilibrium, QLE), or that there are very large numbers of loci (e.g. the infinitesimal model).

There is a very large literature on the population genetics of multiple loci. We can only cover a small fraction of this; so we emphasise the basic principles, and analytical results, rather than simulation. We also emphasise general results that apply to large numbers of loci, rather than analyses of specific models with just a few loci.

In Section 4.3.1 we summarise results that assume linkage equilibrium. In Section 4.3.2, we then show how populations can be represented in terms of multi-locus linkage disequilibria, and review results derived assuming QLE. We discuss the limit of very many loci, and its relation with quantitative genetics. Finally, we consider a quite different approach, which treats the genome as continuous, and follows the blocks of genome that are inherited from different ancestors.

4.3.1 Linkage Equilibrium

4.3.1.1 Selection Gradients

We suppose an infinite population and no mutation. Under linkage equilibrium, different loci evolve independently. For our single-locus diploid Wright–Fisher model with selection (and mutation rates that we set to zero), with just two alleles, the change in allele frequency over a single generation will be

$$\Delta p = p^* - p = \frac{s_0 p(1-p)}{1 + s_0 p} \approx s_0 p(1-p) = \frac{1}{2} p(1-p) \frac{\partial \bar{W}(p)}{\partial p},$$

where $\bar{W}(p) = e^{2s_0 p}$ appeared in equation (4.9). For a haploid population, $\bar{W}(p) = e^{s_0 p}$ and the factor of 1/2 disappears. A similar expression holds when there are multiple alleles.

If diploid genotype frequencies are given by the products of the frequencies of their alleles at different loci, then the change in allele frequencies due to selection will be

$$\Delta p_l = \frac{1}{2} p_l(1-p_l) \frac{\partial \log(\bar{W})}{\partial p_l} \quad (4.16)$$

for two alleles at locus l , and

$$\Delta p_{l,j} = \frac{1}{2} \sum_k p_{l,j} (\delta_{j,k} - p_{l,k}) \frac{\partial \log(\bar{W})}{\partial p_{l,k}} \quad (4.17)$$

for multiple alleles, with frequencies $p_{l,j}$, where $\delta_{j,k} = 1$ if $j = k$, 0 otherwise. We reserve the subscripts l, m for loci and j, k for alleles. Thus $p_{l,j}$ is the proportion of the alleles of the

j th type at locus l . As in the single-locus case, the factors of $1/2$ arise from diploidy, and would be omitted for haploids. These formulae apply with arbitrary epistasis.

If selection is weak, change can be approximated as continuous in time, so that $(\Delta p) \sim \partial_t p$. We can write the resulting system of differential equations in matrix form:

$$\partial_t p = G \cdot \partial_p \log(\bar{W}),$$

the matrix G having elements

$$G_{\{l,j\},\{m,k\}} = \frac{1}{2} \delta_{l,m} p_{l,j} (\delta_{j,k} - p_{l,k}). \quad (4.18)$$

More generally, in an infinite population at linkage equilibrium, we can find an equation of this form for any vector $\underline{x} = (x_1, x_2, \dots)$ of variables that suffices to describe the evolution of the population. We replace G by $\mathcal{G} = AGA^T$ where $A = (\partial x_\alpha / \partial p_\beta)$ is the Jacobian matrix of the coordinate transformation from p to x and A^T is its transpose. In particular, rather than seeing the population as evolving through the space of allele frequencies, we can think of the evolution of the *distribution* of the trait, as represented by (say) its moments. Suppose that a trait Z is some function $Z(X)$ of genotype X ; we ignore the contribution of the environment to the trait value. If fitness depends only on this trait, then the mean fitness is a function of the distribution of the trait, which in turn can be described by its mean, \bar{Z} , variance, V , third moment, M_3 , and so on; these moments are themselves functions of allele frequencies. For an additive trait, that is, one whose value can be written as a sum over contributions from each locus, and assuming linkage equilibrium, the moments of the trait depend on the k th central moments of the contributions of the l th locus, $m_{l,k}$. Thus, the mean is the sum of the first moments $\bar{Z} = 2 \sum_l m_{l,1}$ and the variance is $V = 2 \sum_l m_{l,2}$. (Barton and Turelli, 1987, equation 5.1) show that

$$\begin{aligned} \partial_t \begin{pmatrix} \bar{Z} \\ V \\ M_3 \\ \vdots \end{pmatrix} &= 2 \sum_l \\ &\left(\begin{pmatrix} m_{l,2} & m_{l,3} & (m_{l,4} - 3m_{l,2}^2) & \cdots \\ m_{l,3} & (m_{l,4} - 3m_{l,2}^2) & (m_{l,5} - 4m_{l,2}m_{l,3}) & \\ (m_{l,4} - 3m_{l,2}^2) & (m_{l,5} - 4m_{l,2}m_{l,3}) & (m_{l,6} - m_{l,3}^2 - 6m_{l,2}m_{l,4} + 9m_{l,2}^3) & \vdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \right. \\ &\times \left. \begin{pmatrix} \partial_{\bar{Z}} \log(\bar{W}) \\ \partial_V \log(\bar{W}) \\ \partial_{M_3} \log(\bar{W}) \\ \vdots \end{pmatrix} \right) \quad (4.19) \end{aligned}$$

This change of coordinates, from allele frequencies to trait moments, makes several points. First, if the trait distribution is Gaussian, then the rate of change of the mean is simply $V \partial_{\bar{Z}} \log(\bar{W})$ (Lande, 1976). This is true even if distributions at individual loci are not Gaussian (e.g. if there are only a few alleles), because the sums over loci of the first row of the matrix equal the cumulants of the trait distribution. Second, selection on the mean changes the variance in proportion to the third moment of the trait distribution, and conversely, selection on the variance changes the mean in proportion to the third moment. Third, because we assume linkage equilibrium, the trait distribution must become close to Gaussian as the number of loci, n , increases. Thus, the variance and higher moments change more slowly than the mean, by order $1/n$. Finally, this is not a closed dynamical system: changes in each moment

depend on the higher moments. More seriously, the elements of the matrix G do not depend only on the trait distribution, but rather on the distributions of effects of each locus. Restrictive assumptions, such as equal distributions across loci, are needed to obtain equations that depend only on the trait distribution; even then, one must approximate higher moments by some function of the lower moments in order to arrive at a closed system of equations.

The idea of representing trait evolution in terms of moments was developed independently by Bürger (1991, 2000), Turelli and Barton (1994), and Prügel-Bennett and Shapiro (1994); it can be applied to both asexual populations (Gerrish & Sniegowski, 2012), and to traits determined by multiple loci at linkage equilibrium. However, in our view this approach has not led to useful approximations for trait evolution, beyond the Gaussian framework of classical quantitative genetics. A simple thought experiment suggests why: if favourable alleles are initially at very low frequency in a large population, they will take a long time to become common enough to influence the trait distribution. Thus, the long-term evolution of the trait cannot be predicted from its current distribution. In what follows, we shall see that progress can be made in the limit of extremely large numbers of loci, and by including random drift.

4.3.1.2 Random Drift

We have already seen how to model genetic drift at a single locus in a finite population. Under linkage equilibrium, the covariance of fluctuations in allele frequencies across loci and types is given by the same matrix G (equation (4.18)) that determines the response to selection.

With just two alleles at each locus, we have immediately from equation (4.9) that the stationary distribution under the joint influence of selection, mutation and random drift takes the form

$$\psi(p) = C \left(\prod_l p_l^{4N_e v_l - 1} (1 - p_l)^{4N_e \mu_l - 1} \right) \bar{W}^{2N_e}, \quad (4.20)$$

where, at locus l , in each generation, a type P_l mutates to Q_l with probability μ_l , and mutations in the opposite direction happen with probability v_l (Wright, 1937). In order to obtain an explicit formula for the stationary distribution when there are more than two alleles per locus, we have to assume *parent-independent mutation*; that is, any allele at locus l mutates to type $A_{l,i}$ with probability $v_{l,i}$. Then

$$\psi(p) = C \left(\prod_l \prod_i p_{l,i}^{4N_e v_{l,i} - 1} \right) \bar{W}^{2N_e}.$$

This formula applies with arbitrary epistasis, and shows that at equilibrium, populations will cluster around ‘adaptive peaks’; this motivated Wright’s (1931) ‘shifting balance’ model of evolution.

Just as in the single-locus setting, one can consider what happens when the effective population size is large enough that the population clusters close to an adaptive peak. Entirely analogously to (4.11), the reduction in mean fitness due to drift (the drift load) will be $1/(4N_e)$ per ‘degree of freedom’ (Orr, 2000; Barton, 2017). This diffusion approximation extends to show that the rate at which populations cross an ‘adaptive valley’, despite selection, is proportional to $(\bar{W}_{\text{valley}}/\bar{W}_{\text{peak}})^{2N_e}$, which implies that such shifts will be very rare if the product of N_e and the depth of the valley is large (Wright, 1941; Lande, 1979).

In the limit of very low mutation rate ($4N_e \mu \ll 1$), populations will typically be fixed for one or other genotype, with probability proportional to $\prod_{p_l=0} \mu_l \prod_{p_l=1} v_l \bar{W}^{2N_e}$; the two terms involving mutation rates are the stationary distribution in the absence of selection. This formula can be derived by considering the probability of substitutions in either direction; it applies even

with linkage, because if substitutions are rare, they occur one at a time, and so linkage cannot make any difference. This ‘fixed state’ version of Wright’s formula was derived by Sella and Hirsh (2005), but is really a special case of Wright’s formula.

The stationary distribution of allele frequencies is the product of the neutral distribution, ψ_{Neu} , which would obtain with no selection, and the simple term \overline{W}^{2N_e} . As in the previous section, we can change coordinates, to consider the distribution of a quantitative trait. The state of the population is now described by the *distribution* of the trait, and if that is Gaussian, by its mean and variance, $\{\bar{Z}, V\}$. These will have some neutral distribution, $\phi_{\text{Neu}}(\bar{Z}, V)$, and if selection acts only via the trait, this is simply multiplied by \overline{W}^{2N_e} , so that $\phi(\bar{Z}, V) = \phi_{\text{Neu}}(\bar{Z}, V)\overline{W}^{2N_e}$, where \overline{W} depends only on $\{\bar{Z}, V\}$. This is a very general result, since we have not assumed that the trait is additive: there can be arbitrary epistasis for the trait as well as for fitness (Barton, 1989).

It is difficult to find how the joint distribution of allele frequencies, or of trait means and variances, changes through time, even numerically. An intriguing approximation derives from the observation that the stationary distribution maximises entropy, subject to constraints on the expected value of certain observables (Prügel-Bennett and Shapiro, 1994; Barton and Vladar, 2009). For example, directional selection on a quantitative trait can be represented by maximising entropy, but constraining the expected trait mean, $\mathbb{E}[\bar{Z}]$, leading to a distribution proportional to $e^{2N_e\beta\bar{Z}}$, where $\beta = \partial_{\bar{Z}} \log(\overline{W})$. One can now approximate the dynamics of the expected trait mean by assuming that the distribution of allele frequencies takes the stationary form that yields the current $\mathbb{E}[\bar{Z}]$. This corresponds to a ‘separation of time-scales’: the mean trait value evolves much more slowly than the allele frequencies. This reduces the problem from following the dynamics of the full allele frequency distribution, $\psi(p)$, to the dynamics of a single quantity, $\mathbb{E}[\bar{Z}]$, and yet gives accurate approximations (Barton and Vladar, 2009).

4.3.2 Beyond Linkage Equilibrium

When frequencies of alleles at different loci cannot be treated as independent random variables, the loci are said to be in linkage *disequilibrium* (see Chapter 2). In this regime, population genetics becomes very much harder. Even in the simplest case of n bi-allelic loci in a haploid population, one must follow 2^n haplotype frequencies. Moreover, because there are now so many possible genotypes, most will be rare or absent, even in a very large population. This makes it hard to avoid stochastic effects, whether in analysis or simulation. We first explain how (discrete) genotype frequencies can be represented in a general way, and then outline some of the applications of this fully multi-locus theory to evolutionary questions. We then describe some approximations that allow progress, even with very many loci, and finally, discuss approaches that treat the genome as continuous rather than discrete.

4.3.2.1 Representing Genotype Frequencies

We focus on the simplest case, with two alleles per locus, and a haploid population. However, all the methods we describe extend to multiple alleles, or even a continuum of alleles, in a fairly obvious way. In principle, we could represent genotypes as bit strings, and track their individual frequencies, for example $(x_{000}, x_{001}, \dots, x_{111})$ for three loci; indeed, early work on multi-locus theory took this route (for a review, see Christiansen, 1999). However, this is hard to extend to large numbers of loci, and does not give much insight; more specifically, it does not distinguish the roles of the various evolutionary processes. Typically, and especially close to linkage equilibrium, it is better to represent the population by its allele frequencies, and by measures of association among sets of alleles.

Even for two haploid loci, each with two alleles, there are several ways to represent their association. The most important, introduced by Lewontin and Kojima (1960), is the difference between the frequency of a haplotype, and the product of the component allele frequencies ($D = x_{11} - p_1 p_2$). If we label alleles as $X = 0$ or 1 , then this is just the covariance between the allelic states ($D_{l,m} = \mathbb{E}[(X_l - p_l)(X_m - p_m)]$) for loci l, m). Other measures have been proposed – for example, the correlation $r = D/\sqrt{p_1 q_1 p_2 q_2}$ (where $q = 1 - p$) or the normalised $D' = D/D_{\max}$ (Lewontin, 1964) – with the aim of ‘correcting’ for differences in allele frequencies, by rescaling it to between -1 and 1 . However, the relation between D and allele frequency depends on the processes involved, and so such corrections cannot be general.

It is not obvious how best to generalise measures of linkage disequilibrium to more than two loci, or to describe diploid genotype frequencies. The pairwise measure, $D_{l,m}$, decays by a factor $1 - r_{l,m}$ as a result of recombination at rate $r_{l,m}$ between loci l and m . Bennett (1954), building on work of Geiringer (1944), showed how to transform from gamete frequencies to a system consisting of allele frequencies and ‘principal components’ that measure deviations from linkage equilibrium, in such a way that the disequilibria decouple from one another and decay geometrically in recombination rate. If more than three loci are involved, the multilinear transformation to these coordinates is a hierarchical system of equations that themselves depend on the recombination rates and must be solved in a recursive manner. Baake (2001) considered a continuous time model, assuming that crossovers occur singly, and showed that the corresponding recursions take a simpler form that is independent of recombination rates. More recently, Baake *et al.* (2016) extend this to multiple crossovers and Baake and Baake (2016) show that by taking a ‘backwards in time’ perspective, in which recombination becomes a stochastic fragmentation process (Hudson’s coalescent with recombination, but without coalescence as the population is infinite), one can find an explicit solution to the recombination dynamics in both discrete and continuous time.

However, although such measures of disequilibrium behave in a simple way under recombination, selection changes them in a more complicated way. The simplest generalisation is to label alleles as $X_l = 0$ or 1 , and express their state relative to an arbitrary reference point, which may conveniently be taken as equal to the allele frequency: $\zeta_l = X_l - p_l$. Then associations can be defined as moments of the X_l :

$$D_U = \mathbb{E} \left[\prod_{l \in U} \zeta_l \right], \quad (4.21)$$

where U is a set of loci. Now 2^n genotype frequencies are defined by the n allele frequencies, p_l , the $\binom{n}{2}$ pairwise associations, $D_{l,m}$, up to the single n -way association $D_{1,2,\dots,n}$. There are still 2^n equations, but we can associate meaning to these quantities. (With more than two alleles per locus, we label the alleles by multiple values, and define higher-order moments with repeated indices.) With this representation, the effect of recombination on the association between the set of loci U is a sum over all possible partitions of that set:

$$D'_U = \sum_{S \cup T = U} r_{S,T} D_S D_T, \quad (4.22)$$

where $r_{S,T}$ is the frequency of recombinations that partition U into S, T . The effect of selection is also simple, if fitness, W , is defined as a polynomial function of the ζ_l – which is always possible with discrete genotypes:

$$D'_U = D_U + \sum_S a_S (D_{U \setminus S} - D_U D_S), \quad \text{where } \frac{W}{\bar{W}} = 1 + \sum_S a_S \left(\left(\prod_{l \in S} \zeta_l \right) - D_S \right). \quad (4.23)$$

This is closely related to the selection gradient approach discussed above, for linkage equilibrium: the selection coefficients α_S can be written in terms of gradients of mean fitness with respect to allele frequencies and linkage disequilibria (Turelli and Barton, 1994). The approach generalises naturally to diploids, sex-linkage, and interactions between individuals (Kirkpatrick *et al.*, 2002); it lends itself to symbolic computation, so that exact analytical expressions can be calculated automatically. However, as we shall see, this approach is most useful when combined with approximations that apply when populations are close to linkage equilibrium, so that the D_{UU} are small.

Quantitative traits can be defined as polynomial functions of genotype, as is done for fitness in (4.23); the labels, X_l , attached to alleles are arbitrary, and not necessarily related to the effects of those alleles on quantitative traits (though it may be convenient to make them so). In the special case where the trait is additive (i.e. a linear function of genotype), the effects of selection may be best modelled using cumulants rather than moments (Bürger, 1991; Turelli and Barton, 1994); the effects of recombination on cumulants are more complicated, however.

4.3.2.2 Applications

Early work on multiple loci was stimulated by the discovery in the 1960s of abundant molecular variation, which suggested that selection might act on tightly linked loci, thus involving linkage disequilibria. In diploids, selection for heterozygotes maintains balanced polymorphism; with multiple loci in diploids, there may be multiple equilibria, and complex dynamics (Christiansen, 1999). In an infinite population, fluctuating selection on linked loci in haploids can also maintain diversity (Kirzhner *et al.*, 1994). It is believed that fluctuating selection on haploids cannot maintain polymorphism in the absence of linkage disequilibrium, and conversely, that constant selection on haploids cannot maintain polymorphism even with linkage disequilibrium; however, both these conjectures remain open (Novak and Barton, 2017). Analysis is difficult: although multi-locus selection increases mean fitness, recombination may not, and no quantity necessarily increases – indeed, stable limit cycles may occur (Hastings, 1981). However, the examples of Hastings are somewhat contrived; whereas chaotic dynamics may be common in ecological models, the constraints imposed by Mendelian genetics make complex dynamics less common in models of population genetics (at least when fitness depends only on genotype).

A substantial part of population genetics is devoted to inferring population structure from genetic data (see **Chapter 8**). Many methods rely on the linkage disequilibrium that is generated by mixing of populations. The widely used STRUCTURE software fits a model in which individuals derive from one of a number sources, each well mixed (Pritchard *et al.*, 2000); the signal comes from associations among loci in the population as a whole. When divergent populations meet, an equilibrium is quickly reached between gene flow and recombination, and the rate of gene flow can be estimated from the strength of linkage disequilibrium (Li and Nei, 1974; Szymura and Barton, 1986). Within a single population, the product of effective population size and recombination rate can be estimated from the strength of random linkage disequilibria generated by drift (Hill and Robertson, 1968); when applied to genome-wide data, this method can identify fine-scale recombination hotspots, an approach pioneered by McVean *et al.* (2004).

In abundant species, genetic diversity is much lower than expected under a naive neutral theory (Lewontin, 1974). ‘Lewontin’s paradox’ can be resolved by a combination of population structure and selection on linked loci, and disentangling these remains the central problem in understanding sequence diversity. ‘Background selection’ against rare deleterious alleles reduces diversity at linked neutral loci (Charlesworth *et al.*, 1993); the effects of multiple selected loci can be multiplied, giving a simple expression for the net effect of background selection across the genome, which depends primarily on the total mutation rate per map length (Hudson and Kaplan, 1995). Favourable mutations also reduce diversity: as they fix, they carry

with them a region of map length of order $1/T$, where T is the time taken to go from one copy to high frequency. Such ‘hard sweeps’ were first analysed by Maynard Smith and Haigh (1974) and have since been elaborated to allow for partial fixation, increase from standing variation or multiple mutations (‘soft sweeps’; Hermisson and Pennings, 2017), spatial and temporal structure (Barton, 2000), and multiple selected loci (Pavlidis *et al.*, 2012).

A key tool in analysing both the effects of linked selection, and population structure, on neutral diversity, is the structured coalescent: neutral genes trace back through the various locations or genetic backgrounds, moving between them by migration or by recombination. For example, under background selection, genes tend to trace back into fitter ancestral backgrounds, while in a hard sweep, genes trace back to coalesce in the small population spawned by the initial mutation, unless they escape by recombination. If selection is strong relative to drift ($N_e s \gg 1$), then the background frequencies can be treated deterministically, and analysis is straightforward. When $N_e s$ is of order 1, random fluctuations in background frequencies can be represented through a diffusion approximation, allowing the distortion due to linked selection to be calculated (Kaplan *et al.*, 1988; Barton and Etheridge, 2004).

4.3.2.3 Approximations

Even when random drift can be neglected, exact analysis, or even deterministic simulation, is impractical with more than a few loci: there are too many possible genotypes to enumerate directly. However, populations can often be accurately approximated by supposing that recombination is fast relative to the processes that generate associations among loci, so that populations approach QLE. This concept was introduced by Kimura (1965) and analysed by Nagylaki (1976). It is often surprisingly accurate, and can lead to simple results that can be expressed in terms of measurable quantities. For example, the barrier to flow of a neutral gene from one multi-locus background to another, across a one-dimensional hybrid zone, depends simply on the mean fitness of the hybrid population (Barton, 1986); the strength of selection for recombination depends on the effect of recombination on the mean and variance of fitness (Barton, 1995); and selection on female preference depends on the correlation between the preference and the genetic component of male fitness (Kirkpatrick and Barton, 1997).

Another approach is to study symmetrical models, and to assume that all loci are equivalent: more precisely, one assumes two alleles per locus, and that all genotypes with the same number of ‘+’ alleles are equally frequent. This requires that the loci be unlinked, since otherwise their position on the genetic map would break the symmetry. The model was introduced by Kondrashov (1984) in a study of sympatric speciation, and was used by Doebeli (1996) in a similar way; it is sometimes known as the ‘hypergeometric model’, since offspring values follow that distribution. A difficulty, however, is that the symmetric solution may be unstable: in particular, under stabilising selection one out of the many genotypes that are close to the optimal value tends to fix. Symmetric solutions tend to be stable under disruptive selection (as in models of sympatric speciation), but their stability needs to be checked (Barton and Shpak, 2000).

A more robust and general approximation is provided by the infinitesimal model (Barton *et al.*, 2017; Walsh and Lynch, 2018, Ch. 24). This was introduced (implicitly) by Fisher (1918) as the limit of a large number of loci with additive effects, and written down precisely by Bulmer (1980). It is the basis of practical animal breeding, and at the phenotypic level, simply states that offspring have genetic values normally distributed around the mean of their parents’ genetic values, and with a variance that depends only on the relatedness between the parents, not on the trait values. It is instructive to relate it to the symmetric model, which also approaches a normal distribution of offspring values when the number of loci is large, but with a variance that depends on the parental values; if either parent is near the edge of the range of possible genotypes, offspring will have low variance. Under the infinitesimal model, one must assume

that individuals are far from the extreme genotypes. Since we know that artificial selection can typically change populations by many standard deviations in either direction, this is often a reasonable assumption. Moreover, the infinitesimal model has much wider applicability than the symmetric model, and does not require that loci be equivalent.

4.3.2.4 Blocks of Genome

Even over long time-scales, genomes are passed on as long blocks: the rate of recombination between adjacent bases is $\sim 10^{-8}$ per generation in humans and *Drosophila*. Thus, it is natural to treat the genome as continuous, rather than as a string of discrete loci. Even before the material basis of heredity was discovered, Fisher (1949, 1954) introduced such an approach. He pointed out that the junctions between segments of genome derived from different ancestors, which are generated by crossing over in meiosis, behave like new mutations, and are propagated following Mendelian rules.

Fisher used this theory of junctions to investigate the variability in inbreeding. Although the average fraction of genome shared between relatives is determined by the pedigree, the actual fraction depends on the random process of meiosis (Figure 4.2a). This distribution soon becomes broad, since fluctuations accumulate in each generation. There is a 90% chance that a human ancestor six generations back will not pass on any part of a 1 morgan long chromosome, and a 91% chance that an ancestor 12 generations back will not pass on any part of their 34 M long autosomal genome. (For calculations in this vein, we refer to Speed and Balding, 2015.) Sequence data are commonly used in animal breeding not to identify and select particular genes, but rather to estimate the fraction of genome shared, and hence the breeding value, more accurately (Meuwissen and Goddard, 2010).

The same process can be viewed backwards in time, yielding the coalescent with recombination (Hudson, 1983, 1990). Tracing back, genomes of map length r may coalesce in a common ancestor, at rate $1/(2N_e)$, and may derive from two ancestral genomes via a recombination, at rate r (Figure 4.2b). Thus, the genome is divided into segments with different genealogies,

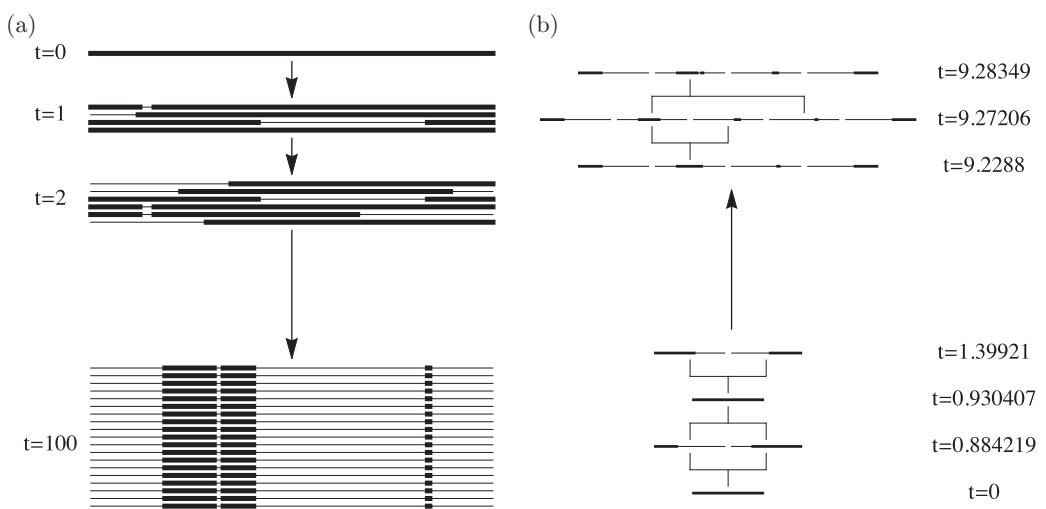


Figure 4.2 (a) Transmission of a single genome forwards in time. (b) Ancestry of a single genome, traced back through the coalescent with recombination. Note that although these reflect the same underlying process, they are quite different. Forwards in time, most of the genome will be lost, but small fragments may fix in the whole population. Backwards in time, the total ancestry must be conserved, and a stationary state will be reached in which this ancestry is scattered over multiple fragments in multiple individuals.

adjacent genealogies differing by a recombination event. This model is the limit of large population size, equivalent to the diffusion approximation, and depends only on the dimensionless parameter $R = 2N_e r$; in this limit, only coalescence events between two genomes, and single crossovers, need be considered. In the distant past, the ancestors of a single genome are scattered over a fluctuating number of ancestral genomes, with mean $\sim R$, and each carrying $\sim R$ ancestral fragments of length $\sim 1/2N_e$ (Wiuf and Hein, 1999). Although this process is easily defined, it is hard to derive explicit mathematical results, and it has only recently been found how to simulate it efficiently (Kelleher *et al.*, 2015). One difficulty is that, moving along the genome, the process is not Markovian: the genealogy at one point of the genome does not depend only on the genealogy immediately preceding. To see this, think of the sequence of coalescence times between two genomes: a single coalescence may bring together several ancestral fragments that are scattered along the genome, so that *precisely* the same coalescence time recurs at well-separated loci. Nevertheless, the sequentially Markov coalescent, which draws genealogies conditional on the immediately preceding genealogy, is a good approximation for many purposes; correspondingly, hidden Markov models are often used to analyse sequence data (e.g. Sheehan *et al.*, 2013). There is a refinement of the original sequentially Markov coalescent (McVean and Cardin, 2005), termed SMC' (Marjoram and Wall, 2006; Wilton *et al.*, 2015).

The population genetics of continuous genomes is relatively unexplored, and results are available only for the simplest cases. Perhaps the most straightforward case is where genomes introgress into another population. Stam and Zeven (1981) find the mean and variance of the amount of genome that introgresses when a single gene is selected in successive backcrosses, and Visscher *et al.* (1996) consider multiple selected loci, by simulation. Barton (1983) considers natural hybridisation, where selection is spread uniformly over the genome. Both for a single-island population and a one-dimensional hybrid zone there is a sharp transition point: if selection is stronger than recombination, then blocks of genome are selected as a whole, whereas if it is weaker, blocks are broken up, and selection becomes ineffective.

The survival of a single genome in a very large population can be approximated using a neutral branching process (Baird *et al.*, 2003). Although its loss is inevitable, some fraction is likely to persist for a very long time: whereas the probability that a single locus persists for t generations declines like $1/t$, the chance that some part of the genome persists declines like $1/\log t$; although any one fragment must eventually be lost, many smaller fragments are generated by recombination. Selection can be included if one assumes it to be either spread uniformly over the genome, or concentrated at discrete loci. If, instead, directional selection acts on a quantitative trait, with additive variance spread evenly over the genome (an extension of the infinitesimal model to include linkage), then a branching process is not quite exact, but gives a good approximation (Sachdeva and Barton, 2018). When a single genome enters a large and homogeneous population, fragments that contribute positively to the trait value may increase and ultimately fix, if their selective advantage outweighs their break-up by recombination.

There has been much recent interest in using shared sequence blocks to identify recent relationships. For example, Ralph and Coop (2013) used a sample of over 2000 Europeans to identify shared blocks of 1 cM or more. The frequency of such blocks declines with distance between individuals, and this rate of decline is steeper for longer (and hence, more recent) blocks. The time back to when two genes in a two-dimensional population coalesce in a common ancestor can be found by assuming that ancestral lineages diffuse (Wright, 1943); the length of the shared block then follows (approximately) an exponential distribution with rate equal to the coalescence time. This allows the rate of diffusion to be estimated, which is not possible from allele frequencies alone (Barton *et al.*, 2013a; Ringbauer *et al.*, 2017). Harris and Nielsen (2013) derive the distribution of shared block lengths, given gene flow into a population of varying size,

and at varying rates, and thus estimate European population history over the past few thousand generations.

4.4 Outlook

Over the past century, population genetics has developed a rich theory, which has both practical application, and inspires intrinsically valuable mathematics. The simple rules of Mendelian genetics lead to results with remarkably wide scope, applying to a wide range of organisms and time-scales. Nevertheless, many challenges remain: even though we are confident in our understanding of the basic evolutionary processes, their consequences cannot be understood by brute simulation. Thus, mathematical analysis will continue to play a central role in helping us understand genetics and evolution, and in making good use of the current abundance of sequence data.

References

- Baake, E. (2001). Mutation and recombination with tight linkage. *Journal of Mathematical Biology* **42**, 455–488.
- Baake, E. and Baake, M. (2016). Haldane linearisation done right: Solving the nonlinear recombination equation the easy way. *Discrete & Continuous Dynamical Systems A* **36**(12), 6645–6656.
- Baake, E., Baake, M. and Salamat, M. (2016). The general recombination equation in continuous time and its solution. *Discrete & Continuous Dynamical Systems A* **36**, 63–95.
- Baird, S.J.E., Barton, N.H. and Etheridge, A.M. (2003). The distribution of surviving blocks of ancestral genome. *Theoretical Population Biology* **64**, 451–471.
- Barton, N.H. (1983). Multilocus clines. *Evolution* **37**, 454–471.
- Barton, N.H. (1986). The effects of linkage and density-dependent regulation on gene flow. *Heredity* **57**, 415–426.
- Barton, N.H. (1989). The divergence of a polygenic system under stabilising selection, mutation and drift. *Genetical Research* **54**, 59–77.
- Barton, N.H. (1995). A general model for the evolution of recombination. *Genetical Research* **65**, 123–144.
- Barton, N.H. (2000). Genetic hitch-hiking. *Philosophical Transactions of the Royal Society of London B* **355**, 1553–1562.
- Barton, N.H. (2017). How does epistasis influence the response to selection? *Heredity* **118**, 96–109.
- Barton, N.H. and Etheridge, A.M. (2004). The effect of selection on genealogies. *Genetics* **166**, 1115–1131.
- Barton, N.H. and Shpak, M. (2000). The stability of symmetrical solutions to polygenic models. *Theoretical Population Biology* **57**, 249–264.
- Barton, N.H. and Turelli, M. (1987). Adaptive landscapes, genetic distance, and the evolution of quantitative characters. *Genetics Research* **49**, 157–174.
- Barton, N.H. and Vladar, H.P. (2009). Statistical mechanics and the evolution of polygenic traits. *Genetics* **181**, 997–1011.
- Barton, N.H., Depaulis, F. and Etheridge, A.M. (2002). Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61**, 31–48.
- Barton, N.H., Etheridge, A.M. and Véber, A. (2010). A new model for evolution in a spatial continuum. *Electronic Journal of Probability* **15**, 162–216.

- Barton, N.H., Etheridge, A.M., Kelleher, J. and Véber, A. (2013a). Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology*.
- Barton, N.H., Etheridge, A.M. and Véber, A. (2013b). Modelling evolution in a spatial continuum. *Journal of Statistical Mechanics: Theory and Experiment* 2013(01), P01002.
- Barton, N.H., Etheridge, A.M. and Véber, A. (2017). The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology* **118**, 50–73.
- Bennett, J.H. (1954). On the theory of random mating. *Annals of Eugenics* **18**, 311–317.
- Berestyski, N. (2009). Recent progress in coalescent theory. *Ensaios Matemáticos*, 16.
- Berestyski, N., Etheridge, A.M. and Hutzenthaler, M. (2009). Survival, extinction and ergodicity in a spatially continuous population model. *Markov Processes and Related Fields* **15**, 265–288.
- Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.
- Bürger, R. (1991). Moments, cumulants, and polygenic dynamics. *Journal of Mathematical Biology* **30**, 199–213.
- Bürger, R. (2000). *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester.
- Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: A new approach I. Haploid models. *Advances in Applied Probability* **6**, 260–290.
- Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* **10**, 195–205.
- Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.
- Christiansen, F.B. (1999). *Population Genetics of Multiple Loci*. Wiley, Chichester.
- Doebeli, M. (1996). A quantitative genetic competition model for sympatric speciation. *Journal of Evolutionary Biology* **9**, 893–910.
- Durrett, R. (1996). *Stochastic Calculus. A Practical Introduction*. CRC Press, Boca Raton, FL.
- Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*, 2nd edition. Springer-Verlag, New York.
- Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172**, 2621–2633.
- Etheridge, A.M. (2008). Drift, draft and structure: Some mathematical models of evolution. *Banach Center Publications* **80**, 121–144.
- Etheridge, A.M. (2011). *Some Mathematical Models from Population Genetics: Ecole d'Eté de Probabilités de Saint Flour XXXIX-2009*, Lecture Notes in Mathematics, 2012. Springer-Verlag, Heidelberg.
- Feller, W. (1951). Diffusion processes in genetics. In J. Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematics, Statistics and Probability*. University of California Press, Berkeley, pp. 227–246.
- Felsenstein, J. (1975). A pain in the torus: Some difficulties with the model of isolation by distance. *American Naturalist* **109**, 359–368.
- Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Proceedings of the Royal Society of Edinburgh* **52**, 399–433.
- Fisher, R.A. (1949). *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.
- Fisher, R.A. (1954). A fuller theory of ‘junctions’ in inbreeding. *Heredity* **8**, 187–198.
- Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics* **15**, 25–57.
- Gerrish, P.J. and Sniegowski, P.D. (2012). Real time forecasting of near-future evolution. *Journal of the Royal Society, Interface* **9**, 2268–2278.
- Haldane, J.B.S. (1932). *The Causes of Evolution*. Longman, Green and Co., New York.

- Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics* **9**(6), e1003521.
- Hastings, A. (1981). Stable cycling in discrete-time genetic models. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 7224–7225.
- Hermission, J. and Pennings, P.S. (2017). Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution* **8**, 700–716.
- Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.
- Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.
- Hudson, R.R. (1990). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44. Oxford University Press, Oxford.
- Hudson, R.R. and Kaplan, N.L. (1995). Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.
- Kaplan, N.L., Darden, T. and Hudson, R.R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.
- Karlin, S. and Taylor, H.M. (1981). *A Second course in Stochastic Processes*. Academic Press, New York.
- Kelleher, J., Etheridge, A.M. and McVean, G. (2015). Efficient simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* **12**, e1004842.
- Kimura, M. (1952). Stepping stone model of population. *Annual Report of the National Institute of Genetics, Japan* **3**, 62–63.
- Kimura, M. (1956). *Stochastic processes in population genetics*. PhD thesis, University of Wisconsin, Madison.
- Kimura, M. (1965). A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proceedings of the National Academy of Sciences of the United States of America* **54**, 731–736.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.
- Kimura, M. and Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.
- Kimura, M. and Weiss, G.H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.
- King, J.L. and Jukes, T.H. (1969). Non-Darwinian evolution: Random fixation of selectively neutral mutations. *Science* **164**, 788–798.
- Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**, 235–248.
- Kirkpatrick, M. and Barton, N.H. (1997). The strength of indirect selection on female mating preferences. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 1282–1286.
- Kirkpatrick, M., Johnson, T. and Barton, N.H. (2002). General models of multilocus evolution. *Genetics* **161**, 1727D1750.
- Kirzhner, V.M., Korol, A.B., Ronin, Y.I. and Nevo, E. (1994). Cyclical behavior of genotype frequencies in a two-locus population under fluctuating haploid selection. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11432–11436.
- Kondrashov, A.S. (1984). On the intensity of selection for reproductive isolation at the beginnings of sympatric speciation. *Genetika* **20**, 408–415.
- Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**, 314–334.

- Lande, R. (1979). Effective deme sizes during long-term evolution estimated from rates of chromosomal rearrangement. *Evolution* **33**, 234–251.
- Lewontin, R.C. (1964). The interaction of selection and linkage I. General considerations; heterotic models. *Genetics* **49**, 49–67.
- Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.
- Lewontin, R.C. and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 450–472.
- Li, W.H. and Nei, M. (1974). Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology* **6**, 173–183.
- Malécot, G. (1948). *Les Mathématiques de l'Hérédité*. Masson, Paris.
- Marjoram, P. and Wall, J.D. (2006). Fast ‘coalescent’ simulation. *BMC Genetics* **7**, 16.
- Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable allele. *Genetics Research* **23**, 23–35.
- McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B* **360**, 1387–1393.
- McVean, G.A.T., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R. and Donnelly, P.J. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.
- Meuwissen, T. and Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome sequencing. *Genetics* **185**, 623–631.
- Möhle, M. (2000). Total variation distances and rates of convergence for ancestral coalescent processes in exchangeable population models. *Advances in Applied Probability* **32**, 983–993.
- Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable models. *Annals of Probability* **29**, 1547–1562.
- Moran, P.A.P. (1958). Random processes in genetics. *Proceedings of the Cambridge Philosophical Society* **54**, 60–71.
- Nagylaki, T. (1976). Evolution of one and two locus systems. *Genetics* **83**, 583–600.
- Nagylaki, T. (1990). Models and approximations for random genetic drift. *Theoretical Population Biology* **37**, 192–212.
- Neuhauser, C. and Krone, S.M. (1997). Genealogies of samples in models with selection. *Genetics* **145**, 519–534.
- Norman, M.F. (1975). Approximations of stochastic processes by Gaussian diffusions, and applications to Wright-Fisher genetic models. *SIAM Journal on Applied Mathematics* **29**, 225–242.
- Novak, S. and Barton, N.H. (2017). When does frequency independent selection maintain genetic variation? *Genetics* **207**, 653–668.
- Orr, H.A. (2000). Adaptation and the cost of complexity. *Evolution* **54**, 13–20.
- Pavlidis, P., Metzler, D. and Stephan, W. (2012). Selective sweeps in multilocus models of quantitative traits. *Genetics* **192**, 225–239.
- Pritchard, J.K., Stephens, M. and Donnelly, P.J. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.
- Prügel-Bennett, A. and Shapiro, J.L. (1994). An analysis of genetic algorithms using statistical mechanics. *Physics Review Letters* **72**, 1305–1309.
- Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biology* **11**, e1001555.
- Ringbauer, H., Coop, G. and Barton, N.H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics* **205**, 1335–1351.
- Robertson, A. (1970). The reduction in fitness from genetic drift at heterotic loci in small populations. *Genetical Research* **15**, 257–259.

- Sachdeva, H. and Barton, N.H. (1303). Introgression of a block of genome under infinitesimal selection. *Genetics* **209**, 1279–1303.
- Sella, G. and Hirsh, A.E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14475–14475.
- Sheehan, S., Harris, K. and Song, Y. (2013). Estimating variable effective population size from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662.
- Speed, D. and Balding, D.J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Review Genetics* **16**, 33–44.
- Stam, P. and Zeven, A.C. (1981). The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* **30**, 227–238.
- Sudbury, A. (1977). Clumping effects in models of isolation by distance. *Journal of Applied Probability* **14**(4), 391–395.
- Szymura, J.M. and Barton, N.H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland. *Evolution* **40**, 1141–1159.
- Turelli, M. and Barton, N.H. (1994). Statistical analyses of strong selection on polygenic traits: What, me normal? *Genetics* **138**(3), 913–941.
- Visscher, P.M., Haley, C.S. and Thompson, R. (1996). Marker-assisted introgression in backcross breeding programs. *Genetics* **144**, 1921–1930.
- Walsh, J.B. and Lynch, M. (2018). *Evolution and Selection of Quantitative Traits*. Oxford University Press, Oxford.
- Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.
- Wilton, P.R., Carmi, S. and Hobolth, A. (2015). The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics* **200**, 343–355.
- Wiuf, C. and Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.
- Wright, S. (1937). The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the United States of America* **23**, 307–320.
- Wright, S. (1941). On the probability of fixation of reciprocal translocations. *American Naturalist* **75**, 513–522.
- Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.
- Wright, S. (1949). Adaptation and selection. In G.L., Jepson, G.G., Simpson and E., Mayr (eds.), *Genetics, Paleontology and Evolution*. Princeton University Press, Princeton, NJ, pp. 365–389.

5

Coalescent Theory

Magnus Nordborg

Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, Vienna, Austria

Abstract

Whereas most of classical population genetics considers the future of a population given a starting point, the coalescent considers the present, while taking the past into account. The pattern of polymorphism, that is, the allelic states of all homologous gene copies in a population, is determined by the genealogical and mutational history of these copies. The coalescent is based on the realization that the genealogy is usually easier to model backwards in time, and that selectively neutral mutations can then be superimposed afterwards. This leads to extremely efficient algorithms for simulating data under a wide variety of models — data that can then be compared with actual observations in order to understand the process that gave rise to the latter.

5.1 Introduction

This is an update of a chapter I wrote 18 years ago (Nordborg, 2001). At that time, there was still no comprehensive textbook treatment of the coalescent, and researchers were forced to rely on reviews (Hudson, 1990, 1993; Donnelly and Tavaré, 1995), unpublished lecture notes, or the primary literature. My chapter was intended to fill this gap. Since then, the coalescent has been covered in several textbooks (Hein *et al.*, 2005; Wakeley, 2009; Charlesworth and Charlesworth, 2010), and the justification for including the present chapter is essentially completeness: it still serves as a concise and informal (but rigorous) introduction to the stochastic process known as ‘the coalescent’. Fortunately for me, basic theory does not change or go out of fashion, and so the changes to the original text are fairly minor. I have added references to some new developments, but, given the existence of textbooks devoted to the subject, a comprehensive review of all theoretical work on the coalescent since 2000 seemed pointless (see Wakeley, 2013, for an overview).

However, while the basic theory has not changed, the applications certainly have. When this chapter was originally written, human geneticists had only recently rediscovered linkage disequilibrium, haplotype blocks had yet to be invented, and single nucleotide polymorphisms were not a household term. Today, modern sequencing technologies are increasingly making it possible to generate genome-wide polymorphism data in large samples for almost any organism. One consequence of this is that it is possible to overcome the huge evolutionary variance at any given locus (which is what the coalescent models) by considering large numbers of loci across the genome. The coalescent provides an extremely powerful tool for exploratory data analysis, both qualitatively, by building intuition, and quantitatively, by simply fitting simulated

data to observations (formalized as approximate Bayesian computation; see **Chapter 1**). Furthermore, while exact likelihood inference based on the coalescent model (**Chapter 2**) has generally proven infeasible for genome-wide data, several powerful algorithms based on various approximations to the full coalescent process have become widely used (as described in several chapters, e.g. **Chapters 2, 3, 8 and 14**). These developments will be alluded to below.

5.2 The Coalescent

The word ‘coalescent’ is used in several ways in the literature, and it will also be used in several ways here. Hopefully, the meaning will be clear from the context. The coalescent, or perhaps more appropriately, the coalescent approach, is based on two fundamental insights, which are the topic of Section 5.2.1. In Section 5.2.2, I then describe the stochastic process known as the coalescent, or sometimes Kingman’s coalescent (Kingman, 1982a,b,c). This process results from combining the two fundamental insights with a convenient limit approximation.

The coalescent will be introduced in the setting of the Wright–Fisher model of neutral evolution, but it applies more generally. This is one of the main topics for the remainder of the chapter. Many different neutral models can be shown to converge to Kingman’s coalescent, and more complex neutral models often converge to coalescent processes analogous to Kingman’s coalescent.

The coalescent was described by Kingman (1982a,b,c), but it was also discovered independently by Hudson (1983) and by Tajima (1983). Indeed, arguments anticipating it had been used several times in population genetics (reviewed by Tavaré, 1984).

5.2.1 The Fundamental Insights

The first insight is that since selectively neutral variants by definition do not affect reproductive success, it is possible to separate the neutral mutation process from the genealogical process. In classical terms, ‘state’ can be separated from ‘descent’.

To see how this works, consider a population of N clonal organisms that reproduce according to the neutral Wright–Fisher model, that is, generations are discrete, and each new generation is formed by randomly sampling N parents with replacement from the current generation. The number of offspring contributed by a particular individual is thus binomially distributed with parameters N (the number of trials) and $1/N$ (the probability of being chosen), and the joint distribution of the numbers of offspring produced by all N individuals is symmetrically multinomial. Now consider the random genealogical relationships (i.e. ‘who begat whom’) that result from reproduction in this setting. These can be represented graphically, as shown in Figure 5.1. Going forward in time, lineages branch whenever an individual produces two or more offspring, and end when there is no offspring. Going backward in time, lineages coalesce whenever two or more individuals were produced by the same parent. They never end. If we trace the ancestry of a group of individuals back through time, the number of distinct lineages will decrease and eventually reach one, when the most recent common ancestor (MRCA) of the individuals in question is encountered. None of this is affected by neutral genetic differences between the individuals.

As a consequence, the evolutionary dynamics of neutral allelic variants can be modeled by superimposing mutations: given a realization of the genealogical process, allelic states are assigned to the original generation in a suitable manner, and the lines of descent then simply followed forward in time, using the rule that offspring inherit the allelic state of their parent unless there is a mutation (which occurs with some probability each generation). In particular, the allelic states of any group of individuals (e.g. all the members of a given generation) can be

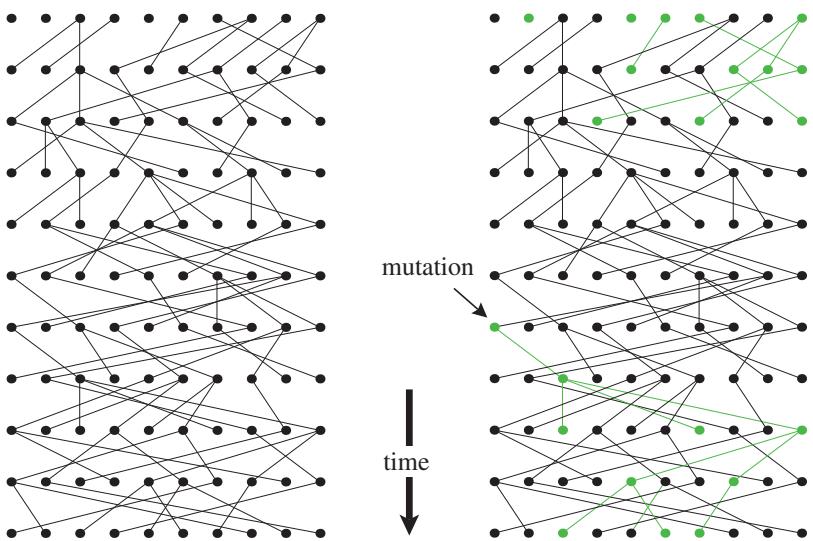


Figure 5.1 The neutral mutation process can be separated from the genealogical process. The genealogical relationships in a particular 10-generation realization of the neutral Wright–Fisher model (with population size $N = 10$) are shown on the left. On the right, allelic states have been superimposed.

generated by assigning an allelic state to their MRCA and then ‘dropping’ mutations along the branches of the genealogical tree that leads to them. Most of the genealogical history of the population is then irrelevant (cf. Figures 5.1 and 5.2).

The second insight is that it is possible to model the genealogy of a group of individuals backward in time without worrying about the rest of the population. It is a general consequence of the assumption of selective neutrality that each individual in a generation can be viewed as ‘picking’ its parent at random from the previous generation. It follows that the genealogy of a

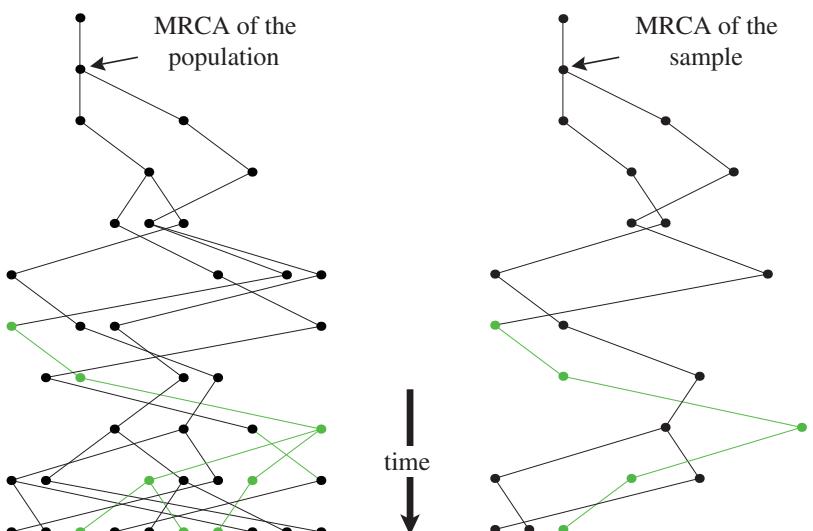


Figure 5.2 The genetic composition of a group of individuals is completely determined by the group’s genealogy and the mutations that occur on it. The genealogy of the final generation in Figure 5.1 is shown on the left, and the genealogy of a sample ($n = 3$) from this generation is shown on the right. In this case, the MRCA of the population and the sample are the same (the probability of this happening is discussed below). These trees could have been generated backward in time without generating the rest of Figure 5.1.

group of individuals may be generated by simply tracing the lineages back in time, generation by generation, keeping track of coalescences between lineages, until eventually the MRCA is found. It is particularly easy to see how this is done for the Wright–Fisher model, where individuals pick their parents independently of each other.

In summary, the joint effects of random reproduction (which causes ‘genetic drift’) and random neutral mutations in determining the genetic composition of a sample (or entire population) of clonal individuals may be modeled by first generating the random genealogy of the individuals backward in time, and then superimposing mutations forward in time. This approach leads directly to extremely efficient computer algorithms (cf. the ‘classical’ approach which is to simulate the *entire*, usually very large, population forward in time for a long period of time, and then to look at the final generation). It is also mathematically elegant, as the next subsection will show. However, its greatest value may be heuristic: the realization that the pattern of neutral variation observed in a population can be viewed as the result of random mutations on a random tree is a powerful one, that profoundly affects the way we think about data.

In particular, we are usually interested in biological phenomena that affect the genealogical process, but do not affect the mutation process (e.g. population subdivision). From the point of view of inference about such phenomena, the observed polymorphisms are only of interest because they contain information about the unobserved underlying genealogy. Furthermore, the underlying genealogy is only of interest because it contains information about the evolutionary process that gave rise to it.

It is crucial to understand this, because no matter how many individuals we sample, there is still only a *single* underlying genealogy to estimate. It could of course be that this single genealogy contains a lot of information about the interesting aspect of the evolutionary process, but if it does not, then our inferences will be as good as one would normally expect from a sample of size 1.

Another consequence of the above is that it is usually possible to understand how model parameters affect polymorphism data by understanding how they affect genealogies. For this reason, I will focus on the genealogical process and only discuss the neutral mutation process briefly toward the end of the chapter.

5.2.2 The Coalescent Approximation

The previous subsection described the conceptual insights behind the coalescent approach. The sample genealogies central to this approach can be conveniently modeled using a continuous-time Markov process known as the coalescent (or Kingman’s coalescent, or sometimes the ‘ n -coalescent’ to emphasize the dependence on the sample size). We will now describe the coalescent and show how it arises naturally as a large-population approximation to the Wright–Fisher model. Its relationship to other models will be discussed later.

Figure 5.2 is needlessly complicated because the identity (i.e. the horizontal position) of all ancestors is maintained. In order to superimpose mutations, all we need to know is which lineage coalesces with which, and when. In other words, we need to know the topology and the branch lengths. The topology is easy to model: because of neutrality, individuals are equally likely to reproduce; therefore all lineages must be equally likely to coalesce. It is convenient to represent the topology as a sequence of coalescing equivalence classes: two members of the original sample are equivalent at a certain point in time if and only if they have a common ancestor at that time (see Figure 5.3). But what about the branch lengths, that is, the coalescence times?

Follow two lineages back in time. We have seen that offspring pick their parents randomly from the previous generation, and that, under the Wright–Fisher model, they do so

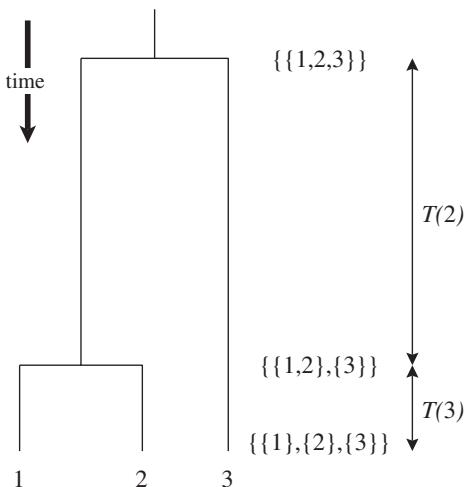


Figure 5.3 The genealogy of a sample can be described in terms of its topology and branch lengths. The topology can be represented using equivalence classes for ancestors. The branch lengths are given by the waiting times between successive coalescence events.

independently of each other. Thus, the probability that the two lineages pick the same parent and coalesce is $1/N$, and the probability that they pick different parents and remain distinct is $1 - 1/N$. Since generations are independent, the probability that they remain distinct more than t generations into the past is $(1 - 1/N)^t$. The expected coalescence time is N generations. This suggests a standard continuous-time diffusion approximation, which is good as long as N is reasonably large (**Chapter 4**). Rescale time so that one unit of scaled time corresponds to N generations. Then the probability that the two lineages remain distinct for more than τ units of scaled time is

$$\left(1 - \frac{1}{N}\right)^{\lfloor N\tau \rfloor} \rightarrow e^{-\tau}, \quad (5.1)$$

as N goes to infinity ($\lfloor N\tau \rfloor$ is the largest integer less than or equal to $N\tau$). Thus, in the limit, the coalescence time for a pair of lineages is exponentially distributed with mean 1.

Now consider k lineages. The probability that none of them coalesce in the previous generation is

$$\prod_{i=0}^{k-1} \frac{N-i}{N} = \prod_{i=1}^{k-1} \left(1 - \frac{i}{N}\right) = 1 - \frac{\binom{k}{2}}{N} + O\left(\frac{1}{N^2}\right), \quad (5.2)$$

and the probability that more than two do so is $O(1/N^2)$. Let $T(k)$ be the (scaled) time till the first coalescence event, given that there are currently k lineages. By the same argument as above, $T(k)$ is in the limit exponentially distributed with mean $2/[(k(k-1))]$. Furthermore, the probability that more than two lineages coalesce in the same generation can be neglected. Thus, under the coalescent approximation, the number of distinct lineages in the ancestry of a sample of (finite) size n decreases in steps of one back in time, so $T(k)$ is the time from k to $k-1$ lineages (see Figure 5.3).

In summary, the coalescent models the genealogy of a sample of n haploid individuals as a random bifurcating tree, where the $n-1$ coalescence times $T(n), T(n-1), \dots, T(2)$ are mutually independent, exponentially distributed random variables. Each pair of lineages coalesces independently at rate 1, so the total rate of coalescence when there are k lineages is ' k choose 2'. A concise (and rather abstract) way of describing the coalescent is as a continuous-time

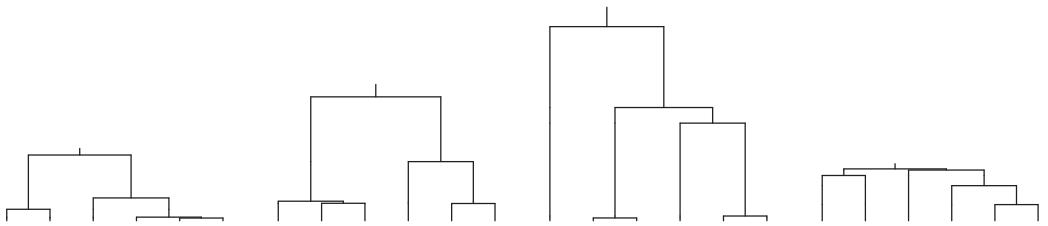


Figure 5.4 Four realizations of the coalescent for $n = 6$, drawn on the same vertical scale (the labels 1–6 should be assigned randomly to the tips).

Markov process with state space \mathcal{E}_n given by the set of all equivalence relations on $\{1, \dots, n\}$, and infinitesimal generator $Q = (q_{\xi\eta})_{\xi, \eta \in \mathcal{E}_n}$ given by

$$q_{\xi\eta} := \begin{cases} -k(k-1)/2 & \text{if } \xi = \eta, \\ 1 & \text{if } \xi < \eta, \\ 0 & \text{otherwise,} \end{cases} \quad (5.3)$$

where $k := |\xi|$ is the number of equivalence classes in ξ , and $\xi < \eta$ if and only if η is obtained from ξ by coalescing two equivalence classes of ξ .

It is worth emphasizing just how efficient the coalescent is as a simulation tool. In order to generate a sample genealogy under the Wright–Fisher model as described in the previous subsection, we would have to go back in time on the order of N generations, checking for coalescences in each of them. Under the coalescent approximation, we simply generate $n - 1$ independent exponential random numbers and, independently of these, a random bifurcating topology.

What do typical coalescence trees look like? Figure 5.4 shows four examples. It is clear that the trees are extremely variable, both with respect to topology and branch lengths. This should come as no surprise considering the description of the coalescent just given: the topology is independent of the branch lengths; the branch lengths are independent, exponential random variables; and the topology is generated by randomly picking lineages to coalesce (in this sense all topologies are equally likely).

Note that the trees tend to be dominated by the deep branches, when there are few ancestors left. Because lineages coalesce at rate proportional to k^2 , coalescence events occur much more rapidly when there are many lineages (intuitively speaking, it is easier for lineages to find each other then). Indeed, the expected time to the MRCA (the height of the tree) is

$$E \left[\sum_{k=2}^n T(k) \right] = \sum_{k=2}^n E[T(k)] = \sum_{k=2}^n \frac{2}{k(k-1)} = 2 \left(1 - \frac{1}{n} \right), \quad (5.4)$$

while $E[T(2)] = 1$, so the expected time during which there are only two branches is greater than half the expected total tree height. Furthermore, the variability in $T(2)$ accounts for most of the variability in tree height. The dependence on the deep branches becomes increasingly apparent as n increases, as can be seen by comparing Figures 5.4 and 5.5.

The importance of realizing that there is only a single underlying genealogy was emphasized above. As a consequence of the single genealogy, sampled gene copies from a population must almost always be treated as dependent, and increasing the sample size is therefore often surprisingly ineffective (the point is well made by Donnelly, 1996). Important examples of this follow directly from the basic properties of the coalescent. Consider first the MRCA of the population. One might think that a large sample is needed to ensure that the deepest split is included, but it can be shown (this and related results can be found in Saunders *et al.*, 1984) that the

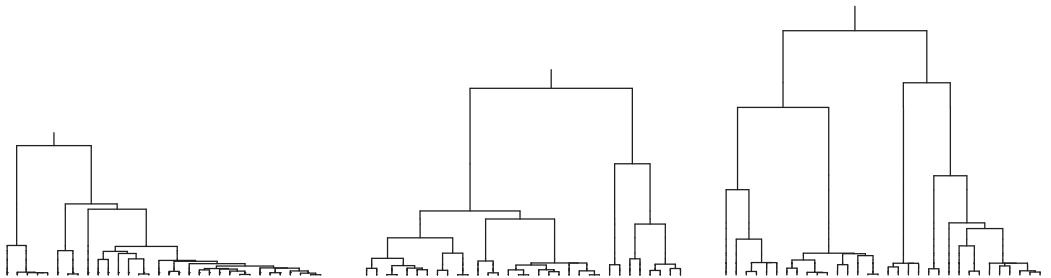


Figure 5.5 Three realizations of the coalescent for $n = 32$, drawn on the same vertical scale (the labels 1–32 should be assigned randomly to the tips).

probability that a sample of size n has the same MRCA as the whole population is $(n - 1)/(n + 1)$. Thus even a small sample is likely to contain it and the total tree height will quickly stop growing as n increases. Second, the number of distinct lineages decreases rapidly as we go back in time. This severely limits inferences about ancient demography (e.g. Nordborg, 1998). Third, since increasing the sample size only adds short twigs to the tree (cf. Figure 5.5), the expected total branch length of the tree, $T_{\text{tot}}(n)$ grows very slowly with n . We have

$$E[T_{\text{tot}}(n)] = E \left[\sum_{k=2}^n kT(k) \right] = \sum_{k=1}^{n-1} \frac{2}{k} \rightarrow 2(\gamma + \log n), \quad (5.5)$$

as $n \rightarrow \infty$ ($\gamma \approx 0.577216$ is Euler's constant). Since the number of mutations that are expected to occur in a tree is proportional to $E[T_{\text{tot}}(n)]$, this has important consequences for estimating the mutation rate, as well as for inferences that depend on estimates of the mutation rate. Loosely speaking, it turns out that a sample of n copies of a gene often has the statistical properties one would expect of an independent sample of size $\log n$, or even of size 1 (which is not much worse than $\log n$ in practice).

5.3 Generalizing the Coalescent

This section will present ideas and concepts that are important for generalizing the coalescent. The following sections will then illustrate how these can be used to incorporate greater biological realism.

5.3.1 Robustness and Scaling

We have seen that the coalescent arises naturally as an approximation to the Wright–Fisher model, and that it has convenient mathematical properties. However, the real importance of the coalescent stems from the fact that it arises as a limiting process for a wide range of neutral models, *provided time is scaled appropriately* (Kingman, 1982b,c; Möhle, 1998b, 1999). It is thus robust in this sense.

This is best explained through an example. Recall that the number offspring contributed by each individual in the Wright–Fisher model is binomially distributed with parameters N and $1/N$. The mean is thus 1, and the variance is $1 - 1/N \rightarrow 1$, as $N \rightarrow \infty$. Now consider a generalized version of this model in which the mean number of offspring is still 1 (as it must be for the population size to remain constant), but the limiting variance is σ^2 , $0 < \sigma^2 < \infty$ (perhaps giants step on 90% of the individuals before they reach reproductive age). It can be shown that this process also converges to the coalescent, provided time is measured in units of N/σ^2 .

generations. We could also measure time in units of N generations as before, but then $E[T(2)] = 1/\sigma^2$ instead of $E[T(2)] = 1$, and so on. Either way, the expected coalescence time for a pair of lineages is N/σ^2 generations. In other words, increased variance in reproductive success causes coalescence to occur faster (at a higher rate). In classical terms, ‘genetic drift’ operates faster. By changing the way we measure time, this can be taken into account, and the standard coalescent process obtained.

The remarkable fact is that a very wide range of biological phenomena (overlapping generations, separate sexes, mating systems – several examples will be given below) can likewise be treated as a simple linear change in the time-scale of the coalescent. This has important implications for data analysis. The good news is that we may often be able to justify using the coalescent process even though ‘our’ species almost certainly does not reproduce according to a Wright–Fisher model (few species do). The bad news is that biological phenomena that can be modeled this way will never be amenable to inference based on polymorphism data alone. For example, σ^2 in the model above could never be estimated from polymorphism data unless we had independent information about N (and vice versa).

Of course, we could not even estimate N/σ^2 without external data. It is important to realize that all parameters in coalescent models are scaled, and that only scaled parameters can be directly estimated from the data. In order to make any kind of statement about unscaled quantities, such as population numbers, or ages in years or generations, external information is needed. This adds considerable uncertainty to the analysis. For example, an often used source of external information is an estimate of the neutral mutation probability per generation. This estimate is often obtained by measuring sequence divergence between species, and dividing by the estimated species divergence time (e.g. Li, 1997). The latter is in turn obtained from the fossil record and a rough guess of the generation length. It should be clear that it is not appropriate to treat such an estimate as a known parameter when analyzing polymorphism data. However, it should be noted that interesting conclusions can often be drawn directly from scaled parameters (e.g. by looking at relative values). Such analyses are likely to be more robust, given the robustness of the coalescent.

Because the generalized model above converges with the same scaling as a Wright–Fisher model with a population size of N/σ^2 , it is sometimes said that it has an ‘effective population size’ $N_e = N/\sigma^2$. Models that scale differently would then have other effective population sizes. Although convenient, this terminology is unfortunate for at least two reasons. First, the classical population genetics literature is full of variously defined ‘effective population sizes’, only some of which are effective population sizes in the sense used here (reviewed in Sjödin *et al.*, 2005). For example, populations that are subdivided or vary in size cannot in general be modeled as a linear change in the time-scale of the coalescent. Even variance in reproductive success cannot be modeled in this way if it is too high – as it may be for organisms with extremely high fecundity (coalescent models with multiple simultaneous coalescences have been developed; see Wakeley, 2013). Second, the term is inevitably associated with real population sizes, even though it is simply a scaling factor. To be sure, N_e is always a function of the real demographic parameters, but there is no direct relationship with the total population size (which may be smaller as well as much, much larger). Indeed, as we shall see in Section 5.7, N_e arguably must vary between chromosomal regions in the same organism!

5.3.2 Variable Population Size

Real populations vary in size over time. Although the coalescent is not robust to variation in the population size in the sense described above (i.e. there is no ‘effective population size’), it is nonetheless easy to incorporate changes in the population size, at least if we are willing to assume that we know what they were – that is, if we assume that the variation can be treated

deterministically. Intuitively, this works as follows (for a rigorous treatment, see Donnelly and Tavaré, 1995).

Imagine a population that evolves according to the Wright–Fisher model, but with a different population size in each generation. If we know how the size has changed over time, we can trace the genealogy of a sample precisely as before. Let $N(t)$ be the population size t generations ago. Going back in time, lineages are more likely to coalesce in generations when the population is small than in generations when the population is large. In order to describe the genealogy by a continuous-time process analogous to the coalescent, we must therefore allow the rate of coalescence to change over time. However, since the time-scale used in the coalescent directly reflects the rate of coalescence, we may instead let this scaling change over time. In the standard coalescent, t generations ago corresponds to t/N units of coalescence time, and τ units of coalescence time ago corresponds to $\lfloor N\tau \rfloor$ generations. When the population size is changing, we find instead that t generations ago corresponds to

$$g(t) := \sum_{i=1}^t \frac{1}{N(i)} \quad (5.6)$$

units of coalescence time, and τ units of coalescence time ago corresponds to $\lfloor g^{-1}(\tau) \rfloor$ generations (g^{-1} denotes the inverse function of g). It is clear from equation (5.6) that many generations go by without much coalescence time passing when the population size is large, and conversely, that much coalescence time passes each generation the population is small. Let $N(0)$ go to infinity, and assume that $N(t)/N(0)$ converges to a finite number for each t , to ensure that the population size remains large in every generation. It can be shown that the variable population size model converges to a coalescent process with a *nonlinear* time-scale in this limit (Griffiths and Tavaré, 1994). The scaling is given by equation (5.6). Thus, a sample genealogy from the coalescent with variable population size can be generated by simply applying g^{-1} to the coalescence times of a genealogy generated under the standard coalescent.

An example will make this clearer. Consider a population that has grown exponentially, so that, backwards in time, it shrinks according to $N(t) = N(0)e^{-\beta t}$ (note that this violates the assumption that the population size be large in every generation – this turns out not to matter greatly). Then

$$g(t) \approx \int_0^t \frac{1}{N(s)} ds = \frac{e^{\beta t} - 1}{N(0)\beta} \quad (5.7)$$

and

$$g^{-1}(\tau) \approx \frac{\log(1 + N(0)\beta\tau)}{\beta}. \quad (5.8)$$

The difference between this model and one with a constant population size is shown in Figure 5.6. When the population size is constant, there is a linear relationship between real and scaled time. The genealogical trees will tend to look like those in Figures 5.4 and 5.5. When the population size is changing, the relationship between real and scaled time is nonlinear, because coalescences occur very slowly when the population was large, and more rapidly when the population was small. Genealogies in an exponentially growing population will tend to have most coalescences early in the history. Since all branches will then be of roughly equal length, the genealogy is said to be ‘star-like’.

5.3.3 Population Structure on Different Time-Scales

Real populations are also often spatially structured, and it is obviously important to be able to incorporate this in our models. However, structured models turn out to be even more important than one might have expected from this, because many biological phenomena can be thought of

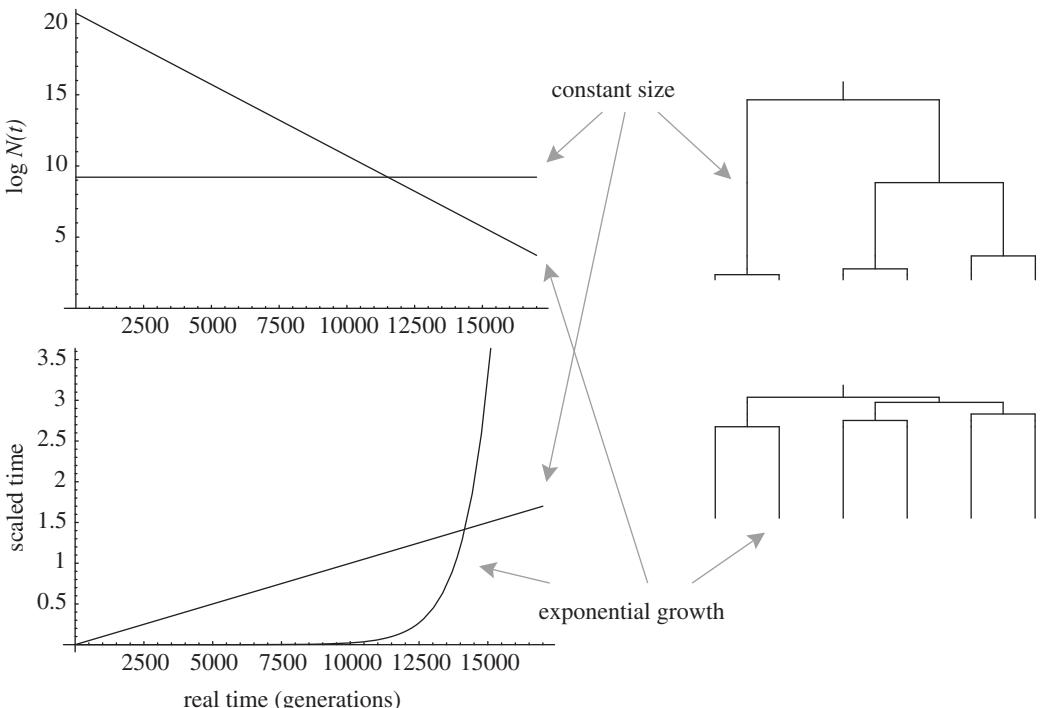


Figure 5.6 Variable population size can be modeled as a standard coalescent with a nonlinear time-scale. Here, a constant population is compared to one that has grown exponentially. As the latter population shrinks backward in time, the scaled time begins to run faster, reflecting the fact that coalescences are more likely to have taken place when the population was small. Note that the trees are topologically equivalent and differ only in the branch lengths.

as analogous to population structure (Nordborg, 1997; Rousset, 1999b). Examples range from the obvious, like age structure, to the more abstract, like diploidy and allelic classes.

The following model, which may be called the ‘structured Wright–Fisher model’, turns out to be very useful in this context. Consider a clonal population of size N , as before, but let it be subdivided into patches of fixed sizes N_i , $i \in \{1, \dots, M\}$, so that $\sum_i N_i = N$. In every generation, each individual produces an effectively infinite number of propagules. These propagules then migrate among the patches independently of each other, so that with probability m_{ij} , $i, j \in \{1, \dots, M\}$, a propagule produced in patch i ends up in patch j . We also define the ‘backward migration’ probability, b_{ij} , $i, j \in \{1, \dots, M\}$, that a randomly chosen propagule in patch i after dispersal was produced in patch j ; it is easy to show that

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}. \quad (5.9)$$

The next generation of adults in each patch is then formed by random sampling from the available propagules.

Thus the number of offspring a particular individual in patch i contributes to the next generation in patch j is binomially distributed with parameters N_j and $b_{ji}N_i^{-1}$. The joint distribution of the numbers of offspring contributed to the next generation in patch j by all individuals in the current generation is multinomial (but no longer symmetric).

Just like the unstructured Wright–Fisher model, the genealogy of a finite sample in this model can be described by a discrete-time Markov process. Lineages coalesce in the previous generation if and only if they pick the same parental patch, and the same parental individual within

that patch. A lineage currently in i and a lineage currently in j ‘migrate’ (backward in time) to k and coalesce there with probability $b_{ik} b_{jk} N_k^{-1}$.

It is also possible to approximate the model by a continuous-time Markov process. The general idea is to let the total population size, N , go to infinity with time scaled appropriately, precisely as before. However, we now also need to decide how M , N_i , and b_{ij} scale with N . Different biological scenarios lead to very different choices in this respect, and it is often possible to utilize convergence results based on separation of time-scales (Möhle, 1998a; Nordborg, 1997, 1999; Nordborg and Donnelly, 1997; Wakeley, 1999). This technique will be illustrated in what follows.

5.4 Geographical Structure

Genealogical models of population structure have a long history. The classical work on identity coefficients concerns genealogies when $n = 2$, and the coalescent was used for this purpose from the outset (e.g. Slatkin, 1987; Strobeck, 1987; Tajima, 1989a; Takahata, 1988). Most coalescent modeling of geographic structure has utilized some version of the basic ‘matrix migration’ model described in Section 5.3.3. Whether this is appropriate is debatable – most populations should arguably be modeled as continuously distributed in space (e.g. Barton *et al.*, 2013) – but the simplicity and flexibility of the matrix migration model makes it appealing. An important variant of the model considers isolation: gene flow that changes over time, for example due to speciation (Wakeley, 2009; Chapter 7).

Here we will mainly use the model to introduce some of the scaling ideas that are central to the coalescent. For time-scale approximations different from the ones discussed below, see Takahata (1991) and Wakeley (1999).

5.4.1 The Structured Coalescent

Assume that M , $c_i := N_i/N$, and $B_{ij} := 2Nb_{ij}$, $i \neq j$, all remain constant as N goes to infinity. Then, with time measured in units of N generations, the process converges to the so-called ‘structured coalescent’, in which each pair of lineages in patch i coalesces independently at rate $1/c_i$, and each lineage in i ‘migrates’ (backward in time) independently to j at rate $B_{ij}/2$ (Herbots, 1994; Notohara, 1990; Wilkinson-Herbots, 1998). The intuition behind this is as follows. By assuming that B_{ij} remains constant, we assure that the backward per-generation probabilities of leaving a patch (b_{ij} , $i \neq j$), are $O(1/N)$. Similarly, by assuming that c_i remains constant, we assure that all per-generation coalescence probabilities are $O(1/N)$. Thus, in any given generation, the probability that all lineages remain in their patch, without coalescing, is $1 - O(1/N)$. Furthermore, the probabilities that more than two lineages coalesce, that more than one lineage migrates, and that lineages both migrate and coalesce, are all $O(1/N^2)$ or smaller. In the limit $N \rightarrow \infty$, the only possible events are pairwise coalescences within patches, and single migrations between patches.

These events occur according to independent Poisson processes, which means the following. Let k_i denote the number of lineages currently in patch i . Then the waiting time till the first event is exponentially distributed with rate given by the sum of the rates of all possible events, that is,

$$h(k_1, \dots, k_M) = \sum_i \left(\frac{\binom{k_i}{2}}{c_i} + \sum_{j \neq i} k_i \frac{B_{ij}}{2} \right). \quad (5.10)$$

When an event occurs, it is a coalescence in patch i with probability

$$\frac{\binom{k_i}{2}/c_i}{h(k_1, \dots, k_M)}, \quad (5.11)$$

and a migration from i to j with probability

$$\frac{k_i B_{ij}/2}{h(k_1, \dots, k_M)}. \quad (5.12)$$

In the former case, a random pair of lineages in i coalesce, and k_i decreases by one. In the latter case, a random lineage moves from i to j , k_i decreases by one, and k_j increases by one. A simulation algorithm would stop when the MRCA is found, but note that this single remaining lineage would continue migrating between patches if followed further back in time.

Structured coalescent trees generally look different from standard coalescent trees. Whereas variable population size only altered the branch lengths of the trees, population structure also affects the topology. If migration rates are low, lineages sampled from the same patch will tend to coalesce with each other, and a substantial amount of time can then pass before migration allows the ancestral lineages to coalesce (see Figure 5.7). Structure will often increase the mean and, equally importantly, the variance in time to the MRCA considerably (discussed in the context of human evolution by Marjoram and Donnelly, 1997).

5.4.2 The Strong-Migration Limit

It is intuitive that weak migration, which corresponds to strong population subdivision, can have a large effect on genealogies. Conversely, we would expect genealogies in models with strong migration to look much like standard coalescent trees. This intuition turns out to be correct, except for one thing: the scaling changes. Strong migration is thus one of the phenomena that can be modeled as a simple linear change in the time-scale of the coalescent. It is important to understand why this happens.

Formally, the strong-migration limit means that $\lim_{N \rightarrow \infty} N b_{ij} = \infty$ because the per-generation migration probabilities, b_{ij} , are not $O(1/N)$. Since the coalescence probabilities are $O(1/N)$, this means that, for large N , migration will be much more likely than coalescence. As $N \rightarrow \infty$, there will in effect be infinitely many migration events between coalescence events.

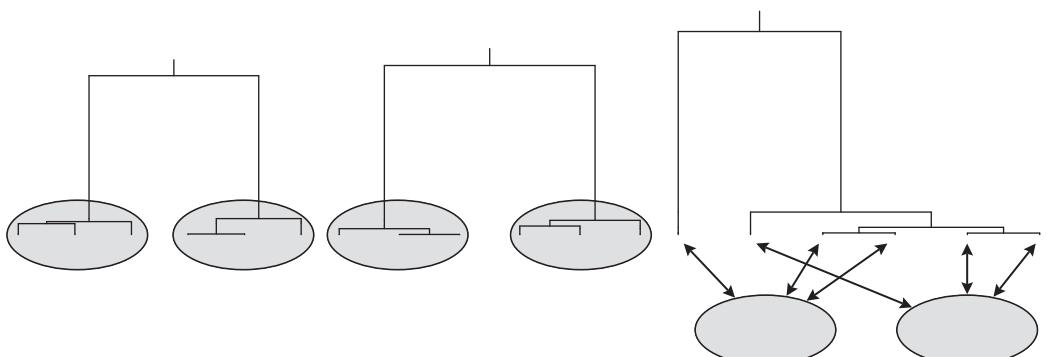


Figure 5.7 Three realizations of the structured coalescent in a symmetric model with two patches, and $n = 3$ in each patch (labels should be assigned randomly within patches). Lineages tend to coalesce within patches – but not always, as shown by the rightmost tree.

This is known as separation of timescales: migration occurs on a faster time-scale than does coalescence. However, coalescences can of course still only occur when two lineages pick a parent in the same patch. How often does this happen? Because lineages jump between patches infinitely fast on the coalescence time-scale, this is determined by the stationary distribution of the migration process (strictly speaking, this assumes that the migration matrix is ergodic). Let π_i be the stationary probability that a lineage is in patch i . A given pair of lineages then co-occur in i a fraction π_i^2 of the time. Coalescence in this patch occurs at rate $1/c_i$. Thus the total rate at which pairs of lineages coalesce is $\alpha := \sum_i \pi_i^2/c_i$. Pairs coalesce independently of each other just like in the standard model, so the total rate when there are k lineages is $\binom{k}{2}\alpha$. If time is measured in units of $N_e = N/\alpha$ generations the standard coalescent is retrieved (Nagylaki, 1980; Notohara, 1993).

It can be shown that $\alpha \geq 1$, with equality if and only if $\sum_{j \neq i} N_i b_{ij} = \sum_{j \neq i} N_j b_{ji}$ for all i . This condition means that, going forward in time, the number of emigrants equals the number of immigrants in all populations, a condition known as ‘conservative migration’ (Nagylaki, 1980). Thus we see that, unless migration is conservative, the effective population size with strong migration is smaller than the total population size. The intuitive reason for this is that when migration is non-conservative, some individuals occupy ‘better’ patches than others, and this increases the variance in reproductive success among individuals. The environment has ‘sources’ and ‘sinks’ (Pulliam, 1988; Rousset, 1999a). Conservative migration models (like Wright’s island model) have many simple properties that do not hold generally (Nagylaki, 1982, 1998; Nordborg, 1997; Rousset, 1999b).

5.5 Diploidy and Segregation

Because everything so far has been done in an asexual setting, it has not been necessary to distinguish between the genealogy of an organism and that of its genome. This becomes necessary in sexual organisms. Most obviously, a diploid organism which was produced sexually has two parents, and each chromosome came from one of them. The genealogy of the genes is thus different from the genealogy (the pedigree) of the individuals: the latter describes the *possible* routes the genes could have taken (and is typically irrelevant; cf. Figure 5.9, below). This is simply Mendelian segregation viewed backwards in time, and it is the topic of this section. It is usually said that diploidy can be taken into account by simply changing the scaling from N to $2N$; it will become clear why, and in what sense, this is true.

The other facet of sexual reproduction, genetic recombination, turns out to have much more important effects. Genetic recombination causes ancestral lineages to branch, so that the genealogy of a sample can no longer be represented by a single tree: instead it becomes a collection of trees, or a single, more general type of graph. Recombination will be ignored until Section 5.6.1.

Sex takes many forms. I will first consider organisms that are hermaphroditic and therefore potentially capable of fertilizing themselves (this includes most higher plants and many mollusks), and thereafter discuss organisms with separate sexes (which includes most animals and many plants).

5.5.1 Hermaphrodites

The key to modeling diploid populations is the realization that a diploid population of size N can be thought of as a haploid population of size $2N$, divided into N patches of size 2. In the

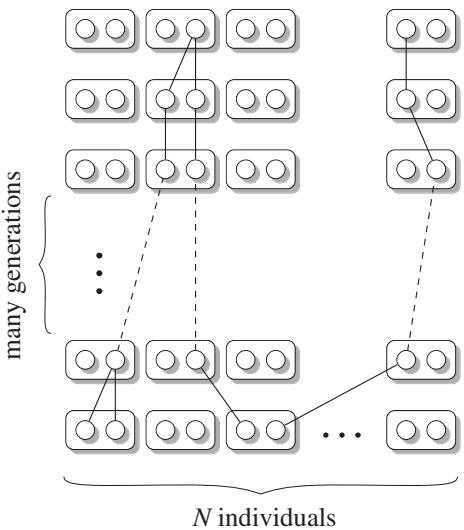


Figure 5.8 The coalescent with selfing. On the coalescent time-scale, lineages within individuals instantaneously coalesce (probability F), or end up in different individuals (probability $1 - F$).

notation of the structured Wright–Fisher model above, $M = N$, $N_i = 2$, and $c_i = 2/N$. Thus, in contrast to the assumptions for the structured coalescent, both M and c_i depend on N . This leads to a convenient convergence result based on separation of time-scales (Nordborg and Donnelly, 1997), that can be described as follows (cf. Figure 5.8).

If time is scaled in units of $2N$ generations, then each pair of lineages ‘coalesces’ into the same individual at rate 2. Whenever this happens, there are two possibilities: either the two lineages pick the same of the two available (haploid) parents, or they pick different ones. The former event, which occurs with probability $1/2$, results in a real coalescence, whereas the latter event, which also occurs with probability $1/2$, simply results in the two distinct lineages temporarily occupying the same individual. Let S be the probability that a fertilization occurs through selfing, and $1 - S$ the probability that it occurs through outcrossing. If the individual harboring two distinct lineages was produced through selfing (probability S), then the two lineages must have come from the same individual in the previous generation, and again pick different parents with probability $1/2$ or coalesce with probability $1/2$. If the individual was produced through outcrossing, the two lineages revert to occupying distinct individuals. Thus the two lineages will rapidly either coalesce or end up in different individuals. The probability of the former outcome is

$$\frac{S/2}{S/2 + 1 - S} = \frac{S}{2 - S} =: F \quad (5.13)$$

and that of the latter $1 - F$. Thus each time a pair of lineages coalesces into the same individual, the total probability that this results in a coalescence event is $1/2 \times 1 + 1/2 \times F = (1 + F)/2$, and since pairs of lineages coalesce into the same individual at rate 2, the rate of coalescence is $1 + F$. On the chosen time-scale, all states that involve two or more pairs occupying the same individual are instantaneous.

Thus, the genealogy of a random sample of gene copies from a population of hermaphrodites can be described by the standard coalescent if time is scaled in units of

$$2N_e = \frac{2N}{1 + F} \quad (5.14)$$

generations (cf. Pollak, 1987). If individuals are obligate outcrossers, $F = 0$, and the correct scaling is $2N$.

Importantly, a sample from a diploid population is not a random sample of gene copies, because both copies in each individual are sampled. This is easily taken into account. It follows from the above that the two copies sampled from the same individual will instantaneously coalesce with probability F , and end up in different individuals with probability $1 - F$. The number of distinct lineages in a sample of $2n$ gene copies from n individuals is thus $2n - X$, where X is a binomially distributed random variable with parameters n and F . This corresponds to the well-known increase in the frequency of homozygous individuals predicted by classical population genetics. This initial ‘instantaneous’ process has much nicer statistical properties than the coalescent, and most of the information about the degree of selfing comes from the distribution of variability within and between individuals (Nordborg and Donnelly, 1997).

5.5.2 Males and Females

Next consider a diploid population that consists of N_m breeding males and N_f breeding females so that $N = N_m + N_f$. The discussion will be limited to autosomal genes, that is, genes that are not sex-linked. With respect to the genealogy of such genes, the total population can be thought of as a haploid population of size $2N$, divided into two patches of size $2N_m$ and $2N_f$, respectively, each of which is further divided into patches of size 2, as in the previous section. Regardless of whether a lineage currently resides in a male or a female, it came from a male or female in the previous generation with equal probability $1/2$. Within a sex, all individuals are equally likely to be chosen. The model looks like a structured Wright–Fisher model with $M = 2$, $c_m = N_m/N$, $c_f = N_f/N$, and $b_{mf} = b_{fm} = 1/2$, the only difference being that two distinct lineages in the same individual must have come from individuals of different sexes in the previous generation, and thus do not migrate independently of each other. However, because states involving two distinct lineages in the same individual are instantaneous, this difference can be shown to be irrelevant. Pairs of lineages in different individuals (regardless of sex) coalesce in the previous generation if and only if both members of the pair came from: (a) the same sex; (b) the same diploid individual within that sex; and (c) the same haploid parent within that individual. This occurs with probability

$$\frac{1}{4} \cdot \frac{1}{N_m} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{N_f} \cdot \frac{1}{2} = \frac{N_m + N_f}{8N_m N_f}, \quad (5.15)$$

or, in the limit $N \rightarrow \infty$, with time measured in units of $2N$ generations, and c_m and c_f held constant, at rate $\alpha = (4c_m c_f)^{-1}$ (in accordance with the strong-migration limit result above). Alternatively, if time is measured in units of

$$2N_e = 2N/\alpha = \frac{8N_m N_f}{N_m + N_f} \quad (5.16)$$

generations, the standard coalescent is obtained (cf. Wright, 1931). Note that if $N_m = N_f = N/2$, the correct scaling is again the standard one of $2N$.

5.6 Recombination

Modeling recombination is essential when analyzing genomic polymorphism data. Viewed backward in time, recombination (in the broad sense that includes phenomena such as gene

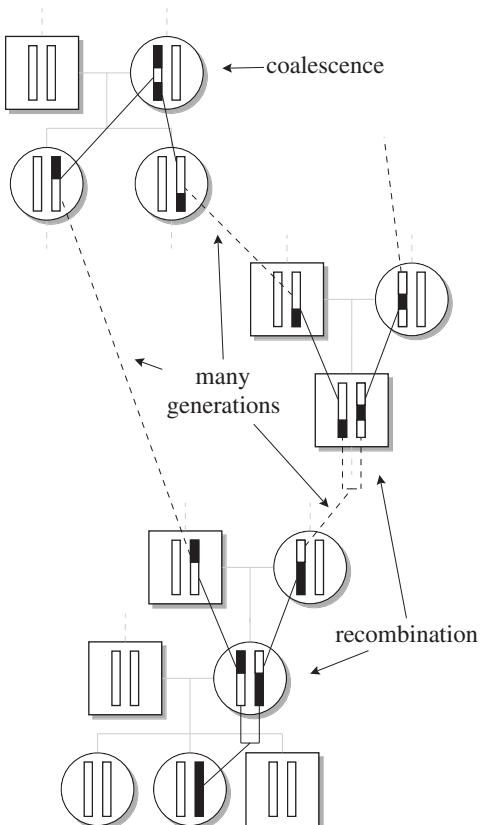


Figure 5.9 The genealogy of a DNA segment (colored black) subject to recombination both branches and coalesces. Note also that the genealogy of the sexually produced *individuals* (the pedigree) is very different from the genealogy of their *genes*.

conversion and bacterial conjugation in addition to crossing over) causes the ancestry of a chromosome to spread out over many chromosomes in many individuals. The lineages branch, as illustrated in Figure 5.9. The genealogy of a sample of recombining DNA sequences can thus no longer be represented by a single tree: it becomes a graph instead. Alternatively, since the genealogy of each point in the genome (each base pair, say) *can* be represented by a tree, the genealogy of a sample of sequences may be envisioned as a ‘walk through tree space’.

5.6.1 The Ancestral Recombination Graph

As was first shown by Hudson (1983), incorporating recombination into the coalescent framework is in principle straightforward. The following description is based on the elegant ‘ancestral recombination graph’ of Griffiths and Marjoram (1996, 1997), which is closely related to Hudson’s original formulation (for different approaches, see Simonsen and Churchill, 1997; Wiuf and Hein, 1999a).

Consider first the ancestry of a single ($n = 1$) chromosomal segment from a diploid species with two sexes and an even sex ratio. As shown in Figure 5.9, each recombination event (depicted here as crossing over at a point – we will return to whether this is reasonable below) in its ancestry means that a lineage splits into two, going backward in time. Recombination spreads the ancestry of the segment over many chromosomes, or rather over many ‘chromosomal lineages’. However, as also shown in Figure 5.9, these lineages will coalesce in the normal fashion,

and this will tend to bring the ancestral material back together on the same chromosome (Wiuf and Hein, 1997).

To model this, let the per-generation probability of recombination in the segment be r , define $\rho := \lim_{N \rightarrow \infty} 4Nr$, and measure time in units of $2N$ generations. Then the (scaled) time till the first recombination event is exponentially distributed with rate $\rho/2$ in the limit as N goes to infinity. Furthermore, once recombination has created two or more lineages, we find that these lineages undergo recombination independently of one another, and that simultaneous events can be neglected. This follows from standard coalescent arguments analogous to those presented for migration above. The only thing that may be slightly non-intuitive about recombination is that *the lineages we follow never recombine with each other* (the probability of such an event is by assumption vanishingly small): they always recombine with the (infinitely many) non-ancestral chromosomes.

Each recombination event increases the number of lineages by one, and because lineages recombine independently, the total rate of recombination when there are k lineages is $k\rho/2$. Each coalescence event decreases the number of lineages by one, and the total rate of coalescence when there are k lineages is $k(k - 1)/2$, as we have seen previously. Since lineages are ‘born’ at a linear rate, and ‘die’ at a quadratic rate, the number of lineages is guaranteed to stay finite and will reach one occasionally (at which point there will temporarily be a single ancestral chromosome again; see Wiuf and Hein, 1997).

A sample of n lineages behaves in the same way. Each lineage recombines independently at rate $\rho/2$, and each pair of lineages coalesces independently at rate 1. The number of lineages *will* hit one, occasionally. The segment in which this first occurs is known as the ‘Ultimate’ MRCA, because, as we shall see, each point in the sample may well have a younger MRCA.

The genealogy of a sample of n lineages back to the Ultimate MRCA can thus be described by a branching and coalescing graph (an ‘ancestral recombination graph’) that is analogous to the standard coalescent. A realization for $n = 6$ is shown in Figure 5.10.

What does a lineage in the graph look like? For each point in the segment under study, it must contain information about *which* (if any) sample members it is ancestral to. It is convenient to represent the segment as a $(0, 1)$ interval (this is just a coordinate system that can be translated into base pairs or whatever is appropriate). An ancestral lineage can then be represented as a set of elements of the form $\{\text{interval}, \text{labels}\}$, where the intervals are those resulting from all recombinational breakpoints in the history of the sample (Fisher’s ‘junctions’ for aficionados of classical population genetics; see Fisher, 1965) and the labels denote the descendants of that segment (using the ‘equivalence class’ notation introduced previously). An example of this notation is given in Figure 5.10. Note that pieces of a given chromosomal lineage will often be ancestral to no one in the sample. Indeed, recombination in a non-ancestral piece may result in an entirely non-ancestral lineage!

So far nothing has been said about where or how recombination breakpoints occur. This has been intentional, to emphasize that the ancestral recombination graph does not depend on (most) details of recombination. It is possible to model almost any kind of recombination (including bacterial transformation and various forms of gene conversion; see: Hudson, 1994; Bahlo, 1998; Wiuf, 2000; Wiuf and Hein, 2000, for example) in this framework. But of course the graph has no meaning unless we interpret the recombination events somehow. To proceed, we will assume that each recombination event results in crossing over at a point, x , somewhere in $(0, 1)$, and ignore the issue that recombination is mechanistically tied to gene conversion (Andolfatto and Nordborg, 1998; Nordborg, 2000). How x is chosen is again up to the modeler: it could be a fixed point; it could be a uniform random variable; or it could be drawn from some other distribution (perhaps centered around hot-spots of recombination). In any case, a breakpoint needs to be generated for each recombination event in the graph. We also need to

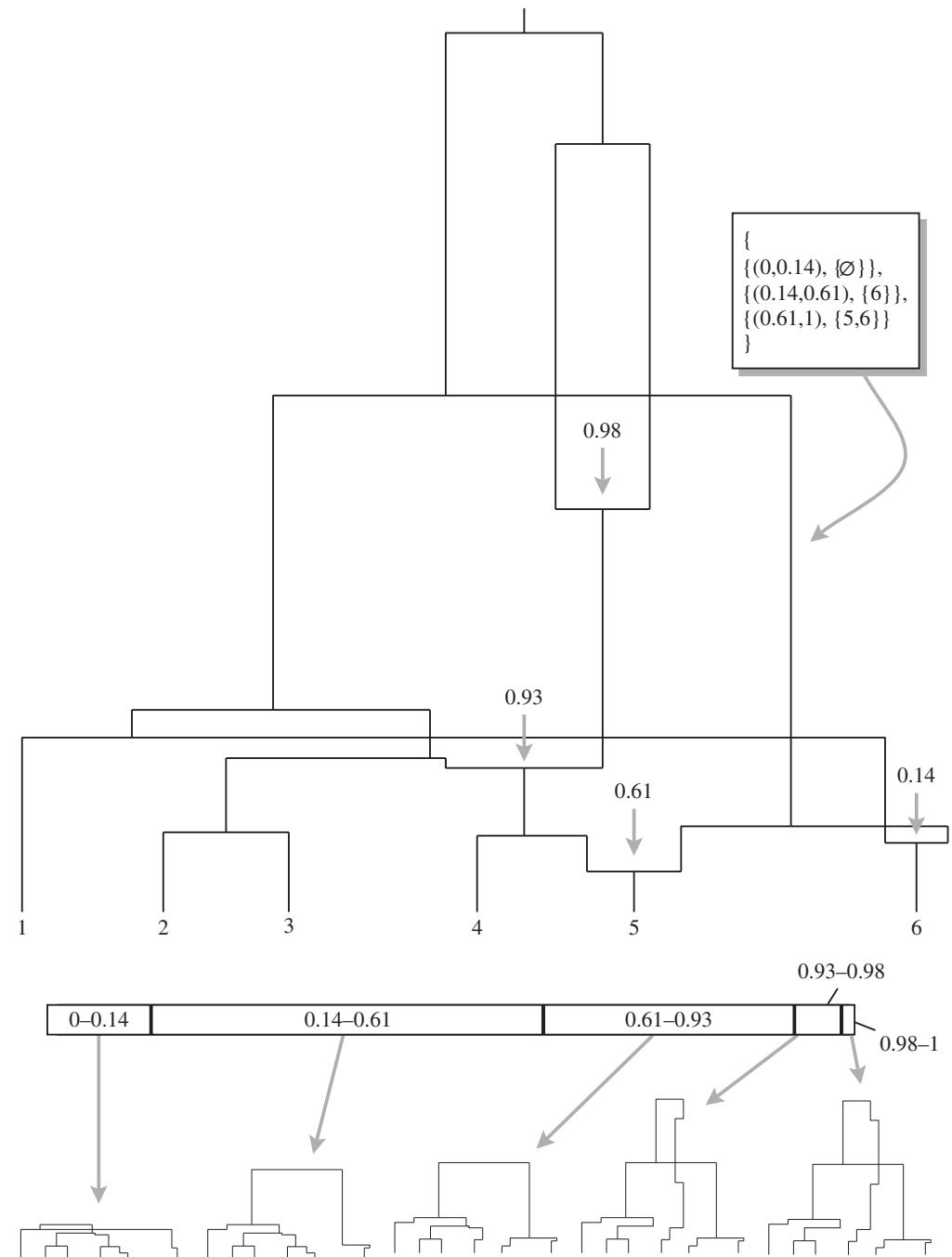


Figure 5.10 A realization of an ancestral recombination graph for $n = 6$. There were four recombination events, which implies $6 + 4 - 1 = 9$ coalescence events. Each recombination was assumed to lead to crossing over at a point, which was chosen randomly in $(0, 1)$. Four breakpoints (or ‘junctions’) implies five embedded trees, which are shown underneath. The tree for a particular chromosomal point is extracted from the graph by choosing the appropriate path at each recombination event. I have followed the convention that one should ‘go left’ if the point is located ‘to the left’ of (is less than) the breakpoint. Note that the two rightmost trees are identical. The box illustrates notation that may be used to represent ancestral lineages in the graph. The lineage pointed to is ancestral to: no (sampled) segment for the interval $(0,0.14)$; segment 6 for the interval $(0.14,0.61)$; and segments 5 and 6 for the interval $(0.61,1)$.

know which branch in the graph carries which recombination ‘product’ (remembering that we are going backward in time). With breaks affecting a point, a suitable rule is that the left branch carries the material to the ‘left’ of the breakpoint (i.e. in $(0, x)$), and the right branch carries the material to the ‘right’ (i.e. in $(x, 1)$).

Once recombination breakpoints have been added to the graph, it becomes possible to extract the genealogy for any given point by simply following the appropriate branches. Figure 5.10 illustrates how this is done. An ancestral recombination graph contains a number of embedded genealogical trees, each of which can be described by the standard coalescent, but which are obviously not independent of each other. An alternative way of viewing this process is thus as a ‘walk through tree space’ along the chromosome (Wiuf and Hein, 1999a), but it is important to note that this process is not Markovian, that is, each tree does not only depend on its immediate neighbor. The strength of the correlation between the genealogies for linked points depends on the scaled genetic distance between them, and goes to zero as this distance goes to infinity. The number of embedded trees equals the number of breakpoints plus one, but many of these trees may (usually will) be identical (cf. the two rightmost trees in Figure 5.10). Note also that the embedded trees vary greatly in height. This means that some pieces will have found their MRCA long before others. Indeed, it is quite possible for every piece to have found its MRCA long before the Ultimate MRCA.

Although many analytical results concerning the number of recombination events and the properties of the embedded trees are available (see, for example, Griffiths and Marjoram, 1996, 1997; Hudson, 1983, 1987; Hudson and Kaplan, 1985; Kaplan and Hudson, 1985; Pluzhnikov and Donnelly, 1996; Simonsen and Churchill, 1997; Wiuf and Hein, 1999a,b; Hudson, 2001), the ancestral recombination graph is an extremely complicated stochastic process. Furthermore, it becomes computationally unwieldy for all but very small recombination rates because the graphs tend to grow very large. However, it turns out that many events have negligible effects on data, intuitively because they mainly concern non-ancestral lineages. Based on this, an efficient approximation to the ancestral recombination graph has been developed that essentially turns the process into a Markovian ‘walk through tree space’ along the chromosome (McVean and Cardin, 2005; Marjoram and Wall, 2006). This approximation underlies the popular PSMC/MSMC algorithm for demographic inference, for example (Li and Durbin, 2011; Chapter 8).

5.6.2 Properties and Effects of Recombination

The chromosomal autocorrelation in genealogies manifests itself as autocorrelation in allelic state, also known as linkage disequilibrium (Nordborg and Tavaré, 2002; McVean, 2002), a crucial aspect of genomic data (Chapter 2).

It is important to recognize that recombination is typically very common. It simply cannot be ignored when analyzing sequence data (unless from non-recombining DNA, like mitochondrial DNA, but then we are looking at a single locus ...). Consider a pair of segments. The probability that they coalesce before either recombines is

$$\frac{1}{1 + 2 \cdot \rho/2} = \frac{1}{1 + \rho} \quad (5.17)$$

(cf. equation (5.11)). In order for recombination not to matter, we would need to have $\rho \approx 0$. It is the *scaled* recombination rate that matters, not the per-generation recombination probability. Estimates based on comparing genetic and physical maps indicate that the average per-generation per-nucleotide probability of recombination is typically as high as the average per-generation per-nucleotide probability of mutation (which can be estimated in various ways).

This means that the scaled mutation and recombination rates will also be of the same order of magnitude, and, thus, that recombination can be ignored only when mutation can be ignored. In other words, as long we restrict our attention to segments short enough not to be polymorphic, we do not need to worry about recombination!

One exception to this is highly selfing organisms. Because recombination only breaks up haplotypes in heterozygous individuals, most recombination in such organisms can effectively be ignored (the separation of time-scales argument used in Section 5.5.1 applies in models with recombination as well, see Nordborg, 2000).

It follows from the above that great care must be used when applying algorithms designed to estimate phylogenetic trees to within-species sequence data, because such algorithms generally output trees regardless of whether the true history of the sample is tree-like or not. To get an idea how common recombination has been in the history of a sample, the ‘four-gamete test’ (Hudson and Kaplan, 1985) is useful: in the absence of repeated mutation events (which are relatively unlikely), the four haplotype configurations AB , Ab , aB , and ab for two linked loci can only arise through recombination.

Although the four-gamete test can be used to demonstrate that recombination must have taken place between two polymorphic sites, it is important to recognize that recombination events can only be detected if there is sufficient polymorphism. Furthermore, many recombination events can *never* be detected even with infinite amounts of polymorphism (Griffiths and Marjoram, 1997; Hudson and Kaplan, 1985; Nordborg, 2000). Consider, for example, the two rightmost trees in Figure 5.10. These trees are identical. This means that the recombination event that gave rise to them cannot possibly leave any trace. This contributes to making estimating recombination rates and detecting hotspots of increased recombination extremely challenging (**Chapter 2**).

5.7 Selection

The coalescent depends crucially on the assumption of selective neutrality, because if the allelic state of a lineage influences its reproductive success, it is not possible to separate ‘descent’ from ‘state’. Nonetheless, it turns out that it is possible to circumvent this problem, and incorporate selection into the coalescent framework. Two distinct approaches have been used. The first is an elegant extension of the coalescent process, known as the ‘ancestral selection graph’ (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). The genealogy is generated backward in time, as in the standard coalescent, but it contains branching as well as coalescence events. The result is a genealogical graph that is superficially similar to the one generated by recombination. Mutations are then superimposed forward in time, and, with knowledge of the state of each branch, the graph is ‘pruned’ to a tree by preferentially removing bad branches (i.e. those carrying selectively inferior alleles). In a sense, the ancestral selection graph allows the separation of descent from state by including ‘potential’ descent: lineages that might have lived, had their state allowed it.

The second approach is based on two insights. First, a polymorphic population may be thought of as subdivided into *allelic classes* within which there is no selection. Second, if we know the historical sizes of these classes, then they may be modeled as analogous to patches, using the machinery described above. Lineages then ‘mutate between classes’ rather than ‘migrate between patches’. This approach, which might be called the ‘conditional structured coalescent’, was pioneered in the context of the coalescent by Kaplan *et al.* (1988). Knowing the past class sizes is the same as knowing the past allele frequencies, so it is obviously not possible to study the dynamics of the selectively different alleles themselves using this approach.

However, it is possible to study the effects of selection on the underlying genealogical structure, which is relevant if we wish to understand how linked neutral variants are affected (e.g. to detect footprints of selection; see **Chapter 14**).

It is not entirely clear how the two approaches relate to each other. Since the second approach requires knowledge of the past allele frequencies, it may be viewed as some kind of limiting (strong selection) or, alternatively, conditional version of the selection graph (Nordborg, 1999). In contrast, the selection graph requires all selection coefficients to be $O(1/N)$. This topic is discussed by Barton and Etheridge (2004) and in **Chapter 4**. The present chapter will focus on the second approach, which will be illustrated through three simple but very different examples (cf. **Chapter 14**).

5.7.1 Balancing Selection

By ‘balancing selection’ is meant any kind of selection that strives to maintain two or more alleles in the population. The effect of such selection on genealogies has been studied by a number of authors (Hey, 1991; Hudson and Kaplan, 1988; Kaplan *et al.*, 1988, 1991; Navarro and Barton, 2002; Nordborg, 1997, 1999; Nordborg and Innan, 2003; Takahata, 1990; Vekemans and Slatkin, 1994). We will limit ourselves to the case of two alleles, A_1 and A_2 , maintained at constant frequencies p_1 and $p_2 = 1 - p_1$ by strong selection. Alleles mutate to the other type with some small probability v per generation, and we define the scaled rate $\nu := 4Nv$. Reproduction occurs according to a diploid Wright–Fisher model, as for the recombination graph above.

Consider a segment of length ρ that contains the selected locus. Depending on the allelic state at the locus, the segment belongs to either the A_1 or the A_2 allelic class. Say that it belongs to the A_1 allelic class. Trace the ancestry of the segment a single generation back in time. It is easy to see that its creation involved an $A_2 \rightarrow A_1$ mutation with probability

$$\frac{\nu p_2}{\nu p_2 + (1 - \nu)p_1} = \frac{\nu}{4N} \cdot \frac{p_2}{p_1} + O\left(\frac{1}{N^2}\right) \quad (5.18)$$

(cf. equation (5.9), and involved recombination with probability $r = \rho/(4N)$). Thus the probability that neither happens is $1 - O(1/N)$, and the probability of two events, for example both mutation and recombination, is $O(1/N^2)$, and can be neglected. If nothing happens, then the lineage remains in the A_1 class. If there was a mutation, the lineage ‘mutates’ to the A_2 allelic class. If there was a recombination event, we have to know the genotype of the individual in which the event took place.

Because the lineage we are following is A_1 , we know that the individual must have been either an A_1A_1 homozygote or an A_1A_2 heterozygote. What fraction of the A_1 alleles was produced by each genotype? In general, this will depend on their relative fitness as well as their frequencies. Let x_{ij} be the frequency of A_iA_j individuals, and w_{ij} their relative fitness. Then the probability that an A_1 lineage was produced in a heterozygote is

$$\frac{w_{12}x_{12}/2}{w_{12}x_{12}/2 + w_{11}x_{11}}. \quad (5.19)$$

If we can ignore the differences in fitness, and assume Hardy–Weinberg equilibrium (see Nordborg (1999) for more on this), equation (5.19) simplifies to

$$\frac{p_1p_2}{p_1p_2 + p_1^2} = p_2. \quad (5.20)$$

Thus the probability that an A_1 lineage ‘meets’ and recombines with an A_2 segment is equal to the frequency of A_2 segments, which is intuitive. The analogous reasoning applies to A_2 lineages, which recombine with A_1 segments with probability p_1 , and with members of their own class with probability p_2 . The above can be made rigorous using a model that treats genotypes as well as individuals as population structure (Nordborg, 1999).

What happens when the lineage undergoes recombination? If it recombines in a homozygote, then both branches remain in the A_1 allelic class. However, if it recombines in a heterozygote, then one of the branches (the one *not* carrying the ancestry of the selected locus) will ‘jump’ to the A_2 allelic class. The other branch remains in the A_1 allelic class.

When more than two lineages exist, coalescences may occur, but only within allelic classes (remember that since mutation is $O(1/N)$ it is impossible for lineages to mutate and coalesce in the same generation).

If time is measured in units of $2N$ generations, and we let N go to infinity, the model converges to a coalescent process with the following types of events:

- each pair of lineages in the A_i allelic class coalesces independently at rate $1/p_i$;
- each lineage in A_i recombines with a segment in class j at rate ρp_j ;
- each lineage in A_i mutates to A_j , $j \neq i$, at rate $\nu p_j/p_i$.

The process may be stopped either when the Ultimate MRCA is reached, or when all points have found their MRCA.

This model has some very interesting properties. Consider a sample that contains both types of alleles. Since coalescence is only possible within allelic classes, the selected locus (in the strict sense of the word, that is, the site in the segment where the selectively important difference lies) cannot coalesce without at least one mutation event. If mutations are rare, then this will typically have occurred a very long time ago. In other words, the polymorphism will be ancient. All coalescences will occur within allelic classes before mutation allows the final two lineages to coalesce. The situation is similar to strong population subdivision (see Figure 5.7). However, this is only true for the locus itself: linked pieces may ‘recombine away’ and coalesce much earlier. This will usually result in a local increase in the time to MRCA centered around the selected locus, as illustrated in Figure 5.11. Because the expected number of mutations is proportional to the height of the tree, this may lead to a ‘peak of polymorphism’ (Hudson and Kaplan, 1988), as well as to detectable distortions in the allele frequency distribution (Tajima, 1989b; Chapter 14).

5.7.2 Selective Sweeps

Next consider a population in which favorable alleles arise infrequently at a locus, and are rapidly driven to fixation by strong selection. Each such fixation is known as a ‘selective sweep’ for reasons that will become apparent. This process can be modeled using the framework developed above, if we know how the allele frequencies have changed over time. Of course we do not know this, but if the selection is strong enough, it may be reasonable to model the increase in frequency of a favorable allele deterministically (Kaplan *et al.*, 1989).

Consider a population that is currently not polymorphic, but in which a selective sweep recently took place. During the sweep, there were two allelic classes just as in the balancing selection model above. The difference is that these classes changed in size over time. In particular, the class corresponding to the allele that is currently fixed in the population will *shrink* rapidly back in time. The genealogy of the selected locus itself (in the ‘point’ sense used above) will therefore behave as if it were part of a population that has expanded from a very small size (cf. Figure 5.6). Indeed, unlike ‘real’ populations, the allelic class *will* have grown from a size of 1.

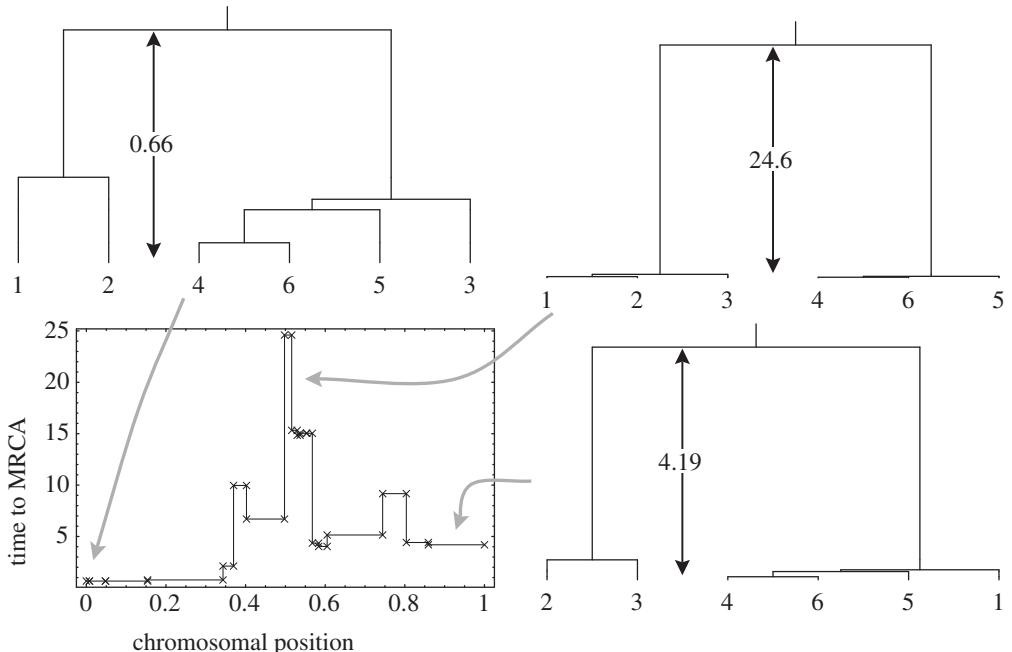


Figure 5.11 Selection will have a local effect on genealogies. A realization of the coalescent with recombination and strong balancing selection is shown. Lineages 1–3 belong to one allelic class, and lineages 4–6 to the other. The selected locus is located in the middle of the region. The plot shows how the time to the MRCA varies along the chromosome (the crosses denote cross-over points). The three extracted trees exemplify how the topology and branch lengths are affected by linkage to the selected locus. Note that the trees are not drawn to scale (the numbers on the arrows give the heights).

A linked point must have grown in the same way, unless recombination in a heterozygote took place between the point and the selected locus. Whether this happens or not will depend on how quickly the new allele increased in frequency. Typically, it depends on the ratio r/s , where s is the selective advantage of the new allele, and r is the relevant recombination probability.

The result of such a fixation event is thus to cause a local ‘genealogical distortion’, just like balancing selection. However, whereas the distortion in the case of balancing selection looks like population subdivision, the distortion caused by a fixation event looks like population growth. Close to the selected site, coalescence times will have a tendency to be short, and the genealogy will have a tendency to be star-like (*cf.* Figure 5.6). Note that a single recombination event in the history of the sample can change this, and that the variance will consequently be large (note the variance in time to MRCA in Figure 5.11). Shorter coalescence times mean less time for mutations to occur, so a local reduction in variability is expected. This is because when the new allele sweeps through the population and fixes, it causes linked neutral alleles to ‘hitch-hike’ along and also fix (Maynard Smith and Haigh, 1974). Repeated selective sweeps can thus decrease the amount of polymorphism in a genomic region (Kaplan *et al.*, 1989; Simonsen *et al.*, 1995), and also cause detectable local distortions in the allele frequency distribution as well as extent of haplotype sharing and linkage disequilibrium (Chapter 14). Because each sweep is expected to affect a bigger region the lower the rate of recombination is, this has been proposed as an explanation for the correlation between polymorphism and local rate of recombination that is observed in many organisms (Begun and Aquadro, 1992; Nachman, 1997; Nachman *et al.*, 1998).

5.7.3 Background Selection

We have seen that selection can distort genealogies in ways reminiscent of strong population subdivision and of population growth. It is often difficult to statistically distinguish between selection and demography for precisely this reason (Tajima, 1989b; Fu and Li, 1993; **Chapter 14**). It is also possible for selection to affect genealogies in a way that is completely indistinguishable from the standard model, that is, as a linear change in time-scale. This appears to be the case for selection against deleterious mutations, at least under some circumstances (Charlesworth *et al.*, 1995; Hudson and Kaplan, 1994, 1995; Nordborg, 1997; Nordborg *et al.*, 1996; **Chapter 4**).

The basic reason for this is the following. Strongly deleterious mutations are rapidly removed by selection. Looking backward in time, this means that each lineage that carries a deleterious mutation must have a non-mutant ancestor in the near past. On the coalescent time-scale, lineages in the deleterious allelic class will ‘mutate’ (backward in time) to the ‘wildtype’ allelic class instantaneously. The process looks like a strong-migration model, with the wildtype class as the source environment, and the deleterious class as the sink environment: the presence of deleterious mutations increases the variance in reproductive success. The resulting reduction in the effective population size is known as ‘background selection’ (Charlesworth *et al.*, 1993).

More realistic models with multiple loci subject to deleterious mutations, recombination, and several mutational classes turn out to behave similarly. The strength of the background selection effect at a given genomic position will depend strongly on the local rate of recombination, which determines how many mutable loci influence a given point. Thus, deleterious mutations have also been proposed as an explanation for the correlation between polymorphism and local rate of recombination referred to above (Charlesworth *et al.*, 1993). The ‘effective population size’ would thus depend on the mutation, selection, and recombination parameters in each genomic region.

It should be pointed out that, unlike the many limit approximations presented in this chapter, the idea that background selection can be modeled as a simple scaling is not mathematically rigorous (discussed in **Chapter 4**). However, we would rather hope that selection against deleterious mutations can be taken care of this way, because given that amino acid sequences are conserved over evolutionary time, practically all of population genetics theory would be in trouble otherwise!

5.8 Neutral Mutations

Not much has been said about the neutral mutation process because it is trivial from a mathematical point of view. Once we know how to generate the genealogy, mutations can be added afterwards according to a Poisson process with rate $\theta/2$, where θ is the scaled per-generation mutation probability. Thus, if a particular branch has length τ units of scaled time, the number of mutations that occur on it will be Poisson distributed with mean $\tau\theta/2$ (and they have equal probability of occurring anywhere on the branch). It is also possible to add mutations while the genealogy is being created, instead of afterwards. This can in some circumstances lead to much more efficient algorithms (see, for example, the ‘urn scheme’ described by Donnelly and Tavaré, 1995), although from the point of view of simulating samples, all coalescent algorithms are so efficient that such fine-tuning does not matter.

It should be noted that the mutation process is just as general as the recombination process. Almost any neutral mutation model can be used. A useful trick is so-called ‘Poissonization’: let mutation events occur according to a simple Poisson process with rate $\theta/2$, but once an event occurs, determine the *type* of event through some kind of transition matrix which includes

mutation back to self (i.e. there was no mutation). This allows models where the mutation probability depends on the current allelic state.

The only restriction is that in order to interpret samples generated by the coalescent as samples from the relevant stationary distribution (that incorporates demography, migration, selection at linked sites, etc.), we need to be able to choose the type of the MRCA from the stationary distribution of the mutation process (alone, since demography *etc.* does not affect samples of size $n = 1$). In many cases, such as the infinite-alleles model (each mutation gives rise to a new allele) or the infinite-sites model (each mutation affects a new site), the state of the MRCA does not matter, since all we are interested in is the number of mutational changes.

5.9 Concluding Remarks

5.9.1 The Coalescent and 'Classical' Population Genetics

The differences between coalescent theory and 'classical' population genetics have often been misunderstood. First, the basic models do not differ. The coalescent is essentially a diffusion model of lines of descent. This can be done forward in time, for the whole population (e.g. Griffiths, 1980), but it was realized in the early 1980s that it is easier to do it backward in time. Second, the coalescent is not limited to finite samples. Everything above has been limited to finite samples because it is mathematically much easier, but it is likely that all of it could be extended to the whole (infinite) population. Of course, it is essential for the independence of events that the number of lineages be finite, but in the whole-population coalescent the number of lineages becomes finite infinitely fast (it is an 'entrance boundary'; see Griffiths, 1984). Third, classical population genetics is not limited to the whole population. A sample of size $n = 6$ from a K -allele model, say, could be obtained either through the coalescent, or by first drawing a population from the stationary distribution found by Wright (1949), and then drawing six alleles conditional on this population. Note, however, that it would be rather more difficult (read 'impossible') to use the second approach for most models. Fourth, the coalescent is not tied to sequence data: any mutation model can be used. The impression that it is no doubt came about because models for sequence evolution such as the infinite-sites model are indeed impossibly hard to analyze using classical methods (Ethier and Griffiths, 1987).

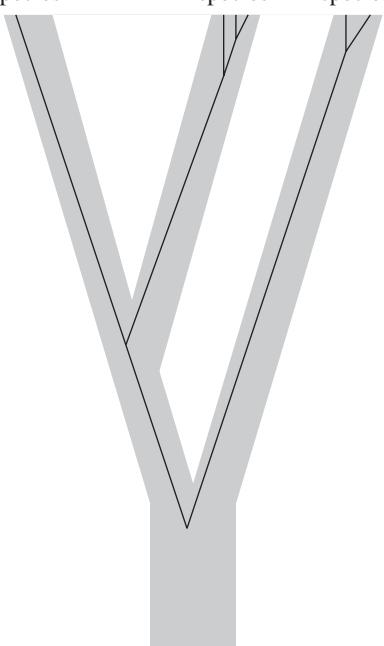
The real difference is rather that (essentially) all of classical population genetics is 'prospective', looking forward in time (Ewens, 1979, 1990). Another way of saying this is that it is conditional: given the state in a particular generation, what will happen? This approach is fine when modeling is done to determine 'how evolution might work' (which is what most classical population genetics was about). It is usually not suitable for statistical analysis of data, however. Wright considered how 'heterozygosity' would decay from the same starting point in infinitely many identical populations, that is, he took the expectation over evolutionary realizations. Data, alas, come from one such realization, and the coalescent provides a very convenient way of taking this into account. Importantly, this difference becomes less important when looking at genomic averages, in which case the evolutionary variance can often be ignored.

5.9.2 The Coalescent and Phylogenetics

The relationship between coalescent theory and phylogenetics remains a source of confusion. The central role played by trees in both turns out to be very misleading (in particular, the purpose of the coalescent is *not* to estimate gene trees, but rather to model the process that gave rise to them, see Rosenberg and Nordborg, 2002). To see this, we need to model speciation.

species A species B species C

Figure 5.12 A gene tree within a species tree.



This has usually been done using an ‘isolation’ model in which randomly mating populations split into two completely isolated ones at fixed times in the past. The result is a ‘species tree’, within which we find ‘gene trees’ (see Figure 5.12). The model is quite simple: lineages will tend to coalesce within their species, and can only coalesce with lineages from other species back in the ancestral species (the multi-species coalescent is described in **Chapter 7**).

Molecular phylogenetics attempts to estimate the species tree by estimating the genealogy of homologous sequences from the different species, that is, by estimating the gene tree. The species tree is assumed to exist and is treated as a model parameter.

In addition, traditional phylogenetic methods rely on all branches in the species tree being very long compared to within-species coalescence times. This means that the coalescent can be ignored: regardless of how we sample, all (neutral) gene genealogies will rapidly coalesce within their species, and thereafter have the same topology as the species tree. Furthermore, the variation in the branch lengths caused by different coalescence times in the ancestral species will be negligible compared to the lengths of the interspecific branches. There is no need to sample more than one individual per species, and recombination is completely irrelevant. Gene trees perfectly reflect the species tree, and tend to become viewed as parameters (rather than random variables) as well. Indeed, in many situations, the problem is the opposite: the branches are so long that repeated mutations have erased much of the phylogenetic information (**Chapter 6**).

When this ‘long-branch’ assumption is not met (i.e. for closely related species), gene trees may differ from species tree due to coalescence variation (so-called ‘deep coalescence’ or ‘lineage sorting’; see Nei, 1987; Takahata, 1989; Hudson, 1992; Maddison, 1997). Attempts to infer the species tree must take this into account. Phylogenetic inference then becomes a special case of population genetics inference in which the goal is to estimate historical population structure. In fact, it is crucial to employ this framework, because it allows a much wider range of model, for example ones that include gene flow (**Chapters 7 and 8**). Perhaps most importantly, this makes it possible to assess whether a bifurcating species tree (like the one in Figure 5.12) is

indeed an appropriate model. Traditional phylogenetic methods do not do this: a hierarchical clustering algorithm will generate a hierarchical clustering (i.e. a tree) regardless of whether it makes sense or not.

5.9.3 Prospects

The major change in population genetics since the original version of this chapter is the flood of polymorphism data that began in humans and a few model plants, and will soon be universal. To make sense of these data, models are needed, and the coalescent remains the simplest possible model. It is used, directly or indirectly (e.g. via software that relies on coalescent theory) in almost every paper that analyses such data.

Still missing is better theory for continuous spatial distributions (isolation-by-distance) and weak selection on large numbers of sites. The latter is especially relevant given the increased use of genomics to study selection on quantitative traits (Simons *et al.*, 2018; Field *et al.*, 2016; Turchin *et al.*, 2012). How to model this is not obvious.

Acknowledgements

I wish to thank David Balding, Bengt Olle Bengtsson, Malia Fullerton, Jenny Hagenblad, Maarit Jaarola, Martin Lascoux, Claudia Neuhauser, François Rousset, Matthew Stephens, and Torbjörn Säll for comments on the original version of the manuscript, long ago, and David Balding for further comments on this version.

References

- Andolfatto, P. and Nordborg, M. (1998). The effect of gene conversion on intralocus associations. *Genetics* **148**, 1397–1399.
- Bahlo, M. (1998). Segregating sites in a gene conversion model with mutation. *Theoretical Population Biology* **54**, 243–256.
- Barton, N.H. and Etheridge, A.M. (2004). The effect of selection on genealogies. *Genetics* **166**, 1115–1131.
- Barton, N.H., Etheridge, A.M. and Véber, A. (2013). Modelling evolution in a spatial continuum. *Journal of Statistical Mechanics: Theory and Experiment* **2013**(01), P01002.
- Begun, D.J. and Aquadro, C.F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520.
- Charlesworth, B. and Charlesworth, D. (2010). *Elements of Evolutionary Genetics*. Roberts and Company Publishers, Greenwood Village, CO.
- Charlesworth, B., Morgan, M.T. and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.
- Charlesworth, D., Charlesworth, B. and Morgan, M.T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632.
- Donnelly, P. (1996). Interpreting genetic variability: The effects of shared evolutionary history. In *Variation in the Human Genome*, Ciba Foundation Symposium 197. Chichester, Wiley, pp. 25–50.
- Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.
- Ethier, S.N. and Griffiths, R.C. (1987). The infinitely many sites model as a measure valued diffusion. *Annals of Probability* **5**, 515–545.

- Ewens, W.J. (1979). *Mathematical Population Genetics*. Springer-Verlag, Berlin.
- Ewens, W.J. (1990). Population genetics theory – the past and the future. In S. Lessard (ed.), *Mathematical and Statistical Developments of Evolutionary Theory*. Kluwer Academic, Dordrecht, pp. 177–227.
- Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. and Pritchard, J.K. (2016). Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764.
- Fisher, R.A. (1965). *Theory of Inbreeding*. Oliver and Boyd, Edinburgh, 2nd edition.
- Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.
- Griffiths, R.C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology* **17**, 37–50.
- Griffiths, R.C. (1984). Asymptotic line-of-descent distributions. *Journal of Mathematical Biology* **21**, 67–75.
- Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.
- Griffiths, R.C. and Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly and S. Tavaré (eds.), *Progress in Population Genetics and Human Evolution*. Springer-Verlag, New York, pp. 257–270.
- Griffiths, R.C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London, Series B* **344**, 403–10.
- Hein, J., Schierup, M.H. and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.
- Herbots, H.M. (1994). *Stochastic models in population genetics: Genealogy and genetic differentiation in structured populations*. PhD thesis, University of London.
- Hey, J. (1991). A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.
- Hudson, R.R. and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.
- Hudson, R.R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.
- Hudson, R.R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetical Research* **50**, 245–250.
- Hudson, R.R. (1990). Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics (eds.), *Oxford Surveys in Evolutionary Biology*, volume 7. Oxford University Press, Oxford, pp. 1–43.
- Hudson, R.R. (1992). Gene trees, species trees and the segregation of ancestral alleles. *Genetics* **131**, 509–512.
- Hudson, R.R. (1993). The how and why of generating gene genealogies. In N. Takahata and A.G. Clark (eds.), *Mechanisms of Molecular Evolution*. Japan Scientific Societies Press, Tokyo, pp. 23–36.
- Hudson, R.R. (1994). Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *Journal of Evolutionary Biology* **7**, 535–548.
- Hudson, R.R. (2001). Two-locus sample distributions and their applications. *Genetics* **159**, 1805–1817.
- Hudson, R.R. and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.
- Hudson, R.R. and Kaplan, N.L. (1994). Gene trees with background selection. In G.B. Golding (ed.), *Non-Neutral Evolution: Theories and Molecular Data*. Chapman & Hall, New York, pp. 140–153.

- Hudson, R.R. and Kaplan, N.L. (1995). Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.
- Kaplan, N.L. and Hudson, R.R. (1985). The use of sample genealogies for studying a selectively neutral m -loci model with recombination. *Theoretical Population Biology* **28**, 382–396.
- Kaplan, N.L., Darden, T. and Hudson, R.R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.
- Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989). The ‘hitch-hiking’ effect revisited. *Genetics* **123**, 887–899.
- Kaplan, N.L., Hudson, R.R. and Iizuka, M. (1991). The coalescent process in models with selection, recombination and geographic subdivision. *Genetics Research* **57**, 83–91.
- Kingman, J.F.C. (1982a). The coalescent. *Stochastic Processes and Their Applications* **13**, 235–248.
- Kingman, J.F.C. (1982b). Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino (eds.), *Exchangeability in Probability and Statistics*. North-Holland, Amsterdam, pp. 97–112.
- Kingman, J.F.C. (1982c). On the genealogy of large populations. In J. Gani and E.J. Hannan (eds.), *Essays in Statistical Science: Papers in Honour of P.A.P. Moran*. Applied Probability Trust, Sheffield. Journal of Applied Probability, special volume 19A, pp. 27–43.
- Krone, S.M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**, 210–237.
- Li, H. and Durbin, R.M. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496.
- Li, W.-H. (1997). *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology* **46**(3), 523–536.
- Marjoram, P. and Donnelly, P. (1997). Human demography and the time since mitochondrial Eve. In P. Donnelly and S. Tavaré (eds.), *Progress in Population Genetics and Human Evolution*. Springer-Verlag, New York, pp. 107–131.
- Marjoram, P. and Wall, J.D. (2006). Fast ‘coalescent’ simulation. *BMC Genetics* **7**, 16.
- McVean, G.A.T (2002). A genealogical interpretation of linkage disequilibrium. *Genetics* **162**, 987–991.
- McVean, G.A.T and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London, Series B* **360**(1459), 1387–1393.
- Möhle, M. (1998a). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Advances in Applied Probability* **30**, 493–512.
- Möhle, M. (1998b). Robustness results for the coalescent. *Journal of Applied Probability* **35**, 438–447.
- Möhle, M. (1999). Weak convergence to the coalescent in neutral population models. *Journal of Applied Probability* **36**, 446–460.
- Nachman, M.W. (1997). Patterns of DNA variability at X -linked loci in *Mus domesticus*. *Genetics* **147**, 1303–1316.
- Nachman, M.W., Bauer, V.L., Crowell, S.L. and Aquadro, C.F. (1998). DNA variability and recombination rates at X -linked loci in humans. *Genetics* **150**, 1133–1141.
- Nagylaki, T. (1980). The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology* **9**, 101–114.
- Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.
- Nagylaki, T. (1998). The expected number of heterozygous sites in a subdivided population. *Genetics* **149**, 1599–1604.
- Navarro, A. and Barton, N.H. (2002). The effects of multilocus balancing selection on neutral variability. *Genetics* **161**, 849–863.

- Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Neuhauser, C. and Krone, S.M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.
- Noah, A. (2002). Rosenberg and Magnus Nordborg. Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **3**, 380–390.
- Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.
- Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics* **63**, 1237–1240.
- Nordborg, M. (1999). The coalescent with partial selfing and balancing selection: An application of structured coalescent processes. In F. Seillier-Moiseiwitsch (ed.), *Statistics in Molecular Biology and Genetics*, volume 33 of *IMS Lecture Notes-Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, pp. 56–76.
- Nordborg, M. (2000). Linkage disequilibrium, gene trees, and selfing: An ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929.
- Nordborg, M. (2001). Coalescent theory. In D.J. Balding, M.J. Bishop and C. Cannings (eds.), *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, pp. 179–212.
- Nordborg, M. and Donnelly, P. (1997). The coalescent process with selfing. *Genetics* **146**, 1185–1195.
- Nordborg, M. and Innan, H. (2003). The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. *Genetics* **163**, 1201–1213.
- Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: What history has to tell us. *Trends in Genetics* **18**, 83–90.
- Nordborg, M., Charlesworth, B. and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research* **67**, 159–174.
- Notohara, M. (1990). The coalescent and the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **29**, 59–75.
- Notohara, M. (1993). The strong-migration limit for the genealogical process in geographically structured populations. *Journal of Mathematical Biology* **31**, 115–122.
- Pluzhnikov, A. and Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–1262.
- Pollak, E. (1987). On the theory of partially inbreeding finite populations I. Partial selfing. *Genetics* **117**, 353–360.
- Pulliam, H.R. (1988). Sources, sinks, and population regulation. *American Naturalist* **132**, 652–661.
- Rousset, F. (1999a). Genetic differentiation within and between two habitats. *Genetics* **151**, 397–407.
- Rousset, F. (1999b). Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55**, 297–308.
- Saunders, I.W., Tavaré, S. and Watterson, G.A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.
- Simons, Y.B., Bullaughey, K., Hudson, R.R. and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biology*, 16:e2002985.
- Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.
- Simonsen, K.L. and Churchill, G.A. (1997). A Markov chain model of coalescence with recombination. *Theoretical Population Biology* **52**, 43–59.
- Sjödin, P., Kaj, I., Krone, S.M., Lascoux, M. and Nordborg, M. (2005). On the meaning and existence of an effective population size. *Genetics* **169**, 1061–1070.

- Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.
- Smith, J.M. and Haigh, J. (1974). The hitchhiking effect of a favourable gene. *Genetics Research* **23**, 23–35.
- Strobeck, C. (1987). Average number of nucleotide differences in a sample from a single subpopulation: A test for population subdivision. *Genetics* **117**, 149–153.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.
- Tajima, F. (1989a). DNA polymorphism in a subdivided population: The expected number of segregating sites in the two-subpopulation model. *Genetics* **123**, 229–240.
- Tajima, F. (1989b). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetic Research* **52**, 213–222.
- Takahata, N. (1989). Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics* **122**, 957–966.
- Takahata, N. (1990). A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 2419–2423.
- Takahata, N. (1991). Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129**, 585–595.
- Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.
- Turchin, M.C., Chiang, C.W.K., Palmer, C.D., Sankararaman, S., Reich, D. (2012). Genetic Investigation of Anthropometric Traits (GIANT) Consortium, and Joel N Hirschhorn. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44**, 1015–1019.
- Vekemans, X. and Slatkin, M. (1994). Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* **137**, 1157–1165.
- Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.
- Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, CO.
- Wakeley, J. (2013). Coalescent theory has many new branches. *Theoretical Population Biology* **87**, 1–4.
- Wilkinson-Herbots, H.M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology* **37**, 535–585.
- Wiuf, C. (2000). A coalescence approach to gene conversion. *Theoretical Population Biology* **57**, 357–367.
- Wiuf, C. and Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics* **147**, 1459–1468.
- Wiuf, C. and Hein, J. (1999a). Recombination as a point process along sequences. *Theoretical Population Biology* **55**(3), 248–259.
- Wiuf, C. and Hein, J. (1999b). The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228.
- Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics* **155**, 451–462.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.
- Wright, S. (1949). Adaptation and selection. In G.L. Jepson, G.G. Simpson and E. Mayr (eds.), *Genetics, Palaeontology, and Evolution*. Princeton University Press, Princeton, NJ, pp. 365–389.

6

Phylogeny Estimation Using Likelihood-Based Methods

John P. Huelsenbeck

Department of Integrative Biology, University of California, Berkeley, CA, USA

Abstract

The likelihood of a phylogenetic tree is proportional to the probability of observing the comparative data (such as aligned DNA sequences) conditional on the tree. The likelihood function is important because it is the vehicle that carries the observations. The likelihood function can be used in two ways to infer phylogeny. First, the tree that maximizes the likelihood can be chosen as the best estimate of phylogeny; this is the method of maximum likelihood. Second, a prior probability distribution on trees can be specified and inferences based upon the posterior probability distribution of trees; this is the approach taken by Bayesians. Although maximum likelihood and Bayesian inference are similar in that the same models of DNA substitution can be used to calculate the likelihood function, they differ in their interpretation of probability. Markov chain Monte Carlo (MCMC) can be used to approximate the posterior probabilities of trees. MCMC also makes it possible to perform comparative analyses that accommodate phylogenetic uncertainty.

6.1 Introduction

Biologists began to construct phylogenies – branching diagrams depicting the genealogical relationships of a group of species – shortly after the publication of Darwin's (1859) *On the Origin of Species* (see, for example, Haeckel, 1866). Early phylogenies were often constructed by paleontologists and incorporated information on the stratigraphic position of fossils as well as on the characteristics of the organisms. Unfortunately, the entire process of phylogeny reconstruction occurred in the biologist's head. The expert on a group of species would simply publish a diagram depicting the evolutionary history, often with question marks to denote uncertainty. How could others challenge the phylogeny if the process of constructing a phylogeny was vaguely, if at all, described? This situation changed with the publication of two remarkable works. The first, Hennig's *Grundzüge einer Theorie der Phylogenetischen Systematik* (*Phylogenetic Systematics*; Hennig, 1950, 1966), argued that phylogenies should be constructed using characters that are shared among species and derived uniquely in those species. The other, Sokal and Sneath's *Principles of Numerical Taxonomy* (Sokal and Sneath, 1963) developed methods to produce taxonomies based on overall similarity. Importantly, these works liberalized the study of phylogeny. With the conceptual tools provided by these authors, anyone possessing the tenacity to learn the biology of a group of species could produce a phylogeny.

For the next twenty years, the intellectual heirs of Hennig (the ‘cladists’) and Sokal and Sneath (the ‘pheneticists’) waged what was often an acrimonious war that was fought out in the literature. Cladists were interested in reconstructing phylogeny and developed tools, such as the parsimony method, to accomplish this goal. Pheneticists, on the other hand, were interested in producing stable taxonomies based on overall similarity, using distance-based methods such as the unweighted pair-group method with arithmetic mean (Sokal and Michener, 1958) to do this. The eventual victor of this war was perhaps predictable: the cladists handily won because their goal of reconstructing phylogeny was of broad interest to biologists. Unfortunately, the victory by the cladists was so complete that not only was the goal of reconstructing phylogeny accepted, but the cladists’ preferred method for reconstructing phylogeny was also accepted by many as the only logically sound method to reconstruct phylogeny.

6.1.1 Statistical Phylogenetics

While war waged between the cladists and pheneticists, another school was developing that took an unabashedly statistical view of the problem. Statistical phylogenetics took the goal of the cladists as granted, but applied standard statistical methods, such as maximum likelihood and Bayesian inference, to estimate phylogeny. The statistical approach to phylogeny estimation was pioneered by L. L. Cavalli-Sforza and A. W. F. Edwards who introduced the least squares, minimum evolution (or parsimony), and maximum likelihood methods of phylogenetic inference (Edwards and Cavalli-Sforza, 1964; Cavalli-Sforza and Edwards, 1967). Perhaps their most important contribution, however, was the early realization that the phylogeny problem was one of statistical inference (Edwards, 1966). Both Edwards and Cavalli-Sforza had ‘sat at the feet of R.A. Fisher’ (Edwards, 1998) – the inventor and proponent of the method of maximum likelihood to estimate statistical parameters (Fisher, 1912, 1921, 1922) – so it was natural for them to consider maximum likelihood to infer phylogeny.

Likelihood-based phylogenetic inference became practical when J. Felsenstein (1981) introduced a ‘pruning algorithm’ (now referred to as the ‘Felsenstein pruning algorithm’; see also Gallager, 1962, 1963) to compute the likelihood. The Felsenstein pruning algorithm makes possible efficient calculation of the likelihood for a large number of taxa by taking advantage of the form of the tree topology; today, all computer programs for calculating likelihoods on phylogenies use Felsenstein’s pruning algorithm.

Since about 1990, the statistical approach to the phylogeny problem has made rapid progress. Specifically, advances in three areas have propelled the field of statistical phylogenetics:

1. Bayesian estimation was applied to phylogeny problems (Rannala and Yang, 1996; Mau, 1996; Mau and Newton, 1997; Li *et al.*, 2000).
2. Computer programs implementing algorithms to maximize likelihoods or to approximate posterior probabilities are freely available for anyone to use. These programs are sophisticated and fast.
3. The probability models applied to the problem have expanded to allow a large variety of questions to be addressed.

6.1.2 Chapter Outline

In this chapter, I address application of the likelihood function in phylogenetics. To do this, I will start by providing an introduction to maximum likelihood and Bayesian estimation (Section 6.2) and to model comparison (Section 6.3) to give the statistical basis needed to understand the rest of the chapter (for more on details on statistical inference, see **Chapter 1**). If you are already familiar with these concepts feel free to skip ahead to Section 6.4, in which I will describe how to

calculate the likelihood for a phylogenetic model. Next, I will describe how this likelihood can be used as a basis for phylogenetic inference (Section 6.5) and then go through some applications of such inference methods in molecular evolution (Section 6.6). The latter part serves the purpose both of showing what the inference methods have been used for so far and of highlighting some of the more recent exciting expansions of the basic models and methods introduced in Sections 6.4 and 6.5. Finally, I will conclude in Section 6.7 with a few remarks on where I think the field will be moving in the future.

6.2 Maximum Likelihood and Bayesian Estimation

The likelihood of a hypothesis is proportional to the probability of observing the data given the hypothesis,

$$\mathcal{L}(\text{Hypothesis}) = C \times f(\text{Observations} \mid \text{Hypothesis}), \quad (6.1)$$

where $f(\text{Observations} \mid \text{Hypothesis})$ is the probability of observing the data given a hypothesis. The constant, C , is arbitrary, and often taken to be 1. (Throughout this chapter, I follow the convention of using $f(\cdot \mid \cdot)$ to denote a conditional probability.)

6.2.1 Maximum Likelihood

Imagine tossing a coin with the goal of determining if the coin is fair. You toss the coin n times, and record heads facing up x of those times. The substantive question is this: what is the probability of the coin landing heads up on a single toss? We will call this probability θ . If θ is indistinguishable from 0.50, then we would argue that the coin is fair. We need to estimate θ from the data, and the method of maximum likelihood is one way to do this.

The first step is to construct a statistical model that explains the potential variability in our observations. In this case in which coins are tossed, the probability model should contain the parameter θ . Here, we can use the binomial probability distribution, which states that the probability of observing x heads on n tosses is

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \quad (6.2)$$

(The binomial coefficient, $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, is the number of ways to choose x objects from a total of n , and accounts for the different ways in which we could have observed x heads.) This is a conditional probability – the probability of the observations conditional on the parameter θ taking some specific value – and should look suspiciously like the description of the likelihood function given earlier. The likelihood, in fact, is

$$\mathcal{L}(\theta) = f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \quad (6.3)$$

Figure 6.1 shows the likelihood surface for the coin-tossing example for several different scenarios. Each graph in Figure 6.1 depicts the likelihood surface for a different experiment in which the number of heads observed (x) and the number of tosses (n) differed. First note that the likelihood surface is more sharply curved when the number of tosses is large ($n = 100$) than when it is relatively small ($n = 10$). This is to be expected; the data contain more information on the value of the parameter θ when the results of many coin tosses have been observed. In fact, the curvature of the likelihood around the maximum value is related to a measure called the Fisher information, which can be used to construct confidence intervals for the maximum

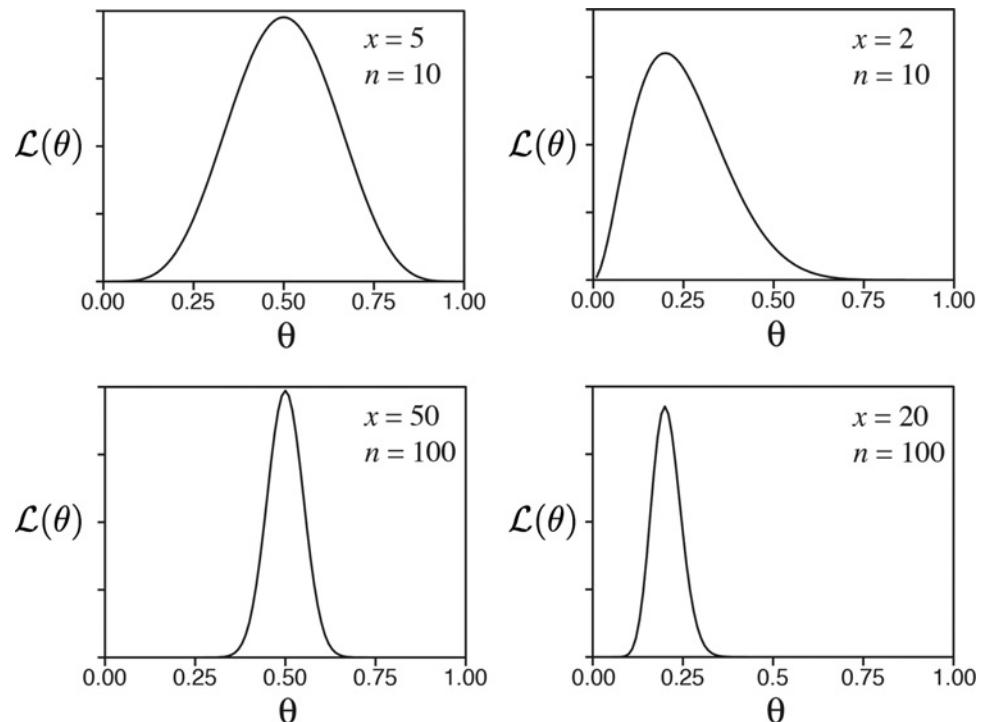


Figure 6.1 The likelihood surface for the case in which x heads are observed on n tosses of a coin. The top two graphs show the likelihood surface when the total number of tosses is $n = 10$. The bottom two graphs show the likelihood when the total number of tosses is $n = 100$.

likelihood estimate (see **Chapter 1**). Also note that, sensibly enough, the likelihood appears to be maximized when θ is equal to the proportion of heads observed.

The maximum likelihood estimate (MLE) of a parameter is the value of the parameter that maximizes the likelihood function (and thus the probability of observing the data); the MLE of θ is $\hat{\theta} = \max_{\theta} L(\theta)$. One can show with minimal calculus that the maximum likelihood estimate of θ for the coin-tossing problem is $\hat{\theta} = x/n$. In words, the MLE of θ for the coin-tossing problem is equal to the fraction of the time heads was observed.

6.2.2 Bayesian Inference

Thomas Bayes was an English clergyman who lived in the early eighteenth century (1701–1761) and published (posthumously) an interesting, but relatively minor, extension of the definition of conditional probability, today known as ‘Bayes’ theorem’ (Bayes, 1763):

$$f(A|B) = \frac{f(A, B)}{f(B)} = \frac{f(B|A)f(A)}{f(B)}. \quad (6.4)$$

In words, the probability of the event A given that the event B has occurred is equal to the joint probability of A and B divided by the probability of B . Bayes’ theorem takes on rather profound implications for statistics, however, suggesting a method of inference and reasoning that is philosophically quite different from many methods of statistical inference with which the reader may be familiar.

The implications of Bayes' theorem become clearer when one considers the events (A and B above) to be the observable and unobservable parts of a statistical model. For example, let us denote the observations as the data, X . Moreover, let us consider n potential discrete hypotheses that could explain the observations, $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_n$. These n different hypotheses are the unobservable part of the statistical model; presumably, deciding which of these n hypotheses best explains the data will go a long way towards addressing the substantive question faced by the statistician. The probability of the i th hypothesis becomes, according to Bayes' theorem,

$$f(\mathcal{H}_i|X) = \frac{f(X|\mathcal{H}_i)f(\mathcal{H}_i)}{\sum_{j=1}^n f(X|\mathcal{H}_j)f(\mathcal{H}_j)}. \quad (6.5)$$

The different parts of this equation are all referred to by specific names. First, Bayesian statisticians base all inferences upon the posterior probability distribution of a hypothesis,

$$f(\mathcal{H}_i|X) \quad (6.6)$$

For example, the hypothesis having the greatest posterior probability might be taken as the best estimate of which hypothesis is correct (this is called the maximum posterior probability (MAP) estimate). The second part of Bayes' theorem,

$$f(X|\mathcal{H}_i), \quad (6.7)$$

is referred to as the 'likelihood function'. The likelihood function is important as it is the vehicle that carries the information contained in the data about the different hypotheses. As we have seen, the likelihood function plays a central role not only in Bayesian inference, but also in the method of maximum likelihood. The third part of the equation,

$$f(\mathcal{H}_i) \quad (6.8)$$

is the prior probability of hypothesis i . The final part of Bayes' theorem, the summation over all hypotheses in the denominator, normalizes the posterior probability distribution and is referred to as the 'marginal likelihood'.

The explicit incorporation of prior information is the main manner in which Bayesian inference differs from other statistical methods. The explicit incorporation of prior information can be viewed as a strength or a weakness of Bayesian inference, depending upon one's viewpoint. One of the main impediments to Bayesian analysis is that the prior probability distribution of a parameter can be quite difficult to specify. Moreover, there is nothing to prohibit different investigators from using different prior information. On the other hand, the explicit incorporation of prior information is a strength of a Bayesian analysis inasmuch as a Bayesian analysis allows any prior information that is available to the investigator to play a role in the analysis. The results of a Bayesian analysis are also quite easy to interpret, as the values of competing hypotheses are judged by their posterior probabilities. Although specifying prior probabilities can often be tricky, typically statisticians will examine the robustness of a conclusion to the use of different prior probability distributions. The usual case is that the prior probability becomes less important as more data are collected; that is, the prior information becomes swamped by the data.

Bayesian analysis also differs from other types of statistical methods in the concept of probability used. Under the frequentist concept of probability, one imagines that an experiment can be infinitely repeated. The probability of any specific event is then the fraction of the time the event occurs in the long run.

Subjective probabilities, used in Bayesian analysis, are interpreted as a person's measure of belief that some event will occur. Even though the investigator may feel that some unknown

model parameter takes one fixed value in reality, the investigator's uncertainty about that parameter can be adequately described by a probability distribution for the parameter. Hence, before an experiment an investigator may give each of the n hypotheses equal weight. This would be equivalent to giving the n hypotheses an equal prior probability ($f(\mathcal{H}_i) = 1/n$). After performing an experiment and making new observations, the investigator's opinions about the relative merit of the n hypotheses should change. The updated beliefs about the hypotheses are contained in the posterior probability of the hypotheses [$f(\mathcal{H}_i|X)$]. Bayes' theorem guides how the new information should update belief about a hypothesis.

Let us reconsider the coin-tossing experiment where the objective is to estimate the probability that heads appears on a single toss of the coin, the parameter θ . As before, we observe x heads on n tosses of the coin. In a Bayesian analysis, the objective is to calculate the posterior probability of the parameter, which for coin tossing is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_0^1 f(x|\theta)f(\theta) d\theta}. \quad (6.9)$$

The likelihood, $f(x|\theta)$, is given by the binomial probability distribution, just as was the case when we performed maximum likelihood estimation. In addition to the likelihood function, however, we now must also specify a prior probability distribution for the parameter, $f(\theta)$. This prior distribution should describe the investigator's beliefs about the hypothesis before the experiment was performed. In this case, it makes sense to use a prior distribution that is flexible, allowing different people to specify different prior probability distributions and also allowing for an easy investigation of the sensitivity of the results to the prior assumptions. A beta distribution is often used as a prior probability distribution for the binomial parameter. The beta distribution has two parameters, α and β . Depending upon the specific values chosen for α and β , one can generate a large number of different prior probability distributions for the parameter θ . Figure 6.2 shows several possible prior distributions for coin tossing. Figure 6.2(a) shows a uniform prior distribution for the probability of heads appearing on a single toss of a coin. In effect, a person who adopts this prior distribution is claiming total ignorance of the dynamics of coin tossing. Figure 6.2(b) shows a prior distribution for a person who has some experience tossing coins; anyone who has tossed a coin realizes that it is impossible to predict which side will face up when tossed, but that heads appears about as frequently as tails, suggesting more prior weight on values around $\theta = 0.5$ than on values near $\theta = 0$ or $\theta = 1$. Lastly, Figure 6.2(c) shows a prior distribution for a person who suspects he is being tricked. Perhaps the coin that is being tossed is from a friend with a long history of practical jokes, or perhaps this friend has tricked the investigator with a two-headed coin in the past. Figure 6.2(c), then, might represent the 'trick-coin' prior distribution.

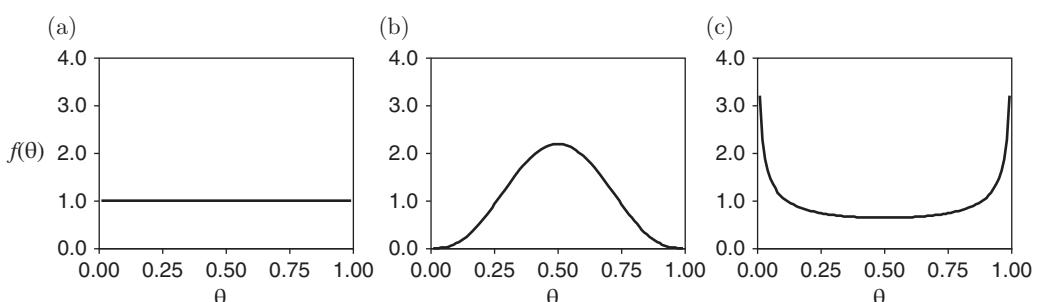


Figure 6.2 The beta distribution can take a variety of shapes depending on the values of the parameters α and β . Here, (a) $\alpha = \beta = 1$, (b) $\alpha = \beta = 4$, and (c) $\alpha = \beta = 1/2$.

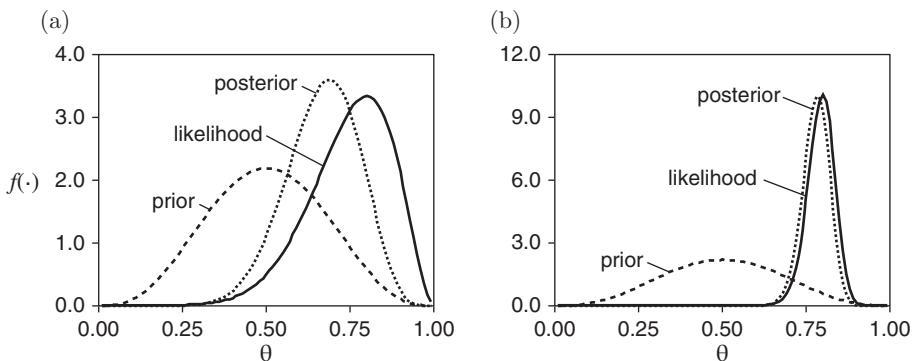


Figure 6.3 As more data are collected for the coin-tossing example, the investigator's prior opinions play a smaller role in the conclusions. (a) The prior distribution, likelihood function, and posterior probability density when $\alpha = \beta = 4$, $n = 10$ and $x = 8$. (b) The prior distribution, likelihood function, and posterior probability density when $\alpha = \beta = 4$, $n = 100$ and $x = 80$.

Besides being flexible, the beta prior probability distribution has one other admirable property: when combined with a binomial likelihood, the posterior distribution also has a beta probability distribution (but with the parameters changed). Prior distributions that have this property – that is, the posterior probability distribution has the same functional form as the prior distribution – are called conjugate priors in the Bayesian literature. The Bayesian treatment of the coin tossing experiment can be summarized as follows:

Prior $[f(\theta), \text{beta}]$	Likelihood $[f(x \theta), \text{binomial}]$	Posterior $[f(\theta x), \text{beta}]$
$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$	$\binom{n}{x} \theta^x (1-\theta)^{n-x}$	$\frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + x)\Gamma(\beta + n - x)} \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1}$

(The gamma function, not to be confused with the gamma probability distribution, is defined as $\Gamma(y) = \int_0^\infty u^{y-1} e^{-u} du$, $\Gamma(n) = (n-1)!$ for integer $n = 1, 2, 3, \dots$, and $\Gamma(\frac{1}{2}) = \pi$.) We started with a beta prior distribution with parameters α and β . We used a binomial likelihood, and after a modest amount of calculus we calculated the posterior probability distribution, which is also a beta distribution but with parameters $\alpha + x$ and $\beta + n - x$.

Figure 6.3 shows the relationship between the prior probability distribution, likelihood, and posterior probability distribution of θ for two different cases, differing only in the number of coin tosses. The posterior probability of a parameter is a compromise between the prior probability and the likelihood. When the number of observations is small, as is the case for Figure 6.3(a), the posterior probability distribution is similar to the prior distribution. However, when the number of observations is large, as is the case for Figure 6.3(b), the posterior probability distribution is dominated by the likelihood.

Bayesian statisticians base inferences upon the posterior probability distribution of a parameter, but are rather agnostic concerning ways to summarize the posterior probability. One way to make a point estimate is to take the mean of the posterior distribution as the best estimate of the parameter. The mean of the posterior probability distribution for the coin-tossing example can be written as

$$\left(\frac{n}{\alpha + \beta + n} \right) \frac{x}{n} + \left(\frac{\alpha + \beta}{\alpha + \beta + n} \right) \frac{\alpha}{\alpha + \beta}. \quad (6.10)$$

Hence, the point estimate is a weighted average of the maximum likelihood estimate (x/n) and the mean of the prior probability distribution ($\alpha/(\alpha + \beta)$), with the respective weights being $n/(\alpha + \beta + n)$ and $(\alpha + \beta)/(\alpha + \beta + n)$. This drives home the point that the posterior probability is a compromise between the prior distribution and the likelihood. As more information is added (i.e. as n increases) the prior distribution has a smaller effect on the inferences about the parameter θ . In fact, the prior probability distribution only counts as $\alpha + \beta$ observations.

6.3 Choosing among Models Using Likelihood Ratio Tests and Bayes Factors

To be able to perform likelihood-based inference of phylogenies, whether frequentist or Bayesian, a model that describes how the characters, such as nucleotides in DNA, evolve along the branches of the phylogenetic tree is needed. Such models are called substitution models. Unfortunately, the researcher new to the field is confronted by a dizzying number of obscurely named models, such as the JC69, K80, HKY85, TN93, GTR, and N98, to name a few. We will return to how these models are defined later, but for now we will focus on a specific related problem: how do we choose between them? Importantly, choosing among the various models is straightforward if one adopts maximum likelihood or Bayesian analysis as the framework for phylogenetic analysis.

In general, model choice involves a tradeoff between explanatory power and model complexity. A complex model with many parameters will do a better job of fitting the data than a simpler, less parameter-rich, model but will do so at the risk of introducing superfluous parameters. Methods for choosing among models penalize complex models in some way. The penalty allows a less parameter-rich model a fighting chance to be the best-fitting model for a particular data set.

Likelihood ratio tests compare the maximum likelihoods obtained under two competing models,

$$\Lambda = \frac{\max_{\theta_0} \mathcal{L}(\theta_0 | M_0)}{\max_{\theta_1} \mathcal{L}(\theta_1 | M_1)} \quad (6.11)$$

where θ_0 and θ_1 are the parameters of models M_0 and M_1 , respectively. Model M_0 is favored when $\Lambda > 1$, whereas the opposite is true for $\Lambda < 1$. The null distribution for Λ can be computed using computer simulation under one of the models (the model denoted as the null model; Goldman, 1993). In the special case in which M_0 is nested within M_1 , the statistic $-2 \ln \Lambda$ asymptotically follows a χ^2 distribution, with the degrees of freedom of the χ^2 distribution being the difference in the number of free parameters between M_1 and M_0 . A model is nested in another model when it can be specified by restricting one or more of the parameters in the more parameter-rich model. For the coin-tossing example, one could consider two models: M_0 in which $\theta = 0.5$ and M_1 in which $0 \leq \theta \leq 1$. Here, M_0 is a special case of M_1 with the parameter θ set equal to 0.5. Although nesting of models appears to be an unusual situation, many of the models used in phylogenetics exhibit nesting behavior. For a more complex model to be considered better using likelihood ratio tests, the likelihood ratio test statistic, $-2 \ln \Lambda$, must exceed a predetermined critical value from the χ^2 distribution. In the case where the two models differ by one free parameter, as would be the case for test of a fair coin, there is 1 degrees of freedom and the test statistic must exceed 3.841 (for a test at the 95% level).

The Akaike information criterion (AIC) represents another method for choosing among a set of models (Akaike, 1973). The i th model, M_i , has K_i parameters. For each model, we calculate its AIC value,

$$AIC_i = -2 \ln \left[\max_{\theta_i} \mathcal{L}(\theta_i | M_i) \right] + 2K_i. \quad (6.12)$$

The model with the minimum AIC value is chosen as the best from the set of models. The penalty for model complexity is explicit in the formula for the AIC. Moreover, use of the AIC does not require nesting of models.

In a Bayesian framework, model choice often relies on the Bayes factor (BF), which is the ratio of the marginal likelihoods of two models,

$$BF = \frac{f(\mathbf{X} | M_1)}{f(\mathbf{X} | M_2)}. \quad (6.13)$$

The marginal likelihood for the i th model,

$$f(\mathbf{X} | M_i) = \int_{\theta_i} f(\mathbf{X} | \theta_i) f(\theta_i) d\theta_i, \quad (6.14)$$

involves integration over the parameters of the model, θ_i . When the $BF > 1$, model M_1 is favored; the opposite is true for $BF < 1$. The penalty for model complexity is built into the Bayes factor calculation. More complex models involve summing and/or integrating over a larger parameter space. The prior probability of any particular combination of parameter values is lower for a complex model than it is for a simpler model. By analogy, the butter from a butter pat will be more thickly spread over a small slice of bread than it will be if spread over a larger slice of bread. Similarly, prior probability mass must be distributed over a larger space for a complex model. Kass and Raftery (1995) provided a table to interpret values of the Bayes factors, measured on a log scale:

$\ln BF$	Interpretation
0–1	Evidence against M_2 is hardly worth mentioning
1–3	Positive evidence against M_2
3–5	Strong evidence against M_2
> 5	Very strong evidence against M_2

Interpreting Bayes factors seems arbitrary. However, the reader should keep in mind that choosing cutoff critical values from a null distribution also involves making an arbitrary decision.

Computing the marginal likelihood for a model can be computationally expensive. The state-of-the art methods for computing the marginal likelihood, such as thermodynamic integration (Lartillot and Philippe, 2006) or stepping-stone integration (Xie *et al.*, 2011), involve running many independent Markov chain Monte Carlo (MCMC) analyses. (MCMC will be described in more detail later in this chapter. You can also read a description of the method in **Chapter 1**.) Programs such as RevBayes implement the path sampling methods for approximating the marginal likelihood.

6.4 Calculating the Likelihood for a Phylogenetic Model

Now we have the basic background to discuss likelihood-based inference in phylogenetics. To this end, in this section I will describe how to calculate the likelihood for a phylogenetic model. Specifically, I will first describe what the observations and the model are in a phylogenetic analysis. Then with this as basis, I will explain how the likelihood can be conceived using character histories. Finally, I will describe a method that is universally used to speed up the likelihood calculations, making maximum likelihood and Bayesian estimation feasible for real-world problems.

6.4.1 Character Matrices and Alignments

For the phylogeny problem, the observations are taken to be a character matrix. For DNA sequence data, the character matrix is called an ‘alignment’, which I will denote \mathbf{X} . As an example of aligned sequences, consider the first and last 15 aligned sites of the *replicase* gene from nine bacteriophage species of the family Leviviridae:

PP7	GACAGC--CGGUUC...CGGAUCCUGACACG
FR	GGCAACGGU---GUG...GCAGACCCACGCCUC
MS2	GGGAACGGA---GUG...UCAGAUCCACGCCUC
GA	GGCAACGGU---UUG...UCAGAUCCGCGACUC
SP	UCA---AAUAAAAGCA...UGGGAUCCUAGAGCA
NL95	UCG---AAUAAAAGCA...UGGGAUCCUAGGGUA
M11	CCUUUCAAUAAAAGCA...UGGGAUCCUAGGGUA
MX1	CCUUUCAAUAAAAGCA...UGGGAUCCUAGGGUA
Q β	CCUUUUAAAAGCA...UGGGAUCCUAGGGCC

The core region of the replicase gene corresponds to amino acid residue numbers 205 to 443, referenced to Q β (GenBank accession numbers: FR, X15031; MS2, J02467; GA, X03869; SP, X07489; NL95, AF059243; M11, AF052431; MX1, AF059242; Q β , X14764; PP7, X80191; Bollback and Huelsenbeck, 2001). Dashes in the alignment denote insertion or deletions in the sequence and are referred to as ‘indels’ or ‘gaps’. The phage PP7 is treated as the outgroup.

The individual observations are the sites (i.e. alignment columns). For example, the first observation in the bacteriophage matrix is the site $\mathbf{x}_1 = (G, G, G, G, U, U, C, C, C)^T$. There are a total of $c = 720$ sites in this example data set.

6.4.2 The Phylogenetic Model

The likelihood is calculated under a phylogenetic model consisting of two parts: a tree (consisting of a topology and set of branch lengths) and a continuous-time Markov model that describes how the characters, in this case DNA sequences, change along the branches of the tree. (We will call the continuous-time Markov model a ‘substitution model’.)

The tree topology, τ , describes the relationships among the species. We imagine that every possible tree topology is labeled, $\tau_1, \tau_2, \dots, \tau_{B(N)}$, where $B(N)$ is the number of trees that are possible for N species. For unrooted trees, this number is the product of the odd numbers up to $2N - 5$,

$$B(N) = 1 \times 3 \times \dots \times (2N - 5) = (2N - 5)!! , \quad (6.15)$$

whereas the number of possible rooted trees is the product of the odd numbers up to $(2N - 3)$. Note that the number of possible trees can become very large for even a moderately sized phylogenetic problem. An unrooted tree has $\mathcal{V}(N) = 2N - 3$ branches whereas a rooted tree has $\mathcal{V}(N) = 2N - 2$ branches.

Each tree has an associated set of branch lengths. Ideally, the branch lengths for a tree would be in terms of time units, such as millions of years, in which case the branch lengths of the i th tree would be $\mathbf{t}_i = (t_1, t_2, \dots, t_{\mathcal{V}(N)})$. Usually, however, branch lengths are not in terms of time, but rather in terms of the expected number of character changes per character (e.g. the expected number of substitutions per site for nucleotide alignments), which is the product of the rate of substitution (u) and time, $v = ut$. The set of branch lengths for the i th tree is then designated $\mathbf{v}_i = (v_1, v_2, \dots, v_{\mathcal{V}(N)})$. The joint tree topology with associated branch lengths is often denoted $\psi_i = (\tau_i, \mathbf{v}_i)$.

The probability of the alignment, conditioned on a specific tree topology and set of branch lengths, is the product of the site probabilities,

$$\mathcal{L}(\psi_i, \theta) = f(\mathbf{X} | \psi_i, \theta) = \prod_{j=1}^c f(\mathbf{x}_j | \psi_i, \theta) \quad (6.16)$$

which is also referred to as the likelihood. This formula for the likelihood assumes that substitutions are independent of one another at different sites. Later, I will describe how the independence assumption can be relaxed for different biological situations. The likelihood also depends on a number of parameters of the substitution model, which for convenience have all been contained in the vector θ . I will return to those and the substitution model itself in Section 6.4.4. But before that, let us take a look at how to calculate the probability of a character history, which is key to understanding how to calculate the site probabilities $f(\mathbf{X} | \psi_i, \theta)$ and thus in turn $\mathcal{L}(\psi_i, \theta)$. Also it motivates the need for a substitution model.

6.4.3 Calculating the Probability of a Character History

The site probabilities for likelihood-based phylogenetic methods account for all of the ways one could observe the data at the tips of the tree. Figure 6.4 shows several examples of character histories for a site in which the observations at the tips of the tree are $\mathbf{x} = (G, C, A, A)^T$. Note that a character history includes not only the changes on the tree, but also the times of the changes; two histories are different if the times of the changes are different, even if they both involve

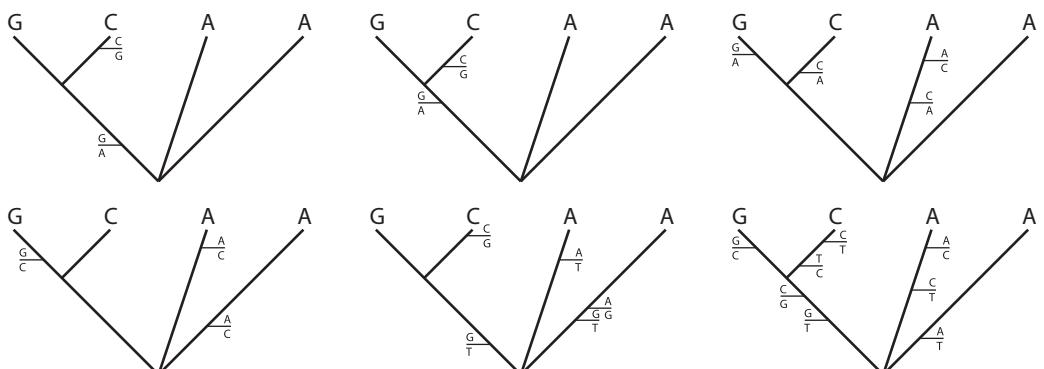


Figure 6.4 A few of the possible character histories that are concordant with the observations at the tips of the tree.

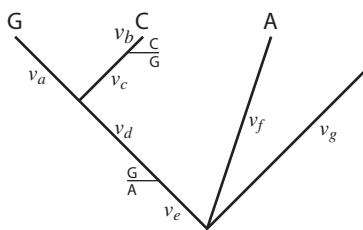


Figure 6.5 A single character history.

changes of the same type along the same branches. (The first two character histories shown in Figure 6.4 have changes of the same type occurring on the same branches, but at different times.) There are an infinite number of character histories that could explain the observations at the tips of the tree.

How can the probability of a character history be calculated? Figure 6.5 shows the first character history in Figure 6.4. The lengths of the intervals between nodes and substitutions are labeled v_a, v_b, \dots, v_g . These lengths are in terms of expected number of substitutions per site. The probability density of the character history shown in Figure 6.5 is

$$\pi_A \times e^{q_{AA}v_e} \times q_{AA}\Delta v \times \left(-\frac{q_{AG}}{q_{AA}}\right) \times e^{q_{GG}v_d} \times e^{q_{GG}v_a} \times e^{q_{GG}v_c} \times q_{GG}\Delta v \times \left(-\frac{q_{GC}}{q_{GG}}\right) \\ \times e^{q_{CC}v_b} \times e^{q_{AA}v_f} \times e^{q_{AA}v_g},$$

which requires some deconstruction: the factor π_A is the probability of starting the process in nucleotide A at the root of the tree; the exponential factors, $e^{q_{ii}v}$, are the probability of waiting v without having a change when the rate of change is $-q_{ii}$; the probability density of a change occurring exactly at a position v units up the branch is $q_{ii}\Delta v$; and the probability of the change from state i to state j is $-q_{ij}/q_{ii}$.

6.4.4 Continuous-Time Markov Model

For a model of nucleotide substitution, there are a total of 12 rates to consider ($q_{AC}, q_{AG}, q_{AT}, q_{CA}, q_{CG}, q_{CT}, q_{GA}, q_{GC}, q_{GT}, q_{TA}, q_{TC}$, and q_{TG}) which are the rates of change between all pairs of nucleotide states. These rates of change are contained in a ‘rate matrix’,

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} q_{AA} & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & q_{CC} & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & q_{GG} & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & q_{TT} \end{pmatrix},$$

where the diagonal elements of the matrix (q_{AA}, q_{CC}, q_{GG} , and q_{TT}) are specified such that each row sums to zero. So, for example, $q_{AA} = -(q_{AC} + q_{AG} + q_{AT})$. The off-diagonal elements are all greater than or equal to zero, whereas the diagonal elements are all negative. The rate matrix can be considered the heart of the continuous-time Markov model for discrete-character likelihood calculations. (Different types of models, such as Brownian motion models, are often used for continuous traits.)

The rate matrix allows for a clear mechanistic view of the substitution process. When in state i , the process waits an exponentially distributed time until the next substitution occurs. The parameter of the exponential is $-q_{ii}$. When a change occurs, it is to state j with probability $-\frac{q_{ij}}{q_{ii}}$.

Two important quantities can be calculated from the rate matrix of the continuous-time Markov chain: the transition probabilities and the stationary probability distribution. The transition probability, $p_{ij}(v)$, is the probability of changing to nucleotide j conditioned on starting in nucleotide i over a branch of length v . The matrix of transition probabilities can be calculated using matrix exponentiation,

$$\mathbf{P}(v) = \{p_{ij}(v)\} = e^{\mathbf{Q}v}. \quad (6.17)$$

Importantly, the transition probability accounts for all of the ways the process can end in nucleotide j after starting in nucleotide i over a branch of length v , and thus allows for the change to happen via more than one substitution. The stationary probability is the probability of capturing the process in state i after a very long time has elapsed. (Formally, after an infinite time has elapsed.) The stationary probability for nucleotide i is usually denoted π_i . The stationary probabilities are a property of the rate matrix. Different rate matrices will usually have different stationary probabilities. The set of stationary probabilities for a specific rate matrix can be calculated by solving the equation $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$. Note that the stationary probabilities enter the likelihood calculation as a factor for the probability of the nucleotide at the root of the tree. This is the probability of starting the process off in a particular state. You should imagine that the process of nucleotide substitution has been operating for a very long time by the time it reaches the root of the tree of interest. (One could imagine a very long branch attached to the root of the tree. One could start the process in any nucleotide. If the root branch is long enough, then the probability that the process is in state i by the time it reaches the root of the tree of interest is π_i .) For now, I will assume that a substitution model, specified through a rate matrix \mathbf{Q} , has been specified, and thus that transition probabilities can be calculated. For more details on specific substitution models, see Section 6.6.1.

6.4.5 Marginalizing over Character Histories

Character histories cannot be directly observed; there are an infinite number of them; and it is disquieting to condition inferences on any one of them, or even a class of them such as those requiring the minimum number of changes, because the probability of any single history seems small. The standard approach in statistics for dealing with situations in which the parameter is not of direct interest and cannot be directly observed is to marginalize (sum and/or integrate the probabilities) over all of the possibilities. In this case, our probability should account for all of the possible character histories. Remarkably, we can accomplish the seemingly impossible task of summing the probabilities of the infinite number of character histories that are concordant with the observations for a site. The task requires two key calculations: the transition probabilities over the branches; and the Felsenstein pruning algorithm for summing over the possible assignments of nucleotides to the interior nodes of the tree.

The Felsenstein pruning algorithm solves an important problem: how to account for the different possible nucleotides at the internal nodes of a tree? Consider the tree shown in Figure 6.6.

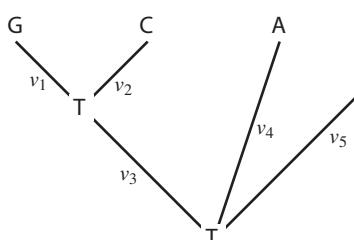


Figure 6.6 A tree with the observations at the tips ($\mathbf{x} = (G, C, A, A)^T$) as well as proposed states at the interior nodes ($\mathbf{y} = (T, T)^T$).

Along with the observations at the tips of the tree, this tree also assumes that the interior nodes of the tree both have the nucleotide T. We denote the tip states, in this case, as $\mathbf{x} = (G, C, A, A)'$ and the states for the interior nodes as $\mathbf{y} = (T, T)'$. The probability of this configuration is

$$f(\mathbf{x}, \mathbf{y} | \psi, \theta) = \pi_T \times p_{TT}(v_3) \times p_{TG}(v_1) \times p_{TC}(v_2) \times p_{TA}(v_4) \times p_{TA}(v_5). \quad (6.18)$$

This equation looks different from the one shown when we knew the character history. Here, we account for all of the possible ways we can have changes along the branches – the transition probabilities accomplish this – but condition on character histories having the nucleotides T at both of the internal nodes. Clearly, we do not want to condition our probability calculations on the (unattainable) knowledge of the ancestral condition at the interior nodes of the tree. To account for all of the possible ways we could have observed the states at the tips of the tree, we sum over all possible configurations of assignments of nucleotides at the interior nodes of the tree, which for the tree of Figure 6.6 is

$$f(\mathbf{x} | \psi, \theta) = \sum_{i \in (A, C, G, T)} \sum_{j \in (A, C, G, T)} \pi_i \times p_{ij}(v_3) \times p_{jG}(v_1) \times p_{jC}(v_2) \times p_{iA}(v_4) \times p_{iA}(v_5) \quad (6.19)$$

(a sum over the $4^2 = 16$ possible configurations). Note that the summation over all possible nucleotides can become quite cumbersome. In general, an unrooted tree of N tips will have $N - 2$ interior nodes and involve summing over 4^{N-2} possible configurations. So, the site likelihood for a tree of $N = 100$ tips will involve a sum with $4^{98} = 1.0043 \times 10^{59}$ terms. Even if a computer could evaluate a billion of these terms every second, it would still take 3.18×10^{42} years to calculate the likelihood for a single site! The reader is probably aware that the likelihood for phylogenetic problems consisting of hundreds of taxa is possible; after all, computer programs compute likelihoods on such large problems every day. Clearly, these programs are doing something other than explicitly enumerating all of the possible combinations of nucleotide assignments to the interior nodes of the tree. What are these programs doing? The answer: these programs use the Felsenstein pruning algorithm to compute the likelihood.

The Felsenstein (1981) pruning algorithm allows the summation over all possible configurations at the interior nodes of the tree to be computed in linear time. The idea is to compute the probability of the observations above specific nodes on the tree. Starting with the tips of the tree, where this probability is easy to compute, one moves down the tree calculating the probability of the observations of successively bigger parts of the tree. When the root of the tree is reached, the algorithm has computed the probability of all of the observations.

Figure 6.7 shows an example of how computations are performed on a tree of $N = 5$ species. Some time will be taken to get the notation right for this tree.

1. This tree has a set of branch lengths denoted v_1, v_2, \dots, v_7 .
2. The nodes on this tree are labeled 1–8 according to a scheme: the node label for the i th branch with length v_i is i . Hence, the tip nodes are labeled 1, 2, 3, 4, 5 and the interior nodes have the labels 6, 7, 8. The root node is labeled 8.
3. We will denote the descendants of node i as the set $\sigma(i)$. For example, the set of descendants for node 6 is $\sigma(6) = (2, 3)$, the set of descendants for node 8 is $\sigma(8) = (7, 4, 5)$, and the set of descendants for node 2, as well as for the other tip nodes, is the empty set, $\sigma(2) = (\emptyset)$.
4. The four boxes at each node in Figure 6.7 contain the four conditional probabilities of the data above the node. The probabilities are conditioned on the state at that node being a particular nucleotide (the nucleotides are in alphabetical order). So, for example, the conditional probabilities assigned to node 1, which has the nucleotide T assigned, are $\varepsilon^{(1)} = (0, 0, 0, 1)$.

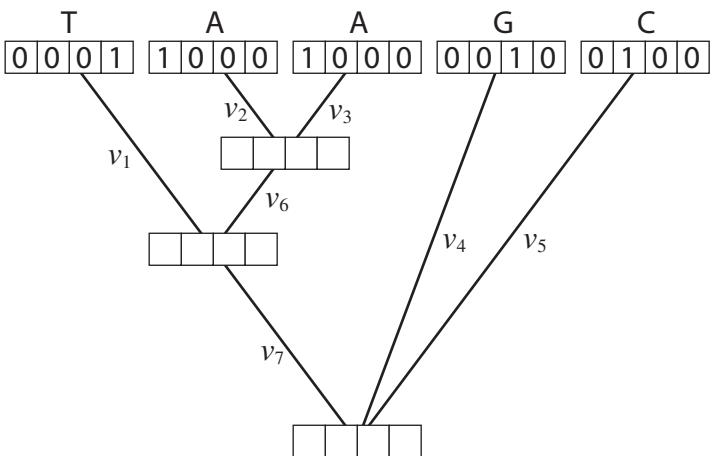


Figure 6.7 A tree of $N = 5$ species with the information for the site at the tips of the tree being $\mathbf{x} = (T, A, A, G, C)'$. The branch lengths of the tree are v_1, v_2, \dots, v_7 . The node above the i th branch (with branch length v_i) is labelled i . The root node is labeled $2N - 2$.

The rule for initializing the tip conditional likelihoods is to assign a probability of 1 to the conditional likelihood corresponding to the observation and 0 otherwise.

The Felsenstein pruning algorithm begins by initializing the conditional probabilities (called ‘conditional likelihoods’) at the tips of the tree. After the conditional likelihoods are initialized, the pruning algorithm visits the interior nodes of the tree in such a way that the conditional likelihoods of the descendant nodes have already been calculated. The conditional likelihood for nucleotide i at node k is

$$\ell_i^{(k)} = \prod_{n \in \sigma(k)} \left(\sum_{j \in (A, C, G, T)} p_{ij}(v_n) \ell_j^{(n)} \right). \quad (6.20)$$

(The superscript only marks which node the conditional likelihood corresponds to.) This equation represents the vast majority of time computer programs take when performing maximum likelihood or Bayesian phylogenetic analyses. The equation also makes clear why the Felsenstein pruning algorithm is referred to as the ‘sum-product algorithm’ in the graphical models literature (Gallager, 1962, 1963); the conditional likelihood is a product of sums.

The conditional likelihoods computed at the root of the tree represent the probability of all the observations at the site, conditional on the root node being in a particular nucleotide state. The final step for computing the site likelihood is to perform a weighted average of the conditional likelihoods at the root of the tree,

$$f(\mathbf{x} | \psi, \theta) = \sum_{i \in (A, C, G, T)} \pi_i \ell_i^{(\text{Root})}, \quad (6.21)$$

with the weights being the stationary probabilities.

Importantly, by combining matrix exponentiation to calculate the transition probabilities with the Felsenstein pruning algorithm, the site likelihoods account for all of the possible ways substitutions could have occurred in the past and led to the current observations.

6.5 The Mechanics of Maximum Likelihood and Bayesian Inference

6.5.1 Maximum Likelihood

Computer programs that implement the maximum likelihood method of phylogenetic inference, such as PHYLP, PAUP*, Garli, and RaxML (Felsenstein, 1993; Stamatakis *et al.*, 2005; Swofford, 1998; Zwickl, 2006), face the formidable task of finding the combination of model parameters (the tree topology with associated branch lengths and substitution model parameters) that maximizes the likelihood. A variety of approaches has been taken. In fact, an entire chapter of this volume could easily have been dedicated to a discussion of the variety of methods used to find optimal trees in phylogenetics. Here, I will discuss only a few of them to give the reader an idea of how tree searches work. For the ‘ordinary’ parameters, such as the branch lengths and substitution parameters, optimization methods that take advantage of the slope and curvature of the likelihood surface as well as derivative-free methods have both been successfully used. The tree topology is typically more difficult to optimize. The topology of a tree cannot be ordered as an ordinary parameter can be ordered. (As an example, it is easy to order people by their heights. Trees, on the other hand, have no such natural way to be ordered because there is no sense in which one tree is greater or less than another.) The general approach in phylogenetics is to impose an order on trees by defining a tree perturbation. The space of trees, then, can be visualized as a graph with the vertices of the graph being the trees and the edges of the graph showing trees that are one perturbation away from each other (Charleston, 1995). Trees that are one perturbation away are considered ‘neighbors’. The most common perturbations that are used are nearest-neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR). Figure 6.8 shows how NNI orders trees.

The general idea is to start with a good tree and explore all the neighbors of this tree (as defined under the perturbation, such as TBR). If one of the neighboring trees has a higher likelihood than the current tree, then the algorithm moves to that tree. The optimization algorithm continues to search for neighbors with better likelihoods until a tree is found in which all of the neighboring trees have lower likelihoods. That tree is considered to be the maximum likelihood tree. The reader should be aware, however, that there may be other trees that have a higher likelihood, in which case the tree that was found is a local optimum, not the global optimum. Typically, tree-search algorithms start from multiple different trees, with the idea that if the same optimal tree is found that it is more likely to be the tree with the highest likelihood (i.e. a global optimum).

Figure 6.9 shows the maximum likelihood tree of nine bacteriophage species based on the *replicase* gene alignment discussed earlier. The ML topology was found using the TBR perturbation method. The details of the substitution model that was used will be described in later sections. However, the model used in the analysis allows for a transition/transversion rate bias, differences in nucleotide composition, and for rate variation across sites.

Numerous methods can be used to establish confidence intervals for the ‘ordinary’ parameters of the model. These methods take advantage of information on the curvature of the likelihood surface around the maximum likelihood value(s). The bootstrap method is a nonparametric method for constructing confidence intervals (Efron, 1979) that has been successfully used to evaluate the uncertainty of phylogenetic trees (Felsenstein, 1985). The bootstrap method works by creating new ‘bootstrapped’ alignments of the same size as the original. Each bootstrap alignment is created by randomly sampling the columns (sites) of the original alignment with replacement. By chance, some sites will be represented multiple times in the bootstrap alignments whereas other sites may not be represented. Each bootstrap alignment is analyzed in the same way the original alignment was analyzed (e.g. using maximum likelihood if the

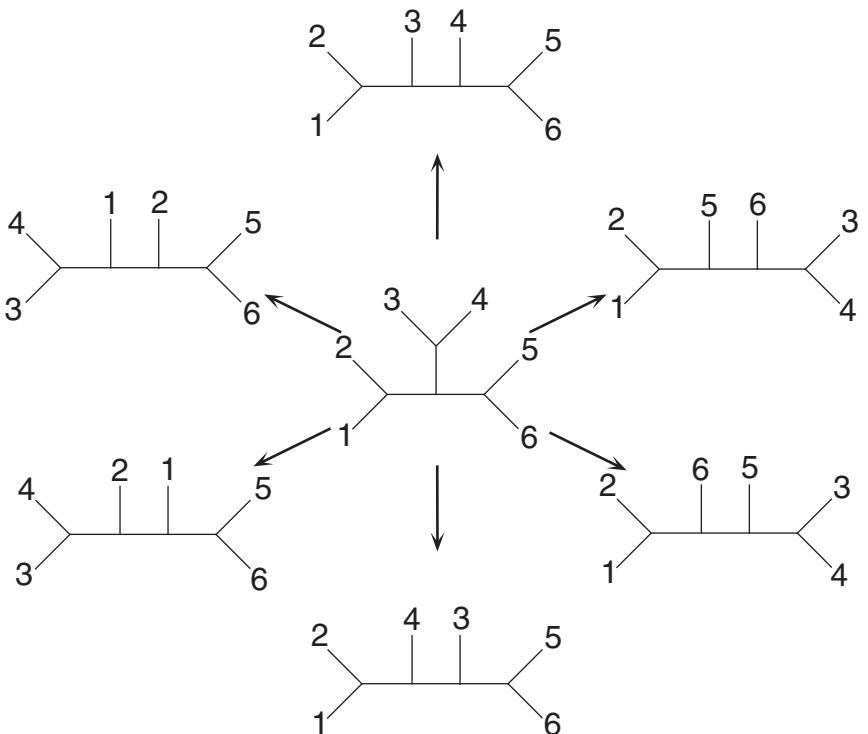


Figure 6.8 The neighboring trees under the NNI perturbation. Each of the six neighbors to the central tree also has six neighbors.

original analysis used maximum likelihood). The tree of Figure 6.9 is not only the maximum likelihood tree, but also the majority-rule tree summarizing the trees obtained from 1000 bootstrap analyses of the *replicase* gene alignment. The numbers on the branches of the tree represent the fraction of the time that each clade was found in analysis of the bootstrapped alignments. Although it is tempting to interpret the bootstrap proportions as probabilities, they are not the probability of the clade conditioned on the data (Hillis and Bull, 1993); the only way to compute such a quantity is to use Bayes' formula, which means the adoption of a prior probability distribution for the parameters of the model. The bootstrap can also be used to formulate confidence intervals for the other parameters of the phylogenetic model. Figure 6.10 shows the sampling distributions of the transition/transversion rate parameter (κ) and the shape parameter of the gamma distribution for among-site rate variation (α).

6.5.2 Bayesian Inference and Markov Chain Monte Carlo

The posterior probability of the i th tree, τ_i , is obtained using Bayes' rule,

$$f(\tau_i | \mathbf{X}) = \frac{f(\mathbf{X} | \tau_i) f(\tau_i)}{\sum_{j=1}^{B(N)} f(\mathbf{X} | \tau_j) f(\tau_j)}, \quad (6.22)$$

where the likelihood involves evaluation of a multi-dimensional integral over all possible combinations of branch lengths (ν) and substitution model parameters (θ),

$$f(\mathbf{X} | \tau_i) = \int f(\mathbf{X} | \tau_i, \nu, \theta) f(\nu, \theta) d\nu d\theta. \quad (6.23)$$

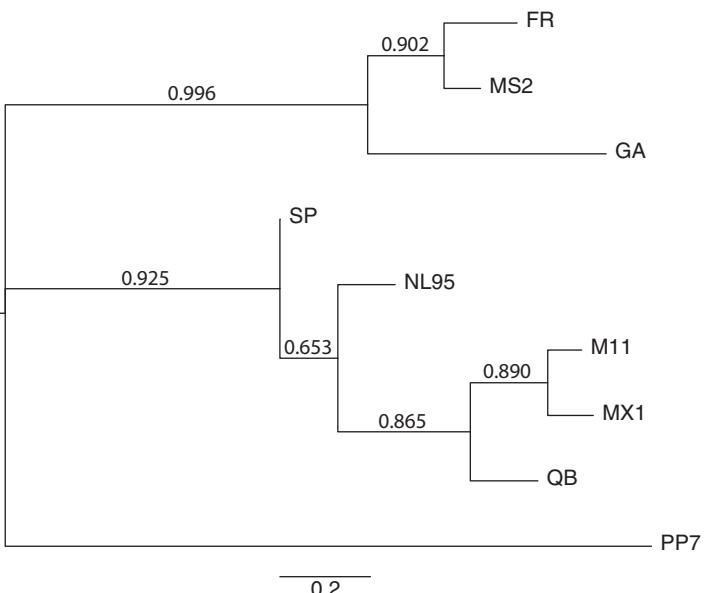


Figure 6.9 The maximum likelihood estimate of phylogeny for the *replicase* alignment. The branch lengths are drawn in proportion to their estimated (via maximum likelihood) lengths. The bootstrap proportions on the interior branches of the tree are based on 1000 bootstrap replicates. I assumed that substitutions occurred under the model described by Hasegawa *et al.* (1985) with gamma-distributed rate variation (Yang, 1993, 1994a). The program PAUP* (Swofford, 1998), which was used to find the maximum likelihood tree, also estimates the substitution model parameters using maximum likelihood. The maximum likelihood estimates of the substitution model parameters are: $\hat{\pi}_A = 0.238$, $\hat{\pi}_C = 0.257$, $\hat{\pi}_G = 0.235$, $\hat{\pi}_T = 0.270$, $\hat{\kappa} = 2.33$, and $\hat{\alpha} = 0.691$.

A Bayesian analysis that marginalizes over the parameters is referred to as a hierarchical Bayesian analysis. Priors are placed on all of the parameters of the model. Such an analysis naturally accounts for uncertainty in all of the model parameters. An alternative method substitutes estimates for the parameters. For example, the MLEs for the branch lengths and substitution model parameters might be used. This is called an empirical Bayes analysis. Most Bayesian

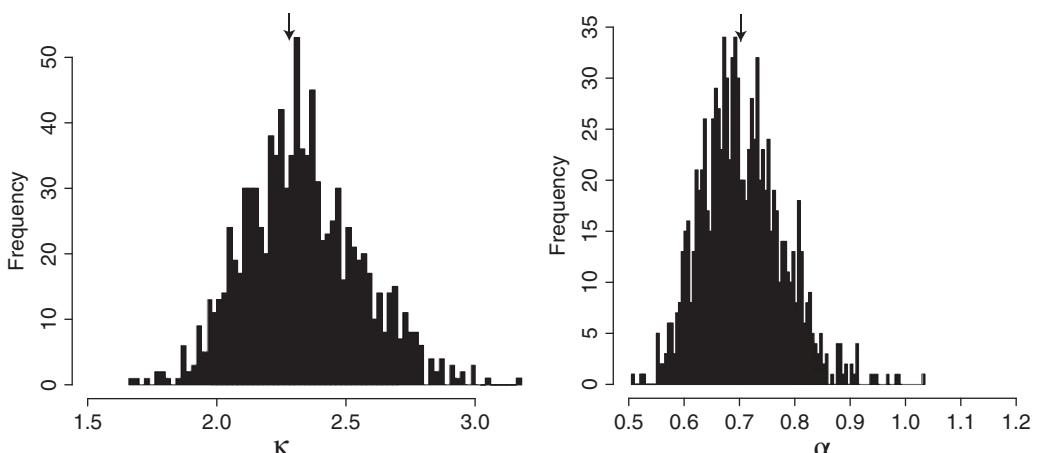


Figure 6.10 The sampling distributions for the transition/transversion rate (κ) and gamma shape (α) parameters created using the bootstrap method. The arrows indicate the MLEs.

analyses in phylogenetics are hierarchical – though the program PAML makes extensive use of empirical Bayesian analysis for detecting evidence of natural selection in protein-coding DNA sequences (Yang, 1997; Yang *et al.*, 2000).

A Bayesian analysis of phylogeny requires that the scientist specify his or her beliefs about the phylogeny before observing any data. Two different types of priors have been used on trees. The first was specified by Rannala and Yang (1996, see also Yang and Rannala, 1997; Thompson, 1975), who used a birth–death process of cladogenesis to specify the prior probabilities of phylogeny and branch lengths. Under a random branching model of cladogenesis, such as the birth–death process, all labeled histories have equal probability. Labeled histories are distinguished from one another not only by their topology but also by the relative speciation times or branch lengths (Edwards, 1970). The birth–death process has two parameters, the speciation and extinction rates. These can be fixed, or priors (called hyperpriors) can be placed on the rates of speciation and extinction. Mau (1996), Mau and Newton (1997), and Newton *et al.* (1999) place uninformative priors on topology and branch lengths. Equal weight is placed on all trees, and uniform priors (from 0 to a large number) are placed on branch lengths. The posterior probability will be mainly determined by the likelihood function when uninformative priors are used.

6.5.2.1 Markov Chain Monte Carlo

The posterior probability of a tree involves, minimally, a summation over all possible trees and integration over branch lengths and other parameters of the substitution model. In short, the posterior probability of a tree cannot be evaluated analytically, and must be approximated. Markov chain Monte Carlo (MCMC; Metropolis *et al.*, 1953; Hastings, 1970) is the method of choice for approximating the posterior probability of phylogenies (and other model parameters).

MCMC is a method that makes valid, but dependent, draws from the probability distribution of interest. The general idea is to construct a Markov chain that has as its state space the parameter(s) of interest and a stationary distribution which is the posterior probability distribution of the parameters. A simple example will illustrate MCMC as applied to the coin-tossing problem (for a good introduction to MCMC, see Gilks *et al.*, 1996). Of course, for the problem of tossing coins with the goal of estimating the probability of heads, we do not need to perform MCMC as we can calculate the posterior probability analytically. For this problem, we will be able to compare the MCMC approximation to the true posterior probability distribution.

Most programs for phylogeny estimation implement a variant of MCMC called the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Green, 1995). The steps of the Metropolis–Hastings algorithm are as follows:

1. The current state of the chain will be denoted θ . If this is the first generation of the chain, then θ is initialized to be some value, perhaps by drawing the initial value from the prior probability distribution.
2. A new state, θ' , is proposed. For the coin-tossing example, the proposal mechanism will be to pick a uniformly distributed random number in the interval $(\theta - \varepsilon, \theta + \varepsilon)$; this means that the next state of the chain will be similar to the current state (ε might be 0.05, for example, meaning that the proposed state of the chain will be within 0.05 of the current state). The proposal mechanism is largely at the discretion of the programmer, though several rules must be followed. For example, the proposal mechanism must result in an aperiodic and irreducible Markov chain (for a detailed description of these terms, see **Chapter 1**). Moreover, the probability of proposing the new state under the proposal mechanism, $q(\theta \rightarrow \theta')$, and the probability of the imagined reverse move which is never actually made in computer memory,

$q(\theta' \rightarrow \theta)$, must be calculable. This is not normally a problem because the programmer has control over the details of the proposal mechanism.

3. Calculate the probability of accepting the proposed state as the next state of the Markov chain

$$\begin{aligned}
 R &= \min \left(1, \frac{f(\theta' | x) \times q(\theta' \rightarrow \theta)}{f(\theta | x) \times q(\theta \rightarrow \theta')} \right) \\
 &= \min \left(1, \frac{f(x | \theta') f(\theta') / f(x) \times q(\theta' \rightarrow \theta)}{f(x | \theta) f(\theta) / f(x) \times q(\theta \rightarrow \theta')} \right) \\
 &= \min \left(1, \frac{f(x | \theta')}{f(x | \theta)} \times \frac{f(\theta')}{f(\theta)} \times \frac{q(\theta' \rightarrow \theta)}{q(\theta \rightarrow \theta')} \right) \\
 &= \min (\text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}). \tag{6.24}
 \end{aligned}$$

Note that the most difficult part of the posterior probability to calculate, $f(x) = \int_0^1 f(x | \theta) f(\theta) d\theta$, cancels.

4. Generate a uniformly distributed random number on the interval (0, 1) called u . If $u < R$, then accept the proposed state as the next state of the chain and set $\theta = \theta'$. Otherwise, the chain remains in state θ .
5. Return to step 2.

This process of proposing a new state and then accepting or rejecting it is repeated a large number of times. The sequence of states (values of θ) visited over the course of the analysis forms a Markov chain. The states visited during the analysis (*i.e.*, different values of θ) are valid, albeit dependent, draws from the posterior probability; a Markov chain that was run for 100 steps and sampled at every step does not represent 100 independent draws from the posterior probability. Regardless, the proportion of the time that different values of θ were visited is a valid approximation of the posterior probability. Figure 6.11 shows an example of an MCMC analysis for the coin-tossing example where the data are $x = 5$ heads in $n = 10$ tosses of a coin. Note that as the length of the chain is increased, that the MCMC approximation converges to the true (analytically calculated) posterior probability.

Although MCMC was illustrated for a Bayesian analysis of coin tossing, the method can be extended to much more complex problems involving many parameters. Moreover, the method has been applied to maximum likelihood analysis where it has been used to integrate over uncertainty in coalescence histories when estimating population parameters (Kuhner and Felsenstein, 1994; Beerli and Felsenstein, 1999).

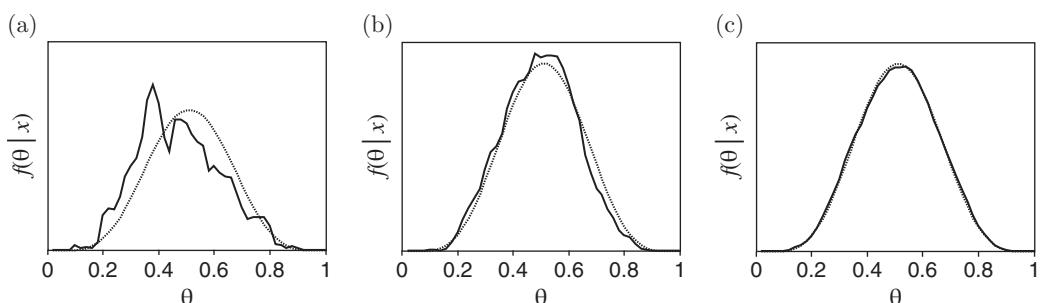


Figure 6.11 The posterior probability of θ for the coin-tossing problem. The dotted line is the posterior probability calculated analytically. The solid line is the approximation of the posterior probability obtained using Markov chain Monte Carlo. The chain was run for (a) 5000 generations, (b) 50,000 generations, and (c) 500,000 generations.

A limitation of MCMC is that it is not clear when the chain has been run long enough to produce a reliable approximation of the distribution of interest. The Markov chain law of large numbers guarantees that posterior probabilities can be validly estimated from long-run samples of the chain (Tierney, 1994). However, for any given analysis, it is difficult to determine if the chain has converged to the desired distribution. Several different heuristics are available to determine if the chain has converged (see Gelman, 1996). These involve running several independent chains starting from different (perhaps random) trees.

Programs that implement MCMC to infer phylogeny, such as MrBayes, Beast, RevBayes, and PhyloBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Ronquist *et al.*, 2012b; Drummond and Rambaut, 2007; Höhna *et al.*, 2016; Lartillot *et al.*, 2009), implement numerous proposal mechanisms to update the parameters of the phylogenetic model. The parameters are not updated all at once. Rather, the programs typically choose one, or a few, of the parameters to update in each cycle of the MCMC. A full description of the proposal mechanisms that are used is beyond the scope of this chapter. However, as with maximum likelihood, the tree topology and associated branch lengths are the most difficult parameters to sum/integrate over. Programs use stochastic variants of the tree perturbations, such as NNI, SPR, or TBR, to update the tree topology and branch lengths.

To illustrate Bayesian inference of phylogeny using MCMC, we analyzed the bacteriophage *replicase* data under a commonly used model of DNA substitution (the HKY85 + Γ model; Hasegawa *et al.*, 1984, 1985; Yang, 1994b). We used MCMC to approximate the posterior probabilities of trees using the program MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003; Ronquist *et al.*, 2012b). The Markov chain was run for 1 million iterations. Figure 6.12 shows the log of the likelihood function through time for the chain. Note that at first the log-likelihood was low and that it quickly reached a plateau near the maximum likelihood value (indicated by a gray horizontal line). The initial state chosen for the chain poorly explained the data because the starting values were chosen at random from the prior probability distribution. Inferences should be based on the portion of the chain at stationarity. Hence, we discarded the first 100,000 steps in the chain as ‘burn-in’. Figure 6.13 shows the six most probable trees that were found. The first two trees account for approximately 0.983 probability. The remaining 0.017 posterior probability is distributed among the remaining 135,133 trees. Typically, the posterior probabilities of individual trees are not shown, as was done in Figure 6.13. Rather, a summary of the sampled trees is used. Most commonly, a majority-rule consensus tree

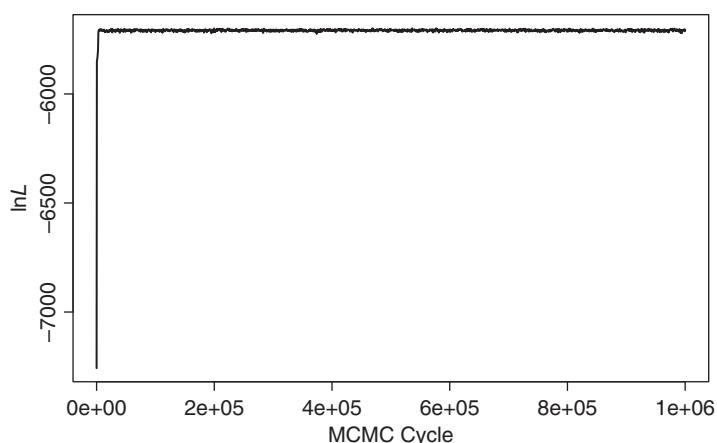


Figure 6.12 The trace of the log-likelihood over time for the MCMC analysis of the *replicase* gene.

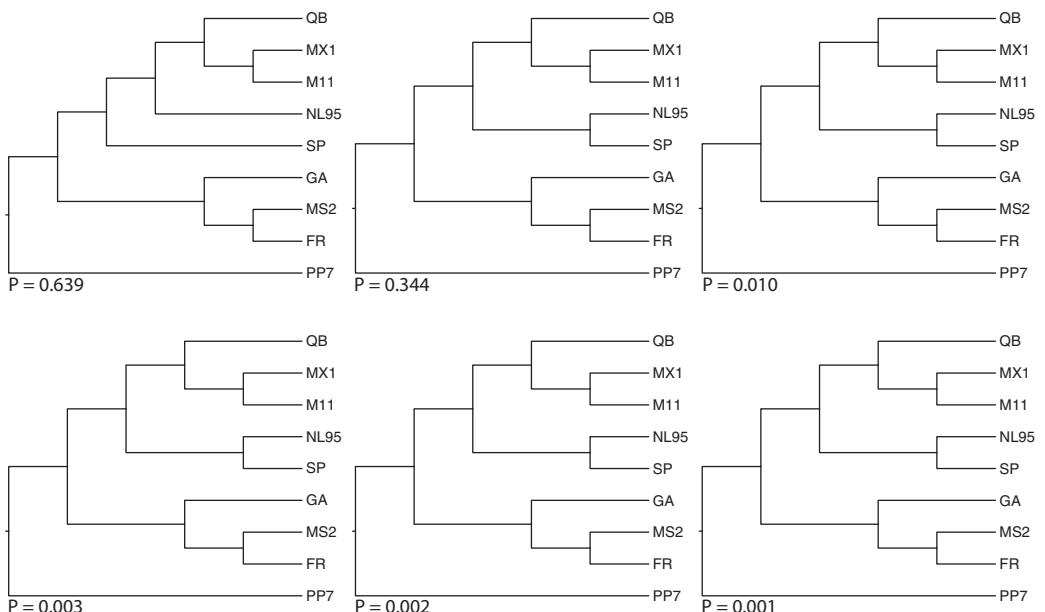


Figure 6.13 The six most probable trees for the replicase alignment. The posterior probability of each tree is shown next to each tree. The first tree, with posterior probability 0.639, is also the majority-rule tree. The posterior probabilities of individual clades are shown on the branches of that tree.

summarizes the results. The first tree in Figure 6.13 is also the majority-rule tree. The posterior probabilities of the individual clades are shown on that tree. The numbers at the interior nodes of the tree do not represent bootstrap support values but rather the posterior probability that the clade is true. Just as with maximum likelihood, the parameters of the substitution model can be estimated using Bayesian inference. Figure 6.14 shows the posterior probability distributions for the transition/transversion rate ratio (κ) and the shape parameter of the gamma distribution (α) for among-site variation.

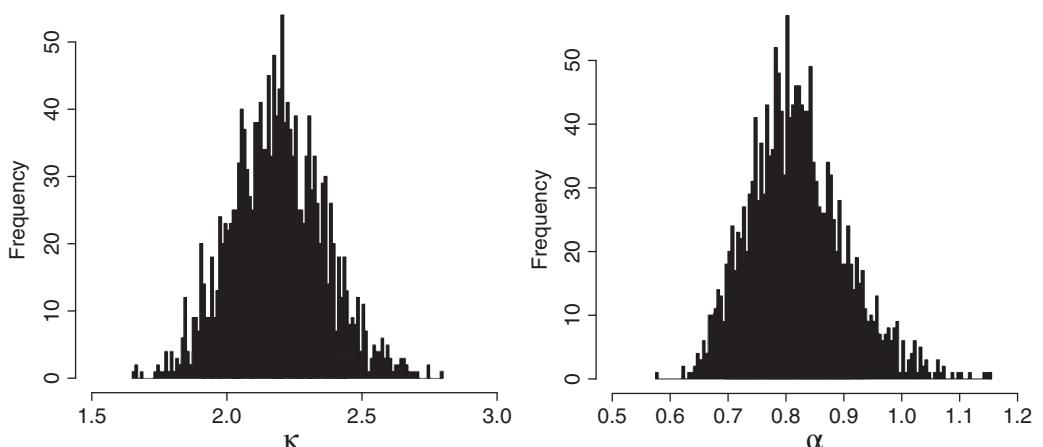


Figure 6.14 The posterior probability distribution of the transition/transversion rate ratio (κ) and the gamma shape parameter for among-site rate variation (α).

6.6 Applications of Likelihood-Based Methods in Molecular Evolution

Now that I have described the basic methodology for estimating trees using likelihood-based methods, I will turn to some of the model details that I have been postponing. More specifically, I will describe some of the substitution models for DNA that are often used in phylogenetic inference. Additionally, I will describe a range of different extensions of the basic methods and models described so far, providing examples of how they can be used to answer evolutionary questions.

6.6.1 A Taxonomy of Commonly Used Substitution Models

A large number of substitution models have been proposed since the late 1960s. The key to understanding these models is to remember that they are all continuous-time Markov models. As previously described, a continuous-time Markov model can be understood through its rate matrix, \mathbf{Q} , which has *states* and *rates* between the states. Here, I will provide the details of several of the best-known models, starting with the model first described by Jukes and Cantor (1969), which is the simplest model of DNA substitution as well as the earliest model proposed. The Jukes and Cantor (1969) model has rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}, \quad (6.25)$$

in which all 12 rates are the same. If the meaning of the rates in the matrix \mathbf{Q} is not clear, consider the table

		To				
		A	C	G	T	
From		A	-1	1/3	1/3	1/3
		C	1/3	-1	1/3	1/3
		G	1/3	1/3	-1	1/3
		T	1/3	1/3	1/3	-1

which contains the rates of the Jukes and Cantor (1969) model in a friendlier form. The states of the Jukes and Cantor (1969) model are the nucleotides A, C, G, and T. The rates between all pairs of states are contained in the rate matrix. Because there are four states for a nucleotide model, there will be a total of $2\binom{4}{2} = 12$ possible rates, which are contained in the off-diagonal portions of the rate matrix. For the Jukes and Cantor (1969) model, all 12 rates are equal.

Note that each row of the rate matrix sums to zero. The (negative) diagonal elements of the rate matrix are determined by the off-diagonal elements. The negative rate can be interpreted as the rate at which the process changes away from the nucleotide. The off-diagonal elements of the rate matrix are all $1/3$, which seems to be a peculiar number. However, in phylogenetics, rate matrices are typically rescaled such that the average rate of change is 1; a value of $1/3$ for all 12 rates accomplishes this for the Jukes and Cantor (1969) model. How are the rates chosen such that the average is 1? One would think that to calculate the average rate of change, one would simply take the average of the off-diagonal elements of the rate matrix. After all, the off-diagonal parts of the rate matrix represent the changes. If one were to do this for the Jukes and Cantor (1969) model, one would obtain an average of $(12 \times \frac{1}{3}) \div 12 = \frac{1}{3}$, which is clearly not 1. However, this naive calculation does not take into account the prior probability of being in each of the four nucleotide states. The correct way to calculate the average rate of change is to realize

that the stationary probabilities of the four states under the Jukes and Cantor (1969) model are $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$. The average rate is a weighted one, with the weights coming from the stationary distribution: $\frac{1}{4} \times \frac{1}{3} \times 12 = 1$.

Why specify the rates of the matrix \mathbf{Q} in such a way that the average rate of substitution is 1? The reason is rather prosaic. Normally, rates and time cannot be disentangled in a phylogenetic analysis. Branch lengths, then, are in terms of expected number of substitutions per site, which is the product of rate and time, $v = u \times t$. The programs that implement maximum likelihood or Bayesian analysis let the rate of substitution be 1 so that the branch lengths on the phylogeny have the intended meaning (expected number of changes per character).

The Jukes and Cantor (1969) model gets little respect among evolutionary biologists. However, even this simple model makes the sensible prediction that sequence divergence should increase with increasing substitution rate and/or time. The weakness of the Jukes and Cantor (1969) model, of course, is its assumption that the rates of change among all pairs of nucleotides are equal. Some of the earliest comparisons between DNA sequences revealed that some sorts of changes occurred more frequently than others. van Ooyen *et al.* (1979), for example, found that transitions (substitutions between A \leftrightarrow G and C \leftrightarrow T) occurred more frequently than the other substitutions, termed transversions. This led to the development of models that allowed transitions to occur at a different rate than transversions (Kimura, 1980; Hasegawa *et al.*, 1984, 1985). The Kimura (1980) model (also known as the 'K80' model), for example, has rate matrix

$$\mathbf{Q} = \begin{pmatrix} -1 & 1/(\kappa + 2) & \kappa/(\kappa + 2) & 1/(\kappa + 2) \\ 1/(\kappa + 2) & -1 & 1/(\kappa + 2) & \kappa/(\kappa + 2) \\ \kappa/(\kappa + 2) & 1/(\kappa + 2) & -1 & 1/(\kappa + 2) \\ 1/(\kappa + 2) & \kappa/(\kappa + 2) & 1/(\kappa + 2) & -1 \end{pmatrix}. \quad (6.26)$$

If $\kappa = 1$, the Kimura (1980) model becomes equivalent to the Jukes and Cantor (1969) model. The Jukes and Cantor (1969) model is nested in (is a special case of) the Kimura (1980) model. The transition/transversion rate parameter, κ , allows the model to fit data in which transitions have a different rate. In fact, Figures 6.10 and 6.14 show the estimates of the transition/transversion rate ratio for the *replicase* alignment. Note that $\kappa \gg 1$ for that gene, whether maximum likelihood or Bayesian inference is used to estimate the parameter, suggesting that the rate bias is a real one. As might be expected, the Kimura (1980) model typically fits data better than the Jukes and Cantor (1969) model when the models are compared using likelihood ratio tests or Bayes factors.

Both the Jukes and Cantor (1969) and Kimura (1980) models predict that the nucleotide frequencies should be approximately equal; the stationary probability distributions for both models are 1/4 for all four nucleotides. In the early days of DNA sequencing, it was noted that the nucleotide frequencies of observed DNA sequences could be quite different from one another. Clearly, both the Jukes and Cantor (1969) and Kimura (1980) models were missing something. In the 1980s several models were introduced to accommodate additional observations about the pattern of nucleotide substitution (Felsenstein, 1981; Hasegawa *et al.*, 1984, 1985; Tavaré, 1986). The most general model was proposed by Tavaré (1986), called the general time-reversible (GTR) model. The GTR model has rate matrix

$$\mathbf{Q} = \begin{pmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{pmatrix} \mu. \quad (6.27)$$

The GTR model has two sets of parameters: the exchangeability parameters, $\mathbf{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$, and the stationary frequencies, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$. The factor μ is

chosen such that the average rate of substitution is 1. The exchangeability parameters allow the GTR model to account not only for a transition/transversion rate bias, but also for other biases in the rates of substitution among pairs of nucleotides. The stationary frequency parameters allow the GTR model to fit data in which the nucleotide frequencies are not equal.

Earlier, when discussing the transition probability matrix and stationary frequencies of continuous-time Markov models, I pointed out that the stationary frequencies are a property of the rate matrix. The stationary frequencies are the probabilities of capturing the process in a particular state after a very long (formally, infinitely long) time has passed. However, the stationary frequencies are built in as parameters of the GTR model, and other models of DNA substitution, too (Felsenstein, 1981; Hasegawa *et al.*, 1984, 1985; Tamura and Nei, 1993). The stationary frequency parameters in these models are specified in a way that ensures that the rate matrix has the specified stationary probabilities.

The GTR model is the most general time-reversible model of DNA substitution. A time-reversible model must satisfy the requirement that $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i \neq j$. Time-reversible models have the property that when at stationarity, the process looks the same when time is reversed. Almost all models applied in phylogenetics are time-reversible. Just how many such models are there? Huelsenbeck *et al.* (2004) pointed out that time-reversible models can be understood as partitions of the exchangeability parameters. So, for example, the Jukes and Cantor (1969) model has all six exchangeability parameters in the same set, $r_{AC} = r_{AG} = r_{AT} = r_{CG} = r_{CT} = r_{GT}$, whereas the Kimura (1980) model has two sets of rates, $r_{AG} = r_{CT}$ and $r_{AC} = r_{AT} = r_{CG} = r_{GT}$. As partitions, they can be described through the restricted growth function notation for a partition (Stanton and White, 1986):

AC	AG	AT	CG	CT	GT
1	1	1	1	1	1
1	2	1	1	2	1

where the partition $(1, 1, 1, 1, 1, 1)$ denotes the Jukes and Cantor (1969) model and the partition $(1, 2, 1, 1, 2, 1)$ denotes the Kimura (1980) model. Other described models are as follows:

		AC	AG	AT	CG	CT	GT
JC69, F81	Jukes and Cantor (1969); Felsenstein (2001)	1	1	1	1	1	1
K80, HKY85	Kimura (1980); Hasegawa <i>et al.</i> (1985)	1	2	1	1	2	1
TN93	Tamura and Nei (1993)	1	2	1	1	3	1
K81	Kimura (1981)	1	2	3	3	2	1
	Posada (2003)	1	2	3	3	4	1
	Posada (2003)	1	2	3	4	2	5
SYM, GTR	Tavaré (1986)	1	2	3	4	5	6

When substitution models are described as partitions of a set of objects (in this case rates), the number of possible models can be determined. In fact, there are a total of 203 time-reversible substitution models, all of which are shown in Huelsenbeck *et al.* (2004); the number of

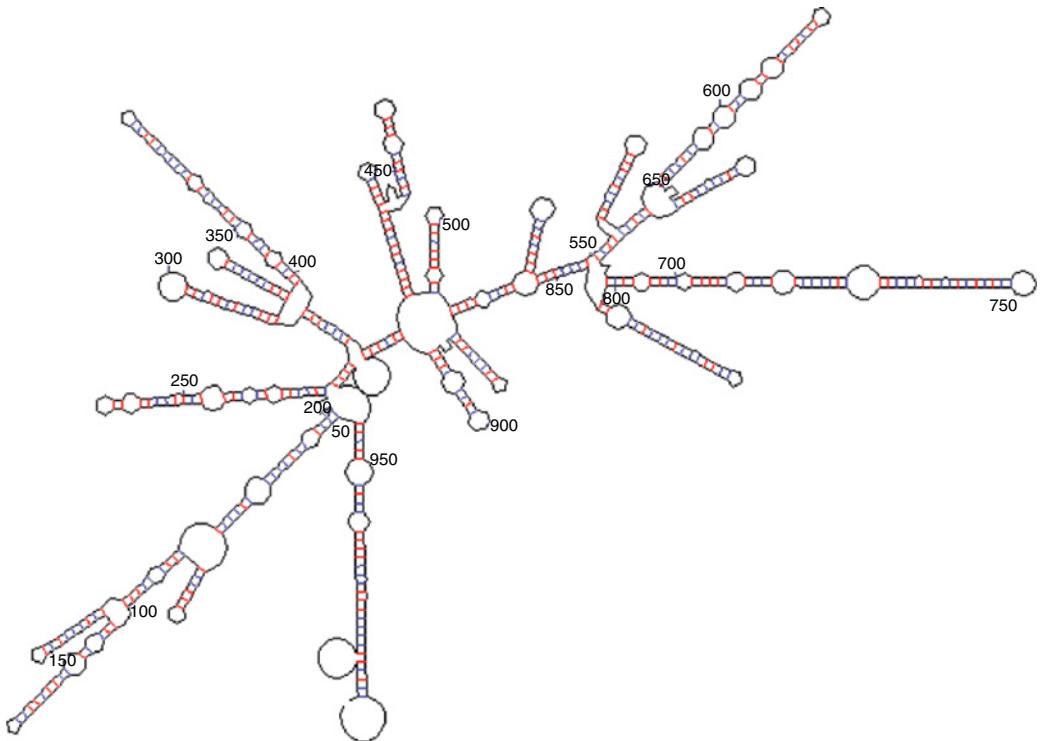


Figure 6.15 The RNA secondary structure of the + strand of the bacteriophage NL95 (part of the coat and read-through proteins).

partitions of six objects is the Bell (1934) number for 6. Huelsenbeck *et al.* (2004) used a variant of MCMC called reversible-jump MCMC (Green, 1995) to explore the full space of substitution models. The method is implemented in the MrBayes program. Interestingly, Huelsenbeck *et al.* (2004) found that many unnamed substitution models had high posterior probability. That said, the named substitution models (i.e. those models in the table above) had more posterior probability than prior probability, reaffirming that the named models were described for a reason: they account for some of the observed vagaries of the substitution process.

6.6.2 Expanding the Model around Groups of Sites

Figure 6.15 shows the RNA secondary structure of the + strand of the bacteriophage NL95 (part of the coat and read-through proteins). The secondary structure in this phage plays an important role in regulating the translation of the four proteins that compose the phage. One would expect natural selection to maintain the Watson–Crick pairing of the stem regions. If a mutation occurs that disrupts the pairing, natural selection should favor compensatory mutations that restore it. This type of non-independence of substitutions is difficult for phylogenetic methods, at least as described so far in this chapter, to accommodate.

Schöniger and von Haeseler (1994) expanded the continuous-time Markov model of substitution around the pair of interacting nucleotide sites to account for the non-independence of substitutions in the stem regions. The so-called doublet model relies on the idea that the researcher can identify the interacting sites. The states of the doublet model are all 16 pairs

of nucleotides. The 16×16 rate matrix for the doublet model of Schöniger and von Haeseler (1994) is specified using a short-hand notation:

$$q_{ij} = \begin{cases} \kappa\pi_j & \text{if the change is a transition,} \\ \pi_j & \text{if the change is a transversion,} \\ 0, & \text{if the change involves two changes.} \end{cases} \quad (6.28)$$

Instead of showing the full 16×16 rate matrix, this equation simply specifies the rules for constructing the rate matrix. Note that this rate matrix will have many entries that are zero. The zero rates represent cases in which two substitutions must occur in an instant of time (e.g. CG → AT). This model only allows substitutions to occur one at a time. This is not to say that changes involving two (or more) substitutions in a doublet pair are impossible; rather, such changes must go through intermediate steps and occur over evolutionary time scales.

Goldman and Yang (1994) and Muse and Gaut (1994) took a similar approach to account for non-independent substitutions in protein-coding DNA sequences. They expanded the continuous-time Markov model around the triplet of nucleotides that compose a codon. The continuous-time Markov model, then, has $4^3 = 64$ possible states. Typically, the three stop codons (for the universal genetic code) are removed from the state space, leaving a total of 61 possible states. The rate matrix for a codon model can be devised in numerous ways. For example, the model devised by Nielsen and Yang (1998) has rate matrix

$$q_{ij} = \begin{cases} \kappa\omega\pi_j & \text{if the change is a non-synonymous transition,} \\ \kappa\pi_j & \text{if the change is a synonymous transition,} \\ \omega\pi_j & \text{if the change is a non-synonymous transversion,} \\ \pi_j & \text{if the change is a synonymous transversion,} \\ 0, & \text{if the change involves two or three changes,} \end{cases} \quad (6.29)$$

and distinguishes changes between codons that involve transitions versus transversions and those that are synonymous versus non-synonymous. The change between codons ACC → ATC is both a transition (the codons differ by a C → T change at the second position) and non-synonymous because the codons code for different amino acids (threonine and isoleucine).

The Nielsen and Yang (1998) formulation of the codon model allows phylogenetic methods to detect the footprint of natural selection in protein-coding DNA sequence alignments. The parameter ω is interpreted as the non-synonymous/synonymous rate ratio. The neutral expectation is that $\omega = 1$. The typical pattern found in alignments of protein-coding sequences is for $\omega < 1$. This is to be expected. After all, mutations are likely to break an already functioning protein. Individuals bearing such mutations will be removed from the population through purifying natural selection, resulting in a pattern of substitution in which non-synonymous changes are underrepresented. The alternative pattern of positive selection, in which $\omega > 1$, occurs less frequently. Occasionally, non-synonymous mutations will confer a fitness advantage to the individuals bearing them. For example, perhaps a mutation in the coat protein of a virus better allows individuals bearing the mutation to escape the host's immune system.

A naive application of the codon model proposed by Nielsen and Yang (1998) is unlikely to find evidence of positive selection ($\omega > 1$); most sites are likely under purifying selection and only a few, if any, under positive selection. Hence, the overall signature will be of purifying selection because the sites under purifying selection will overwhelm the signal from the handful of positively selected sites. Nielsen and Yang (1998) devised a model in which ω can vary across the sequence. A site is assumed to be in one of three classes: $\omega = 0$ with probability p_1 , $\omega = 1$ with probability p_2 , and $\omega > 1$ with probability p_3 . The likelihood for a site is calculated three

times, once for each selection category, and the final likelihood is averaged over the three categories, with the weights being p_1 , p_2 , and p_3 . The probability that the i th codon site is under the influence of positive selection is

$$f(K = 3|x_i) = \frac{f(x_i|K = 3)p_3}{f(x_i|K = 1)p_1 + f(x_i|K = 2)p_2 + f(x_i|K = 3)p_3}, \quad (6.30)$$

where K is the selection category and $K = 3$ corresponds to the positive selection category. One simply calculates the probability that each codon site is in the category corresponding to positive selection. Sites with high posterior probabilities of being in category $K = 3$ are likely under positive selection. Later, Yang *et al.* (2000) devised numerous additional models for allowing the non-synonymous/synonymous rate ratio to vary across the sequence that are less restrictive than the original Nielsen and Yang (1998) model. For more on codon models, see **Chapter 13**.

6.6.3 Rate Variation across Sites

One should keep in mind that the DNA sequences used in molecular phylogenetics are not evolving neutrally. If they were, it would be difficult or impossible to identify the sites as homologous. Mutation and genetic drift work to erode the signature of homology that sequence alignment programs use to identify the fine-scale homology in an alignment. Moreover, different sites in a gene alignment are unlikely to experience natural selection in the same way. Some sites are more conserved than others. The description of the likelihood calculations has assumed that every site in an alignment has the same rate of substitution. This is equivalent to saying that all sites experience natural selection in the same way.

Models for relaxing the equal rates assumption assume that each site has a rate multiplier, r , that acts to expand or contract the lengths of the branches on the tree (Figure 6.16). Unfortunately, there is not enough information at a single site to reliably estimate the rate for each site. Rather, the likelihood calculation for a site marginalizes over all possible rates. For example, one of the earliest models for among-site rate variation was the proportion of invariable sites model. This model assumes that a site can be in one of two rate classes: the site is invariable,

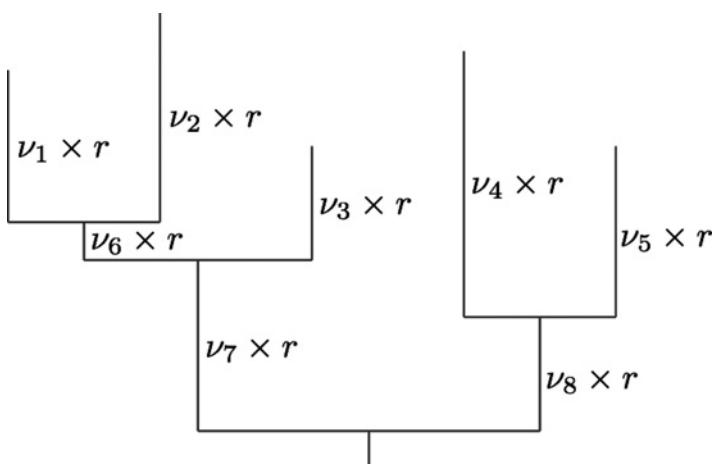


Figure 6.16 The branch lengths of a tree are changed proportionally for each site through a rate multiplier, r , unique to each site.

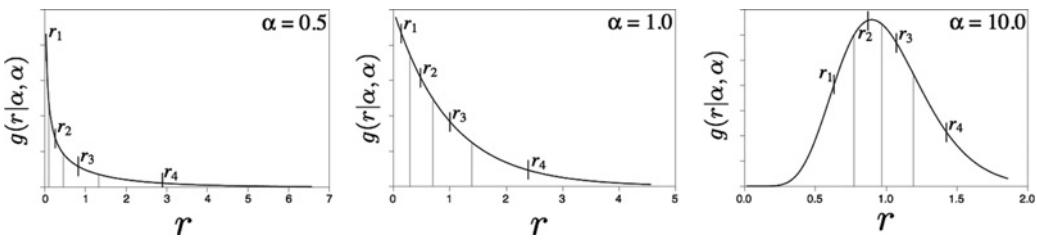


Figure 6.17 The distribution of rates under the gamma model of rate variation across sites. The continuous distribution is broken into four categories and the mean rate for each illustrated.

with rate $r = 0$, with probability p ; and potentially variable, with rate $r = \frac{1}{1-p}$, with probability $1 - p$. The likelihood is calculated twice for each site, once for each rate class, and the likelihood is a weighted average over both rate categories:

$$f(\mathbf{x} | \psi, p, \theta) = f(\mathbf{x} | \psi, r = 0, \theta)p + f(\mathbf{x} | \psi, r = 1/(1-p), \theta)(1-p).$$

A frequently used model for accommodating among-site rate variation assumes that the rate for a site is drawn from a gamma distribution with mean 1. The gamma distribution has two parameters, the shape and scale parameters. For the gamma rate variation model, both the shape and scale parameters take the same value, α . This means that the mean rate of change is 1 and the variance of rates across sites is $1/\alpha$. Figure 6.17 shows the distribution of rates for several different values of α . Marginalizing over possible rates for the gamma model of among-site rate variation involves integrating over all possible rates,

$$f(\mathbf{x} | \psi, \alpha, \theta) = \int_0^\infty f(\mathbf{x} | \psi, r, \theta)g(r|\alpha)dr, \quad (6.31)$$

where $g(r|\alpha)$ is the gamma probability density. It is not feasible to calculate this integral, except for cases in which the number of species is small (Yang, 1993). The usual approach is to approximate the integral by dividing the continuous gamma distribution into some number of discrete categories. The mean (or median) rate for each category is used to represent all of the rates for the category. Because the gamma distribution has been divided into categories of equal size, the prior probability of each category is simply $1/(\text{number of categories})$. Figure 6.17 shows the case in which the gamma distribution is broken into four categories. The mean rate for the i th category is denoted r_i . The probability for a site, then, becomes

$$f(\mathbf{x} | \psi, \alpha, \theta) = \sum_{k=1}^K f(\mathbf{x} | \psi, r_k, \theta) \frac{1}{K} \quad (6.32)$$

for the case in which the gamma distribution is broken into K categories.

Both the proportion of invariant sites and gamma models for among-site rate variation assume that all of the branches on the phylogeny experience the same rate. A site is either a high- or low-rate site. These models do not allow some branches on the tree to experience a high rate while other branches for the same site experience a low rate. The concomitantly variable codons, or covarion, hypothesis of substitution (Fitch and Markowitz, 1970) states that the rate of substitution should vary over the evolutionary history of a group. As mutations are fixed at some sites in a gene, the functional constraints at other positions may change. Hence, the rate at a site should vary over the evolutionary history of a group of organisms. Tuffley and Steel (1998) made explicit the covarion model by proposing a continuous-time Markov model of substitution that captures the spirit of the Fitch and Markowitz (1970) description of varying

constraints. They assumed that a site could either be off, in which case the rate of substitution is zero, or on, in which case substitutions can occur. The continuous-time Markov model they devised has eight states: A_0 , C_0 , G_0 , T_0 , A_1 , C_1 , G_1 , T_1 . Nucleotides that cannot experience substitution have the subscript 0, whereas nucleotides capable of change have the subscript 1. The Tuffley and Steel (1998) model allows switching between the off and on state. The full rate matrix is

$$\mathbf{Q} = \begin{pmatrix} - & 0 & 0 & 0 & \lambda q & 0 & 0 & 0 \\ 0 & - & 0 & 0 & 0 & \lambda q & 0 & 0 \\ 0 & 0 & - & 0 & 0 & 0 & \lambda q & 0 \\ 0 & 0 & 0 & - & 0 & 0 & 0 & \lambda q \\ \lambda p & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda p & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda p & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda p & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (6.33)$$

with $q = 1 - p$ and $0 \leq p \leq 1$. (The scaling factor, μ , only affects the portion of the rate matrix representing nucleotide substitutions.) Many of the rates are zero either because the process is 'off' and incapable of change (i.e. the upper left 4×4 area of the matrix) or because the entry in the matrix would require a nucleotide substitution and a rate class switch. The Tuffley and Steel (1998) model, like the doublet and codon models, only allows one event at a time to occur. The bottom right 4×4 area of the rate matrix represents substitutions. Any of the substitution models discussed earlier, such as the GTR model, can be used to model substitutions.

The covarion model typically fits data better than models that do not allow rates at a site to vary over the tree (Huelsenbeck, 2002). The covarion model of Tuffley and Steel (1998) becomes equivalent to the proportion of invariable sites as $\lambda \rightarrow 0$. Galtier (2001) implemented an alternative model which reduces to a discrete gamma model of rate variation as the category-switching rate approaches zero.

6.6.4 Divergence Time Estimation

The idea that substitutions accumulate at a constant rate – the molecular clock hypothesis – is among the oldest ideas in molecular evolution (Zuckerkandl and Pauling, 1962). The idea is of great practical importance because the molecular clock, if true, suggests a way not only to disentangle rate and time but also to date speciation events on a tree. The idea is simple. If one of the speciation times is assumed to be known, then the other times on the tree can be determined. Simultaneously, one can determine the rate of substitution. Figure 6.18 illustrates the idea. One of the speciation events depicted on the tree is assumed to have occurred 10 million years ago (MYA). This node is referred to as the 'calibration' node. Note that the sum of the branch lengths from any of the extant descendants of the calibration node, to the calibration node itself, is 0.2 measured in expected number of substitutions per site. This means that 0.2 substitutions/site = 10 MYA. The root of the tree, which requires 0.3 substitutions to reach from any tip, must have diverged 15 MYA. One simply solves the following equation for the time of the root node:

$$\frac{0.2 \text{ substitutions}}{10 \text{ MY}} = \frac{0.3 \text{ substitutions}}{? \text{ MY}}$$

Similarly, the times of the other nodes on the tree must be 5 MYA and 7.5 MYA.

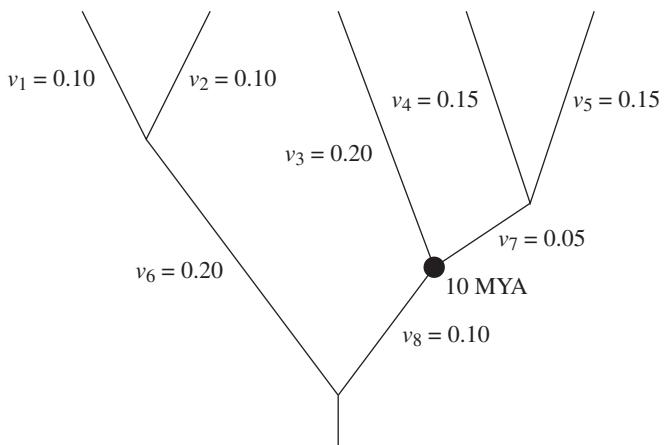


Figure 6.18 A clock-constrained tree with the branch lengths specified. One of the speciation events on this tree (the one with the black circle) is assumed to have diverged 10 million years ago (MYA).

The problem with this exercise is that the molecular clock rarely holds and that it is difficult to know with any certainty the time of any particular node to use as a calibration. Indeed, divergence time estimation is probably the messiest problem in the field of molecular evolution, with many questionable procedures having been applied (see Graur and Martin, 2004).

Over the past twenty years, some of the shortcomings of the molecular clock hypothesis have been addressed using ‘relaxed clock’ models. More recently, considerable progress has been made in producing calibrations that are more faithful to the observed fossil record.

6.6.4.1 Relaxed Molecular Clocks

How can the molecular clock be relaxed in such a way that we can still estimate divergence times? Figure 6.19(a) shows an example of a tree with branch substitution rates (r_1, r_2, \dots, r_6) and divergence times (t_1, t_2, t_3). Remember that the length of each branch, in terms of expected number of substitutions per site, is the product of the rate and time on the branch. So, for example, the length of branch 6 in Figure 6.19(a), measured as expected number of substitutions per site, is $v_6 = r_6 \times (t_1 - t_3)$. The strict molecular clock hypothesis results if all of the rates on the branches are constrained to be equal. At the other extreme, if each of the rates on the

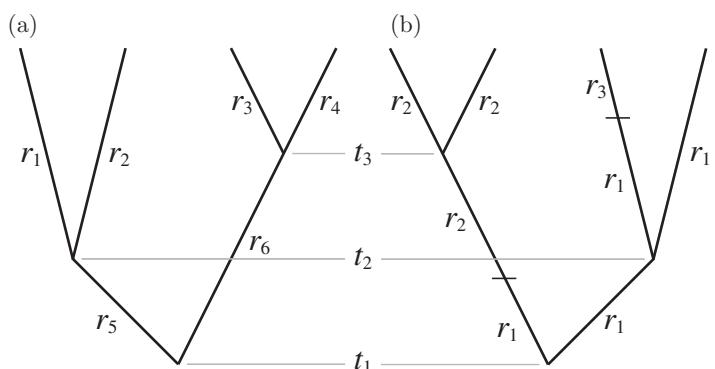


Figure 6.19 Rate multipliers for the branch lengths for trees in which (a) every branch has its own multiplier and (b) rates change episodically.

tree is completely unconstrained, one obtains the parameterization of branch lengths used by many computer programs. In the extreme, unconstrained rates, situation, it is impossible to disentangle rates and time. In fact, the usual solution in phylogenetics is to simply parameterize the branch lengths on the tree as the product of rate and time ($v = ut$) and give up on any idea of estimating the times on the tree. Relaxed-clock models allow rates to vary on the tree in a constrained way, which maintains information on the time component of evolution.

Since the pioneering work of Thorne *et al.* (1998), numerous relaxed-clock models have been proposed. The relaxed-clock models fall into two classes. The autocorrelated-rate models assume that the rate for a branch is similar to the rate of the ancestor of the branch. Uncorrelated models, on the other hand, treat the rate for each branch independently.

Autocorrelated models Under the model proposed by Kishino *et al.* (2001), rates are assigned to the nodes of a tree from a log-normal probability distribution, with the rate for each node centered on the rate of its ancestor. This is equivalent to saying that the log of the rate at node i is drawn from a normal distribution centered on the log of the rate for the ancestor. (The original paper by Thorne *et al.* (1998) considered the rate at the midpoint of a branch.) We set up the problem by defining some terminology for a branch: the i th node on the tree has a branch of length, in units of time, of t_i ; the ancestor of the i th node is denoted $\sigma_A(i)$. For the Thorne *et al.* (1998) model, then, the rate assigned to the i th node is $r_i \sim LN(\mu_{\sigma_A(i)}, \rho t_i)$. Here, $\mu_{\sigma_A(i)} = \ln(r_{\sigma_A(i)}) - \rho t_i/2$ and the parameter ρ is the rate parameter of a Brownian motion model. When the rate parameter of the Brownian motion model (describing the rate at which the substitution rate changes) is equal to zero, then the model of Kishino *et al.* (2001) becomes equivalent to a strict molecular clock.

What is the overall rate for a branch? Thorne *et al.* (1998) and Kishino *et al.* (2001) use the average rate for a branch,

$$\bar{r}_i = \frac{r_i + r_{\sigma_A(i)}}{2}.$$

The expected number of substitutions per site for the i th branch, then, is $v_i = \bar{r}_i t_i$. Likelihood computations then proceed as described earlier in this chapter.

Other autocorrelated models either extend the work of Thorne *et al.* (1998) or assume that rates change episodically:

- Aris-Brosou and Yang (2002) take an approach similar to that of Kishino *et al.* (2001), but draw the rate for a node from a gamma distribution with parameters chosen such that the mean rate is equal to the rate of the ancestor of node i ($r_{\sigma_A(i)}$) and a variance that is proportional to the branch length, in terms of time (ρt_i).
- The rate of substitution for the models proposed by Thorne *et al.* (1998) and Aris-Brosou and Yang (2002) has no stationary distribution. Aris-Brosou and Yang (2002) suggested a model in which the underlying process is not Brownian motion, but rather follows an Ornstein–Uhlenbeck (OU) process. An OU process has a tendency to return to the mean. This modification does lead to a model of rate change that has a stationary distribution. Similarly, Lepage *et al.* (2006) used a Cox–Ingersoll–Ross process (a squared OU model) to model the rate at which substitution rates change. The advantage of their formulation is that the rate of substitution does not decrease through time, as is the case for the OU model.
- Huelsenbeck *et al.* (2000) assume that the rate of substitution changes episodically on the tree. They used reversible-jump MCMC to insert or delete events of rate change on the tree. (See

Figure 6.19(b) for an example in which there are two events of rate change, leading to three different rates on the tree.) When an event of rate change occurs, the new rate, r' , is equal to the old rate (r) multiplied by a factor that is drawn from a modified gamma probability distribution.

Independent branch-rate models Under an independent-rates model, the rates of substitution for a branch are not dependent on its ancestral rate. The general idea is to draw the rates for each branch from an underlying probability distribution. Because the rate for each branch is an independent draw from the underlying distribution, the rate of a branch will not depend on the rate of its ancestor. For example, Drummond *et al.* (2006) assume that the rate parameters of the tree are drawn from a log-normal distribution. When the variance parameter of the log-normal distribution is very small, the rates on the tree more closely correspond to a strict molecular clock. The branch rates can also be drawn from other probability distributions, such as a gamma distribution (Lepage *et al.*, 2007) or exponential distribution (Drummond *et al.*, 2006). Independent-rate models for rate variation have been implemented in the programs BEAST (Drummond and Rambaut, 2007), RevBayes (Höhna *et al.*, 2016), and PhyloBayes (Lartillot *et al.*, 2009).

Grouping branch rates A final approach to relaxing the molecular clock is to group the branches on the tree such that they share rates. If all of the branches are grouped together, for example, they will all share the same rate. This is equivalent to the strict molecular clock. Of course, other groupings of branches are possible, which results in multiple ‘local clocks’ on a tree.

Yoder and Yang (2000) and Yang and Yoder (2003) group branches into local clocks in a supervised manner. That is to say, they use other information to divide the tree into some number of local clocks. The rates of the local clocks are then estimated using maximum likelihood or Bayesian inference. The disadvantage of this approach is that it is not clear exactly how the branches should be grouped together.

Unsupervised methods use a model to specify a probability for each possible grouping of the branches into local clocks. MCMC is then used to explore local clocks in proportion to their posterior probabilities. Drummond and Suchard (2010) developed a model in which a branch either inherits the rate of its ancestral branch, or draws a new rate which is equal to the ancestral rate times a factor drawn from a mean-one gamma distribution. The underlying process is considered to be an episodic clock, in which the rate of change of the rate of change is small. The number of rate change events on the tree follows a Poisson probability distribution. If there are no rate change events, all of the branches have the same rate; the model collapses to the strict molecular clock.

The branch rates under the Drummond and Suchard (2010) model are grouped together such that contiguous branches may share the same rate. Heath *et al.* (2012) proposed a local clock model in which branches that are distantly separated on the tree may share the same rates. They assumed that the branch rates are partitioned, with the prior probability of different rate partitions coming from a Dirichlet process prior model (Antoniak, 1974; Ferguson, 1973). MCMC is used to explore the space of rate partitions and other model parameters.

6.6.4.2 Calibrating Phylogenetic Trees

Besides requiring an assumption of constancy of substitution rates – whether the rates adhere to a strict molecular clock or to a relaxed molecular clock – divergence time estimation requires that the time of at least one node on the tree is known. The process of calibrating the tree is

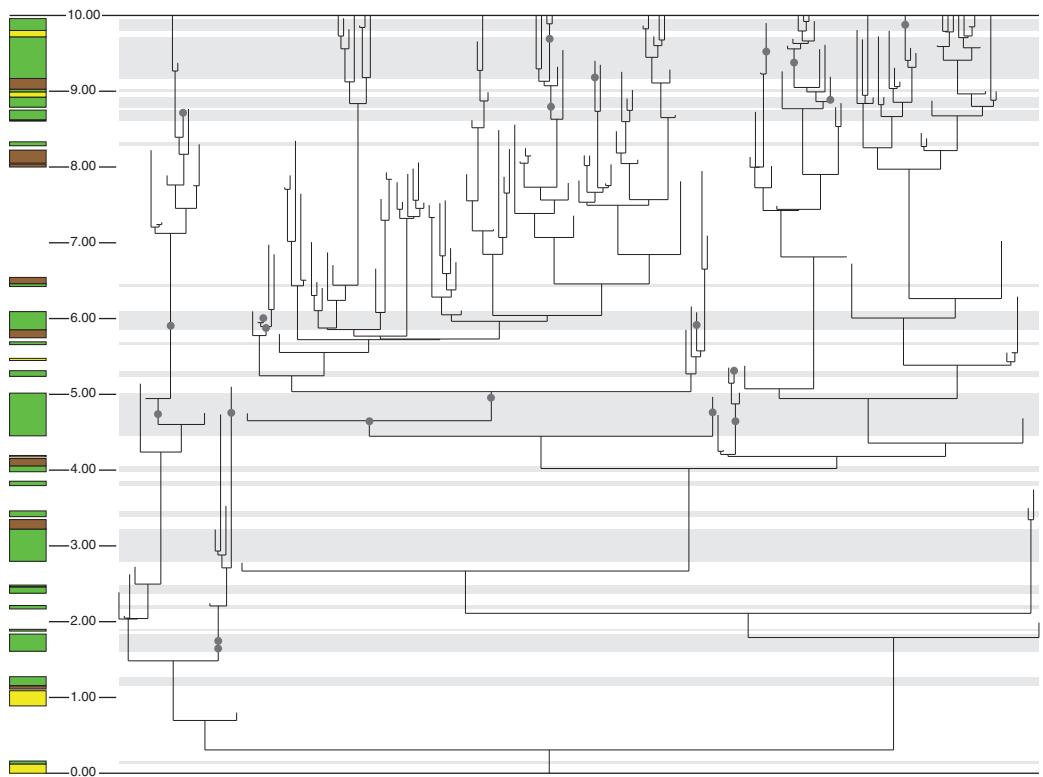


Figure 6.20 An example of a phylogenetic tree with fossils (black dots). This tree was generated under the birth–death process of cladogenesis. The column on the left shows the rocks available: sedimentary (green), igneous (brown), and metamorphic (yellow) rocks. Fossils can only occur in sedimentary rocks. Note that the rock record has numerous gaps during which no rocks were deposited.

difficult because it requires information from the fossil record to be incorporated into a phylogenetic analysis in which the fossils are typically absent. The process is further complicated by the incompleteness of the fossil record and the difficulty of assigning fossils to an extant group for the tree of interest.

Figure 6.20 illustrates some of the complexities of the calibration problem. The tree was generated under a birth–death process of cladogenesis. The tree of Figure 6.20 also has fossils, represented as dots on the tree, which were also generated under a stochastic process. In the simulation of Figure 6.20, fossils are constrained to occur only in sedimentary rocks. The types of rocks are shown in the stratigraphic column on the left of Figure 6.20. Note that there are periods of time during which no rocks of any sort are represented; fossils are not possible for those periods of time. The simulation could have been further complicated by adding the possibility of mass extinction and changes in the rate of diversification.

How can one use the fossils on the tree to calibrate one or more of the nodes of the reconstructed tree? Some of the fossils are direct ancestors of living species on the tree of Figure 6.20. Others are found on lineages that are extinct cousins of living taxa. The difficulty, of course, is that in reality the person does not get to see whether the fossil is a direct ancestor or cousin.

The earliest method for calibrating nodes on the reconstructed tree could hardly be called a ‘method’ because the procedure is not described in any coherent manner. Calibration times

were produced by paleontologists who incorporated the information at hand in some way to produce a point estimate of a divergence time, such as a 56.5 million year time for the divergence of cows and whales (Arnason *et al.*, 1997) or a 310 million year divergence time between birds or mammals (Nei *et al.*, 2001). Later authors became confused as to which nodes on a study were calibrations and which were inferred times. Embarrassingly, some later studies used the inferred times from one study as the calibration times (see Graur and Martin, 2004)!

Early methods for estimating divergence times did not account for uncertainty in the calibration times on the tree; the calibration time was considered to be *precisely* known. The earliest implementation of MrBayes (Huelsenbeck and Ronquist, 2001) allowed the user to specify different probability distributions on the calibration time(s). Probability distributions such as the uniform distribution, with an upper or lower bound, or an offset exponential distribution were implemented in the program. Yang and Rannala (2006) also used arbitrary probability distributions to model the calibration time. Unlike the methods implemented in MrBayes which did not allow the calibration time to occur outside of the bounds of the probability distribution, they allowed for this possibility. Unfortunately, it is not clear how to convert information on the fossils to a uniform or exponential probability distribution.

Explicitly model-based approaches have been developed that obviate the necessity of making arbitrary choices for calibration times. The ‘tip dating’ approach (Ronquist *et al.*, 2012a) takes as its motivation the use of viruses to calibrate trees (Rambaut, 2000). A viral sample that was stored in a freezer decades ago can act as a calibration in a phylogenetic analysis of living viruses. The time at which the virus was frozen is considered known. When combined with the molecular clock assumption, or one of the relaxed-clock models, the other times on the tree can be inferred. Ronquist *et al.* (2012a) applied this idea to fossils. Of course, DNA sequences are not available for fossils. However, fossils preserve information on many morphological characters, at least those characters that can fossilize. The fossil acts as the time calibration and, when combined with a clock-like assumption for the evolution of morphological characters, the divergence times of the nodes on the tree can be estimated. This method is considered a ‘total evidence’ approach because it combines information on the morphological and molecular characters that are available to the researcher. Importantly, the tip dating approach outlined by Ronquist *et al.* (2012a) does not force the researcher into making the arbitrary decisions about the probability distribution for node times.

Heath *et al.* (2014) developed a prior for the tree topology that directly incorporates fossil information – the ‘fossilized birth–death process’. Specifically, they modeled the combined speciation, extinction, and fossilization process. The process can be described as follows: in a small interval of time, Δt , a lineage speciates with probability $\lambda\Delta t$, goes extinct with probability $\mu\Delta t$, or produces a fossil with probability $\phi\Delta t$. Fossils, then, are included in the analysis, even if the paleontologist has not described any of the morphological characters of the fossils. Although the fossils do not contribute to the phylogeny through characters, it is assumed that the researcher knows their group membership. To calibrate the phylogeny, the method relies on two assumptions: that the time of the fossil is known; and that the fossil is a descendant of some group on the tree. The statement, ‘*Tiktaalik* is a tetrapod that lived 375 million years ago’, for example, is sufficient to calibrate a tree of vertebrate taxa under the fossilized birth–death process. Parameters are estimated using Bayesian methods, with MCMC used to approximate the posterior probability distribution of the parameters. In this application, the MCMC procedure must consider scenarios in which the fossil is a direct ancestor of taxa in the phylogeny. The fossilized birth–death process can be combined with morphological character information from the fossils, in which case the resulting analysis combines the best of both the tip dating method of Ronquist *et al.* (2012a) and the coherent prior on speciation times of Heath *et al.* (2014).

6.7 Conclusions

This chapter surveyed only a fraction of the exciting developments that have occurred over the past 40 years in the field of statistical phylogenetics. Topics such as species delineation (Yang and Rannala, 2010), species tree–gene tree inference (Maddison, 1997), models of arbitrary non-independence (Robinson *et al.*, 2003), nonparametric Bayesian analysis (Lartillot and Philippe, 2004; Huelsenbeck *et al.*, 2006; Huelsenbeck and Suchard, 2007), and time-heterogenous models of substitution (e.g. Blanquart and Lartillot, 2006) were not even touched upon in this chapter. The adoption of the framework provided by the likelihood-based methods likely accelerated advances in the field. For one, the likelihood-based methods of maximum likelihood and Bayesian inference allow parameters of evolutionary models to be efficiently estimated. Perhaps more importantly, they also allow different models to be compared (using, for example, likelihood ratio tests or Bayes factors). Although the application of tests of competing models in any particular study may seem inconsequential, the cumulative effect of hundreds of studies proposing new models and testing them against observed DNA sequences has resulted in a vast increase in our knowledge of the pattern and process of molecular evolution.

The field of statistical phylogenetics has come a long way in a relatively short period of time. The Felsenstein pruning algorithm was proposed less than 40 years ago. At that time, there were only a few substitution models that had even been described (Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 2001). Since then, numerous other models have been devised to account for new observations as they were made. To name a few: the discovery that transitions occur at a higher rate than transversions motivated the model of Kimura (1980); the observation that nucleotide frequencies are not equal motivated the models proposed by Felsenstein (2001) and Hasegawa *et al.* (1985); the observation that different sites experience different levels of constraint due to natural selection motivated models of among-site rate variation (Yang, 1993); and the observation that substitutions are unlikely to be independent at different sites motivated models that account for non-independence in various ways (Schöniger and von Haeseler, 1994; Muse and Gaut, 1994; Goldman and Yang, 1994; Robinson *et al.*, 2003).

Most of the theory and practice of molecular phylogenetics has been applied to only one or a handful of alignments. Only so much information about the pattern of substitution can be wrung out of a single alignment. Full genome sequences provide exciting opportunities to extend our knowledge about molecular evolution. Genomic sequence data also presents significant challenges to the field. For example, to date, the field has been able to rely on a set of carefully curated genes that many investigators agree to use in phylogenetics (Graybeal, 1994). These genes are often chosen for the desirable properties of ease of alignment, an appropriate level of variation, and low copy number in the genome. However, with genome-scale data, all of the genes are analyzed, not just those that are easily aligned. When hundreds or thousands of genes are considered, many of the gene alignments are inherently uncertain (Wong, 2006). Methods that accommodate uncertainty in alignment are currently difficult to implement in a way that allows for speedy data analysis. However, if successfully implemented, models that simultaneously allow for substitutions, insertions, and deletions will allow alignment uncertainty to be accounted for. They will also allow us to learn about the pattern of insertion and deletion much as we have learned about the pattern of DNA substitution. The tree-like structure underlying the comparison of genes also becomes more complicated because processes such as gene duplication and gene loss must be accounted for (e.g. Bousau *et al.*, 2013). I believe that the most exciting advances lie ahead for the field of statistical phylogenetics.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csáki (eds.), *2nd International Symposium on Information Theory*. Tsahkadsor, Armenia, USSR, Budapest: Akadémiai Kiadó, pp. 267–281.
- Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *Annals of Statistics* **2**, 1152–1174.
- Aris-Brosou, S. and Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology* **51**, 703–714.
- Arnason, U., Gullberg, A. and Janke, A. (1997). Phylogenetic analyses of mitochondrial DNA suggest a sister group relationship between Xenarthra (Edentata) and ferungulates. *Molecular Biology and Evolution* **14**, 762–768.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **53**, 370–418.
- Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–773.
- Bell, E.T. (1934). Exponential numbers. *American Mathematics Monthly* **41**, 411–419.
- Blanquart, S. and Lartillot, N. (2006). A Bayesian compound stochastic process for modeling non-stationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution* **23**, 2058–2071.
- Bollback, J. and Huelsenbeck, J.P. (2001). Phylogenetic relationships, genome evolution, and host specificity of single-stranded RNA bacteriophage (Family Leviviridae). *Journal of Molecular Evolution* **52**, 117–128.
- Boussau, B., Szöllősi, G.J., Duret, L., Gouy, M., Tannier, E. and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research* **23**, 323–330.
- Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* **19**, 233–257.
- Charleston, M.A. (1995). Toward a characterization of landscapes of combinatorial optimization problems, with special attention to the phylogeny problem. *Journal of Computational Biology* **2**, 439–450.
- Darwin, C. (1859). On the Origin of Species. Murray, J., London.
- Drummond, A. and Rambaut, A. (2007). BEAST v1.4. <http://beast.bio.ed.ac.uk>.
- Drummond, A.J., Ho, S.Y., Phillips, M.J. and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88.
- Drummond, A.J. and Suchard, M.A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology* **8**, 114.
- Edwards, A.W.F. (1966). Studying human evolution by computer. *New Scientist* **30**, 438–440.
- Edwards, A.W.F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society, Series B* **32**, 155–174.
- Edwards, A.W.F. (1998). History and philosophy of phylogeny methods. Paper presented to EC Summer School, Methods for Molecular Phylogenies, Isaac Newton Institute, Cambridge.
- Edwards, A.W.F. and Cavalli-Sforza, L.L. (1964). Reconstruction of evolutionary trees. In J. McNeill (ed.), *Phenetic and Phylogenetic Classification*. Systematics Association, London.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791.
- Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package). Distributed by author.
- Felsenstein, J. (2001). The troubled growth of statistical phylogenetics. *Systematic Biology* **50**, 465–467.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.
- Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41**, 155–160.
- Fisher, R.A. (1921). On the ‘probable error’ of a coefficient of correlation deduced from a small sample. *Metron* **I**(4), 3–32.
- Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.
- Fitch, W.M. and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4**, 579–593.
- Gallager, R.G. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory* **8**, 21–28.
- Gallager, R.G. (1963). Low-density parity check codes. MIT Press, Cambridge, MA.
- Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**, 866–873.
- Gelman, A. (1996). Inference and monitoring convergence. In W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, pp. 131–143.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182–198.
- Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.
- Graur, D. and Martin, W. (2004). Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. *Trends in Genetics* **20**, 80–86.
- Graybeal, A. (1994). Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. *Systematic Biology* **43**, 174–193.
- Green, P.J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 771–732.
- Haeckel, E. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Georg Riener, Berlin.
- Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160–174.
- Hasegawa, M., Yano, T. and Kishino, H. (1984). A new molecular clock of mitochondrial DNA and the evolution of Hominoidea. *Proceedings of the Japan Academy Series B* **60**, 95–98.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

- Heath, T.A., Holder, M.T. and Huelsenbeck, J.P. (2012). A Dirichlet process prior for estimating lineage-specific substitution rates. *Molecular Biology and Evolution* **29**, 939–955.
- Heath, T.A., Huelsenbeck, J.P. and Stadler, T.J. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Science* **111**, E2957–E2966.
- Hennig, W. (1950). Grundzüge einer Theorie der Phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.
- Hennig, W. (1966). Phylogenetic Systematics. University of Illinois Press, Urbana.
- Hillis, D.M. and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**, 182–192.
- Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., Huelsenbeck, J.P. and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* **65**, 726–736.
- Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* **19**, 698–707.
- Huelsenbeck, J.P., Jain, S., Frost, S.W.D. and Pond, S.L.K. (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6263–6268.
- Huelsenbeck, J.P., Larget, B. and Alfaro, M.E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution* **21**, 1123–1133.
- Huelsenbeck, J.P., Larget, B. and Swofford, D.L. (2000). A compound Poisson process for relaxing the molecular clock. *Genetics* **154**, 1879–1892.
- Huelsenbeck, J.P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
- Huelsenbeck, J.P. and Suchard, M. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology* **56**, 975–987.
- Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In H.N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–123.
- Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.
- Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 454–458.
- Kishino, H., Thorne, J.L. and Bruno, W. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution* **18**, 352–361.
- Kuhner, M. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution* **11**, 459–468.
- Lartillot, N., LePage, T. and Blanquart, S. (2009). PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288.
- Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, 1095–1109.
- Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195–207.

- Lepage, T., Bryant, D., Philippe, H. and Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution* **24**, 2669–2680.
- Lepage, T., Lawi, S., Tupper, P. and Bryant, D. (2006). Continuous and tractable models for the variation of evolutionary rates. *Mathematical Biosciences* **199**, 216–233.
- Li, S., Doss, H. and Pearl, D. (2000). Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of the American Statistical Society* **95**, 493–508.
- Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology* **46**, 523–536.
- Mau, B. (1996). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph.D. thesis University of Wisconsin.
- Mau, B. and Newton, M.A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6**, 122–131.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.
- Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.
- Nei, M., Xu, P. and Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2497–2502.
- Newton, M., Mau, B. and Larget, B. (1999). Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. In F. Seillier-Moseiwitsch (ed.), *Statistics in Molecular Biology and Genetics*. Lecture Notes-Monograph Series, Volume 33 Institute of Mathematical Statistics, Hayward, CA, pp. 143–162.
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–93.
- Posada, D. (2003). Using Modeltest and PAUP to select a model of nucleotide substitution. In A.D. Baxevanis, D.B. Davison, R.D.M. Page, G.A. Petsko, L.D. Stein and G.D. Stormo (eds.), *Current Protocols in Bioinformatics*. John Wiley & Sons, New York, pp. 6.5.1–6.5.14.
- Rambaut, A. (2000). Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399.
- Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* **43**, 304–311.
- Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N. and Thorne, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* **20**, 1692–1704.
- Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
- Ronquist, F., Klopstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D.L. and Rasnitsyn, A.P. (2012a). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology* **61**, 973–999.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Höhna, A.D.S., Larget, B.B., Liu, L., Suchard, M.A. and Huelsenbeck, J.P. (2012b). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**, 539–542.
- Schöniger, M. and von Haeseler, A. (1994). A stochastic model and the evolution of autocorrelated DNA sequences. *Molecular Phylogenetics and Evolution* **3**, 240–247.
- Sokal, R.R. and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **28**, 1409–1438.

- Sokal, R.R. and Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco.
- Stamatakis, A., Ludwig, T. and Meier, H. (2005). RAxML-III: A fast program for maximum likelihood based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463.
- Stanton, D. and White, D. (1986). *Constructive Combinatorics*. Springer-Verlag, New York.
- Swofford, D.L. (1998). PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods. Sinauer Associates, Sunderland, MA.
- Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**, 512–526.
- Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures in Mathematics in the Life Sciences* **17**, 57–86.
- Thompson, E.A. (1975). *Human Evolutionary Trees*. Cambridge University Press, Cambridge.
- Thorne, J., Kishino, H. and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* **15**, 1647–1657.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701–1762.
- Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.
- van Ooyen, A., van den Berg, J., Mantel, N. and Weissmann, C. (1979). Comparison of total sequence of a cloned rabbit β -globin gene and its flanking regions with a homologous mouse sequence. *Science* **206**, 337–344.
- Wong, K.M. (2006). A systematic analysis of multiple alignment variability using common phylogenetic estimation parameters. Master's thesis University of California, San Diego.
- Xie, W., Lewis, P.O., Fan, Y., Kuo, L. and Chen, M.H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* **60**, 150–160.
- Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10**, 1396–1401.
- Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.
- Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306–314.
- Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* **13**, 555–556.
- Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.M.K. (2000). Codon-substitution models for heterogeneous selection pressure. *Genetics* **155**, 431–449.
- Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution* **14**, 717–724.
- Yang, Z. and Rannala, B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution* **23**, 212–226.
- Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9264–9269.
- Yang, Z. and Yoder, A.D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology* **52**, 705–716.
- Yoder, A.D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**, 1081–1090.

- Zuckerkandl, E. and Pauling, L. (1962). Molecular disease, evolution, and genetic heterogeneity. In M. Kasha and B. Pullman (eds.), *Horizons in Biochemistry*. Academic Press, New York, pp. 189–225.
- Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis University of Texas, Austin.

The Multispecies Coalescent

Laura Kubatko

Ohio State University, Columbus, OH, USA

Abstract

The advent of rapid and inexpensive sequencing technology has led to the widespread use of methods based on the multispecies coalescent for inference of species-level phylogenetic trees. In this chapter, the models underlying the multispecies coalescent are described, and the corresponding likelihood functions are derived for data consisting of a sample of gene trees with or without branch lengths and for a sample of DNA sequences. Methods that employ the multispecies coalescent to infer species trees are then described, and their use is illustrated on two empirical data sets. The chapter concludes with a brief description of the application of the multispecies coalescent to problems related to inferring the species tree, such as the inference of hybridization and species delimitation.

7.1 Introduction

The *multispecies coalescent* is a general term that refers to the use of coalescent theory (see Chapter 5) to model genealogical relationships within a population or species in the context of a phylogeny connecting them. As such, the multispecies coalescent can be thought of as a merger between the fields of population genetics, which seeks to capture population-level dynamics within species, and phylogenetics, which seeks to model relationships among distinct species across evolutionary time-scales. Development of models at the interface of these two fields has been stimulated by the increasingly widespread availability of genomic data, typically in the form of DNA sequences, for samples that include multiple individuals from within the same species for many closely related species. Effective statistical inference based on such data requires models for the variation in the sequence data at both the within-species and the between-species levels that are based on realistic assumptions concerning the processes that generate variation over time at both scales.

It has long been recognized that variation among individuals within a population can lead to evolutionary histories for specific genes, called *gene trees*, that might differ from the phylogeny that represents the evolutionary history of the species, called the *species tree*. Figure 7.1 gives an example of possible relationships between two gene trees and a corresponding species tree for four species. In the figure, the bold lines depict the species tree – the tree that represents the sequence of speciation events (dashed horizontal lines) that gave rise to the present-day

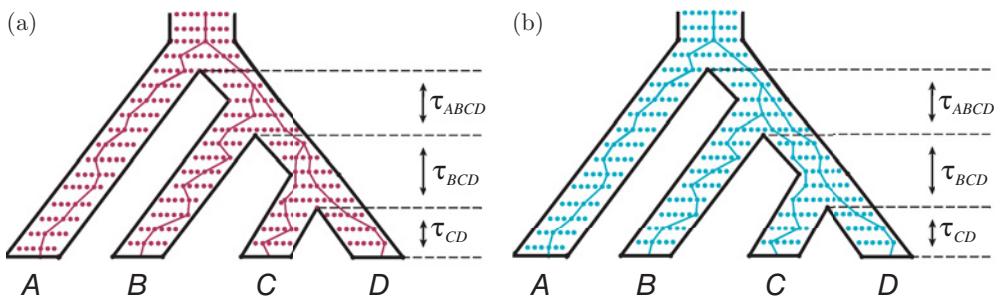


Figure 7.1 Species trees with embedded gene trees. The species tree is depicted by the thicker outline tree (black) with speciation times indicated by the horizontal dotted lines. The lengths of the intervals between speciation events are denoted by τ_X , where X lists the descendant species. The rows of colored dots inside each species tree represent the individuals within each population and the evolution of populations over generations. The gene trees embedded within each species tree show the ancestral relationships among individuals within the populations. In (a) the gene tree topology matches the species tree, while in (b) the gene tree topology differs in that the first coalescent event occurs between the genes sampled in species B , C , and D .

species for which data have been collected (the tips of the tree, denoted by the uppercase letters A , B , C , and D). The branch lengths on the species tree (denoted by τ_X , where the subscript X indicates the descendant species) represent the time between speciation events, and are usually measured in coalescent units. Coalescent units are defined as the number of generations scaled by the population size, N . For example, if a species diverged from its ancestor 10,000 generations ago, and its population size is $N = 20,000$ gene copies, then the time since speciation would be $\tau = 10,000/20,000 = 0.5$ coalescent units.

In the species tree, individual gene copies within each of the species follow their own evolutionary trajectories, subject to the constraints imposed by the species tree. For a particular gene, the multispecies coalescent model specifies that, within each branch of the species tree, a coalescent process operates to generate the genealogical history of the lineages within that branch. Figure 7.1 represents this by the horizontal rows of dots within each branch, which are intended to depict individual gene copies evolving under a genealogical process that is well-approximated by the coalescent model, such as the Wright–Fisher process (see **Chapter 5, Figure 5.1**). Additional assumptions inherent in the multispecies coalescent model are that the coalescent processes that occur in each branch of the species tree are independent of one another, conditional on the number of lineages entering and exiting the branch, and that no gene flow occurs between species (i.e. lineages cannot coalesce, or share a common ancestor, unless they are both within the same species).

With these assumptions, the multispecies coalescent model induces a probability distribution on the set of gene trees that can arise within a given species tree, which in turn induces a probability distribution on the sequence data observed for a sample of individuals from these species once a stochastic model for sequence evolution along a phylogeny is specified (see **Chapter 6, Section 6.4**). When the goal is to infer the species phylogeny and possibly to estimate associated parameters, such as population sizes, it is typical to view sequence data from many loci as having arisen from a random sample of gene trees from this probability distribution. This framework makes the assumption that the genes are unlinked, either because they are found on different chromosomes or because they are located sufficiently far apart on the same chromosome, and are thus conditionally independent given the species tree. Some models do exist for linked gene trees under the coalescent (see, for example, **Section 5.6**), but these will not be discussed further here.

Among the first uses of the multispecies coalescent to evaluate probabilities associated with gene trees are the contributions of Pamilo and Nei (1988) and Takahata (1989), which draw on the earlier work of Kingman (1982a,b,c), Tavaré (1984), and Takahata and Nei (1985), among others. However, these methods were not routinely used for inference of species trees until the early 2000s, following the introduction of a likelihood for sequence data given the species tree by Rannala and Yang (2003). Since then, many methods for inferring species-level phylogenies under the multispecies coalescent model have been proposed, and interesting debates concerning the performance of the various methods continue. In the remainder of this chapter, we provide an overview of species tree inference under the multispecies coalescent. We begin by describing in more detail the various probability distributions induced on gene trees and on sequences by the multispecies coalescent model, leading to the specification of several related likelihood functions that are commonly used for inference. We then describe the classes of methods used for inferring species phylogenies in this framework, highlighting their strengths and weaknesses in terms of accuracy, computational requirements, and robustness to model violations. We discuss estimation of model parameters other than the species tree, and briefly consider extensions of the multispecies coalescent to situations beyond a strictly bifurcating species phylogeny. We conclude with a brief outlook toward future developments in this area.

7.2 Probability Distributions under the Multispecies Coalescent

As described above, specification of a species tree, a corresponding set of speciation times, and population sizes within each branch of the species tree induces probability distributions on gene trees and on sequence data under the multispecies coalescent. In this section, we describe each of these probability distributions in detail and provide some examples.

7.2.1 Gene Tree Probabilities

In **Section 5.2.2** it was shown that, within a population, the time until the first coalescent event among a sample of k lineages follows an exponential distribution with mean $2/[(k(k - 1))]$ under the standard coalescent model. Applying this model within each branch of the species tree and assuming independence between branches allows for the computation of probabilities associated with the gene trees that evolve within the species tree. In what follows, we distinguish between the probability of a gene tree topology, which refers to the specific branching pattern of the gene tree but does not include times associated with coalescent events, and the gene tree itself, which includes both the topology and the associated coalescent times. For example, Figure 7.2 shows a three-taxon species tree with several possible gene trees embedded. Note that the two gene trees in the top row are identical in their topologies, but differ in their branch lengths because of the timing of the coalescent event joining the gene sampled from species B with that sampled from species C .

7.2.1.1 Gene Tree Topology Probabilities

Consider first the case of three species, and suppose that only a single gene is sampled at a locus for each of these species. This is depicted in Figure 7.2, with the species represented by labels A , B , and C , and the lineages sampled within each species represented by a , b , and c . For any species tree, three gene tree topologies are possible: $(a, (b, c))$, $(b, (c, a))$, and $(c, (b, a))$. Figure 7.2 shows the possible sequences of coalescent events that lead to each of the three topologies for this species tree. Note that the topology that matches the species tree $(a, (b, c))$ can result from two possible sequences of events – each of these is called a *coalescent history* (Degnan and Salter,

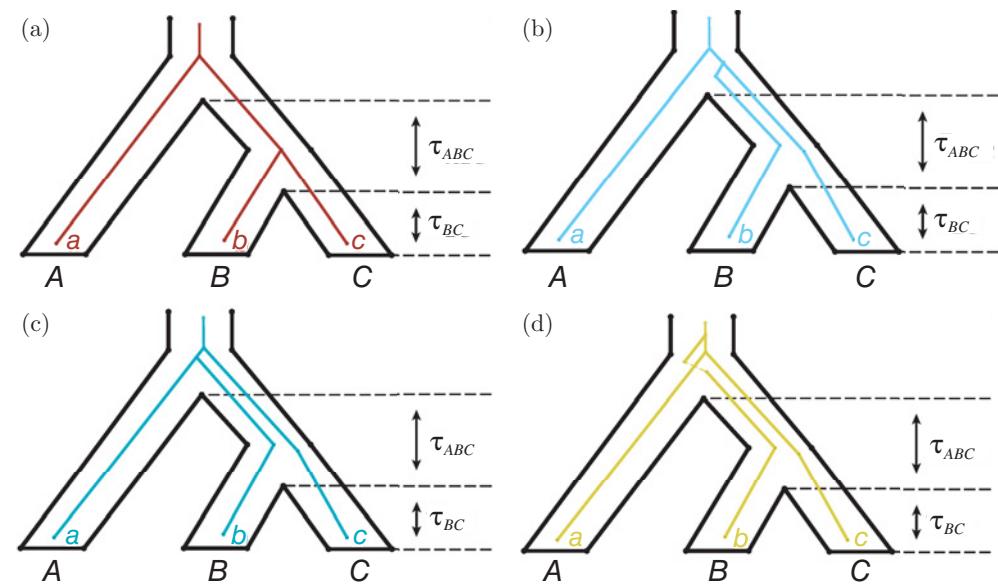


Figure 7.2 The four possible gene tree histories for the three-taxon species tree. The gene trees in (a) and (b) share the same topology as the species tree, but differ from one another in the timing of the first coalescent event. The gene trees in (c) and (d) have topologies that differ from the species tree.

2005). Thus, for three taxa, there are four possible coalescent histories and three possible gene tree topologies.

To derive the probabilities of each coalescent history, consider the population ancestral to species *B* and *C*. Using the exponential distribution with mean 1 (since there are two lineages, *b* and *c*, and hence $k = 2$ gives $2/[(k(k - 1))] = 1$), the probability of a coalescent event occurring within this interval is the probability that the time to coalescence is less than τ_{ABC} , which is $1 - e^{-\tau_{ABC}}$. The probability that lineages *b* and *c* do not coalesce in this interval is then $e^{-\tau_{ABC}}$. Looking at Figure 7.2(a), the embedded gene tree history shows that lineages *b* and *c* coalesce within the population ancestral to species *B* and *C*, and that the ancestor of these lineages coalesces with lineage *a* above the root of the species tree. The total probability for this coalescent history is then $1 - e^{-\tau_{ABC}}$, since the second coalescent event occurs eventually with probability 1. Turning to Figure 7.2(b), there is no coalescent event in the population ancestral to species *B* and *C*, which occurs with probability $e^{-\tau_{ABC}}$, and two coalescent events occur above the root. First, lineages *b* and *c* coalesce, and then the ancestor of *b* and *c* coalesces with *a*. The probability of *b* and *c* being the first two lineages to coalesce in a sample of three lineages is $\frac{1}{3}$, and thus the overall probability of this coalescent history is $\frac{1}{3}e^{-\tau_{ABC}}$ (the probabilities across both populations of the species tree can be multiplied as a consequence of the assumption of independent evolution within branches of the species tree). The gene tree history probabilities for Figure 7.2(c) and (d) are computed similarly. Because the gene tree topologies in Figure 7.2(a) and (b) are the same, the overall probability for the gene tree topology $(a, (b, c))$ will be the sum of the two history probabilities, $1 - e^{-\tau_{ABC}} + \frac{1}{3}e^{-\tau_{ABC}} = 1 - \frac{2}{3}e^{-\tau_{ABC}}$. These probabilities are shown below the corresponding gene tree in Figure 7.3(a). Note that these probabilities sum to 1, that is, this is a valid probability distribution on the discrete set of topologies, as well as on the discrete set of coalescent histories.

Examination of the gene tree topology distribution in the three-taxon case provides several important insights into the multispecies coalescent model. First, consider the effect of the length of the interval between speciation events, τ_{ABC} . Figure 7.3(b) plots the length of this

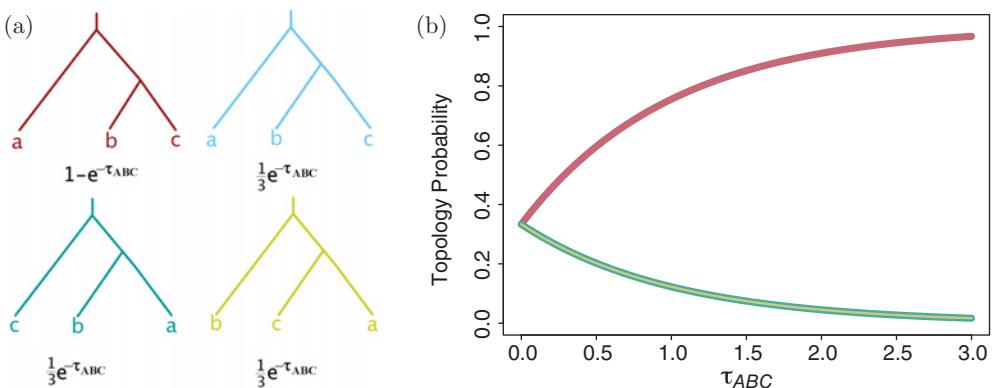


Figure 7.3 Effect of the length of time between speciation events on the gene tree topology probability distribution for three-taxon species trees. In (a) the four coalescent histories are given with their probabilities, while in (b) these probabilities are plotted as a function of the length of time between speciation events, τ_{ABC} . The top, red line gives the topology probability for gene tree $(a, (b, c))$ (the sum of the probabilities of the first two histories in (a)), the bottom, green line gives the topology probability for gene tree $(c, (b, a))$, and the gold line gives the topology probability for gene tree $(b, (c, a))$.

branch (in coalescent units) on the x -axis and the probability of each of the three gene tree topologies (shown in Figure 7.3(a)) on the y -axis. Note that as the length of the branch increases, the probability of the topology that matches the species tree increases as well. This is intuitive, in the sense that a longer interval between speciation events allows more time for the coalescent event between lineages b and c to occur, and coalescence of these lineages within this branch will *always* lead to a gene tree that matches the species tree. This is a desirable situation from an inference standpoint, as agreement between gene trees and the species tree with high probability means that only a handful of loci would be needed to obtain a reliable estimate of the species tree. Second, note that even when the length of the interval between speciation events is small, the gene tree with topology matching the species tree always has probability at least as large as the two non-matching topologies. When this length is 0, the three topologies each have probability $\frac{1}{3}$, which again matches intuition, as a length of 0 would imply simultaneous speciation. These results provide insight into potential methodologies for estimating the species-level phylogeny, as it implies that the most frequently occurring topology across a sample of loci for three species will be the topology that matches the species tree under the assumptions of the multispecies coalescent.

Extending these ideas to the four-taxon setting illustrates some additional principles. The basic method for computing topology probabilities remains the same, namely, coalescent histories are enumerated, probabilities are computed for each history using the relevant exponential distributions, and the history probabilities consistent with each topology are summed to give the overall topology probability. To compute the probabilities of individual histories, the following formula, based on the convolution of exponential densities as described above, can be used to compute the probability that u lineages coalesce into v lineages within a branch of length t (Tavaré, 1984; Watterson, 1984; Takahata and Nei, 1985; Rosenberg, 2002):

$$P_{uv}(t) = \sum_{j=v}^u e^{-j(j-1)t/2} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^{j-1} \frac{(v+y)(u-y)}{u+y}. \quad (7.1)$$

For example, when $u = 2$ and $v = 1$ (meaning that two lineages coalesce to form one lineage within a branch of length t), $P_{21}(t) = 1 - e^{-t}$ and $P_{22}(t) = e^{-t}$, in agreement with the probabilities used in the computations for the three-taxon case. Note that the number of coalescent

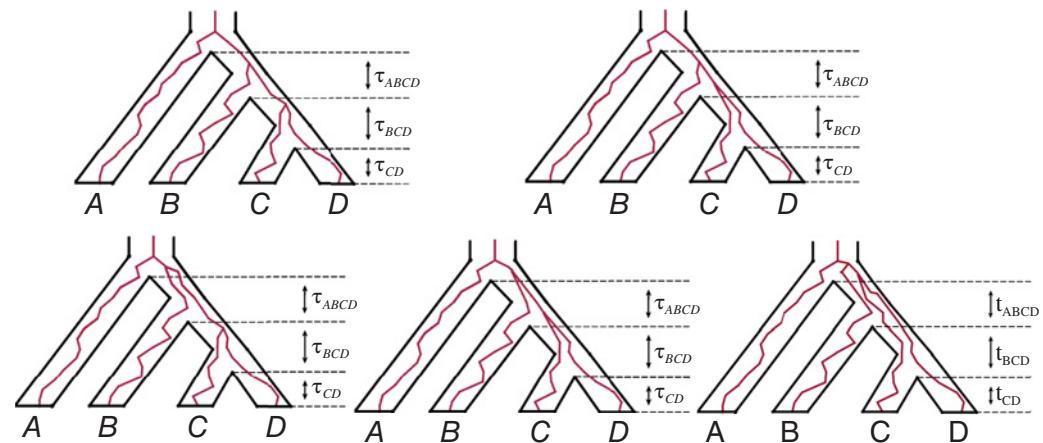


Figure 7.4 Coalescent histories for the asymmetric four-taxon species tree. The gene trees in each panel differ in the timing of their coalescent events.

histories possible for each topology may increase significantly compared to the three-taxon case. For example, in the case of four taxa, the gene tree topology that matches the asymmetric species tree has five distinct histories (see Figure 7.4), and in total there are 31 distinct histories corresponding to the 15 possible gene tree topologies in this case.

Figure 7.5 shows several example probability distributions on the 15 possible gene tree topologies for the species tree in Figure 7.1 for various choices of the branch lengths τ_{BCD} and τ_{ABCD} (measured in coalescent units). When τ_{BCD} and τ_{ABCD} are both fairly long (blue bars; $\tau_{BCD} = \tau_{ABCD} = 2.0$), the gene tree topology that matches the species tree (leftmost bar) has much higher probability than that of all of the other possible gene trees. When these lengths are both shortened (green bars; $\tau_{BCD} = \tau_{ABCD} = 0.5$), the probability is more evenly distributed across topologies, with the topology matching the species tree topology having a probability of only 32%. When branch length τ_{ABCD} is shortened even further (gold bars; $\tau_{BCD} = 1.0$ and $\tau_{ABCD} = 0.01$), a gene tree that differs in topology from the species tree (gene tree $((a, b), (c, d))$) has the highest probability. Such gene trees (i.e. those with topologies that differ from the species tree and that have higher probability than the gene tree with topology

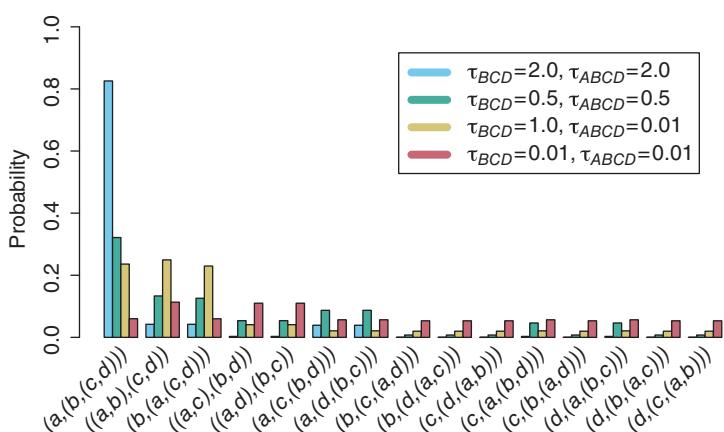


Figure 7.5 Gene tree topology probability distributions for the four-taxon species tree. Heights of the bars give the probabilities for the 15 possible rooted gene tree topologies for four taxa. Bars are colored according to the lengths of the speciation intervals on the species tree in Figure 7.1.

matching the species tree) have been termed *anomalous gene trees*, and the space of species tree branch lengths that give rise to anomalous gene trees is called the *anomaly zone* (Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008). While the three-taxon computations above demonstrate that there are no anomalous gene trees for three-taxon species trees, the work of Degnan and Rosenberg (2006) and Rosenberg and Tao (2008) establishes conditions under which anomalous gene trees exist for four or more species. Interestingly, the choice of branch lengths leading to the distribution displayed by the red bars ($\tau_{BCD} = \tau_{ABCD} = 0.01$) shows that, when both τ_{BCD} and τ_{ABCD} are small, several gene tree topologies may have probability larger than that of the gene tree that matches the species tree, that is, a species tree may have more than one anomalous gene tree. The intuition for this result is straightforward: when branch lengths are short, most coalescent events occur above the root, and there are more orderings of coalescent events leading to symmetric gene trees than to asymmetric gene trees, thus resulting in higher probability assigned to symmetric gene trees.

Several software tools are available for computing gene tree topology distributions in more general cases. Gene tree topology probabilities can be computed in the software Hybrid-Coal (Zhu and Degnan, 2017), which supersedes the previous software COAL. The probabilities of matching gene tree topologies can be computed using tools from Tian and Kubatko (2017), and Tian and Kubatko (2016) provided gene tree topology probability calculations for three-taxon species trees when gene flow between sister taxa following speciation is possible (computations are provided in the software COALGF). Long and Kubatko (2018) noted that anomalous gene trees for the three-taxon case are possible in the presence of gene flow according to the model of Tian and Kubatko (2016).

Given the potential for extensive variation in gene tree topology probabilities predicted by the multispecies coalescent, a natural question is whether empirical data demonstrate the levels of variation that the model predicts. To examine this, consider the work of Ebersberger *et al.* (2007), who collected whole-genome data on five primate species (Human, Chimp, Gorilla, Orangutan, and Rhesus Macaque). Gene trees were estimated for 23,210 distinct loci, and across these loci, 76.6% supported a tree with topology (Gorilla, (Human, Chimp)), 11.4% supported a tree with topology (Chimp, (Human, Gorilla)), and 11.5% supported a tree with topology (Human, (Gorilla, Chimp)). Using the parameters inferred by Rannala and Yang (2003) and the three-taxon species tree calculations above, the predicted proportions of each of the gene trees are 79.1%, 9.9%, and 9.9%, respectively, indicating remarkably good fit of the multispecies coalescent to these data. Further examples can be found in Degnan and Rosenberg (2009).

7.2.1.2 Gene Tree Probability Density

Rather than deriving the probability distribution on the discrete set of gene tree topologies, the probability distribution on the space of gene trees with branch lengths may be desired. This distribution is more complicated to represent, as it involves both a discrete component (the gene tree topology) and a continuous component (the vector of branch lengths, or equivalently, coalescent times). An expression for this probability density was first written explicitly in the phylogenetic setting by Rannala and Yang (2003), though similar versions in population genetics settings had been considered earlier (e.g. Beerli and Felsenstein, 1991). The essential idea in the calculation is to once again consider each species tree branch individually, and to include terms in the density for each ‘event’ that happens within each branch. Possible events include coalescent events, which contribute a term from the corresponding exponential probability density evaluated at the time of the coalescence, and failures to coalesce within branches, which contribute a factor for the probability of not coalescing, again obtained from the corresponding exponential distribution. The reader is referred to Rannala and Yang (2003) for a formal derivation of the probability density. Here, a worked example is presented that will allow the reader to generalize to the arbitrary case.

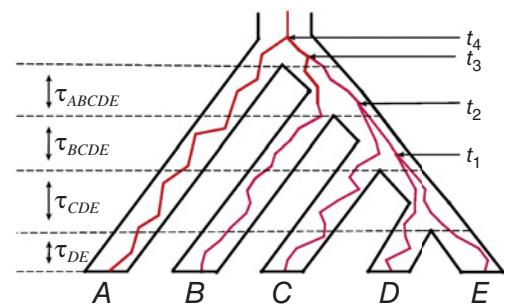


Figure 7.6 Example five-taxon species tree used to demonstrate computation of the gene tree probability density.

Consider the species tree and embedded gene tree in Figure 7.6. We will compute the probability density function for this particular gene tree topology with coalescent times t_1, t_2, t_3 and t_4 . For simplicity of notation, assume that the coalescent times are measured as the time from the beginning of the interval in which they occur, rather than as cumulative time from the tips of the tree. Working from the tips of the tree toward the root, each interval between speciation events is evaluated to determine its contribution to the density. For example, in the interval corresponding to the population ancestral to species D and E , which has length τ_{CDE} , two lineages are available to coalesce, but they do not. Therefore, this interval contributes the probability that two lineages do not coalesce over length of time τ_{CDE} , $P_{22}(\tau_{CDE}) = e^{-\tau_{CDE}}$, to the density. Considering now the interval ancestral to species C, D , and E with length τ_{BCDE} , we see that there is a coalescent event between two of the three lineages at time t_1 , which has probability density $3e^{-3t_1}$, and that the remaining two lineages fail to coalesce, which occurs with probability $e^{-(\tau_{BCDE}-t_1)}$. The probability of the lineages from D and E being the two involved in coalescent event is $\frac{1}{3}$, and thus the total density assigned to this interval is $\frac{1}{3}(3e^{-3t_1})(e^{-(\tau_{BCDE}-t_1)})$. The interval that represents the population ancestral to species B, C, D , and E is similar, with the contribution to the probability density given by the interval equal to $\frac{1}{3}(3e^{-3t_2})(e^{-(\tau_{ABCDE}-t_2)})$. The interval above the root of the species tree includes two coalescence events, and contributes density $\frac{1}{3}(3e^{-3t_3})(e^{-(t_4-t_3)})$.

Thus the overall probability density of this gene tree, denoted by (G, t_1, t_2, t_3, t_4) , given the species tree, denoted by (S, τ) where $\tau = (\tau_{DE}, \tau_{CDE}, \tau_{BCDE}, \tau_{ABCDE})$, is the product of all of these terms, due to the assumption of independence across branches of the species tree,

$$f_G(t_1, t_2, t_3, t_4 | (S, \tau)) = e^{-2t_1} e^{-2t_2} e^{-2t_3} e^{-t_4} e^{-\tau_{ABCDE}} e^{-\tau_{BCDE}} e^{-\tau_{CDE}}, \\ 0 \leq t_1 < \tau_{BCDE}, 0 \leq t_2 < \tau_{ABCDE}, 0 \leq t_3 < t_4 < \infty.$$

It is important to note that the coalescent times, t_1, t_2, t_3, t_4 , are defined specifically for this gene tree topology, and that their range depends on the particular coalescent history. Thus, although the ideas involved in computing the probability density for a gene tree are straightforward, a closed-form expression is not available, and the reader is again referred to Rannala and Yang (2003) for a more general expression.

The gene tree probability density has many uses. The first is as the basis of a likelihood function for species tree inference methods that make use of the likelihood, such as the Bayesian and maximum likelihood (ML) frameworks. This will be discussed more fully in Section 7.2.3 below. In addition, the gene tree density could be used to compute topology probabilities by integrating over coalescent times, or to compute marginal, joint, or conditional distributions of interest. For example, referring again to Figure 7.6, the marginal distribution of t_1 for this gene tree is

given by

$$f_{\mathcal{G}}(t_1|(S, \tau)) = \int_0^{\infty} \int_0^{t_4} \int_0^{\tau_{ABCDE}} f_{\mathcal{G}}(t_1, t_2, t_3, t_4 | (S, \tau_{ABCDE}, \tau_{BCDE}, \tau_{CDE})) dt_2 dt_3 dt_4, \\ 0 < t_1 < \tau_{BCDE}. \quad (7.2)$$

Similarly, the covariance and correlation between collections of coalescent times can be derived. Finally, the gene tree probability density is used to compute site pattern probabilities, as described next.

7.2.2 Site Pattern Probabilities

Thus far, the probability distribution on gene tree topologies and on the overall gene trees (both topology and coalescent times) induced by the species tree have been described. Under the multispecies coalescent model, sequence data arise along each of the gene trees, and thus the species tree also induces a probability distribution on the nucleotides found at the tips of the tree, which we call a site pattern. For example, in Figure 7.6, we could consider the site pattern ‘AACGT’, meaning that species *A* has nucleotide A at the site under consideration, species *B* also has nucleotide A, and species *C, D*, and *E* have nucleotides C, G, and T, respectively. For a species tree with five taxa and one lineage sampled per taxon, there are $4^5 = 1024$ possible site patterns, and one may wish to consider the probability of each these. We refer to this as the site pattern probability distribution. In this section, we provide an expression for a site pattern probability for an arbitrary number of species and of lineages sampled within each species. In practice, however, these probabilities are difficult or impossible to compute for more than four species and for more than one lineage per species.

First, assume that a nucleotide substitution model has been specified to describe the process of evolution along each of the gene trees. Commonly-used nucleotide substitution models are described in **Section 6.6.1**, and the probability of a particular site pattern arising along gene tree $(\mathcal{G}, \mathbf{t})$, where \mathcal{G} denotes the gene tree topology and \mathbf{t} denotes the vector of coalescent times, is given by equation (6.21) (note that in the notation of **Chapter 6** a site pattern is denoted by \mathbf{x} (denoted here by \mathbf{p}), the gene tree topology and branch lengths are denoted by ψ (denoted here by $(\mathcal{G}, \mathbf{t})$), and the substitution model parameters are denoted by θ (these are omitted here)). Next, note that each gene tree is a realization of a random variable with probability density described in Section 7.2.1.2 for the given species tree. Thus, to find the probability of a site pattern given the species tree, the probability distribution corresponding to the gene trees must be integrated over. Let \mathcal{H} denote the set of all possible coalescent histories, and let (S, τ) denote the species tree, where S denotes the species tree topology and τ denotes the vector of lengths of the intervals between speciation events. Let \mathbf{p} represent an arbitrary site pattern probability (i.e. $\mathbf{p} = \text{AACGT}$ in the example above). The probability of site pattern \mathbf{p} is then

$$P(\mathbf{p}|(S, \tau)) = \sum_{\mathcal{H}} \int_{\mathbf{t}} f_h(\mathbf{t}|(S, \tau)) P(\mathbf{p}|(h, \mathbf{t})) d\mathbf{t}, \quad (7.3)$$

where the sum is over the set of all possible coalescent histories, $f_h(\mathbf{t}|(S, \tau))$ represents the probability density for the gene tree with coalescent history h and vector of coalescent times \mathbf{t} , $P(\mathbf{p}|(h, \mathbf{t}))$ is the probability of site pattern \mathbf{p} arising on gene tree history h with vector \mathbf{t} of coalescent times, and the integral is over the allowable range of coalescent times determined by the particular history h under consideration. Equation (7.3) has been called the Felsenstein equation (Hey and Nielsen, 2007) because it was first given by Felsenstein (1988) in the context of estimation of population genetics parameters via integration over gene genealogies.

Table 7.1 Number of coalescent histories when the species tree is symmetric (second column) or asymmetric (third column) and the number of topologies (fourth column) for varying numbers of taxa (reproduced from Degnan and Salter (2005))

Taxa	Number of histories		Number of topologies
	Asymmetric trees	Symmetric trees	
4	5	4	15
5	14	10	105
6	42	25	945
7	132	65	10,395
8	429	169	135,135
9	1,430	481	2,027,025
10	4,862	1,369	34,459,425
12	58,786	11,236	13,749,310,575
16	9,694,845	1,020,100	6.190×10^{15}
20	1,767,263,190	100,360,324	8.201×10^{21}

Careful consideration of equation (7.3) demonstrates the difficulty in computing site pattern probabilities in the general case. First, when n lineages are under consideration, the number of possible site patterns is 4^n , and thus complete computation of the site pattern probability distribution would mean computation of 4^n distinct probabilities. Second, both the number of gene trees and the number of histories for each gene tree grow faster than exponentially in the number of species. Table 7.1 shows the number of possible histories for gene trees matching the species tree for up to 20 taxa. Taking into consideration that each history creates different bounds on possible coalescent times, it is clear that exact computation of the site pattern probability distribution is computationally infeasible for more than a handful of species.

Nonetheless, computation of site pattern probabilities even for a small number of taxa has proven to provide valuable insights into the variation in the distribution of data that can be expected under the multispecies coalescent. For example, consider again the four-taxon species tree in Figure 7.1 and consider the branch lengths used to compute topology probabilities in Figure 7.5. Figure 7.7 shows the corresponding site pattern probability distributions over the

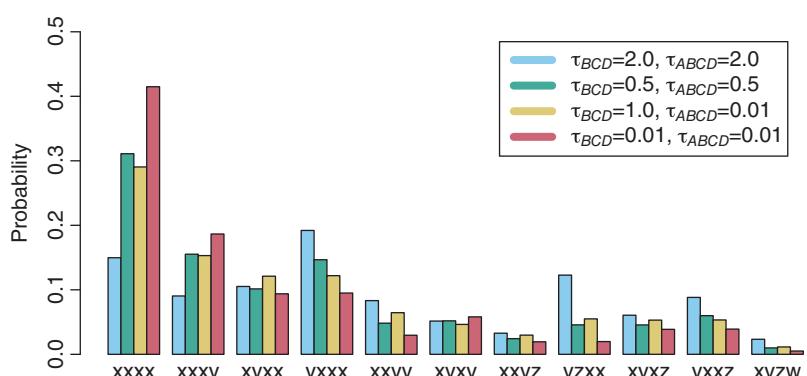


Figure 7.7 Example site pattern probability distributions.

15 distinct site patterns for the Jukes–Cantor model. For example, the first set of bars in Figure 7.7 give the probability of site patterns AAAA, CCCC, GGGG, and TTTT for the same four lengths of speciation intervals considered in Figure 7.5. From this we observe that shorter branch lengths in the species tree increase the probability of the constant site pattern, resulting in less variation in the sequences overall, in spite of substantial variation in the gene tree topologies in this case (see Figure 7.5). As another example, consider the pattern labeled ‘yxxx’, which corresponds to the case in which species *A* has a nucleotide that differs from that of species *B*, *C*, and *D* (e.g. ACCC, TGGG, etc.). In this case, we see that the longest species tree branch lengths lead to the highest probabilities. Overall, Figure 7.7 highlights the fact that variation in the species tree directly impacts the observed data at the sequence level. Furthermore, the site pattern probability distribution is the basis for one class of species tree inference procedures, discussed further in Section 7.3.3.

7.2.3 Species Tree Likelihoods under the Multispecies Coalescent

Having described probability distributions arising under the multispecies coalescent model, we now turn our attention to statistical inference of the species tree under the multispecies coalescent. In either the ML or Bayesian inferential frameworks, successful estimation of the species phylogeny and quantification of uncertainty in the estimate depend on the ability to compute the likelihood function. In the context of species tree inference, at least three distinct likelihood functions have been proposed, depending on what is considered ‘data’. The most natural definition of ‘data’ is to consider the sequence data itself, necessitating use of the site pattern probability distribution described above for computation of the likelihood of a species tree. However, given the complexity in computing site pattern probabilities, inference methods based on first estimating gene trees for each locus have also been proposed. In these cases, the ‘data’ are considered to be the estimated gene trees, either with or without associated coalescent times. Each possible likelihood function will be described carefully below.

7.2.3.1 The Likelihood Based on Gene Tree Topologies

Suppose that one has available a collection of gene tree topologies, $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_L$, for L loci. Then, assuming that the loci are unlinked and thus that the gene tree topologies are independent conditional on the species tree, the likelihood of the species tree, (S, τ) , is given by

$$\mathcal{L}((S, \tau)) = \prod_{l=1}^L P(\mathcal{G}_l | (S, \tau)), \quad (7.4)$$

where $P(\mathcal{G}_l | (S, \tau))$ is the gene tree topology probability for locus l for the given species tree, computed as described in Section 7.2.1.1. The fact that the species tree topology and branch lengths are identifiable from the gene tree topology distribution was established by Allman *et al.* (2011), making likelihood-based estimation of the species tree and corresponding speciation times from the gene tree topology distribution feasible.

7.2.3.2 The Likelihood Based on Gene Trees

If instead one has available a collection of gene trees with coalescent times for the L loci, then a similar expression for the likelihood can be written by replacing the gene tree topology probability in the previous section by the gene tree density from Section 7.2.1.2 to give

$$\mathcal{L}((S, \tau)) = \prod_{l=1}^L f_{\mathcal{G}_l}(\mathbf{t}_l | (S, \tau)), \quad (7.5)$$

where $f_{\mathcal{G}_l}(\mathbf{t}_l | (\mathcal{S}, \tau))$ is the gene tree probability density for locus l with topology \mathcal{G}_l and corresponding coalescent times \mathbf{t}_l . Because the gene trees with coalescent times provide more information beyond simply the topologies, the species tree and speciation times are clearly identifiable using this likelihood.

7.2.3.3 The Likelihood Based on Multilocus Data

Finally, suppose that one has available a collection of sequence alignments arising from L unlinked loci. It is assumed that no recombination occurs within a locus, so that all sites in each alignment evolve along the same gene tree. Given that locus l has gene tree \mathcal{G}_l with associated coalescent times \mathbf{t}_l , the probability of the data in the alignment of length k_l is given by $\prod_{j=1}^{k_l} P(\mathbf{p}_j | (\mathcal{G}_l, \mathbf{t}_l))$, where \mathbf{p}_j is the site pattern in column j of the alignment. To form the overall likelihood, we must integrate over gene trees according to the gene tree probability density, and take the product across loci,

$$\mathcal{L}((\mathcal{S}, \tau) | D_1, D_2, \dots, D_L) = \prod_{l=1}^L \left(\sum_{\mathcal{H}} \int_{\mathbf{t}_h} \left(\prod_{j=1}^{k_l} P(\mathbf{p}_j | (\mathcal{G}_h, \mathbf{t}_h)) \right) f_h(\mathbf{t}_h | (\mathcal{S}, \tau)) d\mathbf{t}_h \right), \quad (7.6)$$

where D_l indicates the alignment for locus l . In the case of single nucleotide polymorphism (SNP) data, each locus has length $k_l = 1$ and the innermost product disappears. Chifman and Kubatko (2015) established identifiability of the species tree topology \mathcal{S} for the case in which $k_l = 1$ for all loci l arising under the GTR + I + Γ substitution model (see Section 6.6.1) and all submodels. Similar arguments can be used to establish identifiability of the speciation times for four taxa. Identifiability of speciation times for species trees of arbitrary size remains an open question.

7.2.4 Model Assumptions and Violations

As described above, the multispecies coalescent provides a model at the interface of population genetics and phylogenetics based upon well-established theoretical considerations from both fields. The various likelihood functions provide a variety of options for statistical inference of the species tree and associated parameters, which will be described in more detail in the next section. Before moving on to this topic, however, it is worth reviewing the assumptions made by the multispecies coalescent, and it is important to evaluate methods for inference in regard to their robustness to violations of these assumptions. First, the multispecies coalescent makes the assumption that all gene flow ceases immediately following a speciation event. This is perhaps the most restrictive assumption of the model, in the sense that speciation is widely believed to be a more gradual process, with the rate of gene flow diminishing over time. Several of the inference methods described below are somewhat robust to this assumption, while others are extremely sensitive to it. A second assumption is that within each ancestral population, evolution proceeds according to a process that is well approximated by the coalescent model, such as the Wright–Fisher model. Because the coalescent model is believed to provide a reasonable approximation in a variety of settings, methods for inferring species trees are believed to be fairly robust to violations of this assumption. Finally, the multispecies coalescent as described above assumes a strictly bifurcating species history. When processes such as hybridization or horizontal gene transfer have played a role in the history of speciation, methods must be designed to accommodate these processes. In Section 7.4.2 some extensions of the techniques described in this chapter are briefly discussed.

7.3 Species Tree Inference under the Multispecies Coalescent

Figure 7.8 summarizes the process by which the multispecies coalescent leads to observed data for species-level phylogenetic inference. In particular, the model specifies that gene trees arise from the species tree according to the gene tree probability density $f((G, t)|(S, \tau))$ described in Section 7.2.1.2, and sequence data arise on the individual gene trees according to the distribution $f(D|(G, t))$. This two-step process is depicted by the arrows in Figure 7.8. In the empirical setting, sequence data at the tips of the tree are collected, and the goal is to use these data to estimate the species tree, (S, τ) . Because computation of the likelihood is intractable for all but the smallest problems, inference in a formal likelihood or Bayesian statistical framework is challenging. For this reason, many methods that simplify the problem in some way have been proposed. Below the various classes of methods that have been developed for species tree estimation under the multispecies coalescent are described in detail, and the advantages and disadvantages of each are presented.

7.3.1 Summary Statistics Methods

The first class of methods has been referred to as summary statistics methods (Liu *et al.*, 2009a) or shortcut methods (Gatesy and Springer, 2014), based on the technique used to simplify the computations. Summary statistics methods are those for which species tree inference is carried out in two distinct steps. In the first step, the alignments for each of the individual loci are used to estimate gene trees for each locus, while in the second step, the estimated gene trees are used as input into a procedure that estimates the species tree. The various methods proposed differ in how they use the gene trees to estimate the species tree.

One group of summary statistics methods for species tree inference uses the ML framework via the likelihood functions defined in equations (7.4) and (7.5). For methods based on only topologies (e.g. Wu, 2012) the only requirement is that rooted gene trees be estimated for each locus, while for methods that use the topologies and branch lengths (e.g. Kubatko *et al.*, 2009), the coalescent times for the estimated phylogenies for each locus must also be provided. The method used to estimate the branch lengths for the gene trees must provide units that can be converted into coalescent units. For example, branch lengths estimated under the standard Markov-based nucleotide substitution models (see **Section 6.6.1**) are said to be in ‘mutation’ units (i.e. the expected number of mutations per site per generation), and can be converted to coalescent units by multiplying by the reciprocal of the scaled mutation rate. It may also be necessary to scale by locus-specific mutation rates, since substantial variation in mutation rates may exist for genome-scale data.

When using only gene tree topologies as input, the ML estimate of the species tree must be obtained using a heuristic search over the space of species trees. However, when the gene

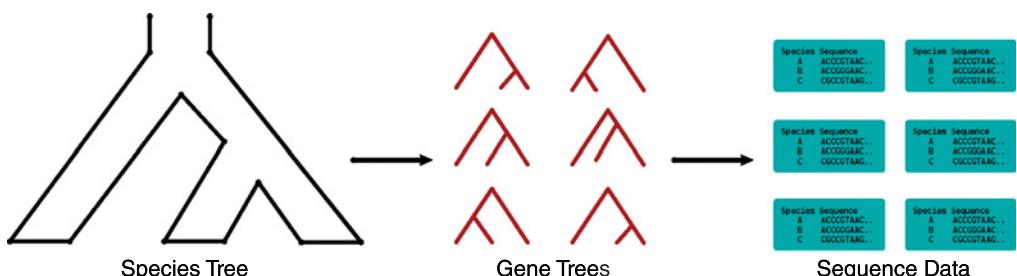


Figure 7.8 Graphical overview of the data generation process under the multispecies coalescent model.

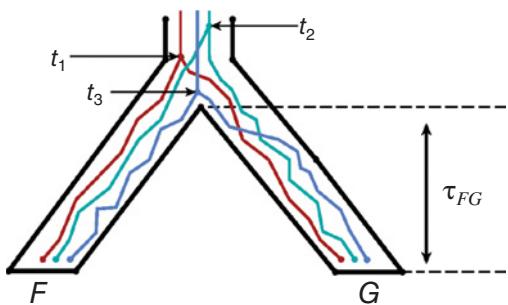


Figure 7.9 Relationship between gene tree coalescent times and the speciation time for two species F and G . Three gene trees with coalescent times t_1 , t_2 , and t_3 (red, green, and blue lines, respectively) are embedded within the species tree, illustrating that all gene coalescent times must pre-date the speciation time separating two distinct lineages under the multispecies coalescent model.

trees used as input include both topology and branch lengths, the ML species tree can be found algorithmically under certain conditions, as shown by Liu and Pearl (2010) and Mossel and Roch (2010). To see this, consider two species F and G , and suppose that for each locus l ($l = 1, 2, \dots, L$) the estimated times to coalescence between two lineages sampled from these species are denoted by t_1, t_2, \dots, t_L . Here the times are viewed as the cumulative time from the tips of the tree to the coalescent event, in contrast to Section 7.2.1.2. Under the assumptions of the multispecies coalescent, all coalescent times between lineages sampled from species F and G must pre-date the speciation time for F and G , and it is intuitive that the ML estimate of the time at which F and G became distinct species is the minimum of (t_1, t_2, \dots, t_L) (see Figure 7.9). To find the ML estimate of the species tree, all pairwise speciation time estimates are obtained, and the two species with the smallest estimated speciation time are joined. New speciation time estimates are then computed for any pair that involved species F or G with a composite node, say FG , that represents their ancestor. The pair with the next smallest estimated speciation time are then joined, and the process continues until a bifurcating tree is obtained. When the population size is constant throughout the tree, the estimate formed in this manner is the ML estimate. Details are given in Liu and Pearl (2010) and Mossel and Roch (2010), and Kubatko *et al.* (2009) showed how to use a similar technique to compute the likelihood of any species tree, which allows for likelihood-based comparisons among competing hypothesized speciation relationships.

The existence of an algorithm for finding the ML estimate of the species tree given gene trees as input leads to a computationally very efficient inference method, as the ML estimate of the species tree can be computed in seconds, even for hundreds of taxa. In practice, however, the resulting estimate is often a poor summary of the species-level information contained in a multilocus data set. When the species under consideration are very closely related, many loci will show little or no variation, leading to estimated coalescent times that are either 0 or extremely small. Because the ML estimate of the speciation time is the minimum across *all* loci, a single non-variable locus will cause the speciation time to be estimated as 0, forcing a multifurcation in the estimated species tree. Use of the average estimated coalescent time across loci or of ranks of estimated coalescent times has been proposed and implemented in software Liu *et al.* (2009b); however, these techniques have not been widely used.

Another group of summary statistics methods uses features of the estimated gene trees as input to the species tree inference step, rather than the estimated trees themselves. Although such methods are often motivated by properties of the multispecies coalescent, several do not explicitly make this assumption. The two most popular of these methods, MP-EST (Maximum Pseudo-likelihood for Estimating Species Trees; Liu *et al.*, 2010) and ASTRAL (Accurate Species Tree ALgorithm; Mirarab *et al.*, 2014; Mirarab and Warnow, 2015), both use the fact that, under the standard multispecies coalescent model, anomalous gene trees do not exist when only three species are considered or when four-taxon trees are viewed as unrooted. This

means that if only these types of relationships were considered, the most frequently occurring gene tree across loci would have the same topology as the species tree.

MP-EST uses this property of the multispecies coalescent to build an inference procedure as follows. From the gene trees estimated in the first step of the procedure, the frequencies of all rooted triple relationships are recorded. For a given set of three taxa, the number of times each of the three possible trees is observed is viewed as a sample from a multinomial distribution. Maximizing the likelihood corresponding to this multinomial distribution over the three possible three-taxon topologies leads to an estimate of the species-level relationships for that set of three taxa. Maximizing over all sets of three taxa leads to an estimate of the overall species-level phylogeny. The method is referred to as a ‘pseudo-likelihood’ method, because the sets of three taxa are not independent of one another because they may include some of the same taxa. For example, species *A*, *B*, *C* and species *A*, *B*, and *D* are both triples used in the calculation, and the overall likelihood that is maximized is the product of the multinomial likelihoods for all such triples. This overall likelihood can also be used to find estimates of the branch lengths in coalescent units. When the gene trees are known without error, MP-EST is statistically consistent since the triple relationships are increasingly likely to be correctly inferred as the number of loci increases, and the species tree is completely determined by the collection of rooted triple relationships (Steel, 1992).

If gene trees are viewed as unrooted, then the most frequently occurring quartet relationships are those that match the species tree under the multispecies coalescent (which can be seen by considering the trees in Figure 7.5 as unrooted and summing the corresponding probabilities). The method used by the software ASTRAL thus involves recording the quartet frequencies in the input set of gene trees, and searching for the species tree that shows maximum compatibility with the observed quartet frequencies (for the specific details of computing the measure of compatibility, see Mirarab *et al.*, 2014; Mirarab and Warnow, 2015). Like MP-EST, ASTRAL is statistically consistent when the gene trees are known without error. The basic method can also be used to derive estimates of the branch lengths and a measure of support on nodes of the estimated tree (Sayyaria and Mirarab, 2016). While the algorithms underlying both MP-EST and ASTRAL are motivated by properties of gene-level relationships among species under the multispecies coalescent, this model is not explicitly assumed for inference. They will be reasonable methods for species tree inference under any model for which the most frequent rooted triple relationships (MP-EST) or most frequent unrooted quartet relationships (ASTRAL) across loci match those found in the species-level phylogeny.

The primary advantage of summary statistics methods is their computational efficiency. While the first step of the inference procedure does involve estimation of gene trees for each of the L loci, which may be computationally intensive, this can be easily parallelized. The second step can generally be carried out rapidly, often requiring only seconds on a standard desktop computer even when the number of species under consideration is large. However, all summary statistics methods rely on accurate estimation of the gene trees for the individual loci. Thus, these methods tend to perform better as loci increase in length (i.e. number of base pairs) and as more loci become available. Even when accurate estimates of the individual gene trees can be obtained, the two-stage procedure for species tree estimation has been criticized because it fails to account for variation in the individual gene tree estimates, instead treating them as fixed and known in the second step of the estimation procedure. Some attempts to remedy this have been made (e.g. one could think of bootstrapping individual genes and re-estimating gene trees for the bootstrap samples, then repeating the species tree estimation procedure for the bootstrapped samples), but the methods then become computationally intensive, losing their advantage over the full-data methods discussed below. In addition, none of the methods incorporate a natural measure of variability in the estimated species phylogeny. This shortcoming is easily overcome by implementing a bootstrap procedure to allow the estimation of bootstrap

support on individual nodes of the species tree, again at the expense of increased computational requirements.

7.3.2 Bayesian Full-Data Methods

Because computation of the likelihood of the species tree from a sample of DNA sequence alignments using equation (7.6) is computationally infeasible, maximum likelihood inference of species trees using the sequence data is not generally possible. However, model-based inference in a Bayesian setting is possible when Markov chain Monte Carlo (MCMC) techniques are used to approximate the posterior distribution (see **Chapter 1**). While several distinct software packages have been proposed for MCMC-based species tree inference, each utilizes the same basic idea to circumvent computation of the full likelihood in equation (7.6). Rather than estimating the posterior distribution of the species tree only, the methods estimate the joint posterior distribution of the species tree and the individual gene trees, and all model parameters associated with these. This allows for computations of acceptance probabilities needed as part of the MCMC algorithm to be based on current values of these parameters in such a way that computation of the full likelihood is not needed. For example, in determining whether a newly proposed gene tree for a locus should be accepted, the only quantities that are relevant are the species tree and the data for that gene. Similarly, given a set of gene trees, the likelihood of the species tree can be computed directly from the gene trees using equation (7.5). The precise way in which these calculations are carried out varies from method to method, and the reader is referred to the description of each specific method for details – for example, MrBayes/BEST (Liu and Pearl, 2007), *BEAST (Heled and Drummond, 2010), and BPP (Yang, 2015)).

The advantages of use of a Bayesian framework for species tree inference are many. First, the original data – the DNA sequence alignments – are used directly for inference, without need for estimation of individual gene trees as in the summary statistics methods. This means that variation in the alignments themselves is directly incorporated into the inference procedure. Second, these methods provide a fully model-based inference framework, incorporating not only the multispecies coalescent but also variation in the substitution process assumed to govern the evolution of the sequences at each of the loci. Most implementations also allow for models of rate variation across loci, and for the selection of a range of prior distributions on all of the parameters of interest, allowing the researchers to incorporate any *a priori* assumptions. Finally, the output of such methods is a sample from the joint posterior distribution of species trees, gene trees, and associated model parameters, which allows summary and reporting of any characteristic of interest from the posterior distribution. Generally, when estimation of a species-level phylogeny is of interest, a summary of the posterior distribution of species trees, such as the maximum clade credibility (MCC) tree or the maximum *a posteriori* (MAP) tree is reported, along with the posterior probability associated with each clade in the tree. Density plots of the sampled trees (Bouckaert, 2010) are also a useful means of visualizing a posterior distribution on the space of phylogenetic trees.

The primary disadvantage of Bayesian methods for inferring species trees is the required computational effort. This is because the space of parameters over which the MCMC algorithm must operate is very large. For example, consider a data set of 20 species with a single individual sampled for each species for 500 loci. For each locus, suppose that a separate GTR + I + Γ model (see **Section 6.6.1**) is specified, resulting in seven substitution model parameters for each of the 500 loci, a total of 3500 parameters. For 20 taxa, there are more than 8.2×10^{21} possible tree topologies, each with 37 branches whose lengths must be estimated. The general goal of inference in a Bayesian framework is to find the posterior probability distribution assigned to all possible combinations of parameters, which is clearly daunting in this case as the space of

parameters that includes substitution model parameters, gene trees, and the species tree is very large. In practice, however, it will be enough to approximate the posterior probability in regions of the parameter space that have high posterior density. Still, this requires a method of traversing the parameter space such that high-probability regions are very likely to be identified but so that the algorithm can run in reasonable time. At the time of this writing, Bayesian inference of species phylogenies is generally limited to data sets containing tens of taxa and hundreds of genes, and such methods cannot generally handle hundreds of taxa and thousands of genes simultaneously. However, much effort is being devoted to the development of faster algorithms that utilize new computational resources (see, for example, Ogilvie *et al.*, 2017; Rannala and Yang, 2017), such as graphical processing units and other methods of parallelization, so that the computational efficiency of such methods is expected to continue to improve.

7.3.3 Site Pattern-Based Methods

Although the likelihood in equation (7.6) is computationally intractable in general, when the number of species is small and only a single site in a sequence alignment is considered, this expression can be used to compute the probability distribution on possible sites patterns. For example, Chifman and Kubatko (2015) provide expressions for the site pattern probabilities under several simple nucleotide substitution models for four species with one individual sampled per species. The observation that the multispecies coalescent induces a particular structure on this probability distribution has led to the development of a computationally efficient method for species tree inference based on the inference of quartet trees from a collection of observed site pattern probabilities.

Consider a four-taxon species tree, as in Figure 7.1, and let p_{ijkl} denote the probability that at a particular site in the genome, species A has nucleotide i , species B has nucleotide j , species C has nucleotide k , and species D has nucleotide l . This probability can be computed using equation (7.6) once the speciation times and population sizes are specified. Consider arranging the $4^4 = 256$ possible site pattern probabilities that result into a 16×16 matrix, called a *flattening*, where the rows of the matrix correspond to the possible nucleotides observed for species A and B , and the columns correspond to possible nucleotides observed for species C and D , as follows:

$$Flat_{AB|CD} = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots & [TT] \\ [AA] & p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AAC} & \cdots & p_{AATT} \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots & p_{ACTT} \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots & p_{AGTT} \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots & p_{ATTT} \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CAC} & \cdots & p_{CATT} \\ & \cdots \\ [TT] & p_{TTAA} & p_{TTAC} & p_{TTAG} & p_{TTAT} & p_{TTCA} & \cdots & p_{TTTT} \end{pmatrix}.$$

For example, the (3, 2)th entry, p_{AGAC} , corresponds to the observation of nucleotide A for taxon A , nucleotide G for taxon B , nucleotide A for taxon C , and nucleotide C for taxon D .

When species A and B are sister taxa in the true species tree, the matrix above has rank 10, while if either A and C or A and D are sister taxa, the matrix will be full rank (i.e. rank 16). This result holds whenever sites arise according to the multispecies coalescent with any submodel of GTR + I + Γ for species trees satisfying the molecular clock assumption (Chifman and Kubatko, 2015), and can be extended to the case in which the clock does not hold for GTR + I and all submodels (Long and Kubatko, 2018a), as well as to the case in which there is gene flow between sister taxa (Long and Kubatko, 2018b).

This result suggests that the species-level relationships for the quartet of taxa A , B , C , and D can be inferred by evaluating the observed collection of site pattern probabilities. The SVDQuartets method (Chifman and Kubatko, 2014) does this as follows. First, the observed probabilities of each site pattern in a data set are used to form estimates of the three matrices $Flat_{AB|CD}$, $Flat_{AC|BD}$, and $Flat_{AD|BC}$. Then, for each matrix, a score based on the singular value decomposition is computed to measure how close each flattening is to the nearest rank-10 matrix. The tree that corresponds to the flattening with the lowest score is selected. To form the overall estimate of the species tree for any number of taxa, all possible quartet relationships are considered (or a random sample is used, if the number of possible quartets is too large), and a quartet assembly algorithm is used to construct the species tree estimate from the collection of inferred quartet trees.

Because computation of singular values for a 16×16 matrix is computationally straightforward, the inference of quartet trees can be done in a computationally efficient manner. Efficient algorithms also exist for the assembly of the quartet trees to form the overall species tree estimate, though the computational cost increases as the number of taxa increases. However, the method is known to be very efficient overall, with feasible species tree estimation for hundreds of taxa and thousands of genes. Moreover, the method performs better as more genes are added with little increase in computational cost, because the flattening matrices are better approximated with more data while the operation being performed on them (computation of singular values) does not increase in complexity. In addition, the method is model-based, in the sense that the matrix rank results are derived under the probability distribution arising from the multispecies coalescent and associated nucleotide substitution models. Finally, the fact that the matrix rank results hold in the presence of gene flow between sister taxa means that the method can be used to relax the assumption of the standard multispecies coalescent of complete cessation of gene flow following speciation.

As in the case of the summary statistics methods, SVDQuartets is not naturally equipped with a method for assessing uncertainty in the estimated species tree. However, the bootstrap can be used to quantify uncertainty, typically represented by placing bootstrap support values on the nodes of the estimated tree. The bootstrap procedure is readily parallelized, and since each replicate can be carried out in a computationally efficient manner, complete inference can be carried out with relative ease. At present, the method does not provide estimates of additional parameters associated with the species tree. SVDQuartets is implemented within the PAUP* software (Swofford, 1998).

7.3.4 Multilocus versus SNP Data

Inference of a species-level phylogenetic tree may be desired for data arising from numerous different sequencing technologies, including those that produce true multilocus sequence data and those that produce SNPs. The methods described above are all appropriate for multilocus data, and indeed the multispecies coalescent was introduced here as a model that generates multilocus data. However, it is also natural to think of loci of length 1, and data arising in this manner have been termed *coalescent independent sites* (CIS; Tian and Kubatko, 2017a; Long and Kubatko, 2018b). If only those sites that are variable in a sample of coalescent independent sites are retained, these data are generally referred to as SNPs.

The SVDQuartets method was derived for CIS data, but it has been applied successfully to multilocus data as well (for an argument justifying this procedure, see Long and Kubatko, 2018a). The other methods described above are not easily adapted to CIS data, as they rely on gene trees at some level. For example, the summary statistics methods require that a gene tree be estimated for each individual locus, and thus are infeasible for either SNP or CIS data.

Table 7.2 Table showing the numbers of individuals sampled within each subspecies and geographic location for the rattlesnake example from Kubatko *et al.* (2011)

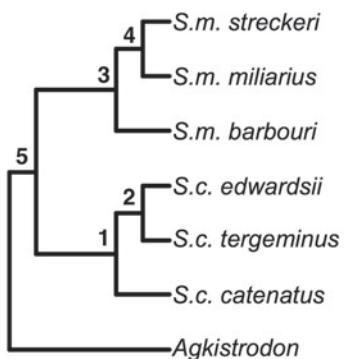
Species	Location	No. of individuals per gene
<i>S. catenatus catenatus</i>	Eastern USA and Canada	9
<i>S. c. edwardsii</i>	Western USA	4
<i>S. c. tergeminus</i>	Western and Central USA	5
<i>S. miliaris miliaris</i>	Southeastern USA	1
<i>S. m. barbouri</i>	Southeastern USA	3
<i>S. m. streckerii</i>	Southeastern USA	2
<i>Agkistrodon</i> sp. (outgroup)	USA	2

The Bayesian methods also use gene trees to form a portion of the parameter space for which the posterior distribution is estimated, and thus as described above, these methods are also not applicable to SNP or CIS data. However, a Bayesian method for estimation of species trees, called SNAPP (Bryant *et al.*, 2012), has been developed that employs a computationally efficient strategy based on the idea of peeling (Felsenstein, 1981) for computation of a likelihood similar to that of equation (7.6) when each site consists of a bi-allelic SNP. This method generally shares the advantages and disadvantages of the other Bayesian approaches described above.

7.3.5 Empirical Examples

Two empirical data sets will be used to illustrate the performance of several of the species tree inference methods described above. The first data set consists of samples from 26 North American rattlesnakes, with sampling from three species (*Sistrurus catenatus*, *S. miliaris*, and *Agkistrodon*), including individuals selected from each of the six subspecies within *Sistrurus* (Kubatko *et al.*, 2011). The number of individuals sampled from each subspecies are shown in Table 7.2. The data set includes 19 genes for each individual sampled, for a total of approximately 8500 bp. The second data set is from Rokas *et al.* (2003) and includes samples from eight yeast species for 106 genes, for a total of approximately 127,000 bp. For each data set, we apply several methods of species tree inference, and discuss their relative performance.

The rattlesnake data set is not very large in terms of the number of genes, and thus both full-data methods and summary statistics methods can be applied in this case. The summary statistic method ASTRAL (Mirarab *et al.*, 2014; Mirarab and Warnow, 2015), the Bayesian full-data methods *BEAST (Heled and Drummond, 2010) and BPP (Yang, 2015), and the SVDQuartets method (Chifman and Kubatko, 2014, 2015) were all used to obtain species tree estimates with corresponding measures of uncertainty for these data. The results are summarized in Figure 7.10, where entries in the table to the right of the estimated species tree give the relevant measure of support for the corresponding node. For ASTRAL, the support values correspond to the local posterior probability (shown as a percentage), while for SVDQuartets, bootstrap support (over 100 replicates) is shown. Posterior probabilities (as percentages) are shown for *BEAST and BPP. Overall, the results are consistent across methods, with strong support for species-level relationships (clades labeled 1 and 3 in Figure 7.10). Within *S. catenatus*, there is generally strong support for subspecies *S. c. edwardsii* and *S. c. tergeminus* as sister taxa, while within *S. miliaris*, no method provides strong support for any of the possible relationships among sister taxa. SVDQuartets provides the strongest support to *S. m. streckeri* and



Node	1	2	3	4	5
ASTRAL	80	62	100	12*	100
BEAST	100	100	100	46	100
BPP	100	99	100	33*	100
SVDQ	93	100	100	46	100

* This clade did not have the highest support value of the three possibilities.

Figure 7.10 Species tree estimated using SVDQuartets for the rattlesnake data described in Table 7.2 (left) and summary of support values estimated by several species tree inference methods across species tree nodes (right). The support values reported for ASTRAL are local posterior probabilities, those for SVDQuartets are bootstrap support values, and those for *BEAST and BPP are posterior probabilities. Node 4 received highest support of the three possible resolutions of subspecies within *S. miliaris* using SVDQuartets, but other methods provided higher support for *S. m. miliaris* and *S. m. barbouri* as sister taxa (see text).

S. m. miliaris, while other methods favor *S. m. miliaris* and *S. m. barbouri*, though none very strongly: ASTRAL gives local posterior probability 69% to this clade, *BEAST assigns posterior probability 49%, and BPP gives posterior probability 59%. It is worth noting that this data set is small in terms of the number of genes, which may affect the performance of the summary statistics methods in particular, since these methods form the species tree estimate from only 19 estimated gene trees.

The results of analyzing the yeast data set with the summary statistics method ASTRAL and with SVDQuartets are shown in Figure 7.11. Support values above the nodes are the local posterior probability for ASTRAL and bootstrap support across 100 replicates for SVDQuartets. As was the case for the rattlesnake data, the methods show strong agreement in terms of the inferred topology, and similar levels of support across nodes. The two methods differ only in the level of support assigned to the clade containing *Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. kudriavzevii*, with SVDQuartets finding bootstrap support for this clade in 53 of 100 replicates and ASTRAL providing local posterior probability 1. Interestingly, previous analyses have found support for a sister relationship between *S. kudriavzevii* and *S. bayanus* (Edwards *et al.*, 2007) and for horizontal transfer within these groups (Wen and Nakhleh, 2018).

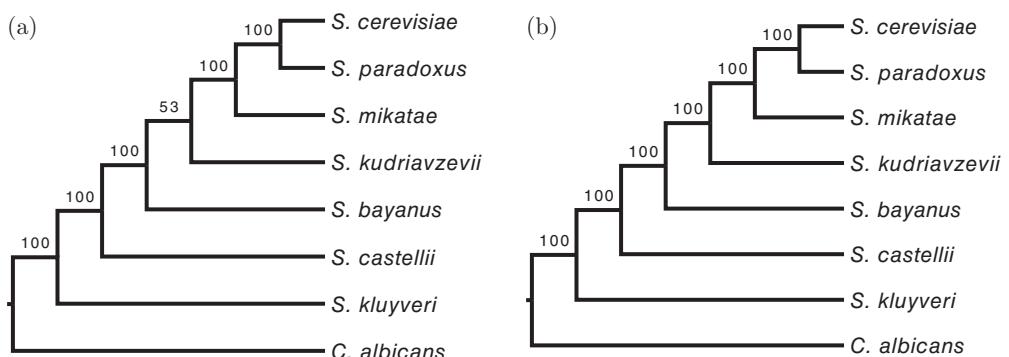


Figure 7.11 Species tree estimates for the yeast data set using SVDQuartets (a) and ASTRAL (b).

7.4 Coalescent-Based Estimation of Parameters at the Population and Species Levels

The focus in the previous section was on inference of the species-level topology, though the species tree includes other parameters of interest as well. For example, estimates of the speciation times in the phylogeny as well as the effective population sizes may be desired. The methods described above vary in the ease with which estimates of the associated parameters can be obtained. In addition, other interesting questions arise. For example, samples from a collection of closely related species may contain species that have a history of hybridization or species that may have been subject to gene flow with related species at various times along their evolutionary trajectory. Thus, the desired inference under the multispecies coalescent may include methods that identify hybrid species and that estimate the extent of past gene flow. Finally, all of the discussion in this chapter has assumed that species are well defined before inference proceeds. However, defining species and determining membership of sampled taxa within species, a problem known as *species delimitation*, is clearly a crucial step in understanding the history of a set of recently diverged species. Within the last decade, methods for species delimitation have increasingly cast the problem in the framework of the multispecies coalescent. After briefly reviewing these three areas in this section, we conclude with a look to the future of inference under the multispecies coalescent.

7.4.1 Speciation Times and Population Sizes

Among the most important parameters associated with a species phylogeny are the times of the speciation events, or analogously, the lengths of the branches within the tree, and the population sizes for the current and ancestral populations. As mentioned above, these parameters are often combined in the multispecies coalescent model when coalescent units are used to measure time along a species tree. Recall from Section 7.1 that a coalescent unit was defined to be the number of generations scaled by the population size, N , and, referring back to Figure 7.2, recall that when three species are considered, the probability of the gene tree topology that matches the species tree under the multispecies coalescent is $1 - \frac{2}{3}e^{\tau_{ABC}}$. This suggests a natural estimator of the branch length τ_{ABC} whenever a sample of gene tree topologies relating three species is available for inference. Let \hat{p} be the frequency of the most commonly observed tree of the three possible topologies in the sample. Setting $\hat{p} = 1 - \frac{2}{3}e^{\tau_{ABC}}$ and solving for τ_{ABC} gives the estimator $\hat{\tau}_{ABC} = -\ln(\frac{3}{2}(1 - \hat{p}))$ (Allman *et al.*, 2011). This is used as the basis for the estimation of branch lengths in some of the summary statistics methods (e.g. Liu *et al.*, 2010; Sayyari and Mirarab, 2016). A drawback of this method is that whenever the frequency of the dominant three-taxon topology is much larger than that of the other two, the estimate will be imprecise (because the expression for $\hat{\tau}_{ABC}$ goes to ∞ as $\hat{p} \rightarrow 1$).

In the Bayesian framework, estimation of speciation times and population sizes can be carried out separately, as each can be included in the parameter space over which the posterior distribution is obtained. This approach has the advantage that estimates of the posterior distributions of these parameters are easily approximated by marginalizing the posterior samples over the other parameters, thus providing not only an estimate of these parameters but also an estimate of the uncertainty in the inferred values. Methods based on site patterns can also be used to provide natural estimates of species-level branch lengths in coalescent units for simple substitution models. For example, Chifman and Kubatko (2015) provide a system of equations relating site pattern frequencies to branch lengths for fixed four-taxon trees for Jukes–Cantor-like models. Solving this system yields formulas that can be used to estimate branch lengths

by replacing theoretical site pattern probabilities with observed site pattern frequencies. Variances of the estimates and confidence intervals are easily obtained via standard asymptotic procedures for estimation from a multinomial distribution.

In general, estimation of speciation times and effective population sizes, together with an assessment of uncertainty in these quantities, represents an important direction for future work in this area. Genome-scale data certainly contain information about these parameters, and thus providing estimates of these parameters is a crucial step in understanding the evolutionary history of a collection of species.

7.4.2 Hybridization and Gene Flow

Because the processes of hybridization and gene flow between distinct species are known to be ubiquitous in the evolution of species, significant methodological development over the last decade has focused on extending the multispecies coalescent model to capture these processes. The most general of these extensions involves the estimation of general tree-like structures termed *phylogenetic networks* that allow for complex events, such as hybridization and horizontal gene flow, to occur throughout the phylogenetic history of the group under consideration. Most of the model-based methods in common use focus on a particular type of hybridization event in which the ancestry of an individual is shared, possibly unevenly, by two distinct species, as depicted in Figure 7.12. Viewed forward in time, the hybrid species H can be seen as arising via a hybridization event in which parental species P_1 and P_2 interbred to produce a hybrid population. After time τ_H , all three populations became isolated again, leading to the present-day distinct species H, P_1 and P_2 . The parameter γ , which has been called the *hybridization parameter* (Meng and Kubatko, 2009) or *inheritance probability* (Yu *et al.*, 2011, 2012; Wen *et al.* 2016), represents the proportion of the genome of species H that is inherited from P_2 (and thus the remaining proportion $1 - \gamma$ of the genome of species H is inherited from species P_1).

In its originally proposed form, this model was used to modify the probability distributions on gene tree topologies and gene trees described in Sections 7.2.1.1 and 7.2.1.2, as follows. First, for a particular gene, a parental species tree (either parental tree 1 or parental tree 2) is generated, with probabilities γ and $1 - \gamma$, respectively. Then a gene tree evolves along the selected parental species tree according to the multispecies coalescent, as described in Section 7.2.1. Meng and Kubatko (2009) used the topology distribution to assess the evidence for hybridization along a fixed species tree, and Kubatko (2009) used the gene tree probability density in a model-selection framework to infer hybrid speciation. The model has been extended in

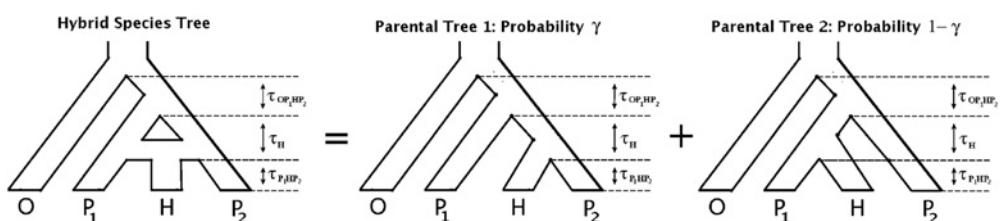


Figure 7.12 A model for hybrid speciation in the framework of the multispecies coalescent. The hybrid species tree (left) contains an outgroup species, O , two parental species, P_1 and P_2 , and a hybrid species H that arises through hybridization between P_1 and P_2 . To model sequence evolution along this hybrid species tree under the multispecies coalescent, two parental species trees that correspond to the two possible alternative placements of hybrid species H are considered, and it is assumed that gene trees arise along the two parental trees with probabilities γ and $1 - \gamma$. Sequence data then evolve along the gene trees according to the standard nucleotide substitution models.

numerous way by Nakhleh and colleagues in a series of papers (Yu *et al.*, 2011, 2012; Wen *et al.* 2016; Zhu *et al.*, 2018), in which formal likelihood inference of the species tree and associated inheritance probabilities was developed for general phylogenetic networks that include more than one hybridization event and that do not require hybridization events to be synchronous in time. These methods have recently been extended to a Bayesian framework (Wen *et al.* 2016; Zhu *et al.*, 2018), providing all of the advantages associated with Bayesian inference in general, as discussed above. Solís-Lemus *et al.* (2016) have considered the consequences of evolution along a species network under similar models, and have shown that for those models, anomalous gene trees can exist. They further examine the effect of gene flow on several of the commonly-used species tree inference methods. Similarly, Zhu *et al.* (2016) provide an alternative definition of an anomalous gene tree in reference to a species network, and provide conditions under which such anomalous gene trees arise. Solís-Lemus and Ané (2016) provide an algorithm and software for estimation of species networks using maximum pseudolikelihood.

Site pattern probabilities have also been used to detect hybrid species, and more generally, species for which hybridization and/or gene flow may have played a role in the evolutionary history of the taxa under consideration. The most popular method for assessing hybridization using site pattern probabilities is the ABBA-BABA method (Green *et al.*, 2010; Durand *et al.*, 2011), which is based on the assumption that for two sister species that have *not* experienced hybridization, site patterns that follow 'ABBA' patterns should occur in equal frequency to those that follow 'BABA' patterns. Chifman and Kubatko (2015) used more general site pattern frequencies to derive statistics and corresponding asymptotic distributions that could be used to formally test hypotheses of hybrid speciation, as depicted in Figure 7.12. In addition to identifying species subject to hybridization, their method can be used to obtain an estimate of the hybridization parameter γ .

More recent work has focused on models that allow gene flow between species in the multispecies coalescent framework in more general settings. For example, Zhu and Yang (2012), Tian and Kubatko (2016), and Long and Kubatko (2018) have considered ongoing gene flow between sister species following speciation for three-taxon species trees, and have derived the resulting probability distributions on gene trees and gene tree topologies. Zhu and Yang (2012) used these calculations to obtain ML estimates of speciation times and effective population sizes in three-taxon phylogenies that were subject to gene flow, and developed a likelihood ratio test of speciation with gene flow. Long and Kubatko (2018b) showed the existence of anomalous gene trees in this case (which generalizes the setting of Solís-Lemus and Ané, 2016) and characterized the space over which anomalous trees exist under their model. Like Solís-Lemus and Ané, they examined the effect of gene flow on the performance of several methods of species tree estimation.

7.4.3 Species Delimitation

All of the models and methods described in this chapter have assumed that species are well defined and that each sample can be unambiguously classified into species. In practice, however, many of the problems for which the multispecies coalescent model is most appropriate (i.e. those for which speciation has occurred rapidly and recently, resulting in short intervals of time between speciation events) will involve cases where the precise number of species is not known and/or for which it may not be clear to which species a particular sampled organism belongs. Thus, significant research effort is currently focused around the problem of *species delimitation* in general, with methods based on the multispecies coalescent playing a prominent role in recent years. Reviews of the use of the multispecies coalescent model in delimiting species are given by Fujita *et al.* (2012), Carstens *et al.* (2013), and Rannala (2015).

7.4.4 Future Prospects

The advent of efficient and inexpensive sequencing technologies has ushered in a new era in the inference of the phylogenetic relationships among a collection of species. The availability of sequence data for hundreds to thousands of loci, of genome-wide SNP data sets, and even of whole genomes has necessitated the development of models that capture variation both within and between species as well as across the genome. The multispecies coalescent is firmly established as a model that realistically captures these processes, and thus much effort has been expended to implement the model in increasingly complex settings in a computationally efficient manner. Methods for inferring bifurcating species tree topologies have continued to advance, and while more progress is to be expected, particularly with respect to computational efficiency, much of the emphasis in the development of new methodology will rightly shift to estimation of related quantities, as well as to the estimation of phylogenetic structures that do not fit the standard model of bifurcation with immediate cessation of gene flow. In addition, methods for detecting and validating incipient species as well as for identifying those existing species that were not carefully studied in the past will continue to be developed.

In terms of mathematical and statistical modeling and algorithm development, the field is at an interesting juncture in which traditional model-based methods that often require extensive computations will increasingly compete with ‘machine learning’ and other feature extraction methods. Such methods are generally capable of rapidly producing estimates of quantities of interest, but this often comes at the expense of an intuitive understanding of how such estimates were obtained and how they should be interpreted. Nonetheless, in the context of an evolutionary process that has unfolded over millions of years and for which any model is clearly a simplified version of the underlying process, such methods will clearly have a role in maximizing the extraction of information from large-scale data sets. The multispecies coalescent will play an important role in providing a context for interpreting inferences obtained from such methods and data in the years to come.

Acknowledgements

I thank James Degnan, Liang Liu and David Balding for helpful comments on an earlier draft of this chapter.

References

- Allman, E.S., Degnan, J.H. and Rhodes, J.A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology* **62**, 833–862.
- Beerli, P. and Felsenstein, J. (1991). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences* **98**(8), 4563–4568.
- Bouckaert, R.R. (2010). DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* **26**(10), 1372–1373.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A. and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* **29**(8), 1917–1932.

- Carstens, B.C., Pelletier, T.A., Reid, N.M. and Satler, J.D. (2013). How to fail at species delimitation. *Molecular Ecology* **22**(17), 4369–4383.
- Chifman, J. and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**(23), 3317–3324.
- Chifman, J. and Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology* **374**, 35–47.
- Degnan, J.H. and Rosenberg, N.A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics* **3**, 762–768.
- Degnan, J.H. and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* **24**(6), 332–340.
- Degnan, J.H. and Salter, L.A. (2005). Gene tree distributions under the coalescent process. *Evolution* **59**, 24–37.
- Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**(8), 2239–2252.
- Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M. and Von Haeseler, A. (2007). Mapping human genetic ancestry. *Molecular Biology and Evolution* **24**(10), 2266–2276.
- Edwards, S.E., Liu, L. and Pearl, D.K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 5936–5941.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**(6), 368–376.
- Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**, 521–565.
- Fujita, M.K., Leaché, A.D., Burbrink, F.T., McGuire, J.A. and Moritz, C. (2012). Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology and Evolution* **27**, 480–488.
- Gatesy, J. and Springer, M.S. (2014). Phylogenetic analysis at deep timescales, unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution* **80**, 231–266.
- Green, R.E., Krause, J., Briggs, A.W., et al. (2010). A draft sequence of the Neandertal genome. *Science* **328**(5979), 710–722.
- Heled, J. and Drummond, A.J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**(3), 570–580.
- Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 2785–2790.
- Kingman, J.F.C. (1982). Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino (eds.), *Exchangeability in Probability and Statistics*. North-Holland, pp. 97–112.
- Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43.
- Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**, 235–248.
- Kubatko, L. and Chifman, J. (2015). An invariants-based method for efficient identification of hybrid species from large-scale genomic data. Preprint, bioRxiv 034348.
- Kubatko, L.S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology* **58**(5), 478–488.
- Kubatko, L.S., Carstens, B.C. and Knolwes, L.L. (2009). STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25**(7), 971–973.

- Kubatko, L.S., Gibbs, H.L. and Bloomquist, E. (2011). Inferring species-level phylogenies using multi-locus data for a recent radiation of *Sistrurus* rattlesnakes. *Systematic Biology* **60**(4), 393–409.
- Liu, L. and Pearl, D.K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* **56**, 504–514.
- Liu, L. and Pearl, D.K. (2010). Maximum tree: A consistent estimator of the species tree. *Journal of Mathematical Biology* **60**(1), 95–106.
- Liu, L., Yu, L. and Edwards, S.V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, **10**(1), 302.
- Liu, L., Yu, L., Kubatko, L., Pearl, D.K. and Edwards, S.V. (2009). Coalescent methods for estimating multilocus phylogenetic trees. *Molecular Phylogenetics and Evolution* **53**, 320–328.
- Liu, L., Yu, L., Pearl, D.K. and Edwards, S.V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* **58**(5), 468–477.
- Long, C.L. and Kubatko, L. (2018). Identifiability and reconstructibility of species phylogenies under a modified coalescent. Preprint, arXiv:1701.06871.
- Long, C.L. and Kubatko, L.S. (2018). The effect of gene flow on coalescent-based species tree inference. Preprint, arXiv:1710.03806.
- Meng, C. and Kubatko, L.S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology* **75**, 35–45.
- Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. and Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**(17), i541–i548.
- Mirarab, S. and Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**(12), i44–i52.
- Mossel, E. and Roch, S. (2010). Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**(1), 166–171.
- Ogilvie, H.A., Bouckaert, R.R. and Drummond, A.J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* **34**, 2101–2114.
- Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**(5), 568–583.
- Rannala, B. (2015). The art and science of species delimitation. *Current Zoology* **61**(5), 846–853.
- Rannala, B. and Yang, Z. (2003). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* **164**, 1645–1656.
- Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology* **66**, 823–842.
- Rokas, A., Williams, B.L., King, N. and Carroll, S.B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**(6960), 798–804.
- Rosenberg, N.A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* **61**, 225–247.
- Rosenberg, N.A. and Tao, R. (2008). Discordance of species trees with their most likely gene trees: The case of five taxa. *Systematic Biology* **57**, 131–140.
- Sayyari, E. and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* **33**, 1654–1668.
- Solís-Lemus, C., Yang, M. and Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology* **65**, 843–851.

- Solís-Lemus, C., Yang, M. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics* **12**(3), e1005896.
- Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* **9**, 91–116.
- Swofford, D. (1998). *PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods)*. Version 4. Sinauer Associates, Sunderland, MA.
- Takahata, N. (1989). Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genetics* **122**, 957–966.
- Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325–344.
- Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26**, 119–164.
- Tian, Y. and Kubatko, L. (2016). Distribution of gene tree histories under the coalescent model with gene flow. *Molecular Phylogenetics and Evolution* **105**, 177–192.
- Tian, Y. and Kubatko, L. (2017). Expected pairwise congruence among gene trees under the coalescent model. *Molecular Phylogenetics and Evolution* **106**, 144–150.
- Tian, Y. and Kubatko, L. (2017). Rooting phylogenetic trees under the coalescent model using site pattern probabilities. *BMC Evolutionary Biology* **17**, 263.
- Watterson, G.A. (1984). Lines of descent and the coalescent. *Theoretical Population Biology* **26**, 77–92.
- Wen, D. and Nakhleh, L. (2018). Coestimating reticulate phylogenies and gene trees using multilocus data. *Systematic Biology* **67**, 439–457.
- Wen, D., Yun, Y. and Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics*, **12**(5), e1006006.
- Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* **66**(3), 763–775.
- Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology* **61**(5), 854–865.
- Yu, Y., Degnan, J.H. and Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, **8**(4), e1002660.
- Yu, Y., Than, C., Degnan, J.H. and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology* **60**(2), 138–149.
- Zhu, J., Wen, D., Yu, Y., Meudt, H.M. and Nakhleh, L. (2018). Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS Computational Biology* **14**(1), e1005932.
- Zhu, J., Yu, Y. and Nakhleh, L. (2016). In the light of deep coalescence: Revisiting trees within networks. *BMC Bioinformatics* **17**(Suppl 14:415), 271–282.
- Zhu, S. and Degnan, J.H. (2017). Displayed trees do not determine distinguishability under the network multispecies coalescent. *Systematic Biology* **66**, 283–298.
- Zhu, T. and Yang, Z. (2012). Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution* **29**(10), 3131–3142.

Population Structure, Demography and Recent Admixture

G. Hellenthal

University College London Genetics Institute (UGI), Department of Genetics, Evolution and Environment, University College London, London, UK

Abstract

The increasing availability of large-scale genetic variation data sampled from world-wide geographic areas, coupled with advances in the statistical methodology to analyse these data, is showcasing the power of DNA as a major tool to gain insights into the history of humans and other organisms. This chapter describes the concepts behind some widely used methods in the field of population genetics applied to whole-genome autosomal data. In particular, the chapter focuses on techniques that analyse genetic data in order to learn about sub-structure among sampled individuals and the dynamics of population size changes and intermixing among genetically different populations. While this is by no means an exhaustive look at the many interesting methods in the field, it provides an overview of some of the demographic signals inherent in genetic data and the challenges in extracting this information. We describe these methods as if they will be applied to data from humans, though note that the concepts here extend to other diploid organisms that experience homologous recombination (with potentially straightforward extensions to organisms of other ploidy as well). In general this chapter illustrates the power of DNA as an important data resource that, when interpreted in the context of knowledge from other data sources (e.g. archaeology, anthropology, linguistics), can resolve existing controversies or unearth previously unknown features of human history.

8.1 Introduction

The ancestral history of anatomically modern human groups is exceedingly complex. Demographic factors affecting genetic variation data include *population splits* where different groups ('populations') of individuals become isolated from one another, after which each group is subjected to independent genetic drift that results in allele frequency differences between them. In addition, changes in the sizes within each population, in terms of the effective number of breeding individuals, can alter the speed at which drift acts, with population expansions and contractions (e.g. 'bottlenecks') decreasing and accelerating the effects of drift, respectively. Another important process is *admixture*, where previously isolated groups intermix. As discussed below, even the concept of a 'population' is not straightforward, as individuals can be grouped together in many different ways.

Each of these processes affects genetic variation patterns in distinct ways, yet disentangling all of the possible processes that can lead to observed variation patterns is an intractable

statistical problem. Nonetheless advances have been made to attempt to distinguish the effects of some of these features on DNA, taking advantage of the vast amounts of genetic information currently available. A key insight is that genetic patterns are correlated among sequences from different individuals, including among individuals unrelated at the familial level (i.e. individuals who are not first or second cousins, etc.), with such unrelated individuals being the focus of this chapter. This correlation carries vital information on the extent of shared recent ancestry among such samples of unrelated individuals. Therefore by studying genetic correlations, we can hope to learn about the degree to which different sets of individuals (e.g. sampled from different geographic regions) are ancestrally related to one another. While every pair of individuals shares a common ancestor at each point of the genome, approaches here attempt to determine which individuals share ancestors who lived more recently than the shared ancestors of other individuals.

This chapter will concentrate on four specific types of inference:

- (1) exploring spatial summaries of genetic variation data;
- (2) classifying individuals into clusters based on genetics;
- (3) inferring population size changes and split times;
- (4) identifying and describing admixture events.

We will not exhaustively explore all methods related to (1)–(4), but instead will provide insights into some commonly used approaches applicable to genome-wide autosomal data that address these questions. We will discuss the patterns in genetic variation data that theoretically allow inference under each approach, providing an overview of some of the mathematical details. We will also highlight some applications and limitations of each.

8.1.1 ‘Admixture’ versus ‘Background’ Linkage Disequilibrium

As described in more detail elsewhere (**Chapter 2**), the non-random association among allelic types at different genetic loci is called *linkage disequilibrium* (LD). As this is a principal feature used for inference in many of the approaches discussed in this chapter, it is helpful to define different types of LD based on the factor driving the association. Consider the admixed population formed as illustrated in Figure 8.1, which shows a single ‘pulse’ of admixture (i.e. admixture occurring over a short time interval, such as one generation) between two populations. Subsequently, individuals in the newly admixed population mate randomly for r generations. As shown at the bottom of Figure 8.1, DNA in these admixed individuals will be mixtures of segments (‘tracts’) inherited intact from individuals from the admixing source populations. Therefore loci within the same block can be correlated due to inheritance from a common recent ancestor from this admixture event. Following (Falush *et al.*, 2003), we will refer to this as *admixture LD*. As discussed below, depending on the date of admixture r , admixture LD can extend over relatively large segments of an autosome (e.g. over megabases).

A second, distinct type of LD we will refer to as *background LD*, again following the terminology of (Falush *et al.*, 2003), which measures the rate of decay of associations among loci *within* a population (e.g. within each solid black and dashed line bar of Figure 8.1). The level of background LD may be different within each admixing population, and reflects a population’s demographic history, including, for example, previous admixture, population size changes and population substructure. As an example, if one of the admixing populations experienced a strong bottleneck (e.g. due to a founder event) and the others did not, we would usually expect its background LD to extend farther than that of the other populations. In general, as the physical distance between two loci increases, the level of background LD between the loci decays at a much faster rate relative to the level of admixture LD between them, with background LD decaying in the order of tens to hundreds of kilobases in humans (Pritchard and Przeworski, 2001). This

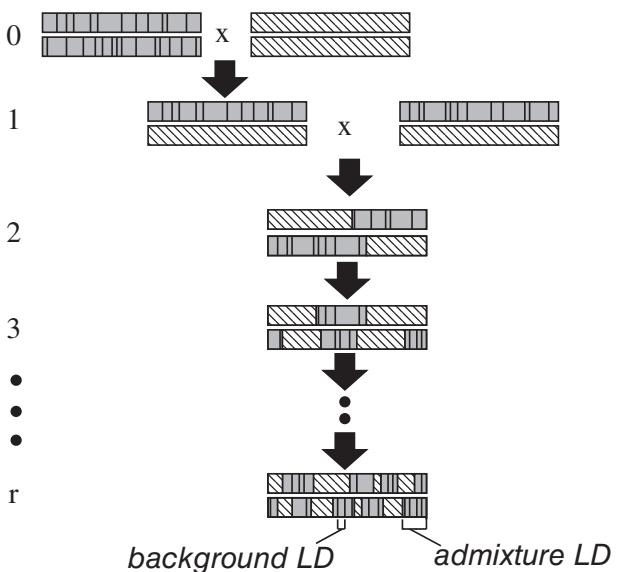


Figure 8.1 Schematic of the effects on an autosomal region of admixture occurring r generations ago between two populations. Two genetically distinct populations, with DNA represented by solid grey and dash-lined bars, admix at generation 0 at top. For simplicity, the two chromosomes for one individual per population are shown at the top. This admixture event is followed by r generations of subsequent random mating among individuals from the admixed population. (For simplicity, after generation 1 we show the two chromosomes of one individual each generation for this autosomal region.) In the first generation following admixture (second row from top), in this genetic region each admixed individual receives a chromosome from an individual representing each population. Subsequent generations inherit ‘tracts’ of contiguous DNA segments from each admixing source, with the lengths of these segments getting smaller each generation due to recombination. As an illustration of ‘background LD’, black vertical lines within each solid black bar reflect boundaries of tracts inherited from ancestors living at a time much earlier than r . Loci within each of these tracts will be in ‘background LD’, while loci inherited within the same contiguously solid grey (or dash-lined) block at the bottom will be in ‘admixture LD’.

is because background LD captures the effects of demographic processes occurring over the entire history of a population, which can span a very large time-frame and hence is affected by many historical recombination events, while admixture LD captures only the effects of more recent specific admixture events. A consequence of its faster rate of decay is that background LD is only potentially significant among loci that are physically quite close to each other. While some approaches described below ignore background LD or attempt to remove it, this chapter also highlights methods that attempt to exploit background LD to increase precision when inferring demographic parameters.

8.2 Spatial Summaries of Genetic Variation Using Principal Components Analysis

A widely used technique to visualize genetic patterns in a data set applies principal components analysis (PCA) to the high-dimensional genetic variation data of sampled individuals, and then plots low-dimensional *principal components* that efficiently summarize these data. First proposed in the context of analysing genetic variation data by Cavalli-Sforza and colleagues (Menozzi *et al.*, 1978), PCA is an algebraic technique for summarizing variation in multi-dimensional data sets in order to potentially highlight interesting patterns.

Assuming each locus is bi-allelic in the sample, such as single nucleotide polymorphism (SNP) data, let $X_i \equiv \{X_{i1}, \dots, X_{iL}\}$ be the genotype for individual $i \in (1, \dots, N)$ at all L sampled bi-allelic loci, with X_{il} the genotype at locus l of individual i . Specifically, X_{il} will be 0, 1, or 2, reflecting the number of copies of a particular allele that diploid individual i carries at locus l . Let X represent the $L \times N$ matrix with columns equal to X_i . Typically the matrix X is standardized in some manner, with different standardizations proposed by (Price *et al.*, 2006) and (Patterson *et al.*, 2006), among others, creating a new $L \times N$ matrix Y that, for example, subtracts the row average so that

$$Y_{il} = X_{il} - \frac{1}{N} \sum_{i=1}^N X_{il},$$

or subtracts the row average and standardizes by the variance so that

$$Y_{il} = \left[X_{il} - \frac{1}{N} \sum_{i=1}^N X_{il} \right] / \sqrt{\nu_l(1 - \nu_l)},$$

where ν_l estimates the allele frequency of locus l . (For example, (Price *et al.*, 2006) use $\nu_l = (1 + \sum_{i=1}^N X_{il})/(2 + 2N)$.) Relative to the former, this latter standardization upweights the relative influence on the PCA from variants with low minor allele frequencies among the sampled individuals. Then the N *eigenvectors* and *eigenvalues* of the $N \times N$ matrix $\Omega = Y^T Y$ can be determined using, for example, singular value decomposition (see Patterson *et al.*, 2006; Galinsky *et al.*, 2016). Each eigenvector contains N values, with the i th value a linear combination of the genotype data at all L loci of individual i , that is, a summary of all data points for individual i . An attractive property of PCA is that the eigenvectors are mutually orthogonal, so that they each capture independent information in the data. Furthermore, the first eigenvector explains more overall variation in the data Y than the second eigenvector, and so on. Therefore, plotting the first two eigenvectors can summarise the strongest signals in the genetic variation data when using only two data points per individual, rather than using the L data points that contain the full genetic information.

As an example, we applied the PCA software EIGENSTRAT (Price *et al.*, 2006) to simulated data from (Lawson *et al.*, 2012). These simulations consist of genetic variation data from five simulated populations that are related according to the tree in Figure 8.2(a), where populations

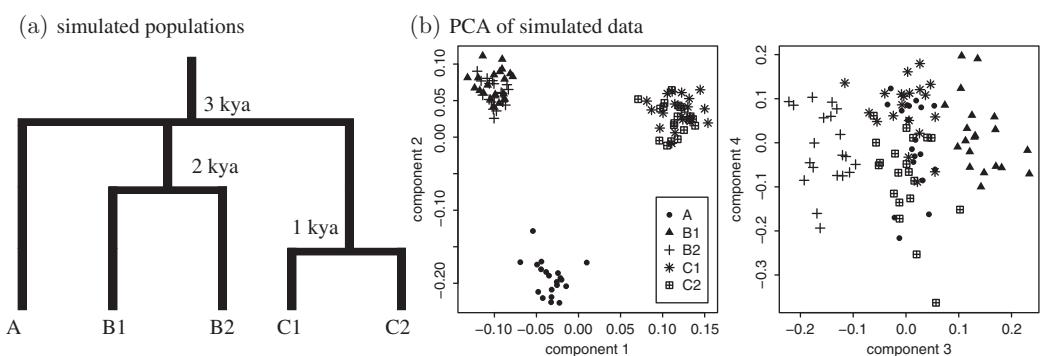


Figure 8.2 (a) Graphical depiction of tree relating simulated populations {A,B1,B2,C1,C2} from (Lawson *et al.*, 2012). (b) First four principal components of a PCA applied to the genotype data for the simulations of populations {A,B1,B2,C1,C2}, as calculated by EIGENSTRAT, with each point a simulated individual from one of the five populations.

$A, B = \{B1, B2\}$ and $C = \{C1, C2\}$ split from each other 3000 years ago (3 kya), followed by sub-populations $B1$ and $B2$ splitting from each other 2 kya and sub-populations $C1$ and $C2$ splitting from each other 1 kya. Twenty individuals were sampled from each population $\{A, B1, B2, C1, C2\}$, each having SNP data for 150 five-megabase regions simulated to mimic sequencing data, with ~24,000 SNPs per region after removing singletons (see (Lawson *et al.*, 2012) for more details). As can be seen in the PCA results of Figure 8.2(b), the first two eigenvectors cleanly separate populations A , B and C , suggesting that the strongest signal in these data are the genetic differences between these three groupings. Eigenvector 3 further separates sub-populations $B1$ and $B2$, while eigenvector 4 partially separates (though with substantial overlap) sub-populations $C1$ and $C2$. In total there are N eigenvectors that can be plotted, though each explains progressively less variation in the total data. Typically the first few eigenvectors are the most informative about underlying population sub-structure, for example separating different populations within continental regions such as Africa, Asia (Patterson *et al.*, 2006) and Europe (Novembre *et al.*, 2008; Lao *et al.*, 2008).

While PCA is a useful means of summarizing large-scale genetic data, interpreting the past demographic processes leading to PCA patterns can be challenging. McVean (McVean *et al.*, 2009) showed how the expectation of Ω can be related to the mean time of coalescence between pairs of samples. However, PCA projection can depend strongly on sample size and the ascertainment of samples, and different ancestral histories can lead to very similar PCA patterns (McVean *et al.*, 2009; Novembre and Stephens, 2008). Related techniques such as multi-dimensional scaling have also been employed on genetic data (Jakobsson *et al.*, 2008), which have similar advantages and disadvantages to PCA.

8.3 Clustering Algorithms

8.3.1 Defining ‘Populations’

Another widely used means of summarizing genetic variation data is to cluster individuals based on associations in their genetic patterns. To some extent, clustering can provide insights into the processes leading to observed genetic patterns. Another motivation for clustering is that researchers often wish to study the demographic processes that have affected a particular ‘population’. While many studies define populations using self-described labels from the individuals sampled or using the geographic sampling location of these individuals, Figure 8.2(b) suggests that there is clear utility for classifying individuals into discrete groups using solely genetic variation data. Indeed this classification may be particularly useful when exploring population demography, as individuals who self-identify with a particular label may sometimes have a very different ancestral history than others who self-identify with that same label. Such ‘outlier’ individuals may complicate interpretation of a labelled population’s history by averaging over individuals with different ancestral backgrounds.

Of course, there are numerous other definitions of population that may be useful depending on the question being asked. For example, an unbiased understanding of the genetic variation within a geographic region in the present day would require a random sample representative of that geographic region, regardless of how many different ancestral populations are represented in such a sample. However, in the following we focus in part on defining populations as groups of genetically homogeneous individuals. We also note that individuals may descend from multiple such populations. Indeed, applications of PCA to data from multiple European populations (Novembre *et al.*, 2008; Lao *et al.*, 2008) show a cline of genetic variation whereby sampled individuals more geographically near one another typically are more genetically related. Therefore a restriction to discrete homogeneous populations may sometimes lead to an incomplete

understanding of population structure (Rosenberg *et al.*, 2005). While this complicates inference, clustering algorithms have been adapted to cope with individuals deriving their ancestry from multiple sources, as we discuss below.

8.3.2 Clustering Based on Allele Frequency Patterns

First, start with the relatively simple scenario where the genome from each of N sampled individuals of ploidy J is derived entirely from one of K genetically distinct populations. At each locus $l \in (1, \dots, L)$, let $p_{kl}(a)$ be the frequency of allele type a in population k , with A_l possible allele types at locus l . In the following, we will let P represent the set of allele frequencies $\{p_{kl}(a)\}$ for each allele type at all L loci in all K populations. Under a simplified model that assumes random mating and ignores LD and admixture, each allele X_{ilj} for $j \in (1, \dots, J)$ (e.g. $J = 2$ in diploids) carried by individual i at locus l is drawn at random according to the allele frequencies at this locus in the population Z_i to which individual i is assigned, that is,

$$\Pr(X_{ilj} = a | Z_i = k, P) = p_{kl}(a), \quad (8.1)$$

for each possible allele type a at locus l . Assuming all loci are independent, the probability of observing the full genetic data $X_i \equiv \{X_{i11}, \dots, X_{iLJ}\}$ for individual i from population Z_i is

$$\Pr(X_i | Z_i, P) = \prod_{l=1}^L \prod_{j=1}^J \prod_{a=1}^{A_l} p_{Z_il}(a)^{[X_{ilj} == a]}, \quad (8.2)$$

where $[X_{ilj} == a]$ equals 1 when X_{ilj} is equal to allele type a and equals 0 otherwise. Assuming sampled individuals are independent, we have

$$\Pr(X_1, \dots, X_N | Z_1, \dots, Z_N, P) = \prod_{i=1}^N \Pr(X_i | Z_i, P). \quad (8.3)$$

We do not observe P or the Z_i in reality, but instead must infer them using the X_i that we do observe. A natural way to do this is using the Bayesian approach taken by the program STRUCTURE (Pritchard *et al.*, 2000), where

$$\Pr(Z_1, \dots, Z_N, \Theta | X_1, \dots, X_N) \propto \Pr(X_1, \dots, X_N | Z_1, \dots, Z_N, \Theta) \Pr(Z_1, \dots, Z_N, \Theta), \quad (8.4)$$

and the $\{Z_1, \dots, Z_N\}$ and Θ are inferred using Markov chain Monte Carlo (MCMC; see **Chapter 1**). In this section $\Theta = \{P\}$, but it will contain additional parameters in the next section. In the STRUCTURE model, the cluster assignments (Z_i) are assumed to be independent across individuals. Therefore,

$$\Pr(Z_1, \dots, Z_N, \Theta) = \left[\prod_{i=1}^N \Pr(Z_i | \Theta) \right] \Pr(\Theta). \quad (8.5)$$

In this section, $\Pr(\Theta) = \Pr(P)$, which can be broken down into the product of probabilities across (assumed independent) loci

$$\Pr(P) = \prod_{l=1}^L \Pr(\vec{p}_{1l}, \dots, \vec{p}_{Kl}), \quad (8.6)$$

where $\vec{p}_{kl} = \{p_{kl}(1), \dots, p_{kl}(A_l)\}$ is the vector of frequencies in population k for each of the A_l allele types at locus l . Following (Balding and Nichols, 1995), the original STRUCTURE model assumes that the \vec{p}_{kl} are independent across clusters, so that equation (8.6) is equal

to $\prod_{l=1}^L \prod_{k=1}^K \vec{p}_{kl}$. They further assume that each \vec{p}_{kl} follows a Dirichlet distribution with A_l parameters that can be either fixed or estimated along with the P and Z_i terms. An alternative formulation for equation (8.6), devised by Falush *et al.* (2003) and based in part on work described in (Nicholson *et al.*, 2002), models correlations among clusters' allele frequencies by assuming each cluster's allele frequency has drifted independently from that of an ancestral population common to all clusters. This new formulation is intuitively attractive in that allele frequencies generally are highly correlated across closely related populations in real life (e.g. across human groups sampled from nearby geographic areas).

Finally, to complete the original formulation of STRUCTURE, each individual is equally likely *a priori* to be assigned to any of the K clusters, so that

$$\Pr(Z_i = k | \Theta) = \frac{1}{K}. \quad (8.7)$$

MCMC is used to sample P and $\{Z_1, \dots, Z_N\}$ conditional on the data. This can be accomplished by first proposing initial values for $\{Z_1, \dots, Z_N\}$, for example by randomly assigning the N individuals to the K clusters with equal probability. Then, at each MCMC iteration, all parameters of P are sampled using the above probabilities conditional on these initial values of $\{Z_1, \dots, Z_N\}$. New values for $\{Z_1, \dots, Z_N\}$ are then sampled conditional on these updated P parameters, and this iteration between sampling P and sampling $\{Z_1, \dots, Z_N\}$ continues until the algorithm converges for these parameters.

The above assumes that K is known. Choosing an appropriate value of K is a notoriously challenging statistical problem (Alexander *et al.*, 2009) that may depend on sampling strategy and other factors. Indeed there is no true value of K , so approaches aim to find a K that scores best according to some intuitive criteria. Several suggestions have been proposed in the literature to choose the 'best' K using heuristics (Pritchard *et al.*, 2000; Alexander *et al.*, 2009; Raj *et al.*, 2014). In contrast, the approach taken in STRUCTURAMA (Huelsenbeck *et al.*, 2011) models K as a random variable using an approach proposed by Pella and Masuda (Pella and Masuda, 2006). However, perhaps the most prevalent usage of STRUCTURE-based clustering algorithms (e.g. Rosenberg *et al.*, 2002) is to cluster using multiple values of K , and then compare the cluster results at each value and attempt to interpret results in light of historical information from other resources such as linguistics, archaeology and anthropology. We discuss limitations of interpreting clustering in Section 8.3.6.

8.3.3 Incorporating Admixture

Many individuals derive from recent mixtures of individuals with genetically different ancestries, a topic we will address further in Section 8.5. Therefore, assigning each individual to a single cluster can miss important information about recent ancestry. To address this, the original STRUCTURE model allows for each allele at each locus within an individual to have its own ancestry. In particular, we now let $Z_i = \{Z_{i1}, \dots, Z_{iL}\}$, with Z_{ilj} the cluster from which individual i derives its j th allele at locus l . The Z_i in equations (8.1) and (8.2) are replaced with Z_{ilj} , and equation (8.7) is replaced with

$$\Pr(Z_{ilj} = k | \Theta) = \Pr(Z_{ilj} = k | Q) = q_{ik}, \quad (8.8)$$

where q_{ik} can be thought of as the proportion of DNA for which individual i is most closely related to that of individuals in cluster k , with Q containing the set of all such proportions across all individuals and clusters. In this admixture model, note that $\Theta = \{P, Q\}$ in equations (8.4) and (8.5), and that now $\{Z_1, \dots, Z_N\}$, P and Q are jointly inferred using MCMC.

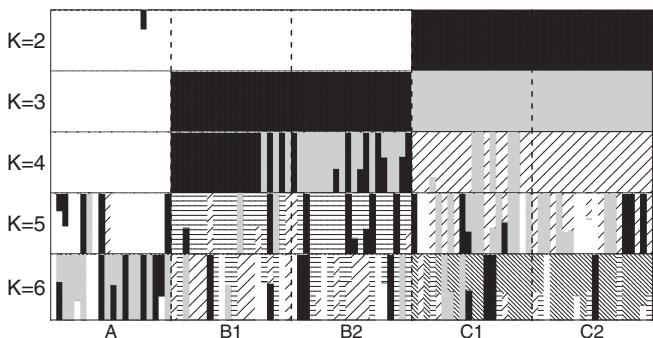


Figure 8.3 Inferred cluster assignments (different patterns/colors) for each individual (column) for the simulations of populations {A,B1,B2,C1,C2} related as shown in Figure 8.2(a), as inferred by ADMIXTURE for $K = 2, 3, \dots, 6$ clusters.

Here $\Pr(\Theta) = \Pr(P)\Pr(Q)$ in equation (8.5) and, assuming that the ancestry proportions are independent across individuals, we have that

$$\Pr(Q) = \prod_{i=1}^N \Pr(q_{i1}, \dots, q_{iK}). \quad (8.9)$$

$\Pr(q_{i1}, \dots, q_{iK})$ is assumed to be Dirichlet distributed, with K parameters that again can be fixed or inferred using MCMC (Pritchard *et al.*, 2000; Falush *et al.*, 2003).

The models implemented in STRUCTURE can carry a relatively high computational burden due to its MCMC sampling scheme. However, subsequent models that are much more computationally efficient, and thus applicable to larger genetic variation data sets, have been developed that infer P and Q using a variational Bayes framework (fastSTRUCTURE, (Raj *et al.*, 2014)) or maximum likelihood techniques calculated using Expectation-Maximisation (FRAPPE, (Tang *et al.*, 2005)) or optimisation techniques (ADMIXTURE, (Alexander *et al.*, 2009)) instead of MCMC. In particular, ADMIXTURE and FRAPPE have been used in a large number of studies to explore the ancestry of worldwide groups (e.g. (Li *et al.*, 2008; Behar *et al.*, 2010; Bryc *et al.*, 2010; Metspalu *et al.*, 2011; Schlebusch *et al.*, 2012; Rasmussen *et al.*, 2014; Jones *et al.*, 2015)), typically by comparing results across various values of K . Figure 8.3 shows results from applying ADMIXTURE with $K = 2\text{--}6$ to the simulated data illustrated in Figure 8.2(a). After cross-validation (Alexander *et al.*, 2009), the best inferred K out of 2–9 is for $K = 2$, which only separates populations {A,B1,B2} from populations {C1,C2}. This highlights the challenges of inferring the ‘best’ K , as $K = 3$ cleanly separates groups A, B and C, and $K = 4$ shows some ability to separate B1 from B2. This clustering reflects, with some noise, the patterns in the PCA plot of Figure 8.2(b), where visual separation of A, B1, B2 and C = {C1,C2} is apparent.

8.3.4 Incorporating Admixture Linkage Disequilibrium

While assuming independence across loci ignores LD, Falush *et al.* (2003) introduced an update to STRUCTURE that allows for correlations in cluster membership among adjacent loci. Their model attempts to capture the mosaic-like pattern of DNA expected in an individual who descends from the admixing of ancestors from genetically distinct populations as illustrated in Figure 8.1. This leads to admixture LD, as an individual’s genome consists of a segment of contiguous loci that all share a common ancestral source, followed by another segment from a different ancestral source, with each of these ancestral source groups (ideally) represented

by a different cluster. To mimic this, in their formulation the cluster membership at a locus depends on that of the previous locus, forming a Markov chain. In particular, equation (8.8), which models each locus l independently of other loci, is replaced by a different model reflecting this dependence structure by using the following formulae:

$$\Pr(Z_{ilj} = k | \Theta) = \Pr(Z_{ilj} = k | r, Q) = q_{ik} \quad (8.10)$$

and

$$\Pr(Z_{i(l+1)j} = k' | Z_{ilj} = k, r, Q) = \begin{cases} \exp(-g_l r) + (1 - \exp(-g_l r))q_{ik} & \text{if } k' = k, \\ (1 - \exp(-g_l r))q_{ik'} & \text{otherwise,} \end{cases} \quad (8.11)$$

where g_l is the genetic distance (in morgans) between loci l and $l + 1$, assumed known, and r can be thought of as a scaling parameter (related to the number of generations ago in which populations admixed) that typically is estimated using the data. Note then that, under this ‘linkage’ model of STRUCTURE, $\Theta = \{P, Q, r\}$ in equations (8.4) and (8.5), with r an additional parameter to be estimated from the data. Though the formulation in equations (8.10) and (8.11) assumes haploid data, Falush *et al.* (2003) describe an approach for dealing with uncertain or unknown phase (see **Chapter 3**).

The model defined by equations (8.10) and (8.11) is equivalent to assuming that the number of switches in cluster membership for haploid genome j of individual i follows a Poisson distribution with mean $g_l r$ between loci l and $l + 1$. Therefore high values of r and/or regions with high rates of recombination (i.e. regions with high g_l) are expected to switch cluster membership between loci more frequently. This formulation mirrors expectations under a simple model of instantaneous admixture between two or more groups (e.g. with each group represented by a distinct cluster) that occurred r generations ago, followed by random mating among individuals from the admixed population (Figure 8.1). Under this setting, each generation of random mating incurs recombinations that break down the lengths of contiguous DNA segments inherited from each admixing group. As a result, within the DNA of descendants of the admixed population living r generations after the admixture event, the boundaries of DNA segments inherited from each admixing group will form a Poisson process of rate r per morgan. The top part of equation (8.11) reflects the probability that either there is no switch between l and $l + 1$, which has probability $\exp(-g_l r)$, or there is at least one switch, which has probability $(1 - \exp(-g_l r))$, with the final switch resulting in a return to the same cluster k with probability q_{ik} . The bottom part of equation (8.11) gives the probability there is at least one switch between l and $l + 1$, and that the final switch results in a switch to cluster k' .

While this new linkage model of (Falush *et al.*, 2003) is a more accurate reflection of the process of admixture in organisms that experience homologous recombination, it can be more computationally expensive. Therefore, more efficient algorithms such as ADMIXTURE and FRAPPE that ignore admixture LD are often applied in practice to cope with the large scales typical of current human data collections. However, computational considerations may be less limiting when considering applications to other organisms.

8.3.5 Incorporating Background Linkage Disequilibrium: Using Haplotypes to Improve Inference

While the linkage model of STRUCTURE models admixture LD, a limitation is that it does not model the background LD occurring *within* each ancestral population. As a result, STRUCTURE and related models assume that each locus is not linked to other neighboring loci, and typically remove loci until all pairwise combinations of nearby loci are not strongly correlated (e.g. have squared correlation coefficient r^2 less than some threshold) to meet this

assumption. This does not take advantage of recent advances in high-throughput genotyping technology that allow the routine generation of high-density SNP data, including sequencing data. The alternative software fineSTRUCTURE (Lawson *et al.*, 2012) accounts for LD when classifying individuals into genetic clusters. Relative to STRUCTURE/ADMIXTURE/FRAPPE and related approaches, fineSTRUCTURE does not require removing data, and furthermore exhibits increased power in some scenarios as a result of modelling the associations among tightly linked loci.

To do so, Lawson *et al.* (2012) first employ the ‘chromosome painting’ technique CHROMOPAINTER that compares genetic variation patterns in one sampled individual to that in all others. Individuals are assumed phased, and each haploid genome of individual i is compared to the $D = 2(N - 1)$ haploid genomes of all $N - 1$ other sampled (diploid) individuals. We will let $Y_{ij} \equiv \{Y_{i1j}, \dots, Y_{iLj}\}$ represent the genetic variation data at all L loci for the (phased) j th haploid genome of individual i (where there are $J = 2$ such haploids in humans and other diploids), with $H_d \equiv \{H_{d1}, \dots, H_{dL}\}$ analogously representing the genetic variation data at all L loci for the (phased) d th haploid genome that individual i 's genome is being compared to. CHROMOPAINTER aims to identify which haploid genome $d \in (1, \dots, D)$ among these other sampled individuals is most recently related to the j th haploid of i at each locus. To do so, CHROMOPAINTER uses an approach derived from the copying model of Li and Stephens (2003), which attempts to capture key features of coalescent modelling while remaining computationally tractable. Two haploids will have relatively similar DNA sequences if they share a recent ancestor, so intuitively CHROMOPAINTER is attempting to identify which DNA sequence(s) among D has allelic patterns that best match that of the j th haploid of i (see Figure 8.6(a)). Therefore, conditional on sharing an inferred most recent ancestor with d at a locus l , CHROMOPAINTER puts high probability on $Y_{ilj} = H_{dl}$. The haploid that is most recently related to the j th haploid of i is expected to switch along the chromosome, because of historical recombination events along the ancestral graph relating the sample. Following Li and Stephens (2003), (Lawson *et al.*, 2012) assume these switches occur as a Poisson process, and that the probability of observing Y_{ij} conditional on H_1, \dots, H_D follows a Markov chain, with

$$\Pr(Y_{i1j} = d | \Phi) = q_{id} \quad (8.12)$$

and

$$\Pr(Y_{i(l+1)j} = d' | Y_{ilj} = d, \Phi) = \begin{cases} \exp(-g_l \rho) + (1 - \exp(-g_l \rho))q_{id} & \text{if } d' = d, \\ (1 - \exp(-g_l \rho))q_{id'} & \text{otherwise.} \end{cases} \quad (8.13)$$

Here $\Phi = \{\rho, q_{i1}, \dots, q_{iD}\}$ are the parameters of the model, with q_{id} the probability of individual i matching to donor d , which can be fixed or estimated (this is fixed to be $1/D$ in fineSTRUCTURE applications), g_l the genetic distance between loci l and $l + 1$ as before, and ρ a scaling constant that is estimated using the data. This is analogous to the linkage STRUCTURE model equations (8.10) and (8.11), but a critical difference is that it captures background LD by comparing genetic variation data among sampled individuals. While there is no straightforward interpretation of ρ here, intuitively ρ summarizes the total amount of diversity among the $2N$ haploid genomes. For example, ρ will generally be lower if comparing genetic variation patterns among individuals sampled within Europe relative to comparing genetic variation patterns among individuals sampled world-wide.

Under the hidden Markov model (HMM) structure defined by equations (8.12) and (8.13), it is straightforward (see, for example, Rabiner, 1989) to calculate $W_i \equiv \{W_{i1}, \dots, W_{iN}\}$, where W_{ih} is the expected number of haplotype segments that the J haploid genomes of individual i match to those of individual $h \in [1, \dots, N]$. Here $W_{ii} = 0$, as individual i 's haploid genomes

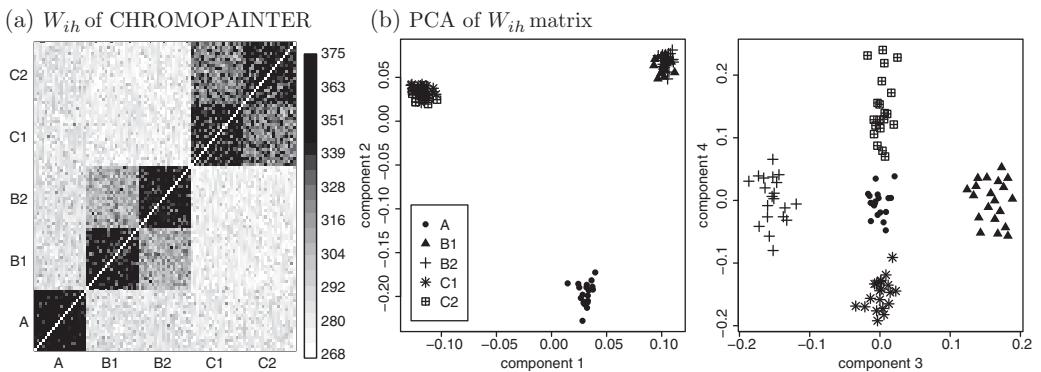


Figure 8.4 (a) Expected number of haplotype segments (W_{ih}) by which each individual i (column) matches to each other individual h (row) for the simulations of populations {A,B1,B2,C1,C2} related as shown in Figure 8.2(a), as inferred by CHROMOPAINTER. (b) The first four principal components of a PCA of the matrix in (a), with each point an individual from one of the five populations, after resetting the diagonal and scaling as described in (Lawson *et al.*, 2012). Note that visualizing the data using the heatmap in (a) shows clear population structure without requiring working through the principal components.

are not matched to each other. An example of the matrix of W_{ih} across all pairings of individuals (i, h) is provided in Figure 8.4(a) for the simulated individuals from Figure 8.2(a). Lawson *et al.* (2012) show that when setting $\rho = \infty$ in equation (8.13), which therefore models loci as unlinked, the matrix containing all W_i closely relates to the genetic information used by STRUCTURE (Pritchard *et al.*, 2000) and by principal components analysis (e.g. using EIGENSTRAT (Price *et al.*, 2006)), motivating use of the W_i as summary statistics in practice.

Roughly speaking, the program fineSTRUCTURE then classifies individuals into K clusters based on which share similar W_i vectors. In particular, fineSTRUCTURE assumes (up to an adjustment factor c below) that W_i follows a multinomial distribution, with $B_{(Z_i)k}/n_{ik}$ the probability that individual i matches a haplotype segment to an individual from cluster k , where Z_i is the cluster assignment of individual i as before and n_{ik} is the number of individuals from cluster k that individual i 's haploid genomes were compared to using equations (8.12) and (8.13). Note that $B_{(Z_i)k}$ is therefore the (unknown) total proportion of haplotype segments that an individual from cluster Z_i matches to all individuals from cluster k . Letting $B_{(Z_i)} = \{B_{(Z_i)1}, \dots, B_{(Z_i)K}\}$ and $W = \{W_1, \dots, W_N\}$, (Lawson *et al.*, 2012) set

$$\Pr(W|Z_1, \dots, Z_N, B_1, \dots, B_K, K) = \prod_{i=1}^N \prod_{h=1}^N \left(\frac{B_{(Z_i)(Z_h)}}{n_{i(Z_h)}} \right)^{W_{ih}/c}, \quad (8.14)$$

where c is a ‘likelihood tuning’ parameter, inferred using the data, to account for the non-independence of the W_{ih} terms violating the assumption of the multinomial distribution. Each individual is assigned to only a single cluster in fineSTRUCTURE, analogous to the model of STRUCTURE that does not allow for admixture. Each $B_m = \{B_{m1}, \dots, B_{mK}\}$ for $m \in [1, \dots, K]$ is assumed to follow a Dirichlet distribution with additional parameters estimated using the model, with these parameters, $\{Z_1, \dots, Z_N\}$, and K estimated using MCMC steps derived in part from (Huelskenbeck and Andolfatto, 2007) and (Pella and Masuda, 2006). FineSTRUCTURE also constructs a tree relating the clusters, by merging pairs of clusters in a greedy fashion until only two remain. This tree can highlight which groups of clusters are most genetically related to one another. However, the authors caution against interpreting the fineSTRUCTURE tree as an ancestral tree relating the groups, as the tree is based on the model and affected by sampling strategy (e.g. sample sizes of different groups) among other factors.

Overall the fineSTRUCTURE model is similar to STRUCTURE, but clusters based on the W_i vectors that capture information about haplotype sharing patterns, rather than clustering based on allele frequencies. Matching haplotype patterns enables an improved inference of shared recent ancestors, which in turn can provide more discriminatory power to distinguish among populations that have only recently become isolated from one another. This is demonstrated in Figure 8.4(b), which applies PCA to the matrix of W_{ih} in Figure 8.4(a). The first four eigenvectors of this PCA show a clear visual separation of all five simulated populations. When applied to these simulated data, fineSTRUCTURE therefore infers the correct number of populations ($K = 5$) and correctly assigns each individual to their true population. In contrast, ADMIXTURE has difficulty distinguishing the DNA of individuals from sub-populations B1 and B2 that split at 2 kya, with little resolution to distinguish the DNA of individuals from sub-populations C1 and C2 that split at 1 kya (Figure 8.3). ADMIXTURE results mimic the PCA analysis of Figure 8.2(b), which similarly uses allele frequencies rather than haplotypes and only partly separates individuals from populations C1 and C2. In other words, using haplotypes appears to make the most difference when trying to capture very subtle genetic variation (e.g. distinguishing C1 from C2), while it may not make much a difference when trying to characterize less subtle variation (e.g. distinguishing A from B = {B1,B2} from C = {C1,C2}).

As a real data example, an application of fineSTRUCTURE to individuals sampled across the United Kingdom elucidated a strong correlation between genetics and geography across England that was not as apparent when analysing the same data using PCA (Leslie *et al.*, 2015). Therefore, among human groups, using haplotypes might show noticeably greater resolution when considering variation within a country, at least for those with predominantly European ancestry. It is worth noting that other definitions of W_{ih} can also be used in equation (8.14) or a related probability function, for example replacing the definition of W_{ih} above with an estimate of the total amount of genome shared identical by descent between individuals i and h as inferred by, for example, BEAGLE (Browning and Browning, 2013). For a comprehensive look at the use of different ‘similarity’ or ‘coancestry’ matrices W , as well as a comparison of different algorithms to cluster individuals using these matrices, see (Lawson and Falush, 2012).

In addition to having increased resolution, both by not having to remove loci *a priori* and from using haplotype information, an additional benefit is that analyses based on haplotypes appear to be less affected by genotype chip ascertainment strategies (Conrad *et al.*, 2006; Hellenthal *et al.*, 2008). However, a disadvantage of CHROMOPAINTER/fineSTRUCTURE is that they require the phase of each individual to be estimated *a priori*. They are also computationally expensive in their current implementation relative to, for example, ADMIXTURE and PCA. Perhaps more importantly, they also currently only allow a model that classifies each individual into a single cluster, rather than the more flexible versions of STRUCTURE that allow individuals to be classified into multiple clusters. In practice, this typically means that individuals with substantial amounts of admixture will be separated into their own clusters rather than cluster with individuals that are closely related to any particular one of the admixing source groups.

8.3.6 Interpreting Genetic Clusters

While the motivation for the mathematical models behind these clustering algorithms reflects plausible biological scenarios, they make a number of simplifying assumptions that may be inappropriate in many applications. For example, the choice of K is often arbitrary. Indeed, note that fineSTRUCTURE selects K automatically in an attempt to parse individuals into groups of genetically indistinguishable individuals, while STRUCTURE-based approaches often assume

fixed smaller values of K (e.g. less than 10) in order to capture a higher level of genetic differentiation among clusters. However, clusters should not necessarily be interpreted as reflecting K genuine ancestral populations that lived in the past (Pritchard *et al.*, 2000). Therefore, while the STRUCTURE-based clustering models can accurately infer proportions of admixture under certain scenarios, the observation that different individuals are classified to be mixtures of different clusters does not necessarily imply admixture.

In particular, individuals who have experienced relatively strong isolation and low effective population size can be assigned to a unique cluster in certain applications with small fixed K . For example, an application of STRUCTURE with $K = 5$ clusters to short tandem repeat data in individuals sampled from 52 world-wide populations largely separates individuals from major worldwide regions, with clusters roughly corresponding to Africa, the Americas, East Asia, West Eurasia, and Oceania (Rosenberg *et al.*, 2002). However, at $K = 6$ clusters, primarily only the Kalash, an isolated group from northwest Pakistan, are separated from these worldwide clusters (Rosenberg *et al.*, 2002). While the clusters at $K = 5$ likely reflect the relatively long time of isolation between these world-wide groups, the separation of the Kalash at $K = 6$ instead likely reflects isolation of this group from the others on a much more recent time-scale (Rosenberg *et al.*, 2002), highlighting the complications in interpreting the demographic meaning of clusters.

This issue was demonstrated by van Dorp *et al.* (2018) in an application to different ethnic and occupational groups from Ethiopia (Pagani *et al.*, 2012). As explained in more detail by (Lawson *et al.*, 2018), applications of ADMIXTURE to these data at small values of K in three different studies classified a particular labelled group, ‘Ari Blacksmiths’, into their own relatively homogeneous cluster, with individuals from other Ethiopian groups showing various degrees of partial assignment to this cluster (Pagani *et al.*, 2012; Hodgson *et al.*, 2014; Dorp *et al.*, 2018). Two of these studies concluded that these observations were consistent with the Ari Blacksmiths reflecting an ancestral source population that subsequently intermixed with the ancestors of some other Ethiopian groups (Pagani *et al.*, 2012; Hodgson *et al.*, 2014). However, the other study performed additional analyses, including those with techniques described in Section 8.5, that instead suggested Ari Blacksmiths are recently related to other Ethiopian groups and have similar admixture histories, with the relatively recent isolation and low effective population size of Ari Blacksmiths from other groups likely driving ADMIXTURE inference (Dorp *et al.*, 2018).

In essence these methods, including CHROMOPAINTER heatmaps as in Figure 8.4(a), are descriptions of the data influenced by demographic processes, for which many processes can lead to the same patterns (Dorp *et al.*, 2018; Lawson *et al.*, 2018). As with all models discussed here, this suggests caution in over-interpreting the results of clustering algorithms. In the next two sections, we discuss methods that instead attempt to infer specific demographic parameters that lead to genetic variation patterns.

8.4 Inferring Population Size Changes and Split Times

In addition to classifying individuals into populations, another major focus of interest is inferring features of populations’ demographic histories, including the timings of when populations separated (or became isolated), and the extent and timing of size changes within each population. There are many techniques to assess the demographic history of a population(s), including those that use derivations of F_{ST} and LD (e.g. (McEvoy *et al.*, 2011)). Here we will focus on two types of demographic inference models that are prevalent in recent literature: techniques that ignore LD by predicting how different demographic models will affect the observed allele

frequency spectrum, and techniques that model LD while analysing whole-genome sequencing data.

8.4.1 Allele Frequency Spectrum Approaches

One technique towards inferring population demography involves predicting its effect on the allele frequency spectrum (AFS), also known as the site frequency spectrum (SFS), which is the distribution of allele counts across loci in a population. The AFS is a complete summary of the data when loci are independent (Adams *et al.*, 2004; Gutenkunst *et al.*, 2009), and several researchers have derived expressions for the AFS under a variety of demographic scenarios involving single or multiple populations (e.g. (Marth *et al.*, 2004; Adams and Hudson, 2004; Gutenkunst *et al.*, 2009; Excoffier *et al.*, 2013)).

As an example, Adams and Hudson (2004) consider the demography of a single population. Let m_i be the number of sampled bi-allelic loci (e.g. SNPs) that have exactly i copies of the derived allele in a sample size of N chromosomes from the population. Given $L = \sum_{i=1}^{N-1} m_i$ total bi-allelic unlinked loci have been sampled, then in the absence of ascertainment bias the AFS (m_1, \dots, m_{N-1}) follows a multinomial distribution,

$$\Pr(m_1, \dots, m_{N-1}) = \binom{L}{m_1 \ m_2 \ \dots \ m_{N-1}} \prod_{i=1}^{N-1} \gamma_i^{m_i}, \quad (8.15)$$

where γ_i is the probability that a locus carries i copies of the derived allele conditional on being polymorphic (Wooding and Rogers, 2002; Polanski and Kimmel, 2003; Adams *et al.*, 2004). Using standard coalescent theory, Adams and Hudson (2004) show that the γ_i terms can be predicted using coalescent simulations under a variety of parameters of interest for a particular demographic model, so that maximum likelihood estimation can be used to identify the parameters that best fit the AFS. In their application, Adams and Hudson (2004) are interested in inferring the time and extent of population size change under a model of exponential population growth following an instantaneous bottleneck, though they note that other demographic models could be considered using equation (8.15).

Other authors have used different derivations of demographic models that predict the AFS based on the questions of interest, often using similar assumptions though attempting to cope with ascertained SNP data. For example, Marth *et al.* (2004) predict the AFS under a scenario where a single population has experienced multiple ‘epochs’ of constant population size, inferring the magnitude of each population size change and the times during which the population remained at each constant size (see Figure 8.5(a)). A model that describes population size fluctuations as periods of constant population size over discrete time intervals, with instantaneous population size changes between intervals, is referred to as ‘piecewise constant population sizes’. Keinan *et al.* (2007) extended the model of Marth *et al.* (2004) to include a second population and predict the split time between the two populations. Meanwhile Gutenkunst *et al.* (2009) used different derivations and considered up to three populations. Using linked loci does not affect the expectation of the AFS under these models but can affect the variance, which the above works address by using simulations with linkage (Adams and Hudson, 2004; Gutenkunst *et al.*, 2009) or bootstrap resampling of linked regions (Keinan *et al.*, 2007) to infer uncertainty around parameters. There are limitations in the amount of information available from the AFS, however, in that very distinct demographic histories can give very similar AFS (Myers *et al.*, 2008), a familiar drawback common to PCA and clustering algorithms, though the ramifications of this in practice are debated (Bhaskar and Song, 2014).

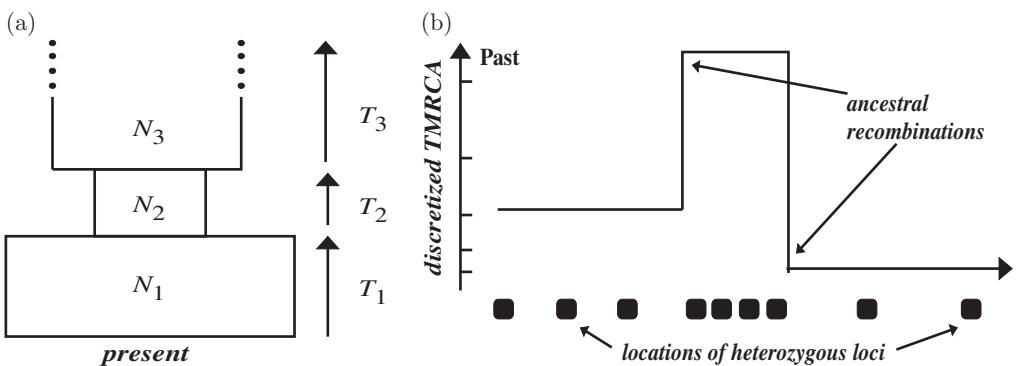


Figure 8.5 (a) Example of the history of a single population divided into epochs (rectangles) of piecewise constant effective population sizes, with an effective population size of N_1 during time period T_1 (leading to the present day at the bottom), size N_2 during time period T_2 , and N_3 during time period T_3 . Adapted from (Marth *et al.*, 2004), with permission. (b) Schematic of the PSMC approach (see main text), illustrating how heterozygous loci (black squares) in a diploid individual inform the TMRCA (lines) of the individuals' two sequences, with the TMRCA categorized into discretized piecewise time periods (tick marks on the y -axis) and changing along the genome due to ancestral recombinations. Adapted from (Li and Durbin, 2011).

8.4.2 Approaches Using Whole-Genome Sequencing

The increasing availability of sequenced whole genomes of humans (e.g. 1000 Genomes Project Consortium, 2012) has encouraged a new suite of techniques to exploit these data to unearth details of human demography with increased precision. Several methods use approximate coalescent models to fit these sequences and infer parameters of interest. For example, Gronau *et al.* (2011) use a Bayesian, coalescent-based approach that assumes unlinked loci and no recombination within loci. In contrast, many whole-genome sequencing approaches explicitly model linkage between loci (Li and Durbin, 2011; Schiffels and Durbin, 2014; Sheehan *et al.*, 2013; Raghavan *et al.*, 2015), thus requiring no *a priori* reduction of data to unlinked loci.

A widely used example of the latter is the pairwise sequentially Markovian coalescent (PSMC) model (Li and Durbin, 2011), which is a specialization of the sequentially Markovian coalescent model of McVean and Cardin (2005), described in **Chapter 2**, to two sampled haploid genomes. In particular, (Li and Durbin, 2011) infers the time to the most recent common ancestor (TMRCA) between the two chromosomes of a diploid individual at each segment of that individual's genome. Assuming a known per-nucleotide mutation rate across the genome, the number of mutations within a genomic segment that has not experienced any historical recombination since the TMRCA can be used as a clock to determine the TMRCA within that segment. Assuming a model of piecewise constant population sizes, the effective population sizes leading to the individual's genetic patterns can be inferred over different time intervals in the past. Due to historical recombination, the TMRCA changes along the individual's genome, with each such genetic segment in theory providing independent information to assist with inferring this demographic history. Changes in the TMRCA are modelled with an HMM that takes into account linkage between loci, with a transition from a segment with time s to a segment with time t modelled as

$$\Pr(t|s) = (1 - e^{-\rho t}) \left[\frac{1}{\lambda(t)} \int_0^{\min(s,t)} \frac{1}{s} e^{-\int_u^t \frac{dv}{\lambda(v)}} du \right] + e^{-\rho t} \delta(t-s), \quad (8.16)$$

with ρ the scaled recombination rate, $\lambda(t)$ the population size during the segment with TMRCA t relative to the present day, and $\delta(t-s)$ the Dirac delta function. Conditional on t , the

emission probabilities of the HMM determine the probability that each observed locus is homozygous (i.e. no historical mutations since the TMRCA), which is modelled as a function of the (unknown) mutation rate and t . The mutation rate, ρ , and the $\lambda(t)$ across discretized time intervals are inferred using an Expectation-Maximization algorithm.

Intuitively, the PSMC model can be thought of as dividing an individual's genome into regions separated by historical recombinations and counting heterozygotes within each region to infer the TMRCA between the individual's chromosomes in that region (see Figure 8.5(b)). Regions with a relatively large number of heterozygotes have a relatively large expected TMRCA. If several regions have their inferred TMRCAs classified into the same discretized time interval, this suggests a smaller effective population size during this time interval. This is because a smaller population size causes an increase in the rate of coalescence events, which is suggested by the large number of regions coalescing during this period.

One attractive feature of PSMC is that inference on population demography can be performed using data from only a single genome, which is of particular interest for analysing DNA from ancient human remains (aDNA) for which relatively few samples have high-coverage DNA representing any particular group. However, this can also be seen as a major limitation, in that PSMC has little power to infer population size changes in recent history within the last 20,000 years (Schiffels and Durbin, 2014). In particular, under standard coalescent theory the mean TMRCA between two sequences is the effective population size N (Hudson, 1991), which among human populations is inferred to be 2000–10,000 generations ago (e.g. (Tenesa *et al.*, 2007)). In contrast, the expected time to the first coalescence among multiple sequences is more recent (e.g. a factor of $\binom{10}{2} = 45$ times more recent with 10 sequences). In part for this reason, various approaches have been proposed that approximate the coalescent when analysing multiple sequences (Schiffels and Durbin, 2014; Sheehan *et al.*, 2013; Raghavan *et al.*, 2015) that have been pre-phased, with the aim of characterizing more recent demography within populations and/or inferring the timing of population splits. In addition, while clean 'splits' of populations seem unlikely in reality, such models can also identify groups instead gradually becoming separated with less frequent intermixing over time (Schiffels and Durbin, 2014). However, these multiple individual methods are computationally demanding, at present only allowing joint analysis of tens of samples at a time, and require phased genomes.

As with the AFS, multiple demographic histories can give similar patterns under these approaches, with Mazet *et al.* (2016) showing how sub-structure within a population can lead to inaccurate inference of population size changes, suggesting caution is warranted when interpreting results. Furthermore, inferring the time in years of population splits and size changes relies on an estimate of the mutation rate in humans, which is a subject of some debate (Mazet *et al.*, 2012).

8.5 Identifying/Dating Admixture Events

While the STRUCTURE-based clustering algorithms described in Section 8.3 can identify whether individuals are admixed and infer the proportions of DNA inherited from each admixing source in some settings, different ancestral histories that either include or exclude admixture can lead to the same patterns in clustering (Lawson *et al.*, 2018). This makes reliable identification of admixture challenging using these approaches. Correlated-drift approaches (Patterson *et al.*, 2012; Pickrell and Pritchard, 2012) can identify admixture and infer proportions with some accuracy, but these approaches cannot determine the time period(s) over which admixture has occurred. In this section we describe approaches that attempt to describe both which groups admixed and when they admixed. These approaches often assume that admixture

happens in pulses (as exemplified in Figure 8.1), but can also infer continuous intermixing (Henn *et al.*, 2012) between sources at a constant rate over a period of time, with some challenges in distinguishing between these scenarios.

As illustrated in Figure 8.1 and described earlier, assume two or more populations experience an instantaneous admixture event r generations ago. Individuals in the admixed population randomly mate for r generations (i.e. until the present day) following this admixture event. Let a ‘tract’ refer to a segment of DNA that has been inherited intact from a haploid genome in one of the admixing populations without any recombination occurring within the segment since r . Recall in Section 8.3.4 that, under some simplifying biological assumptions, such as no crossover interference or impact of gene conversion, the boundaries between tracts of DNA in present-day descendants will follow a Poisson process of rate r per morgan (Falush *et al.*, 2003). On an admixed genome, the probability $\Pr(A \rightarrow A; g)$ that two loci separated by distance g (in morgans) are inherited from the same source A , which has contributed a proportion α of DNA overall to the admixed population, is

$$\Pr(A \rightarrow A; g) = \alpha(e^{-gr} + [1 - e^{-gr}]\alpha) = \alpha^2 + \alpha(1 - \alpha)e^{-gr}. \quad (8.17)$$

This is the probability of either (i) being an intact segment of length g inherited from a single ancestor from population A , which has probability αe^{-gr} , or (ii) both loci being inherited from A despite one or more tract switches between them, which has probability $\alpha^2(1 - e^{-gr})$. In the simple case of admixture between only two populations, with the other population B contributing a proportion $\beta = 1 - \alpha$ of DNA overall to the admixed population, the analogous expression to equation (8.17) for $\Pr(B \rightarrow B; g)$ simply replaces α with β . The probability $\Pr(A \rightarrow B; g)$ that two loci separated by g are inherited from different populations is

$$\Pr(A \rightarrow B; g) = \Pr(B \rightarrow A; g) = \alpha\beta(1 - e^{-gr}). \quad (8.18)$$

Analogous expressions can be derived for more complex cases involving more than two admixing populations and more than one pulse of admixture (Hellenthal *et al.*, 2014).

A key difference between equations (8.17) and (8.18) is that the former decays with increasing g , while the latter increases with g , with both having rate gr . This makes it possible in theory to disentangle which sampled groups best reflect source A versus source B , a feature that is exploited by the program GLOBETROTTER (Hellenthal *et al.*, 2014) described below.

In this section, we describe two different types of approaches to infer the dates and proportions of admixture. The first type explicitly infers local segments of DNA that are inherited from different admixing source groups. The second type considers associations among loci that are attributable to admixture LD, without having to directly identify which DNA segments are inherited from each source group. Figure 8.6(a) illustrates the intuition behind the various approaches mentioned in this section. In all methods that date admixture based on using tracts of DNA segments, typically only admixture within the last few hundred generations can be detected reliably, since for older events most tracts will have decreased to lengths not distinguishable from background noise.

8.5.1 Inferring DNA Segments Inherited from Different Sources

The Poisson process for delineating tracts of DNA from admixing sources provides a means of modelling the distribution of lengths of contiguous tracts inherited from each admixing source. Therefore, accurately identifying which DNA segments in the genomes of present-day admixed individuals are derived from each source enables estimation of the proportions of DNA contributed by each source and how many generations ago the admixture occurred.

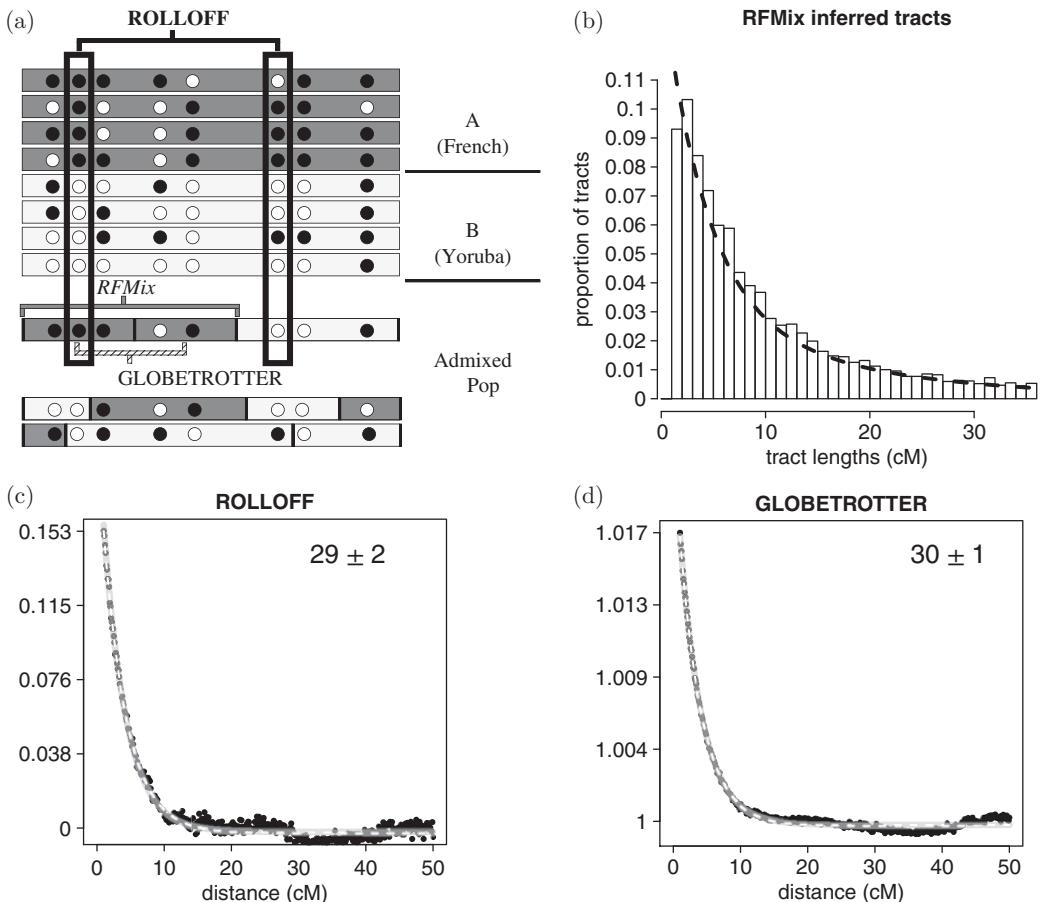


Figure 8.6 (a) Illustration of approaches to infer dates and proportions of admixture. Two source populations *A* and *B* (e.g. French and Yoruba) mix r generations in the past (e.g. $r = 30$) as in Figure 8.1, generating admixed haplotypes at bottom. Circles denote allele types at each bi-allelic SNP; dark vertical bars separate segments in admixed chromosomes that exactly match a region of a depicted source haplotype. RFMix (Maples *et al.*, 2013) attempts to identify contiguous tracts inherited from each of *A* and *B*. ROLLOFF/ALDER (Patterson *et al.*, 2012; Loh *et al.*, 2013) measure associations among SNPs separated by genetic distance that are informative for distinguishing *A* and *B*. GLOBETROTTER (Hellenthal *et al.*, 2014) measures associations among haplotype segments separated by genetic distance that are inferred by CHROMOPAINTER to match to surrogate haploids for *A* and *B*. (b) Distribution of inferred lengths of tracts of at least 1 cM from each source *A* and *B* in simulated admixed individuals using RFMix, with the dotted line giving the predicted distribution under the true dates and proportions of admixture (Wangkumhang and Hellenthal, 2018). (c), (d) Inferred association (black lines) between loci versus genetic distance, ignoring loci separated by less than 1 cM, for ROLLOFF and GLOBETROTTER, respectively, for the same simulations, with the true LD decay curve in light grey and the curve for the best fitting model for each method in dashed white. The legend at top right gives the inferred date and standard error, with the latter determined by jackknifing over chromosomes for ROLLOFF and 100 bootstrap resamples for GLOBETROTTER (Wangkumhang and Hellenthal, 2018).

Typically, statistical models are used to compare the genetic variation data of present-day admixed individuals to that of ‘surrogate’ individuals who are meant to represent genetically each of the admixing source groups, identifying contiguous tracts of DNA inferred to descend from each admixing source as in Figure 8.6(a). The sizes of these inferred tracts are then compared to that expected under different models of migration, to determine, for example, the best-fitting date(s) of admixture and proportions of ancestry inherited from each admixing source

(Pool and Nielsen, 2009; Gravel *et al.*, 2012). Several different algorithms have been proposed to do this matching (e.g. (Tang *et al.*, 2006; Sankararaman *et al.*, 2008; Price *et al.*, 2009; Bryc *et al.*, 2010; Baran *et al.*, 2012; Brisbin *et al.*, 2012; Churchhouse *et al.*, 2013; Maples *et al.*, 2013; Guan, 2014)), with the details of many summarized in (Churchhouse and Marchini, 2013).

As an example, Figure 8.6(b) summarises results from applying RFMix (Maples *et al.*, 2013), which uses linear discriminant analysis to match segments in admixed individuals to those from reference populations, to simulated data from (Hellenthal *et al.*, 2014). Here 20 simulated admixed individuals were composed of tracts of DNA copied intact from individuals randomly chosen from the source populations, with tract sizes sampled according to an exponential distribution as described in (Price *et al.*, 2009). The simulations presented here generated admixed genomes composed of 474,491 autosomal SNPs, each carrying DNA from African and European populations assuming a single instantaneous pulse of admixture between them occurring 30 generations ago. Twenty-one Yoruban individuals from Nigeria were used to represent the African population and 28 individuals from France were used to represent the European population, with 20% of the DNA coming from the French individuals. RFMix was then applied to the admixed genomes using the genomes of 22 Mandenka individuals as surrogates for the African source, and 23 English, Irish, Scottish and Welsh individuals as surrogates for the European source. Figure 8.6(b) shows a histogram of the segment sizes matched entirely to the African or French reference groups, as inferred by RFMix, ignoring tracts smaller than 1 cM. The dashed line in this figure shows the expected distribution of tracts, summed across both sources. While equation (8.17) cannot be used directly for this expectation, since it allows for tracts to be inherited from different sources between the two loci, an expression for $\tilde{\Pr}(A \rightarrow A; g)$, the probability that all tracts between two loci separated by distance g are from A , is

$$\tilde{\Pr}(A \rightarrow A; g) = \sum_{x=0}^{\infty} [\alpha^{x+1} (gr)^x e^{-gr} / x!]. \quad (8.19)$$

Equation (8.19) and its analogue $\tilde{\Pr}(B \rightarrow B; g)$, using $r = 30$ and $\alpha = 0.2$, provide the probability distributions to calculate the expected line in Figure 8.6(b). In general there is good agreement between this expectation and the observed distribution inferred by RFMix, highlighting the utility of these approaches that explicitly infer local ancestry tracts.

A limitation of these ancestry tract inference methods is that many, though not all (Guan, 2014; Sankararaman, 2008), require pre-specified reference individuals to act as surrogates to the admixing sources. Therefore, their accuracy depends on how well these surrogates genetically match the original admixing source groups. An issue for all such approaches is that the admixing sources must be distinguishable using only a few SNPs in a local region. For these reasons, these approaches are typically only applicable in humans to admixture scenarios that involve recent intermixing among different continental groups, as is the case in Latin and African Americans. In such cases, tracts are both easier to identify due to their increased length and the genetic differentiation between admixing sources, and it is more likely that sampled extant populations well reflect the original admixing sources.

8.5.2 Measuring Decay of Linkage Disequilibrium

Various approaches have been proposed to identify and describe admixture without having to directly infer local tracts of DNA inherited from each admixing source. These approaches instead measure the decay of admixture LD versus genetic distance.

For example, the techniques ROLLOFF (Moorjani *et al.*, 2011; Moorjani, 2013; Patterson, 2012) and ALDER (Loh *et al.*, 2013) measure the decay in LD versus genetic distance among

pairs of bi-allelic loci (e.g. SNPs), weighted by each locus's ability to distinguish between the original admixing populations. In particular, again consider an admixed population formed by an instantaneous admixture event between populations A and B occurring r generations ago, with proportion α of the DNA from population A and the rest from B , followed by r generations of random mating (Figure 8.1). Assuming an infinite population size, the covariance $\text{cov}(x, y)$ between two bi-allelic loci x and y , which are separated by distance g (in morgans) and in linkage equilibrium in both A and B , in a diploid individual from this admixed population is:

$$\text{cov}(x, y) = 2\alpha(1 - \alpha)(p_{Ax} - p_{Bx})(p_{Ay} - p_{By})e^{-gr}, \quad (8.20)$$

where p_{Ax} and p_{Bx} are the proportions of a particular allele type in populations A and B , respectively, at locus x (Chakraborty and Weiss, 1988; Loh, 2013). As x and y are in linkage equilibrium in A and B , this covariance between x and y is attributable to the admixture. Only loci with allele frequency differences between A and B contribute to equation (8.20). Therefore, using surrogates for populations A and B , ROLLOFF and ALDER weight the sample covariance (or correlation) between pairs of bi-allelic loci in the admixed population by the sample allele frequency differences at these loci in the two surrogate populations. They then fit the decay in this weighted covariance (or correlation) to an exponential function that decays with rate r per morgan, in order to estimate r . Furthermore, in certain cases the amplitude of these decay curves can be used to infer the proportion of admixture α (Loh *et al.*, 2013). The model has also been extended to consider multiple admixture events at different times, which incorporates additional exponential functions to equation (8.20), each with rate equal to the number of generations since their respective admixture event (Pickrell *et al.*, 2014). Figure 8.6(c) shows the inferred decay in the association between (weighted) pairs of SNPs when applying ROLLOFF to the simulated data described in Section 8.5.1, using the same surrogate groups. Note how well this decay fits an exponential decay with rate 30, which is the true date of admixture.

The program GLOBETROTTER (Hellenthal *et al.*, 2014) takes an alternative approach to date and describe admixture. First CHROMOPAINTER matches segments of DNA within (pre-phased) individuals of the putatively admixed population to that of surrogate populations' (pre-phased) individuals. GLOBETROTTER then measures the association among pairs of segments separated by genetic distance g that are matched to surrogate populations A' and B' , following the theory defined by equations (8.17) and (8.18) and using mixture models to account for biases in the CHROMOPAINTER inference that may be due, for example, to unequal sample size. This approach can increase power by using haplotype information from tightly linked SNPs, rather than allele frequency differences, to discriminate among populations. Figure 8.6(d) shows the inferred correlation among segments matched to a particular surrogate population (Mandenka) when applying GLOBETROTTER to the same simulations and surrogates as above. These illustrate a tighter curve around the true date, increasing the precision slightly even in this relatively easy admixture example.

GLOBETROTTER also leverages information on which pairs of surrogate populations A' and B' show exponentially increasing curves as predicted by equation (8.18), allowing inference of which surrogate populations best represent each admixing source group. For this reason, GLOBETROTTER does not require *a priori* specification of surrogates for each admixing source group, but instead infers each admixing group as a mixture of an unlimited number of sampled surrogate groups. In contrast, ROLLOFF and ALDER infer a single best surrogate group to represent each source by finding the best model fit out of all possible pairings of available surrogates. GLOBETROTTER can also identify multiple admixture events at different times, as well as admixture among three or more sources occurring at approximately the same time. This comes at increased computational expense, in addition to the requirement of phasing the data *a priori*.

In each of ROLLOFF, ALDER and GLOBETROTTER, inferring proportions of admixture from each source typically is more challenging than inferring dates of admixture, with the former often suffering from a lack of identifiability. These models also assume ‘pulses’ of admixture occurring over a short time frame. Continuous admixture occurring over several (perhaps continuous) migrations may be more realistic. In the case of continuous admixture, inferred dates from these approaches typically fall in between the start and end dates of continuous admixture, though may be biased towards the most recent date (Loh *et al.*, 2013; Hellenthal *et al.*, 2014). Also, as in the methods of Section 8.5.1, accuracy still depends on how genetically differentiated the admixing sources are, and how well genetic patterns in each source are captured by sampled surrogate individuals. Nonetheless, decay of LD due to admixture can still be usefully modelled by these approaches in far more subtle cases relative to approaches that directly identify tracts in Section 8.5.1. For example, GLOBETROTTER inferred admixture among European-like source groups dated to over 1 kya when applied to genetic variation data from British individuals (Leslie *et al.*, 2015).

By averaging over information across the genome, these techniques should be relatively robust to occasional genotyping or sequencing errors that affect only a limited number of genetic regions. They also appear to be robust to using different human genetic maps in practice (Hellenthal *et al.*, 2014), despite evidence that recombination hotspot locations can vary among human groups (Crawford *et al.*, 2004). However, this is perhaps not surprising, as admixture LD tends to extend over megabase scales, for which average recombination rates typically are concordant across continental groups (Hinch *et al.*, 2011). Nonetheless, disparities in genetic maps may affect inference of older admixture events, for which a large proportion of admixture LD may decay within, say, a megabase.

8.6 Conclusion

While this chapter attempts to summarize the models and concepts behind presently popular algorithms for exploring large-scale autosomal data, we remind the reader that this is by no means an exhaustive look at the numerous exciting methods to explore these (and other) questions about demography. Each method presented here has limitations, in terms of both modelling assumptions and computational tractability. Not all limitations and assumptions may be listed here. We refer the reader to the cited literature for more details on particular approaches, though note that exploration of how several of these methods behave in practice is ongoing (e.g. see (Liang and Nielsen, 2014) for a discussion on limitations of admixture date inference approaches).

While the approaches outlined here are the current state of the art in the field of population genetics, they all still assume a major over-simplification of history. For example, the approaches described above focus on inferring one or two features of demographic history, such as population substructure or admixture or population size changes, while ignoring the other factors that may have altered observed genetic variation patterns. In contrast, there are also more comprehensive approaches that try to infer jointly the ancestral relationships among dozens of sampled populations, including the order of splits and subsequent drift effects (e.g. related to population size changes), as well as admixture events among the ancestors of these populations (Pickrell *et al.*, 2012; Lipson, 2013). However, exhaustively exploring all possible scenarios of splits, admixture and population trees is intractable, so that simplifying assumptions (such as assuming a fixed tree topology) must be used in practice to reduce the search space. For this reason, techniques in this chapter that simplify the problem by, for example, inferring only one or two of these demographic processes will likely remain important and shed light on the demographic history of humans in a piecemeal fashion.

Furthermore, increasingly larger data sets are becoming available for human populations, from massive whole-genome sequencing studies of different populations such as UK Biobank (<http://www.ukbiobank.ac.uk/>) and China Kadoorie Biobank (<http://www.ckbiobank.org/site/>) and potentially through large databases of customers' genotypes acquired through genetic ancestry testing companies (Bryc *et al.*, 2015). These large data resources will enable better understanding of demography, for example allowing identification of individuals that share ancestors more recently than is typically seen in collections of smaller sample size, but will require substantial computational improvements to fully extract the rich available information. In addition to the rapidly increasing resources from present-day populations, techniques to reliably extract high-quality DNA from aDNA are facilitating the increasing availability of genetic resources from numerous historical cultures and time periods (e.g. (Lazaridis *et al.*, 2016)). These aDNA samples are already proving extremely valuable for our understanding of ancient human history and will continue to do so. Overall these forthcoming data resources will increase the resolution of genetic studies for unearthing details of human history, suggesting that we are just scratching the surface with current applications.

Acknowledgements

I thank David Balding and Mark Beaumont for their helpful comments that improved the chapter considerably. I am supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant no. 098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References

- 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 55–65.
- Adams, A.M. and Hudson, R.R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712.
- Alexander, D.H., Novembre, J. and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664.
- Balding, D.J. and Nichols, R.A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**(1–2), 3–12.
- Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., Rodriguez-Santana, J., Burchard, E.G. and Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**(10), 1359–1367.
- Behar, D.M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Roots, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E.K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M.F., Skorecki, K. and Villems, R. (2010). The genome-wide structure of the Jewish people. *Nature* **466**, 238–242.
- Bhaskar, A. and Song, Y.S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics* **42**(6), 2469–2493.
- Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reyonalds, A., Ostrer, H., Mezey, J.G. and Bustamante, C.D. (2012). PCAdmix: Principal components-based assignment of

- ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology* **84**(4), 343–364.
- Browning, B.L. and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**(2), 459–471.
- Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A. and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America* **107**(2), 786–791.
- Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D. and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American Journal of Human Genetics* **96**, 37–53.
- Chakraborty, R. and Weiss, K. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences of the United States of America* **85**(23), 9119–9123.
- Churchhouse, C. and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic Epidemiology* **37**(1), 1–12.
- Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A. and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics* **38**(11), 1251–1260.
- Crawford, D.C., Bhagale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A. and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics* **36**(7), 700–706.
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9**(10), e1003905.
- Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.
- Galinsky, K.J., Bhatia, G., Loh, P.R., Georgiev, S., Mukherjee, S., Patterson, N.J. and Price, A.L. (2016). Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *American Journal of Human Genetics* **98**(3), 456–472.
- Gravel, G. (2012). Population genetics models of local ancestry. *Genetics* **191**, 607–619.
- Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics* **43**, 1031–1034.
- Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics* **196**, 625–642.
- Gutenkunst, R.N., Hernandez, R., Williamson, S.H. and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**(10), e1000695.
- Hellenthal, G., Auton, A. and Falush, D. (2008). Inferring human colonization history using a copying model. *PLoS Genetics* **4**(5), e1000078.
- Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014). A genetic atlas of human admixture history. *Science* **343**, 747–751.
- Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetti, J., Bustamante, C.D. and Comas, D. (2012). Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genetics* **8**(1), e1002397.
- Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., Bock, C.H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Cupples, L.A., Deming, S.L., Diver, W.R., Divers, J., Fornage, M., Gillanders, E.M., Glessner, J.,

- Harris, C.C., Hu, J.J., Ingles, S.A., Isaacs, W., John, E.M., Kao, W.H., Keating, B., Kittles, R.A., Kolonel, L.N., Larkin, E., Le Marchand, L., McNeill, L.H., Millikan, R.C., Murphy, A., Musani, S., Neslund-Dudas, C., Nyante, S., Papanicolaou, G.J., Press, M.F., Psaty, B.M., Reiner, A.P., Rich, S.S., Rodriguez-Gil, J.L., Rotter, J.I., Rybicki, B.A., Schwartz, A.G., Signorello, L.B., Spitz, M., Strom, S.S., Thun, M.J., Tucker, M.A., Wang, Z., Wiencke, J.K., Witte, J.S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K.A., Zheng, W., Ziegler, R.G., Zhu, X., Redline, S., Hirschhorn, J.N., Henderson, B.E., Taylor, H.A. Jr., Price, A.L., Hakonarson, H., Chanock, S.J., Haiman, C.A., Wilson, J.G., Reich, D. and Myers, S.R. (2011). The landscape of recombination in African Americans. *Nature* **476**, 170–175.
- Hodgson, J.A., Mulligan, C.J., Al-Meeri, A. and Raaum, R.L. (2014). Early back-to-Africa migration in the Horn of Africa. *PLoS Genetics* **10**(6), e1004393.
- Hudson, R.R. (1991) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7**, 1–44.
- Huelsenbeck, J.P., Andolfatto, P. and Huelsenbeck, E.T. (2011). Structurama: Bayesian inference of population structure. *Evolutionary Bioinformatics Online* **7**, 55–59.
- Huelsenbeck, J.P. and Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175**, 1787–1802.
- Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, R.J., Vanlier, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C., Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A. and Singleton, A.B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003.
- Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., Meshveliani, T., Bar-Yosef, O., Muller, W., Belfer-Cohen, A., Matskevich, Z., Jakeli, N., Higham, T.F., Currat, M., Lordkipanidze, D., Hofreiter, M., Manica, A., Pinhasi, R. and Bradley, D.G. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications* **6**, 8912.
- Keinan, A., Mullikin, J.C., Patterson, N. and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics* **39**, 1251–1255.
- Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetti, J., Bindoff, L.A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabruni, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A.G., Gieger, C., Wichmann, H.E., Ruther, A., Schreiber, S., Becker, C., Nurnberg, P., Nelson, M.R., Krawczak, M. and Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology* **18**(16), 1241–1248.
- Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics* **8**(1), e1002453.
- Lawson, D.J. and Falush, D. (2012). Population identification using genetic data. *Annual Review of Genomics and Human Genetics* **13**, 337–361.
- Lawson, D., van Dorp, L. and Falush, D. (2018). A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *Nature Communications* **9**(1), 3258.
- Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E.R., Roodenberg, S.A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J.M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S.M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D.A., O'Reilly, S., Richards, M.B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J.F., Yengo, L., Hovhannisan, N.A.,

- Patterson, N., Pinhasi, R. and Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**(7617), 419–424.
- Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Rorvik, E.C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D.J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P. and Bodmer, W. (2015). The fine scale genetic structure of the British population. *Nature* **519**, 309–314.
- Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**(4), 2213–2233.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496.
- Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics* **197**, 953–967.
- Lipson, M., Loh, P.R., Levin, A., Reich, D., Patterson, N. and Berger, B. (2013). Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution* **30**(8), 1788–1802.
- Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D. and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**(4), 1233–1254.
- Maples, B.K., Gravel, S., Kenny, E.E. and Bustamante, C.D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics* **93**(2), 278–288.
- Marth, G.T., Czabarka, E., Murvai, J. and Sherry, S.T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372.
- Mazet, O., Rodriguez, W., Grusea, S., Boitard, S. and Chikhi, L. (2016). On the importance of being structured: Instantaneous coalescence rates and human evolution – lessons for ancestral population size inference? *Heredity* **116**, 362–371.
- Mazet, O., Rodriguez, W. and Chikhi, L. (2012). Revising the human mutation rate: implications for understanding human variation evolution. *Nature Reviews Genetics* **13**, 745–753.
- McEvoy, B.P., Powell, J.E., Goddard, M.E. and Visscher, P.M. (2011). Human population dispersal ‘out of Africa’ estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Research* **21**(6), 821–829.
- McVean, G. (2009). A genealogical interpretation of principal components. *PLoS Genetics* **5**(10), e1000686.
- McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1387–1393.
- Menozzi, P., Piazza, A. and Cavalli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792.
- Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Magi, R., Metspalu, E., Remm, M., Pitchappan, R., Singh, L., Thangaraj, K., Villemans, R. and Kivisild, T. (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *American Journal of Human Genetics* **89**, 731–744.
- Moorjani, P., Patterson, N., Hirschhorn, J.N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A.L. and Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics* **7**(4), e1001373.

- Moorjani, P., Patterson, N., Loh, P.R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kadasi, L., Riess, O., Berger, B., Reich, D. and Melegh, B. (2013). Reconstructing Roma history from genome-wide data. *PLoS ONE* **8**(3), e58633.
- Myers, S., Fefferman, C. and Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology* **73**, 342–348.
- Nicholson, G., Smith, A.V., Jónsson, F., Gústafsson, Ó., Stefánsson, K. and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society, Series B* **64**, 695–715.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergman, S., Nelson, M.R., Stephens, M. and Bustamante, C.D. (2008). Genes mirror geography within Europe. *Nature* **456**, 98–101.
- Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics* **40**, 646–649.
- Pagani, L., Kivisild, T., Tarekagn, A., Ekong, R., Plaster, C., Gallego-Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., Bekele, E., Bradman, N., Balding, D.J. and Tyler-Smith, C. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *American Journal of Human Genetics* **91**(1), 83–96.
- Patterson, N., Price, A.L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**(12), e190.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. (2012). Ancient admixture in human history. *Genetics* **192**(3), 1065–1093.
- Pella, J. and Masuda, M. (2006). The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences* **63**, 576–596.
- Pickrell, J.K. and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**, e1002967.
- Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences of the United States of America* **111**(7), 2632–2637.
- Polanski, A. and Kimmel, M. (2003). New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165**, 427–436.
- Pool, J.E. and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 504–509.
- Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics* **5**(6), e1000519.
- Pritchard, J.K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**(1), 1–14.
- Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotypes data. *Genetics* **155**, 945–959.
- Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286.
- Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspina, A.S., Eriksson, A., Moltke, I., Metspalu, M.,

- Homburger, J.R., Wall, J., Cornejo, O.E., Moreno-Mayar, J.V., Korneliussen, T.S., Pierre, T., Rasmussen, M., Campos, P.F., Damgaard, P.D.B., Allentoft, M.E., Lindo, J., Metspalu, E., Rodríguez-Varela, R., Mansilla, J., Henrickson, C., Seguin-Orlando, A., Malmström, H., Stafford, T., Shringarpure, S.S., Moreno-Estrada, A., Karmin, M., Tambets, K., Bergström, A., Xue, Y., Warmuth, V., Friend, A.D., Singarayer, J., Valdes, P., Balloux, F., Leboreiro, I., Vera, J.L., Rangel-Villalobos, H., Pettener, D., Luiselli, D., Davis, L.G., Heyer, E., Zollikofer, C.P.E., Ponce de León, M.S., Smith, C.I., Grimes, V., Pike, K.A., Deal, M., Fuller, B.T., Arriaza, B., Standen, V., Luz, M.F., Ricaut, F., Guidon, N., Osipova, L., Voevoda, M.I., Posukh, O.L., Balanovsky, O., Lavryashina, M., Bogunov, Y., Khusnutdinova, E., Gubina, M., Balanovska, E., Fedorova, S., Litvinov, S., Malyarchuk, B., Derenko, M., Mosher, M.J., Archer, D., Cybulski, J., Petzelt, B., Mitchell, J., Worl, R., Norman, P.J., Parham, P., Kemp, B.M., Kivisild, T., Tyler-Smith, C., Sandhu, M.S., Crawford, M., Villemans, R., Smith, D.G., Waters, M.R., Goebel, T., Johnson, J.R., Malhi, R.S., Jakobsson, M., Meltzer, D.J., Manica, A., Durbin, R., Bustamante, C.D., Song, Y.S., Nielsen, R. and Willerslev, E. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, 841.
- Raj, A., Stephens, M. and Pritchard, J.K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**(2), 573–589.
- Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W. Jr., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., Poznik, G.D., Gudmundsdóttir, V., Yadav, R., Malaspinas, A.S., White 5th, S.S., Allentoft, M.E., Cornejo, O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S., Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V.M., Lopes, M.C., Malhi, R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F., Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R. and Willerslev, E. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229.
- Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K. and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**(6), e70.
- Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002). Genetic structure of human populations. *Science* **298**, 2981–2985.
- Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**, 919–925.
- Schlebusch, C.M., Skoglund, P., Sjodin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G., Soodyall, H. and Jakobsson, M. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379.
- Sheehan, S., Harris, K. and Song, Y.S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662.
- Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008). Estimating local ancestry in admixed populations. *American Journal of Human Genetics* **82**(2), 290–303.
- Tang, H., Peng, J., Wang, P. and Risch, N. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology* **28**, 289–301.
- Tang, H., Coram, M., Wang, P., Xhu, X. and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *American Journal of Human Genetics* **79**, 1–12.
- Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E. and Visscher, P.M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**(4), 520–526.

- van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., Tarekegn, A., Thomas, M.G., Bradman, N. and Hellenthal, G. (2015). Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the Last 4500 years: Lessons for clustering-based inference. *PLoS Genetics* **11**(8), e1005397.
- Wangkumhang, P. and Hellenthal, G. (2018). Statistical methods for detecting admixture. *Current Opinion in Genetics & Development* **53**, 121–127.
- Wooding, S. and Rogers, A. (2002). The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics* **161**, 1641–1650.

Statistical Methods to Detect Archaic Admixture and Identify Introgressed Sequences

Liming Li¹ and Joshua M. Akey²

¹Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA

²Department of Ecology and Evolutionary Biology and Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA

Abstract

Genome-scale sequencing data sets from many taxa have shown that hybridization across leaky species boundaries is common in natural populations. This observation extends to our own species, as gene flow between anatomically modern humans and other hominins, such as Neanderthal and Denisovans, has been well established. In this chapter, we describe statistical approaches that have been used to test hypotheses of gene flow between species and identify introgressed sequences. These methods provide powerful tools to investigate population history, understand mechanisms of evolutionary change, and better understand the heritable basis of phenotypic variation.

9.1 Introduction

Recent technological innovations have enabled whole-genome sequencing of ancient DNA (aDNA) from a wide variety of organisms and time-scales (Slatkin and Racimo, 2016; Grealy *et al.*, 2017; Green and Speller, 2017; Warinner *et al.*, 2017). These data are revolutionizing inferences of population history and mechanisms of evolutionary change (Stoneking and Krause, 2011; Racimo *et al.*, 2015; Marciniak and Perry, 2017; Nielsen *et al.*, 2017). For instance, aDNA is revealing fascinating new stories about the history of anatomically modern humans, and how our species interacted with archaic hominins, such as Neanderthals and Denisovans. In particular, sequencing of the Neanderthal (Prüfer *et al.*, 2014, 2017) and Denisovan (Meyer *et al.*, 2012) genomes, combined with the development of novel statistical methods, provided the resources to infer that hybridization occurred between these now extinct hominin species and modern human ancestors (Lafferty *et al.*, 2001; Plagnol and Wall, 2006; Reich *et al.*, 2010; Durand *et al.*, 2011; Meyer *et al.*, 2012; Patterson *et al.*, 2012; Prüfer *et al.*, 2014; Seguin-Orlando *et al.*, 2014; Vernot and Akey, 2014; Vernot *et al.*, 2016; Browning *et al.*, 2018). Indeed, it has now been well established that approximately 2% of non-African ancestry is derived from Neanderthal ancestors and an additional 4–6% of ancestry in Melanesians was inherited from Denisovans (Vernot *et al.*, 2016). Thus, admixture has occurred throughout human evolutionary history, both anciently between archaic hominin species and more recently among modern human populations (Xu and Jin, 2008; Reich *et al.*, 2009).

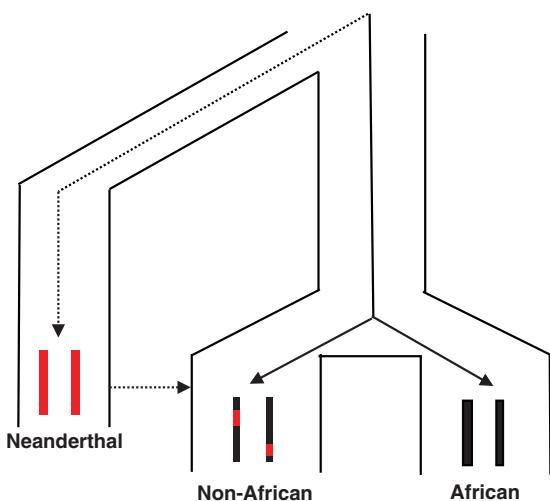


Figure 9.1 Schematic of Neanderthal gene flow into non-African modern humans. Bars represent genomes from Neanderthals, non-African modern humans, and African modern humans. The red regions in the non-African modern human genomes depict genomic regions with introgressed Neanderthal sequence that are present as the result of gene flow (horizontal dotted arrow). Note that the introgressed sequences tend to have an older most recent common ancestor (MRCA) with non-introgressed modern human sequences (the time when the dotted and the solid lineages coalesce) – and hence higher divergence – compared to two modern human sequences (the time when the two solid lineages coalesce). Note also that the introgressed sequences, and thus the resulting regions with higher divergence, tend to be sizable as hybridization occurred relatively recently.

In addition to demonstrating that hybridization occurred and estimating admixture proportions, statistical methods have also been developed to identify sequences inherited from archaic hominin ancestors (also referred to as introgressed sequences) (Sankararaman *et al.*, 2014, 2016; Vernot and Akey, 2014; Vernot *et al.*, 2016). The goal of these methods is to identify DNA sequences that were inherited by ancient gene flow in individual genomes (Figure 9.1). These approaches take advantage of certain population genetics features that introgressed sequences are expected to harbor, such as haplotypes that are unusually divergent and extend over large genomic regions (Figure 9.1). The resulting maps of introgressed sequence hold considerable information about population history and the evolutionary consequences of hybridization (Vernot and Akey, 2014; Gittelman *et al.*, 2016; Juric *et al.*, 2016; Lu *et al.*, 2016; Simonti *et al.*, 2016; Vernot *et al.*, 2016; Dannemann and Kelso, 2017; Martin and Jiggins, 2017; McCoy *et al.*, 2017; Racimo *et al.*, 2017b). For example, studies to identify sequences that modern humans have inherited from Neanderthal and Denisovan ancestors have shown that introgression was a potent source for acquiring beneficial mutations (Vernot and Akey, 2014; Gittelman *et al.*, 2016; Lu *et al.*, 2016; Vernot *et al.*, 2016; Dannemann and Kelso, 2017; Racimo *et al.*, 2017b), was subject to widespread purifying selection (Juric *et al.*, 2016), and continues to have widespread impacts on phenotypic variation and disease susceptibility (Simonti *et al.*, 2016; Dannemann and Kelso, 2017; McCoy *et al.*, 2017). A number of recent reviews on studies of archaic hominin admixture are available for the interested reader (Stoneking and Krause, 2011; Pääbo, 2014; Racimo *et al.*, 2015; Vattathil and Akey, 2015; Nielsen *et al.*, 2017).

In this chapter, we review statistical methods commonly used to test hypotheses about hybridization and identify introgressed sequences. Note that statistical tools to study more recent admixture among present-day populations will be discussed elsewhere in this book (see **Chapter 8**). Here, we focus on approaches that have been developed for studying admixture between genetically divergent lineages. However, some of these methods can be (and have been) used for more recent admixture as well. Furthermore, we focus on approaches that have been used in the context of archaic hominin admixture, but in general such approaches can be used in other organisms as well (Beddows *et al.*, 2017; Svardal *et al.*, 2017; Zeng *et al.*, 2017). We also discuss the advantages and disadvantages of the various statistical tools described and outline important gaps in knowledge and opportunities for future methodological development.

9.2 Methods to Test Hypotheses of Archaic Admixture and Infer Admixture Proportions

In this section, we summarize methods for testing whether archaic admixture occurred and for estimating admixture proportions. These methods are often applied to data sets before specifically identifying introgressed sequences, as they formally test whether there is statistically significant evidence that gene flow has occurred and estimate its magnitude. We primarily focus on a class of methods that are referred to as *D*- and *F*-statistics, and their earlier predecessors (see also Peter, 2016, for an excellent overview). Note that, although we focus specifically on archaic admixture, these methods can be used more generally to test hypotheses of gene flow within species. We start with a brief review of how the evolutionary force referred to as genetic drift leads to allele frequency divergence in genetically structured populations, as *D*- and *F*-statistics can be understood in terms of quantifying levels of shared and lineage-specific amounts of genetic drift among populations.

9.2.1 Genetic Drift and Allele Frequency Divergence in Genetically Structured Populations

Most natural populations exhibit some degree of population structure, which arises when individuals do not mate at random. Many factors can cause departures from random mating, such as inbreeding, assortative mating, and geography. Indeed, many natural populations occupy broad geographic ranges such that random mating is impossible, and, rather than one single panmictic group, the population consists of various demes, or sub-populations. The essential consequence of population structure is that genetic differences can accumulate between sub-populations, due to the stochastic variation in allele frequency (referred to as genetic drift) that occurs from one generation to the next (Masel, 2011). As genetic drift occurs in each sub-population, their allele frequencies diverge, with the rate and magnitude of allele frequency divergence being a function of how long the sub-populations have been separated (more time allows more genetic drift to occur), population sizes (genetic drift happens faster in smaller populations), and how much migration occurs among sub-populations (migration rate is inversely proportional to levels of divergence; Hartl and Clark, 2006).

9.2.2 Three-Population Test

The three-population test (Reich *et al.*, 2009) is a formal test of admixture based on a sample of three populations. Although it was originally developed to detect admixture between populations within a species (Reich *et al.*, 2009), it is conceptually similar to subsequent statistics used to detect gene flow between species. Thus, it is useful to briefly summarize the rationale of the three-population test in order to facilitate understanding of the statistics discussed in this chapter.

Consider the tree shown in Figure 9.2 where R is the ancestral population of populations A , B , and C , and X is the ancestral population of populations A and C . Define a' , b' , c' , x' and r' as the allele frequencies of population A , B , C , X and R , respectively. For simplicity, we will ignore the possibility of recurrent or back mutations (this assumption is approximately true for most genomic sites except for hypermutable sites such as CpGs, which are typically excluded from the analysis). Assuming we know a' , b' and c' , we define the statistic

$$F_3(C; A, B) = E[(c' - a')(c' - b')], \quad (9.1)$$

where $F_3(C; A, B)$ is one type of *F*-statistic (Reich *et al.*, 2009; Patterson *et al.*, 2012). In words, $F_3(C; A, B)$ is the expected product of allele frequency differences between populations C and A

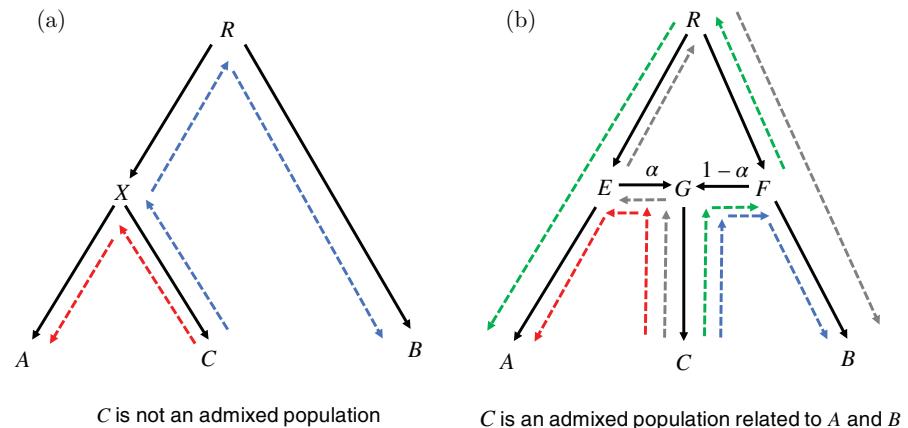


Figure 9.2 Schematic illustration of detecting admixture with the three-population test. In (a), population C is not an admixed population, and red and blue dotted lines represent the drift paths from $C \rightarrow A$ and $C \rightarrow B$, respectively. In (b), population C is descended from admixture between ancestral populations E and F . Red and green dotted lines represent two possible drift paths from $C \rightarrow A$ with probability α and $1 - \alpha$, respectively, whereas blue and gray dotted lines represent two possible drift paths from $C \rightarrow B$ with probability $1 - \alpha$ and α , respectively. Adapted from Patterson *et al.* (2012).

and C and B , which is proportional to the amount of genetic drift that has occurred between C and A and C and B , respectively. If the tree is true and population C is not an admixed population of A and B , then it can be shown (Patterson *et al.*, 2012) that

$$F_3(C; A, B) = E[(c' - a')(c' - b')] = E[(c' - x')^2] \geq 0. \quad (9.2)$$

Similarly it can be shown that $F_3(C; A, B)$ can be negative if C has ancestry from populations related to both A and B .

The intuition behind this statistic can be explained using a concept called a ‘drift path’, representing the magnitude and ‘route(s)’ of drift between two populations. More specifically, F_3 can be thought of as the overlap of the drift path(s) from C to A and the drift path(s) from C to B , which – if the model shown in Figure 9.2(a) (with no admixture) is true – is the genetic drift path between C and X (the overlap of the red and blue dotted lines in Figure 9.2(a)). Thus, if the model in Figure 9.2(a) is correct (i.e. there has not been any admixture), $F_3(C; A, B)$ should always be greater than or equal to 0. However, if population C has ancestry from populations related to both A and B (i.e. there has been admixture), then $F_3(C; A, B)$ can be negative due to drift paths that run in the opposing direction relative to what is expected if the tree topology in Figure 9.2(a) is correct (Figure 9.2(b)). Specifically, as shown in Figure 9.2(b), there are two possible drift paths from $C \rightarrow A$ (red and green dotted lines) and two possible drift paths from $C \rightarrow B$ (blue and gray dotted lines), respectively. The overlap of $C \rightarrow A$ and $C \rightarrow B$ (i.e. $F_3(C; A, B)$) can thus be summarized by four components:

- overlap of red and blue dotted lines: $C - G$ with probability $\alpha(1 - \alpha)$;
- overlap of red and grey dotted lines: $C - G$ plus $G - E$ with probability α^2 ;
- overlap of green and blue dotted lines: $C - G$ plus $G - F$ with probability $(1 - \alpha)^2$;
- overlap of green and grey dotted lines: $C - G$ minus $R - F$ (opposite direction) and $R - E$ (opposite direction) with probability $(1 - \alpha)\alpha$.

Note that the last component contains negative values, which would be the only source contributing to a negative value of $F_3(C; A, B)$.

Importantly, the consequence of these observations is that one can test if C is admixed by estimating $F_3(C; A, B)$ from data and testing if the value is significantly negative. If it is, this is

evidence that admixture occurred in the history of population C (Patterson *et al.*, 2012). This test is referred to as the three-population test. It is interesting to note that the three-population test can also be shown to be related to earlier phylogenetics concepts. Specifically, define $F_2(P_1, P_2)$ as

$$F_2(P_1, P_2) = E(p'_1 - p'_2)^2. \quad (9.3)$$

By performing the Gromov product in phylogenetics (Semple and Steele, 2003), $F_3(C; A, B)$ can also be represented by $F_2(P_1, P_2)$ (Reich *et al.*, 2009):

$$F_3(C; A, B) = \frac{1}{2}(F_2(P_C, P_A) + F_2(P_C, P_B) - F_2(P_A, P_B)). \quad (9.4)$$

In practice, unbiased estimates of $F_3(C; A, B)$ can be obtained based on sample frequencies (given in Patterson *et al.*, 2012), and the test of whether it is a significantly negative value can be performed by estimating standard errors using a resampling approach such as the jackknife procedure (see Reich *et al.*, 2009).

9.2.3 D-Statistic

The *D-statistic*, also referred to as the four-population test or the ABBA-BABA test (an admittedly obscure name that will become clear below), is a test for admixture based on data from four populations and was originally designed to test the hypothesis of gene flow between Neanderthals and modern humans (Reich *et al.*, 2010; Durand *et al.*, 2011). *D*-statistics are typically calculated on genome-scale data (either whole-genome sequences or single nucleotide polymorphism (SNP) chip data), and require data from two sister populations, the hypothesized introgressing population, and an outgroup.

We will first introduce the original version of the *D*-statistic that was developed for the situation where one has sequencing data from a single genome from each population. To understand this *D*-statistic and how it can be used to test for admixture, consider the phylogenetic tree shown in Figure 9.3, and assume we have sampled one individual from each population and at each bi-allelic site with two possible alleles A (ancestral) and B (derived) sample one allele from each individual at random. The null hypothesis that we wish to test is that there has been no gene flow from P_3 (Neanderthal) to P_2 (non-African modern humans), and it is assumed that no admixture has occurred between P_3 (Neanderthal) and P_1 (African modern humans). To simplify interpretation, we only consider sites where the allele sampled from P_4 is the ancestral allele (A), the allele sampled from P_3 is the derived allele (B) and where we have sampled one A allele and one B allele from the two remaining populations. Furthermore, it is also assumed that there are no sequencing errors or recurrent mutation and that populations are randomly mating and of constant size (Green *et al.*, 2010; Durand *et al.*, 2011). What is important to note is that if these assumptions hold, then under the null hypothesis of no admixture (Figure 9.3), we expect to see two site patterns that occur equally frequently:

ABBA: P_2 and P_3 are B, and P_1 and P_4 are A;

BABA: P_1 and P_3 are B, and P_2 and P_4 are A.

The reason why ABBA and BABA are expected to occur with equal frequency under the null hypothesis (no archaic admixture) is that if the null hypothesis is true then in all the sites considered the mutation must have happened in the ancestral population to modern humans and Neanderthals (red stars in Figure 9.3) and the phenomenon known as incomplete lineage sorting (ILS), where gene and species trees are discordant, must have taken place. For example, consider the ABBA pattern shown in Figure 9.3(a). Here, a lineage from population P_2 must have coalesced with a lineage from P_3 first, whereas in the species tree lineages from populations

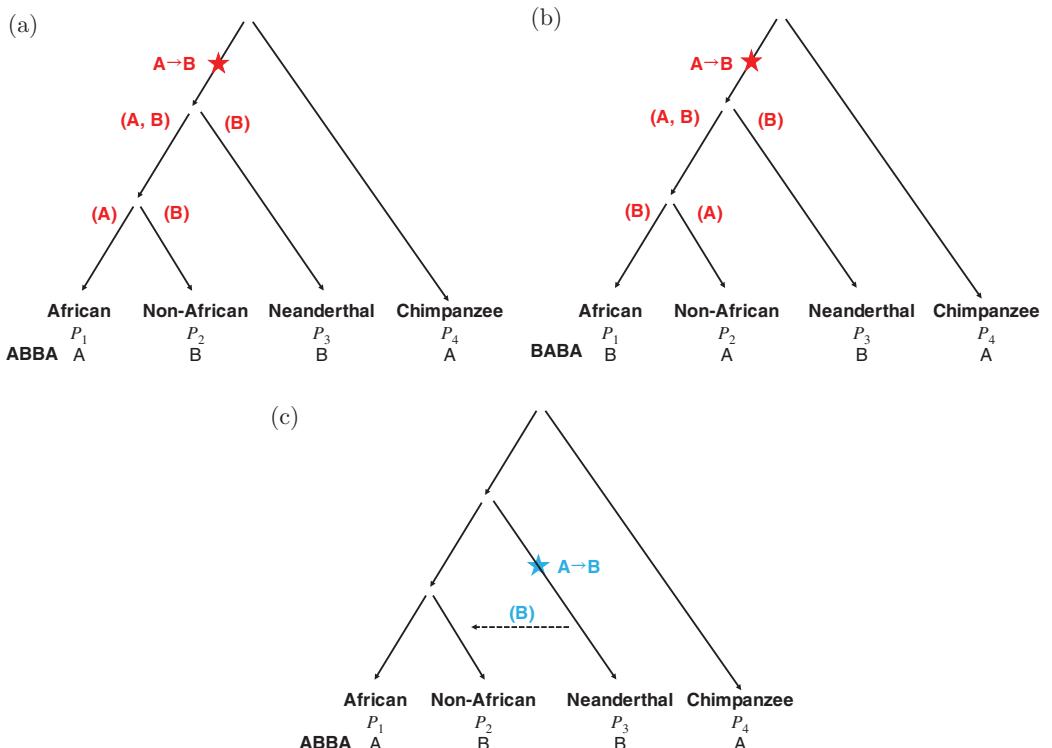


Figure 9.3 A phylogenetic tree to test for Neanderthal admixture in non-African populations. Here we consider a single bi-allelic site where the outgroup (P_4) contains the ancestral (A) allele and the putative introgressing population (P_3) contains the derived (B) allele. The two site patterns of interest are 'ABBA' and 'BABA'. In panels (a) and (b), the mutation occurred in the common ancestral population of P_1 , P_2 , and P_3 (red), and thus because of ILS (see text) the site patterns 'ABBA' and 'BABA' are expected to occur with equal frequency under the null hypothesis of no admixture. In panel (c), the alternative hypothesis of gene flow from P_3 to P_2 is shown with the mutation occurring on the branch leading to P_3 (blue). If gene flow occurs from P_3 to P_2 , it will create an excess of ABBA site patterns.

P_1 and P_2 always coalesce first. Similarly, to get the BABA pattern, a lineage from population P_1 must have coalesced with a lineage from P_3 first (Figure 9.3(b)). Importantly, if the null hypothesis is true these two ILS events are equally likely to occur and thus ABBA and BABA patterns are expected to occur with equal frequency. In contrast, under the alternative hypothesis of gene flow from population P_3 to P_2 there is an excess of the ABBA site patterns compared to the BABA patterns, because P_2 inherits the derived allele B from P_3 (blue star in Figure 9.3(c)).

With these considerations in mind, the D -statistic, summed across n bi-allelic sites, can be defined as

$$D = \frac{\sum_{i=1}^n (C_{ABBA}(i) - C_{BABA}(i))}{\sum_{i=1}^n (C_{ABBA}(i) + C_{BABA}(i))} \quad (9.5)$$

where $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables that take on the values of 0 or 1 if the ABBA or BABA pattern is observed, respectively. Under the null hypothesis of no gene flow, the numerator should be zero, and thus evidence of admixture is tested by asking whether D is significantly different than zero – often by a standard z -test with the standard error estimated by the jackknife procedure (Busing *et al.*, 1999). For example, if D is significantly different from zero for the data shown in Figure 9.3, and the number of ABBA sites is higher than the number

of BABA sites, it would provide evidence of Neanderthal introgression into the non-African population.

Note that we have assumed the effects of sequencing errors are negligible, which is reasonable for high-coverage carefully filtered data from contemporary individuals. However, caution should be taken to mitigate the potential effects of sequencing errors on the D -statistic in ancient DNA samples, which may have higher error rates (Rasmussen *et al.*, 2011). Furthermore, a significant D -statistic can also result from evolutionary forces unrelated to admixture, such as the deviations from the assumed tree topology.

Another useful way of thinking about the D -statistic is through population allele frequencies (Durand *et al.*, 2011). Let p_{i1} , p_{i2} , p_{i3} , and p_{i4} denote frequencies of the A allele at the i th site in populations P_1 (African), P_2 (non-African), P_3 (Neanderthal), and P_4 (chimpanzee), respectively. Then when we sample an allele at this i th site from each population the probabilities of ABBA and BABA pattern are:

$$\Pr(\text{ABBA}) = E[p_{i1}(1-p_{i2})(1-p_{i3})p_{i4}], \quad (9.6)$$

$$\Pr(\text{BABA}) = E[(1-p_{i1})p_{i2}(1-p_{i3})p_{i4}]. \quad (9.7)$$

Consequently, we can also estimate the D -statistic as

$$D(P_1, P_2, P_3, P_4) = \frac{\sum_{i=1}^n [\hat{p}_{i1}(1-\hat{p}_{i2})(1-\hat{p}_{i3})\hat{p}_{i4} - (1-\hat{p}_{i1})\hat{p}_{i2}(1-\hat{p}_{i3})\hat{p}_{i4}]}{\sum_{i=1}^n [\hat{p}_{i1}(1-\hat{p}_{i2})(1-\hat{p}_{i3})\hat{p}_{i4} + (1-\hat{p}_{i1})\hat{p}_{i2}(1-\hat{p}_{i3})\hat{p}_{i4}]}, \quad (9.8)$$

where \hat{p}_{i1} , \hat{p}_{i2} , \hat{p}_{i3} , and \hat{p}_{i4} are unbiased estimates of p_{i1} , p_{i2} , p_{i3} and p_{i4} , respectively. This is useful because such estimates are straightforward to obtain if one has high-quality genotype data from numerous samples from each of the four populations. This allele frequency based version of the D -statistic is therefore often applied to large SNP chip data sets. It also works if you have high-quality genotypes from numerous whole genomes from each population.

When only one or a few low-depth genomes are available from each population the sampling-based version of the D -statistics described above is typically used instead. However, because the sampling-based version of the D -statistic only uses a single sample allele from each population it does not fully exploit all data when there is more than one low-depth sample available from each population. It is therefore worth noticing that a new version of the D -statistic, which allows all data from several low-depth sequenced genome to be used, has been proposed by Soraggi *et al.* (2018). This method considers ABBA patterns to be all patterns where P_1 and P_4 carry one allele and P_2 and P_3 another allele. Similarly, it considers BABA patterns to be all patterns where P_1 and P_3 carry one allele and P_2 and P_4 another allele. When doing this

$$\Pr(\text{ABBA}) = E[p_{i1}(1-p_{i2})(1-p_{i3})p_{i4} + (1-p_{i1})p_{i2}p_{i3}(1-p_{i4})], \quad (9.9)$$

$$\Pr(\text{BABA}) = E[(1-p_{i1})p_{i2}(1-p_{i3})p_{i4} + p_{i1}(1-p_{i2})p_{i3}(1-p_{i4})]. \quad (9.10)$$

Consequently,

$$\Pr(\text{ABBA}) - \Pr(\text{BABA}) = E[(p_{i1} - p_{i2})(p_{i4} - p_{i3})], \quad (9.11)$$

$$\Pr(\text{ABBA}) + \Pr(\text{BABA}) = E[(p_{i1} + p_{i2} - 2p_{i1}p_{i2})(p_{i4} + p_{i3} - 2p_{i3}p_{i4})], \quad (9.12)$$

which leads to the following estimate of the D -statistic:

$$D = \frac{\sum_{i=1}^n (\hat{p}_{i1} - \hat{p}_{i2})(\hat{p}_{i4} - \hat{p}_{i3})}{\sum_{i=1}^n (\hat{p}_{i1} + \hat{p}_{i2} - 2\hat{p}_{i1}\hat{p}_{i2})(\hat{p}_{i3} + \hat{p}_{i4} - 2\hat{p}_{i3}\hat{p}_{i4})}, \quad (9.13)$$

where \hat{p}_{i1} , \hat{p}_{i2} , \hat{p}_{i3} , and \hat{p}_{i4} again are unbiased estimators of p_{i1} , p_{i2} , p_{i3} and p_{i4} , respectively. Soraggi *et al.* use this formula to estimate D with the unbiased estimate of p_{ij} given as

$$\hat{p}_{ijl} \times \sum_{l=1}^{N_j} w_{ijl}, \quad (9.14)$$

where $j = 1, 2, 3, 4, N_j$ is the number of individuals in population j , \hat{p}_{ijl} is the estimated allele frequency of the A allele at site i for individual l in population j , and w_{ijl} is a weight defined as

$$w_{ijl} \propto \frac{2n_{ijl}}{n_{ijl} + 1}, \quad (9.15)$$

where n_{ijl} is the number of base pair observed in data (i.e. base depth or coverage). This estimator takes uneven depths of different genomes into account, and Soraggi *et al.* show that this is the estimator for the j th population frequency at locus i with minimal variance. Furthermore, they show that when more than one sample is available from each population their version of the D -statistic is more powerful than the version based on a single sampled base from each population. Hence this may become a useful tool in the future.

9.2.4 F_4 -Statistic

Similar to the D -statistic, F_4 -statistics (Reich *et al.* 2009; Patterson *et al.*, 2012) typically are applied to genome-scale data and require an outgroup, but in addition also use data from a taxon closely related to the introgressing group. As shown below, F_4 -statistics allow one to infer the admixture proportion (typically denoted as α). For example, consider the tree shown in Figure 9.4(a), where we want to estimate the proportion of P_3 (Neanderthal) ancestry, α , in the

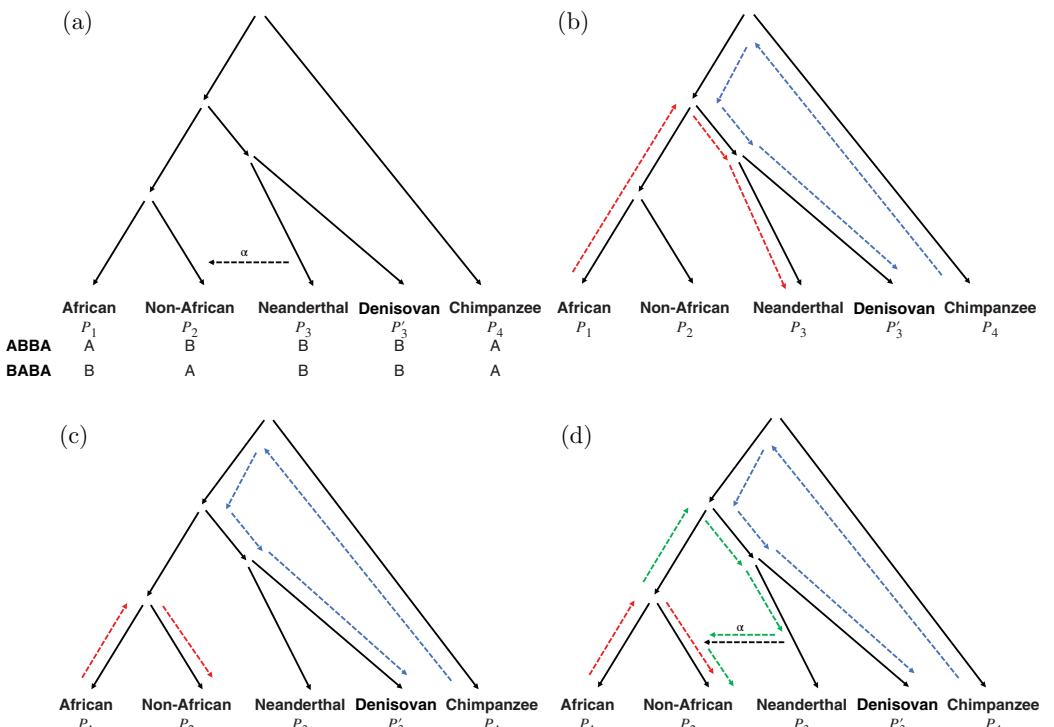


Figure 9.4 Schematic representation showing that population P_2 (non-African) derives a certain amount of ancestry (α) from Neanderthals due to hybridization. (a) Interpretation based on ABBA and BABA patterns, (b) interpretation of $F_4(P_1, P_3; P_4, P'_3)$ based on overlap of genetic drift paths, no admixture model, (c) interpretation of $F_4(P_1, P_2; P_4, P'_3)$ based on overlap of genetic drift paths, no admixture model, (d) interpretation of $F_4(P_1, P_2; P_4, P'_3)$ based on overlap of genetic drift paths, admixture model.

(non-African) population P_2 . We will continue to use the notation introduced above, but now define another archaic population (such as Denisovans) as P'_3 , and its frequency at a bi-allelic locus as p'_3 . Using this notation, we define the F_4 -statistics as $F_4(P_X, P_Y, P_Z, P_W) = E(p_x - p_y)(p_z - p_w)$ for any populations X, Y, Z and W . This is useful because if gene flow occurred from P_3 to P_2 as showed in Figure 9.4(a) then

$$p_2 = \alpha p_3 + (1 - \alpha)p_1, \quad (9.16)$$

which means that

$$F_4(P_1, P_2; P_4, P'_3) = E[(p_1 - p_2)(p_4 - p'_3)] = \alpha E[(p_1 - p_3)(p_4 - p'_3)] = \alpha F_4(P_1, P_3; P_4, P'_3).$$

Consequently, we can estimate α by the following ratio of two F_4 -statistics:

$$\alpha = \frac{F(P_1, P_2; P_4, P'_3)}{F(P_1, P_3; P_4, P'_3)}. \quad (9.17)$$

One way to interpret F_4 is to note that $F_4(P_1, P_3; P_4, P'_3)$ is the numerator of the D -statistic $D(P_1, P_3, P'_3, P_4)$ defined in equation (9.5) and is thus the expected difference in ABBA and BABA patterns for the populations P_1, P_3, P'_3 and P_4 (equation (9.11)). Similarly, $F_4(P_1, P_2; P_4, P'_3)$ is the expected difference in ABBA and BABA patterns for the populations P_1, P_2, P'_3 and P_4 (Figure 9.4(a)).

In Patterson *et al.* (2012), $F_4(P_1, P_3; P_4, P'_3)$ was interpreted as the overlap of genetic drift paths $P_1 \rightarrow P_3$ and $P_4 \rightarrow P'_3$, which can also be represented as the drift path between the common ancestor of P_1 and P_3 and the common ancestor of P_3 and P'_3 (Figure 9.4(b)) (Peter, 2016). Similarly, if there is no admixture $F_4(P_1, P_2; P_4, P'_3)$ represents the overlap of genetic drift paths $P_1 \rightarrow P_2$ (red dotted line in Figure 9.4(c)) and $P_4 \rightarrow P'_3$ (blue dotted line in Figure 9.4(c)). And since there is no overlap between $P_1 \rightarrow P_2$ and $P_4 \rightarrow P'_3$, it follows that $F_4(P_1, P_2; P_4, P'_3) = 0$.

Conversely, if there has been admixture and thus $\alpha \neq 0$ (Figure 9.4(d)) there are two possible paths for $P_1 \rightarrow P_2$ (red and green dotted lines in Figure 9.4(d)), and one of these – the path following $P_1 \rightarrow P_3 \rightarrow P_2$ (green dotted line in Figure 9.4(d)), which has probability α – overlaps with the genetic drift path $P_4 \rightarrow P'_3$. Hence, if there is admixture, $F_4(P_1, P_2; P_4, P'_3)$ is not 0 but instead α times this overlap. And importantly, this overlap is exactly the same as the overlap of the genetic drift paths $P_1 \rightarrow P_3$ and $P_4 \rightarrow P'_3$, so $F_4(P_1, P_3; P_4, P'_3)$ is equal to $\alpha F_4(P_1, P_2; P_4, P'_3)$, making it intuitively clear why α can be estimated as the ratio of these to F_4 -statistics. An alternative interpretation is that α measures how much closer P_3 and P_2 (Neanderthal and non-African) are to the common ancestor relative to the common ancestor of P'_3 and P_3 (Denisovan and Neanderthal) (Peter, 2016). Readers interested in more technical details about the F_4 -statistic can find them in Peter (2016).

9.3 Methods to Identify Introgressed Sequences

In this section we will describe statistical approaches that have been developed to identify introgressed sequences (i.e. sequences inherited from an archaic ancestor through gene flow). In general, two main classes of approaches have been described to date. The first uses a summary statistic designed to capture the expected population genetics features of introgressed sequences and has the advantage that it can be used without explicit information from an archaic reference genome. The second class of approaches are based on rigorous probabilistic models that explicitly use an archaic reference genome. Below we describe each class of approaches and discuss their relative advantages and disadvantages. It is important to note that

although *D*-statistics have been proposed as a means to scan along a genome and identify introgressed sequences, this approach is not advised as it is subject to high false positive rates. The high rates of false positives are a consequence of *D*-statistics being designed to detect deviations from the aggregate composition of ABBA and BABA site patterns across many loci, and thus the observation of a site pattern at any particular locus is difficult to interpret (Martin *et al.*, 2015). For these reasons, we recommend using the methods presented in this section to identify introgressed sequences at particular loci.

9.3.1 *S**-Statistic

A variety of methods based on the *S**-statistic are used to identify archaic introgressed segments using only contemporary human DNA sequences. A number of different implementations of *S** (Plagnol and Wall, 2006; Vernot and Akey, 2014; Vernot *et al.*, 2016; Browning *et al.*, 2018) have been described but are closely related, as we will see below. In particular, high values of *S** capture population genetic signatures expected of introgressed sequences (Figure 9.1). Specifically, *S** was designed to identify divergent haplotypes that extend over sizable genomic regions. These population genetic features are due to the fact that the time to the most recent common ancestor (TMRCA) between an introgressed and non-introgressed haplotype (i.e. Neanderthal versus modern human) is on average much longer than the TMRCA between two non-introgressed haplotypes at a locus. Thus, mutations have more time to accumulate resulting in higher divergence. Moreover, assuming hybridization happened relatively recently – for instance, admixture between modern and archaic humans occurred ~50–76 kya (Sankararaman *et al.*, 2012; Fu *et al.*, 2014; Seguin-Orlando *et al.*, 2014; Yang and Fu, 2018) – introgressed haplotypes are expected to extend ~40–50 kb (Plagnol and Wall, 2006; Vernot and Akey, 2014; Vernot *et al.*, 2016). Therefore, introgressed segments can be distinguished from other divergent sequences that extend only short distances, which can be present simply because of the considerable variation in TMRCA expected along the genome. Readers interested in more details about the expected length of archaic sequences from different models can find them elsewhere (Plagnol and Wall, 2006; Racimo *et al.*, 2015).

In the following, we will describe three different implementations of the *S** (Vernot and Akey, 2014; Vernot *et al.*, 2016; Browning *et al.*, 2018). Let us define a ‘target population’ and a ‘reference population’. The target population is the population that we hypothesize segregates introgressed segments and the reference population is a population we expect not to have experienced hybridization (i.e. is non-admixed), used to mitigate spurious signals of introgression due to ancestral polymorphism shared between the target and reference populations and ILS (Plagnol and Wall, 2006). In applications to archaic hominin admixture, where we are interested in inferring segments of Neanderthal or Denisovan introgressed into non-Africans, we would use a non-African population as the target populations and an African population as a reference population (typically Yoruba, as they have the smallest estimated levels of Neanderthal ancestry, although what constitutes the best reference population remains an open question).

Before *S** is calculated, variants shared between the target and reference population are removed to mitigate the effects of ILS as noted above (i.e. by removing potential sites that represent ancestral polymorphism and thus subject to ILS). Formally, define f as the derived allele frequency in the reference population. Variants with a derived allele frequency $f \geq x$ in the target population are removed from the reference population. In some analyses, x is set to zero (Vernot and Akey, 2014; Vernot *et al.*, 2016) – that is, all shared variants are removed – whereas in others it is set to a small number such as 0.01 (Browning *et al.*, 2018). As recent gene flow likely introduced some Neanderthal variants back into African populations (Vernot *et al.*, 2016),

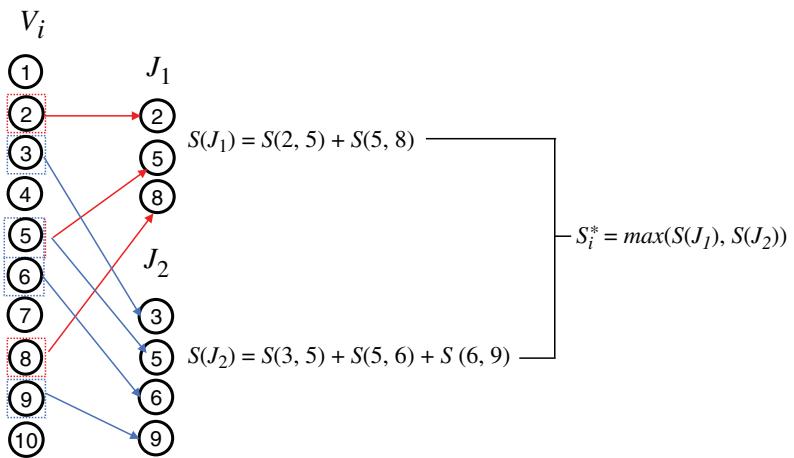


Figure 9.5 Schematic illustration of calculating S_i^* for individual i . In this example, V_i contains ten variants numbered 1 to 10. J_1 (red) and J_2 (blue) are two subsets of V_i , containing three and four variants, respectively. If we for simplicity ignore all other possible subsets of V_i , then S_i^* is the maximum value of $S(J_1)$ and $S(J_2)$.

there is a tradeoff in sensitivity and specificity when choosing x that has not been extensively evaluated.

Given a genomic region (e.g. a 50 kb window) and for the i th individual in the target population, the statistic is defined as $S_i^* = \max_{J \subseteq V_i} S(J)$, where

$$S(J) = \sum_{j=1}^{\max(J)-1} S(j, j+1). \quad (9.18)$$

Here V_i is the set of all variants in this genomic region, J is a subset of V_i , $S(j, j+1)$ is the pairwise score function of two adjacent variants j and $j+1$ in J , and $\max(J)$ represents the last variant in J . The task is to find some subset of variants (J) from all variants (V_i) that maximizes $S(J)$. The variants of J that maximize $S(J)$ will be those that exhibit the strongest linkage disequilibrium (LD) that persists over the largest distance (measured in base pairs). As we will see below, the score for adjacent variants, $S(j, j+1)$, will be positive if they share the same genotypes and will be larger the greater the distance is between them. A hypothetical example shown in Figure 9.5 illustrates the process of calculating S_i^* . V_i includes ten variants in this genomic region. Suppose that we select two sets of variants from V_i , defined as J_1 and J_2 , which include three and four variants, respectively (note that there are more than two subsets that can be selected, but for simplicity, here we only show two subsets). According to equation (9.18), we calculate $S(J_1)$ and $S(J_2)$, and then S_i^* is equal to the maximum value of $S(J_1)$ and $S(J_2)$.

The major difference among S^* -statistics are the definition of the pairwise score function $S(j, j+1)$. For example, in Vernot and Akey (2014) the scoring function is

$$S(j, j+1) = \begin{cases} -\infty, & d(j, j+1) > 5, \\ -10000, & d(j, j+1) \in \{1, \dots, 5\}, \\ 5000 + bp(j, j+1), & d(j, j+1) = 0, \\ 0, & j = \max(J), \end{cases} \quad (9.19)$$

where $bp(j, j+1)$ is the distance in base pairs between two variants j and $j+1$, whereas $d(j, j+1)$ is the genotype difference between these two variants, which is defined as

$$d(j, j+1) = \sum_i |GT(i, j) - GT(i, j+1)|, \quad (9.20)$$

where $GT(i, j)$ and $GT(i, j + 1)$ are the genotype values of the variants j and $j + 1$ for individual i . The genotype values are coded as 0, 1 and 2 and the sum runs over all individuals. According to this equation, if there is no genotype difference between j and $j + 1$ (two variants are congruent sites and in perfect LD), the score is 5000 plus the distance in base pairs between them, which will give a higher score if this distance in base pairs is larger. If the genotype difference is larger than 0 but smaller than 6, a penalty of $-10,000$ will be scored. $-\infty$ means that it does not allow consecutive variants in J to have more than five genotype differences. A smaller genotype difference represents stronger LD. Again, the term $\max(J)$ represents the last variant in J .

This scoring function works best for small sample sizes (which is a consequence of the allowed genotype differences), and in Vernot and Akey (2014) they analyzed subsets of 20 individuals at a time. Subsequent extensions of S^* proposed a way to circumvent this difficulty by analyzing individuals one at a time and not summarizing the genotypic differences of other individuals in the target population (Vernot *et al.*, 2016). Specifically, for individual i , the score $S(j, j + 1)$ is simplified to

$$S(j, j + 1) = \begin{cases} -10000, & d(j, j + 1) > 0, \\ 5000 + bp(j, j + 1), & d(j, j + 1) = 0, \\ 0, & j = \max(J), \end{cases} \quad (9.21)$$

where $bp(j, j + 1)$ is the same as the 2014 version, but $d(j, j + 1)$ is changed as

$$d(j, j + 1) = |GT(i, j) - GT(i, j + 1)|, \quad (9.22)$$

which means $d(j, j + 1)$ is equal to the absolute genotype difference between two variants for individual i .

A more recently developed version of S^* has been proposed called Sprime (Browning *et al.*, 2018). It is conceptually similar to previous S^* -statistics, but uses a different pairwise score function to enable analysis of arbitrarily large target samples. In addition, it incorporates information for local mutation (Scally and Durbin, 2012) and recombination rates (International HapMap Consortium, 2007) into the model.

Consider a pair of variants j and $j + 1$. For individual i in the target population, the definition of $d(j, j + 1)$ is the same as in the 2014 version of S^* , except it is summarized by genotype differences between j and $j + 1$ of all target samples. We also denote the value of $d(j, j + 1)$ as 0, 1 and 2 for homozygous ancestral, heterozygous and homozygous derived allele, respectively.

The score function is defined as

$$S(j, j + 1) = 6000w \frac{1 - \exp(-0.01/m)}{1 - \exp(-1)} - 25000 \frac{d(j, j + 1)}{n}, \quad (9.23)$$

where w represents different weights for rare and absent alleles in the reference population such that

$$w = \begin{cases} 1, & f = 0, \\ 0.8, & 0 < f \leq 0.01, \end{cases} \quad (9.24)$$

n is the smaller number of $\sum_i GT(i, j)$ and $\sum_i GT(i, j + 1)$, which can be written as

$$n = \min \left(\sum_i GT(i, j), \sum_i GT(i, j + 1) \right) \quad (9.25)$$

m represents the local rate of mutation per centimorgan per meiosis, and can be estimated as:

$$\text{local rate of mutations per bp per meiosis} = \frac{\text{local variant density}}{\text{global variant density}} \times u_g, \quad (9.26)$$

where u_g is the average mutation rate on genome (Scally and Durbin, 2012), and the local and global variant densities are estimated by the number of variant positions divided by the number of base pairs in the region.

Note that the above score function contains a positive term, which is dependent on the local mutation rate, and a negative term that penalizes large values of $d(j, j+1)/n$.

In practice, S^* is typically calculated within sliding windows along the genome (such as a 50 kb window size with a 10 kb step size) using a dynamic programming algorithm; see Vernot and Akey (2014) for a worked example and Plagnol and Wall (2006), Vernot and Akey (2014), Vernot *et al.* (2016) and Browning *et al.* (2018) for general background information on dynamic programming). To determine if a haplotype within a window is introgressed, we evaluate its statistical significance by comparing the observed S^* value estimated in this window to the expected S^* distribution under the null hypothesis through simulations of a demographic model that does not include hybridization (Vernot and Akey, 2014; Hsieh *et al.*, 2016; Vernot *et al.*, 2016). Ideally, a well-calibrated demographic model is available for the population under study and local rates of mutation and recombination are accounted for as these influence the probability density function of a given S^* metric under the null hypothesis. Furthermore, although S^* itself does not need genomic information from the hypothesized introgressing species, when available such data can be used to calibrate false discovery rates and refine the set of putatively introgressed sequences. For example, calculating how well the set of S^* significant sequences match an archaic reference genome relative to what is expected by chance has proven an efficient way to enrich for true positives (Vernot and Akey, 2014; Vernot *et al.*, 2016). Finally, it is worth noting that recent extensions of S^* allow it to be calculated using a novel tiling strategy that avoids analyses focused on fixed window sizes (Browning *et al.*, 2018).

9.3.2 Hidden Markov and Conditional Random Field Models

Approaches based on hidden Markov models (HMMs) or conditional random fields (CRFs) have also been developed to detect introgressed sequences. These models are both formal probabilistic models (Prüfer *et al.*, 2014; Seguin-Orlando *et al.*, 2014; Sankararaman *et al.*, 2016; Racimo *et al.*, 2017a; Steinrücken *et al.*, 2018; for more information on HMMs and inference methods for such models, see **Chapter 1**, this volume). Unlike S^* -based methods, HMMs and CRFs directly use genomic information from the introgressing species. In particular, they compare a haplotype from an individual we want to test for introgression sampled from the target population (i.e. a population that potentially experienced admixture) with archaic genomes of interest (such as Neanderthal or Denisovan), and use a reference population (i.e. a population that is hypothesized not to have experienced admixture) to infer the hidden state of each allele (i.e. archaic or modern).

More specifically, they base their inference on ‘observed’ values defined as follows. Consider a genomic region and denote an archaic haplotype as A , a test haplotype from the target population as T , and a reference haplotype as Y . For a site within T , define f_T , f_A , and f_Y as the frequency of the derived allele in test haplotype T , archaic haplotype A , and reference haplotype Y , respectively. Note that $0 \leq f_A \leq 1$, $0 \leq f_Y \leq 1$, and $f_T = 0$ or $f_T = 1$, depending on whether

the allele in the test haplotype is ancestral or derived. Based on f_T , f_A and f_Y , we can define a site in the test haplotype as consistent or inconsistent with being an introgressed sequence:

$$\begin{aligned} \text{consistent, } f_T - f_Y = 1 \text{ and } |f_A - f_T| < 1, \\ \text{inconsistent, } 0 < f_Y < 1 \text{ and } |f_A - f_T| = 1. \end{aligned}$$

Notably, haplotypes that are consistent (C) indicate that target haplotypes exhibit higher sequence similarity to archaic haplotypes, whereas inconsistent haplotypes (N) indicate that the target and reference haplotypes are closer. Based on these 'observed' values of C and N for the test haplotype at a set of variant sites, the task of the HMM or CRF is to walk along a set of variant sites and infer their hidden state; specifically, whether they are introgressed from an archaic ancestor or not.

A schematic of how an HMM or CRF works to detect introgressed sequences is shown in Figure 9.6. In HMMs, the edges represent emission probabilities (vertical) and transition probabilities (horizontal) for the HMM, but in CRFs represent emission functions (vertical and diagonal) and transition functions (horizontal) (Racimo *et al.*, 2015). Transition probabilities (HMMs) and functions (CRFs) model linkage between sites and emission probabilities (HMMs) and functions (CRFs) link the states and observations. The major difference between HMMs

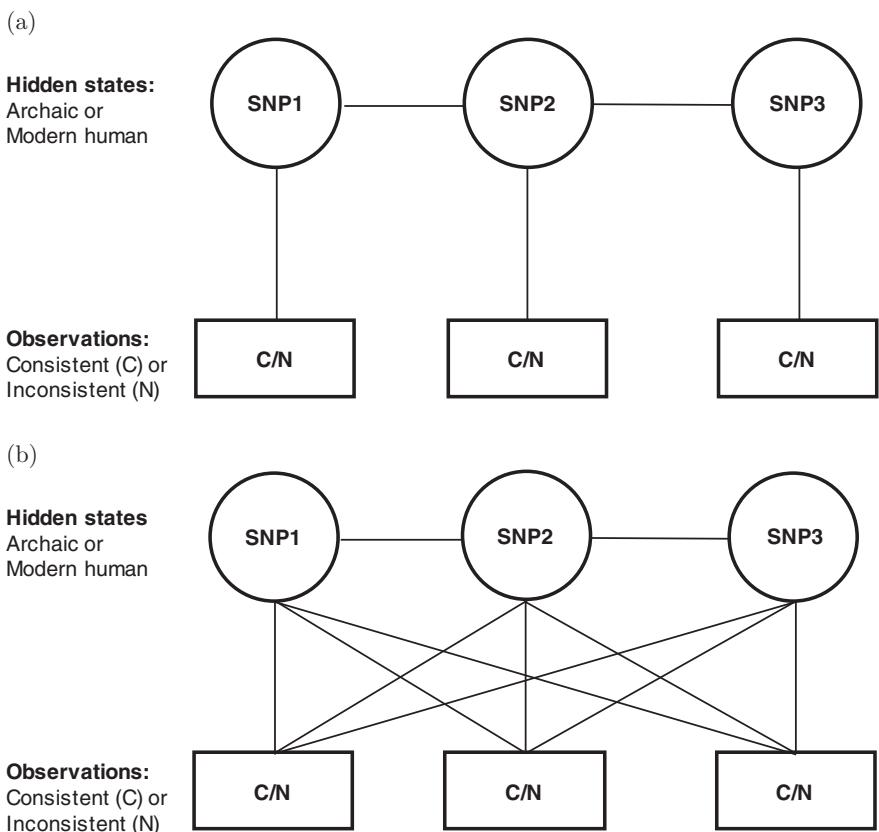


Figure 9.6 Schematic illustration of the probabilistic models developed for identifying introgressed haplotypes. (a) HMM: vertical lines represent emission probabilities, and horizontal lines represent transition probabilities (b) CRF: vertical and diagonal lines represent emission functions, and horizontal lines represent transition functions.

and CRFs is that the emission probabilities in HMMs only depend on the current observations (vertical edges), whereas in CRFs the emission functions also depend on past and future observations (all edges) and would avoid interpreting dependence between observations due to LD. More technical details about the similarities and differences between HMMs and CRFs can be found elsewhere (Lafferty *et al.*, 2001).

So far four HMM models have been proposed, one that used fixed emission probabilities which were chosen *a priori* (Prüfer *et al.*, 2014), and the others where the emission probabilities were estimated from the observed data (Seguin-Orlando *et al.*, 2014; Racimo *et al.*, 2017a; Steinrücken *et al.*, 2018). In comparison one CFR model has been proposed and it used two classes of emission functions: the first one captures information from a single SNP's allelic patterns, and the second uses multiple SNPs to capture the signal of archaic ancestry (Sankararaman *et al.*, 2016).

9.3.3 Relative Advantages and Disadvantages of Approaches to Detect Introgressed Sequences

To our knowledge, the relative power and false positive rates of S^* , HMM, and CRF models to detect introgressed archaic hominin sequences has not been directly compared. Nonetheless, it is possible to make several general inferences on the relative advantages and disadvantages of S^* , HMM, and CRF models. For example, as noted above, a strength of S^* is that it does not need data from an archaic reference genome, and in principle it can be used to search for introgressed sequences from unknown archaic lineages (Hammer *et al.*, 2011; Lachance *et al.*, 2012; Hsieh *et al.*, 2016). However, it should be noted that the null distribution of S^* is a function of demographic history (Vernot and Akey, 2014) and therefore care should be taken in applications to populations that do not have well-calibrated demographic histories. Furthermore, in practice, Neanderthal and Denisovan genomes have been used to refine sets of S^* significant sequences to reduce false positive rates (Vernot and Akey, 2014; Vernot *et al.*, 2016). In contrast, as the HMM and CRF approaches described to date explicitly use archaic reference genomes, it is reasonable to hypothesize that they are more powerful than S^* as they use more available information. Finally, a weakness of all of the methods described in this chapter is the use of a reference population that is hypothesized not to be admixed with an archaic group. In the case of identifying Neanderthal and Denisovan sequences, this is not especially problematic since reasonable reference populations exist. However, this may not always be the case in studies searching for unknown lineages. Moreover, the concept of reference populations may be problematic when applying these methods to non-humans, where such clear hypotheses on which groups have and have not admixed may not be available.

9.4 Summary and Perspective

In this chapter we have summarized the statistical tools that have been used to study archaic hominin admixture. These methods range from testing hypotheses about whether admixture occurred and estimating the admixture proportion (Section 9.2) to statistical approaches that can identify introgressed sequences (Section 9.3). Although these tools have enabled dramatic insights into human evolutionary history, a number of statistical issues remain. For example, as discussed above, a more rigorous statistical comparison of methods to detect introgressed sequences is necessary. In addition, technical issues such as how robust inferences are to sequencing depth and data filtering strategies should also be further explored. It would also be of considerable interest to better characterize how accurately these methods delimit break-points

between introgressed and non-introgressed sequences, as the length distribution of archaic sequences may provide considerable information for population genetics inferences (Gravel, 2012; Sankararaman *et al.*, 2012; Liang and Nielsen, 2014). Moreover, although the methods to identify introgressed sequences described here can be adapted to organisms beyond humans, they should be carefully evaluated in the context of each species to ensure robust inferences are being made.

The ability to build catalogs of introgressed sequences marks an exciting phase for population genetics, as new methodological advances are now needed to capitalize on this wealth of information. To date, population genetics inferences have largely focused on analyzing genome-wide patterns of introgressed sequences, revealing, for instance, the likely contribution of positive and negative selection (Vernot and Akey, 2014; Juric *et al.*, 2016; Sankararaman *et al.*, 2016; Vernot *et al.*, 2016) acting on them. Notable exceptions are an approximate Bayesian model approach that leveraged levels of Neanderthal ancestry among modern populations (Vernot and Akey, 2014) and a probabilistic method based on reciprocal patterns of Neanderthal allele sharing (Vernot *et al.*, 2016) to infer parameters of archaic admixture models. However, a well-defined set of archaic sequences holds a wealth of potential information, including information about the number of archaic ancestors, the number of hybridization events, effective population size of the introgressing population, and whether sex-biased admixture occurred, to name a few. A challenge in developing population genetics methods in this context is that they should account for power and false discovery rates in calls of introgressed sequences, which will influence the observed site frequency spectrum of archaic alleles.

Despite these limitations and challenges, we are optimistic that population genomics will rise to the challenge and develop powerful new inference methods to extract information about population history and evolution from catalogs of introgressed sequences. As ancient DNA becomes more abundant, such methods will enable a deeper understanding of not only human history, but also the frequency and evolutionary consequences of hybridization in the wild (Payseur and Rieseberg, 2016; Martin and Jiggins, 2017).

References

- Beddows, I., Reddy, A., Kloesges, T. and Rose, L.E. (2017). Population genomics in wild tomatoes – the interplay of divergence and admixture. *Genome Biology and Evolution* **9**(11), 3023–3038.
- Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S. and Akey, J.M. (2018). Analysis of human sequence data reveals two pulses of archaic admixture from Denisovans. *Cell* **173**, 53–61.
- Busing, F.M.T.A., Meijer, E. and van der Leeden, R. (1999). Delete-m jackknife for unequal m. *Statistics and Computing* **9**(1), 3–8.
- Dannemann, M. and Kelso, J. (2017). The contribution of Neanderthals to phenotypic variation in modern humans. *American Journal of Human Genetics* **101**(4), 578–589.
- Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239–2252.
- Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Petri, A.A., Prüfer, K., de Filippo, C., Meyer, M., Zwyns, N., Salazar-Garcia, D.C., Kuzmin, Y.V., Keates, S.G., Kosintsev, P.A., Razhev, D.I., Richards, M.P., Peristov, N.V., Lachmann, M., Douka, K., Higham, T.F.G., Slatkin, M., Hublin, J.-J., Reich, D., Kelso, J., Viola, T.B. and Pääbo, S. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**(7523), 445–449.
- Gittelman, R.M., Schraiber, J.G., Vernot, B., Mikacenic, C., Wurfel, M.M. and Akey, J.M. (2016). Archaic hominin admixture facilitated adaptation to out-of-Africa environments. *Current Biology* **26**(24), 3375–3382.

- Gravel, S. (2012). Population genetics models of local ancestry. *Genetics* **191**(2), 607–619.
- Grealy, A., Rawlence, N.J. and Bunce, M. (2017). Time to spread your wings: A review of the avian ancient DNA field. *Genes* **8**(7), 184.
- Green, E.J. and Speller, C.F. (2017). Novel substrates as sources of ancient DNA: Prospects and hurdles. *Genes* **8**(7), 180.
- Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Ž. Kucan, Gušić, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science* **328**(5979), 710–722.
- Hammer, M.F., Woerner, A.E., Mendez, F.L., Watkins, J.C. and Wall, J.D. (2011). Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **108**(37), 15123–15128.
- Hartl, D.L. and Clark, H.G. (2006). *Principles of Population Genetics*, 4th edition. Sinauer Associates, Sunderland, MA.
- Hsieh, P.H., Woerner, A.E., Wall, J.D., Lachance, J., Tishkoff, S.A., Gutenkunst, R.N. and Hammer, M.F. (2016). Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Research* **26**(3), 291–300.
- International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.
- Juric, I., Aeschbacher, S. and Coop, G. (2016). The strength of selection against Neanderthal introgression. *PLOS Genetics* **12**(11), e1006340.
- Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., Zhang, K., Akey, J.M. and Tishkoff, S.A. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**(3), 457–469.
- Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. Kaufmann, San Francisco.
- Liang, M. and Nielsen, R. (2014). The lengths of admixture tracts. *Genetics* **197**(3), 953–967.
- Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., Feng, Q., Hu, Y., Ding, Q., Yang, Y., Li, S., Jin, L., Guan, Y., Su, B., Kang, L. and Xu, S. (2016) Ancestral origins and genetic history of Tibetan highlanders. *American Journal of Human Genetics* **99**(3), 580–594.
- Marciniak, S. and Perry, G. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* **18**, 659–674.
- Martin, S.H. and Jiggins, C.D. (2017). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics and Development* **47**, 69–74.
- Martin, S.H., Davey, J.W. and Jiggins, C.D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* **32**, 244–257.
- Masel, J. (2011). Genetic drift. *Current Biology* **21**, R837–R838.
- McCoy, R.C., Wakefield, J. and Akey, J.M. (2017). Impacts of Neanderthal-introgressed sequences on the landscape of human gene expression. *Cell* **168**(5), 916–927.
- Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A.,

- Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andrés, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226.
- Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature* **541**, 302–310.
- Pääbo, S. (2014). The human condition – a molecular approach. *Cell* **157**, 216–226.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. (2012). Ancient admixture in human history. *Genetics* **92**, 1065–1093.
- Payseur, B.A. and Rieseberg, L.H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology* **25**(11), 2337–2360.
- Peter, B.M. (2016). Admixture, population structure, and *F*-statistics. *Genetics* **202**, 1485–1501.
- Plagnol, V. and Wall, J.D. (2006). Possible ancestral structure in human populations. *PLoS Genetics* **2**(7), e105.
- Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49.
- Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., Dannemann, M., Nelson, B.J., Key, F.M., Rudan, P., Ž. Kućan, Gušić, I., Golovanova, L.V., Doronichev, V.B., Patterson, N., Reich, D., Eichler, E.E., Slatkin, M., Schierup, M.H., Andrés, A.M., Kelso, J., Meyer, M. and Pääbo, S. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658.
- Racimo, F., Sankararaman, S., Nielsen, R. and Huerta-Sánchez, E. (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics* **16**, 359–371.
- Racimo, F., Gokhman, D., Fumagalli, M., Ko, A., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E. and Nielsen, R. (2017a). Archaic adaptive introgression in TBX15/WARS2. *Molecular Biology and Evolution* **34**, 509–524.
- Racimo, F., Marnetto, D. and Huerta-Sánchez, E. (2017b). Signatures of archaic adaptive introgression in present-day human populations. *Molecular Biology and Evolution* **34**, 296–317.
- Rasmussen, M., Guo, X. and Wang, Y. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98.
- Reich, D., Thangaraj, K., Patterson, N., Price, A.L. and Singh, L. (2009). Reconstructing Indian population history. *Nature* **461**, 489–494. 2009.
- Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M. and Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060.
- Sankararaman, S., Patterson, N., Li, H., Pääbo, S. and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* **8**(10), e1002947.
- Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**(7492), 354–357.

- Sankararaman, S., Mallick, S., Patterson, N. and Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology* **26**(9), 1241–1247.
- Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**, 745–753.
- Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspina, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J.D., Nigst, P.R., Foley, R.A., Mirazon Lahr, M., Nielsen, R., Orlando, L. and Willerslev, E. (2014). Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**(6213), 1113–1118.
- Semple, C. and Steele, M.A. (2003). *Phylogenetics*. Oxford University Press, Oxford.
- Simonti, C.N., Vernot, B., Bastarache, L., Bottinger, E., Carrell, D.S., Chisholm, R.L., Crosslin, D.R., Hebbright, S.J., Jarvik, G.P., Kullo, I.J., Li, R., Pathak, J., Ritchie, M.D., Roden, D.M., Verma, S.S., Tromp, G., Prato, J.D., Bush, W.S., Akey, J.M., Denny, J.C. and Capra, J.A. (2016). The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**(6274), 737–741.
- Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America* **113**(23), 6380–6387.
- Soraggi, S., Wiuf, C. and Albrechtsen, A. (2018). Powerful inference with the D-statistic on low-coverage whole-genome data. *G3: Genes, Genomes, Genetics* **8**(2), 551–556.
- Steinrücken, M., Spence, J.P., Kamm, J.A., Wieczorek, E. and Song, Y.S. (2018). Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology* **27**(19), 3873–3888.
- Stoneking, M. and Krause, J. (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics* **12**, 603–614.
- Svardal, H., Jasinska, A.J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C.A., Jacquelin, B., Ramensky, V., Müller-Trutwin, M., Antonio, M., Weinstock, G., Grobler, J.P., Dewar, K., Wilson, R.K., Turner, T.R., Warren, W.C., Freimer, N.B. and Nordborg, M. (2017). Ancient hybridization and strong adaptation to viruses across African velvet monkey populations. *Nature Genetics* **49**, 1705–1713.
- Vattathil, S. and Akey, J.M. (2015). Small amounts of archaic admixture provide big insights into human history. *Cell* **163**, 281–284.
- Vernot, B. and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**(6174), 1017–1021.
- Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., Scheinfeldt, L.B., Merriwether, D.A., Koki, G., Friedlaender, J.S., Wakefield, J., Pääbo, S. and Akey, J.M. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**(6282), 235–239.
- Warinner, C., Herbig, A., Mann, A., Fellows Yates, J.A., Weiß, C.L., Burbano, H.A., Orlando, L. and Krause, J. (2017). A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics* **18**, 321–356.
- Xu, S. and Jin, L. (2008). A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *American Journal of Human Genetics* **83**, 322–336.
- Yang, M.A. and Fu, Q. (2018). Insights into modern human prehistory using ancient genomes. *Trends in Genetics* **34**(3), 184–196.
- Zeng, L., Ming, C., Li, Y., Su, L.Y., Su, Y.H., Otecko, N.O., Liu, H.Q., Wang, M.S., Yao, Y.G., Li, H.P., Wu, D.D. and Zhang, Y.P. (2017). Rapid evolution of genes involved in learning and energy metabolism for domestication of the laboratory rat. *Molecular Biology and Evolution* **34**, 3148–3153.

10

Population Genomic Analyses of DNA from Ancient Remains

Torsten Günther and Mattias Jakobsson

Human Evolution, Department of Organismal Biology, Uppsala University, Sweden

Abstract

The possibility of obtaining genomic data from ancient biological material has opened the time dimension to genetic research. Obtaining genomic data from individuals and populations before, during, and after particular events allows us to study the individuals that were directly involved in these events. Such an approach can overcome the limitations of studying data from present-day individuals and trying to make inferences about past events from these data. Studies on present-day samples do not include potential groups that went extinct or have not contributed substantially to present-day populations, and some of the signals of ancient events may have been erased by later demographic processes. In this chapter we outline some of the specific approaches working with ancient population-genetic data and point to some of the known methodological and statistical specifics for working with empirical genetic data from ancient samples. The chapter mainly focuses on the power and possibilities of genomic approaches, in contrast to single-marker studies. The prime advantage of moving towards genome-wide studies is that data from even single individuals can lead to new insights as the thousands of different loci across the genome can be seen as (relatively) independent samples of the same evolutionary history. After we have described the statistical challenges and opportunities in working with ancient human DNA, we conclude with some examples on topics where this approach has led to new and better understanding of human evolution and prehistory.

10.1 Introduction

Retrieving DNA from ancient organic material was pioneered more than three decades ago (Pääbo, 1985). Initially, the field was plagued by problems such as contamination, post-mortem chemical modifications of the DNA, low levels of endogenous DNA, and molecular methods relying heavily on polymerase chain reaction (PCR) amplification (see, for example, Malmström *et al.*, 2005; Poinar *et al.*, 2006). An early landmark study was the sequencing of the mitochondrial control region of a Neanderthal (Krings *et al.*, 1997), which sparked a greater interest in genetic information from ancient remains, ‘ancient DNA’ (aDNA), among population and statistical geneticists (Nordborg, 1998; Wall, 2000). The aDNA field faced several major challenges at the time: the large amounts of ancient material needed for the PCR techniques to produce data, discussions on best practice for authentication methods (Cooper and Poinar, 2000), which

was particularly problematic for working with prehistoric humans, and the small amount of data produced by the targeted PCR approach. Advancements in sequencing technology (known as ‘next generation sequencing’, NGS) in the early 2000s revolutionized the field in happening to be particularly suitable for working with aDNA, which is typically scarce and highly fragmented in most ancient samples. The NGS techniques specifically work with short DNA fragments at the same time as having the ability to immortalize most of the DNA from an ancient extract in a ‘DNA library’. The techniques were quickly adapted to highly degraded DNA (Poinar *et al.*, 2006; Green *et al.*, 2006) and today, coupled with specific molecular genetics protocols adapted for ancient remains, they allow the generation of genome data from various sources of ancient material. Due to this development, the need for new knowledge about how to properly handle and analyze aDNA has become an important field of research. Fortunately, one of the important insights obtained thus far in the field is that many standard population genetics/genomics methods can be applied to genetic data from ancient remains, if the data is processed properly beforehand, and appropriate care is taken when analyzing the data and interpreting the results. In essence, most statistical and population genetic approaches described in this book are applicable also to investigating ancient population genetic data, with two additional notes: first, that the samples will typically be distributed along the time dimension, which needs to be accounted for in analyzing the data; and second, that the genetic data from ancient material typically has some peculiar characteristics pertaining sequence quality that needs additional consideration.

In this chapter we will first describe some of the challenges with working with aDNA and how they can be handled. Next, we will describe some new methods that have been developed to take advantage of the new opportunities that the availability of aDNA provides. Finally, we will give some examples of the new insights into history that studies of ancient human samples have led to so far. Note that for this chapter, we have chosen a human focus, partly because many of the specific approaches for working with DNA from ancient remains have been pioneered in this research area, and partly to keep the presentation coherent. While the techniques and approaches we describe here obviously can be used to investigate other organisms, we emphasize the benefit of a good genomics infrastructure. It is clear that much of the ancient DNA research on humans and hominids has greatly benefited from high-quality reference genome(s), large-scale population genotype reference data, and bioinformatic tools for handling genome sequence data.

10.2 Challenges of Working with and Analyzing Ancient DNA Data

Working with aDNA is associated with some particular challenges such as contamination, post-mortem damage, small amounts of data, and few samples. Below, we describe these challenges and the general approaches to processing aDNA data in order to handle these specific challenges and avoid potential pitfalls.

10.2.1 Sequence Degradation

After an individual dies, the DNA of the individual starts decomposing. The typical features of degraded DNA include fragmentation of the DNA strands into shorter fragments and cytosine deamination at the ends of the DNA fragments (Lindahl, 1993; Hofreiter *et al.*, 2001; Brotherton *et al.*, 2007; Briggs *et al.*, 2007). Both of these features tend to be correlated mainly with environmental conditions, but also with time – the older the sample, and the warmer and more humid the environment the sample has existed in, the stronger the observed post-mortem damage (Sawyer *et al.*, 2012; Weiß *et al.*, 2016; Wagner *et al.*, 2018). Fragment lengths have become an important sanity check for aDNA sequencing since long fragments are

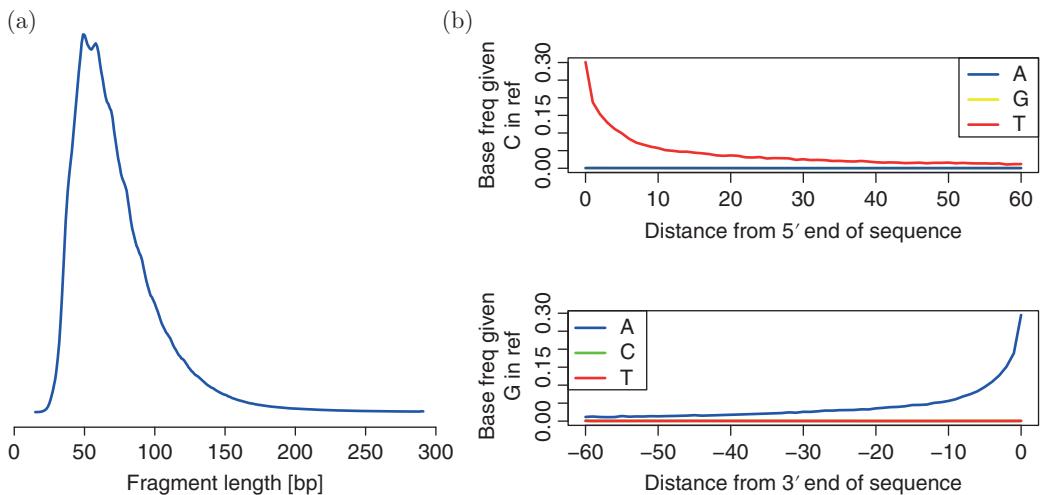


Figure 10.1 Characteristic damage pattern in aDNA: (a) length distribution of sequenced fragments; (b) enrichment of deaminations towards the fragment ends.

typically very rare in old samples. Hence, if long fragments occur in large numbers, or perhaps as a specific mode in a fragment length distribution, that can be an indication of contamination (Wall and Kim, 2007). Therefore, plotting the fragment length distribution is an important initial quality check in aDNA studies (Figure 10.1(a)). Furthermore, it has been shown that fragmentation shows regional patterns connected to nucleosome protection (Pedersen *et al.*, 2014), which has been suggested as an additional means for authentication (Hanghøj *et al.*, 2016). Fragment length also has an important impact on the potential to obtain DNA from very old samples since we rely on the ability to map fragments to genomes. The shorter the fragment, the greater the challenge to uniquely map it to an organism and a genome region (Meyer *et al.*, 2016; Renaud *et al.*, 2017).

The second feature, deamination of the nucleotide cytosine (C) to uracil (U), was first seen as a nuisance since it was read by sequencing machines as thymine (T), mimicking a true polymorphism. However, this feature has since become one of the most important diagnostic characteristics for authenticating DNA as old (Brotherton *et al.*, 2007; Briggs *et al.*, 2007; Ginolhac *et al.*, 2011; Jónsson *et al.*, 2013; Skoglund *et al.*, 2014b), although mixtures of authentic and contaminant DNA would also produce a damage pattern making methods to estimate contamination levels essential (see below). The feature will manifest as a pattern of increasing C → T variants toward the 5' end of the sequence fragments and a complementary increase in guanine (G) → adenine (A) variants at the 3' end (Figure 10.1(b)), although some library preparation strategies produce different patterns (Meyer *et al.*, 2012; Seguin-Orlando *et al.*, 2013) and methods exist to repair these changes before sequencing (Briggs *et al.*, 2010; Rohland *et al.*, 2015).

10.2.2 Contamination

Humans handling ancient remains implies a risk of contamination during excavation, storage and processing of samples. The contaminant DNA would typically be younger, of higher quality, and in greater concentration compared to the authentic DNA in the sample. For non-human/hominid samples, this contamination can be handled by mapping to reference genomes (species-specific and contaminant-specific), but the contamination can easily overwhelm the

sample endogenous DNA content, making sequencing efforts very costly. For ancient human samples, contamination by humans is particularly challenging. Clean lab procedures have substantially reduced the risk of contaminating the sample, but computational methods are required to assess and estimate contamination in sequence data, and such approaches are part of the standard quality checks performed in aDNA studies. Contamination estimation approaches are usually based on observing more than one allele in parts of the genome that should be haploid, which would indicate the presence of sequences from more than one individual. Most methods are based on the mitochondrial genome, which has orders of magnitude more copies per cell than the nuclear genome, allowing precise estimates even for low-coverage data. The procedures usually start by reconstructing the supposed authentic mitochondrial genome followed by estimating the proportion of mitochondrial sequence fragments in the library that belong to this sequence versus sequences stemming from a potential contaminant source (Green *et al.*, 2008; Fu *et al.*, 2013; Renaud *et al.*, 2015). For example, an established frequentist approach first constructs a consensus sequence for the mitochondrial genome, assuming that the majority of reads are authentic (Green *et al.*, 2008). Then the numbers of reads representing the consensus allele n_{cons} and alternative alleles n_{alt} are counted for each of U informative sites (i.e. sites that are nearly private to the consensus sequence compared to a set of reference mitochondria). An estimate for the proportion of contamination c can then be obtained by calculating

$$\hat{c} = \frac{\sum_{i=1}^U n_{\text{alt},i}}{\sum_{i=1}^U (n_{\text{alt},i} + n_{\text{cons},i})}.$$

Mitochondrial contamination estimation approaches usually give good estimates of mitochondrial contamination levels even with limited amounts of data, and they can also give a qualitative indication of autosomal contamination. However, due to different copy numbers of the mitochondrial genome between cells and cell types, the estimate of mitochondrial contamination itself might only be weakly correlated with the proportion of autosomal contamination – the quantity of interest in population genomic studies. More elaborate methods that use the nuclear genome – such as the hemizygote X chromosome in males (Rasmussen *et al.*, 2011) or the autosomes (Jun *et al.*, 2012; Racimo, 2016) – can overcome these limitations, but they require much more sequence data than mitochondrial approaches, and the estimates may be sensitive to the present-day reference populations used as potential contaminants and their evolutionary distance to the studied population (Rasmussen *et al.*, 2011; Jun *et al.*, 2012; Racimo, 2016).

If a substantial level of contamination has been found in the sample, the data need not be discarded completely. Assuming that contamination stems from a modern source, the known patterns of post-mortem DNA fragment degradation towards the fragment ends can be used to separate authentic and contaminant DNA fragments. If concerns about modern contamination exist, the analysis can be restricted to DNA fragments showing C→T or G→A changes at the fragment termini relative to the reference genome. Models of post-mortem degradation (PMD) can also incorporate the distribution of C→T (or G→A) changes along the fragment when determining whether a DNA fragment is authentic or not (Skoglund *et al.*, 2014b). One such approach is to set up a model for DNA fragments with PMD (M_{PMD}) and a model for DNA fragments without PMD (M_{NULL}). The models focus on two categories of sites in a pairwise alignment between a DNA fragment and a reference sequence where: (a) the reference has a C and the aligned fragment has either a C (C–C match) or a T (C→T mismatch); and (b) where the reference base is G and the aligned base is either a G (G–G match) or an A (G→A mismatch). There are three non-mutually exclusive events that can have occurred at these sites:

(i) a true biological polymorphism (occurring at rate π); (ii) a sequence error (occurring at rate ϵ); or (iii) a PMD (occurring at rate D_z). Based on empirical observations of PMD (e.g. Figure 10.1(b)), we can model D_z to follow a (modified) geometric distribution (plus a small constant K) across the DNA fragment with decreasing probability with distance z (measured in base pairs and where the first base has $z = 1$) from the relevant fragment terminus so that

$$D_z = (1 - p)^{z-1}p + K,$$

where p is the probability of a mismatch due to PMD. Observing a C–C match could be due to none of (i), (ii), or (iii) having occurred, but could also be caused by a sequence error having reverted a PMD base or a sequence error having reverted a biological polymorphism. A PMD event reverting a biological polymorphism does not need to be considered here as PMD only results in C→T (or G→A), and not the reverse. If we assume that events (i), (ii), and (iii) are independent processes (in other words, biological mutations, sequencing errors, and PMD occur independently of each other), we have three mutually exclusive possibilities,

$$P_{\text{PMD}}(\text{Match}|z) = (1 - \pi)(1 - \epsilon)(1 - D_z) + (1 - \pi)\epsilon D_z + \pi\epsilon(1 - D_z),$$

under the PMD model. For the null model (without PMD), where $D_z = 0$ for all z , we have

$$P_{\text{NULL}}(\text{Match}|z) = (1 - \pi)(1 - \epsilon) + \epsilon\pi.$$

The probability of a C→T mismatch (or a G→A mismatch) for a particular site is then any other event or combination of events, such that

$$P(\text{Mismatch}|z) = 1 - P(\text{Match}|z).$$

For humans, π can be set to 0.001 to approximate autosomal genetic variation between a pair of chromosomes and ϵ can be obtained from sequence technology-specific error rate – such as the phred-scaled base qualities, but note that only specific mismatches are considered; see Skoglund *et al.* (2014b) for details. For D_z , it has been shown that $p = 0.3$ and $K = 0.01$ are realistic values for many sequence data from remains that are several thousand years old (Skoglund *et al.*, 2014b), but other values may be warranted for sequence data with other characteristics. For a DNA fragment S with PMD, a likelihood function for site i is

$$L(M_{\text{PMD}}|S_i) = X \times P_{\text{PMD}}(S_i = \text{Match}|z) + (1 - X) \times P_{\text{PMD}}(S_i = \text{Mismatch}|z),$$

and for a DNA fragment without PMD,

$$L(M_{\text{NULL}}|S_i) = X \times P_{\text{NULL}}(S_i = \text{Match}|z) + (1 - X) \times P_{\text{NULL}}(S_i = \text{Mismatch}|z),$$

where X is an indicator variable ($X = 1$ if $S_i = \text{Match}$, and $X = 0$ if $S_i = \text{Mismatch}$). To determine whether a DNA fragment is more likely to originate from a PMD model or a model without PMD, a score can be calculated as the natural logarithm of the ratio of the likelihood for each model multiplied over all positions I in the sequence fragment, so that

$$\text{PMD-score} = \log \left(\frac{\prod_i^I L(M_{\text{PMD}}|S_i)}{\prod_i^I L(M_{\text{NULL}}|S_i)} \right).$$

This PMD-score gives information on how likely it is that a sequence fragment originates from a degraded sample; the greater the score, the more likely the fragment has degraded post mortem. In practice, a PMD-score of 2 or 3 typically removes the vast majority of contaminant fragments while retaining a substantial amount of authentic data for population genomic analysis (Skoglund *et al.*, 2014b). Other approaches take contamination estimates into account while performing data analysis (Racimo, 2016; Malaspina *et al.*, 2016), which may be a way to retain more of the data while adjusting the confidence in the results.

10.2.3 Handling Sequence Data from Ancient Material

The characteristic properties of aDNA require some specific steps in the handling of NGS data obtained from ancient remains (Kircher, 2012; Schubert *et al.*, 2014) besides the authenticity investigations already mentioned above. We briefly outline these steps.

10.2.3.1 Preprocessing of NGS Data

Although the development in sequencing technology moves towards longer reads (much longer than 100 bp), the average fragment size is typically less than 100 bp in DNA extracted from ancient material (Figure 10.1). This discrepancy means that while the field of aDNA largely benefits from the general increase in output per sequencing run, it will not directly benefit from the potential to sequence longer reads. For most aDNA NGS data it is important to appropriately trim the raw reads as they are usually longer than the actual DNA fragment and the sequence read will therefore include part of the adapter sequence at the other end of the fragment. If a paired-end strategy is chosen for sequencing, both forward and reverse reads may contain overlapping information on the authentic fragment. To improve the quality of the sequence data, a common approach (after adapter trimming) is to merge the two reads by adding the base quality scores for the overlapping bases in the two reads (Kircher, 2012).

10.2.3.2 Mapping and Identification of Endogenous DNA

A sequencing library of DNA extracted from ancient remains represents a metagenomic mix of DNA from various sources, including soil organisms and (ideally) the individual the remains stem from. In order to separate endogenous DNA from DNA from other sources, all sequence data is usually mapped to a reference genome of the species of interest (or a closely related species). For genomic studies, short fragments (typically <35 bp) are usually discarded as they might be too unspecific to map uniquely to one genome region and/or even be unspecific among organisms (Meyer *et al.*, 2016; Renaud *et al.*, 2017). Since ancient samples often contain limited amounts of DNA, and to avoid losing endogenous DNA, some adjustments in sequence processing are typically made compared to commonly established practices for sequence data from present-day individuals. Mapping the trimmed reads against a reference genome is usually done with different parameters than the ones used in genomic studies of present-day samples and DNA. Mismatch and gap penalties are typically set to lower values, allowing more reads to successfully map to the genome. These relaxations of mapping accuracy increase the total number of mapped reads, resulting in more data available for downstream analysis, which is of importance for low-quality samples where as much data as possible needs to be retained (Schubert *et al.*, 2012). The goal is to keep the false negative rate low, while a modest false positive rate is acceptable. This relaxation of mapping parameters is necessary in all cases where post-mortem damage has not been removed from the sequence reads prior to mapping since damage will lead to increased mismatch rates. Furthermore, this practice is particularly important when no reference genome is available (for the species of interest) and a genome from a closely related species is used instead as the mapping reference (Shapiro and Hofreiter, 2014). In these situations, it is crucial to test and compare different mapping settings in order to minimize reference bias, and simulation suites are available to generate data for testing such scenarios (Renaud *et al.*, 2017).

10.2.3.3 Additional Filtering

Samples with suboptimal conservation especially tend to produce DNA libraries of low complexity, which means that the total number of unique authentic DNA molecules in the DNA extract is low. Sequence data from such libraries (and also from deeply sequenced libraries) can contain many PCR duplicates. As these PCR duplicate sequence reads likely originate from the

same DNA molecule, they do not represent statistically independent observations and should be filtered out. Such duplicates can be identified by searching for reads that have identical orientation, start and end coordinates relative to the reference genome. The duplicates can be used to call a consensus sequence of the particular DNA fragment to reduce sequencing errors, or alternatively the copy with the highest quality can be retained for further analysis.

Post-mortem damage has additional implications for population genomic analysis as C→T or G→A changes due to post-mortem damage mimic true transition polymorphism. A very conservative strategy is to restrict the analysis to transversion sites, which decreases the number of sites available for analysis by about two thirds. As the post-mortem damage tends to occur towards the end of fragments, one can alternatively trim the first (and last) bases of each read (Seguin-Orlando *et al.*, 2014) or rescale base qualities based on a probabilistic model for post-mortem damage (Jónsson *et al.*, 2013). Finally, specifically developed genotype callers explicitly account for post-mortem damage when calling alleles (Hofmanová *et al.*, 2016; Link *et al.*, 2017).

10.2.4 Different Sequencing Approaches and the Limitations in their Resulting Data

In favorable circumstances, the complete genome can be sequenced from the ancient individual (Rasmussen *et al.*, 2010; Prüfer *et al.*, 2014; Meyer *et al.*, 2012; Lazaridis *et al.*, 2014; Gamba *et al.*, 2014; Günther *et al.*, 2018), but for many samples, DNA preservation, library complexity and budget restrictions result in less genomic information available per individual. One way to try to reduce the amount of sequencing (and costs) needed to reach a certain level of coverage is to perform hybridization capture and then sequence those specific captured fragments of interest. This can be conducted as whole-genome capture (Carpenter *et al.*, 2013) or single nucleotide polymorphism (SNP) capture panels that can exceed 1 million SNPs (Mathieson *et al.*, 2015). While these capture approaches can reduce sequencing costs, some implications need to be considered before data generation. Capture approaches may introduce a slight bias towards the allele carried by the sequences used to design the capture probes. This bias can have implications for whole-genome capture as probes are usually designed based on a single individual's genome. SNP capture approaches usually have one probe per allele at each bi-allelic SNP included in the capture array (alleviating the bias towards one individual in a whole-genome capture approach), but instead restrict the data to specific sites, which leads to SNP ascertainment bias that affects the results of standard population genetic analysis such as principal components analysis (PCA) and F_{ST} estimation (Albrechtsen *et al.*, 2010), limiting the options for downstream analyses and inferences to SNP-genotype approaches. Similar to population genomics with present-day samples, whole-genome sequence data can overcome problems with pre-selecting target variants, and in particular high-coverage whole-genome shotgun sequencing can overcome most limitations, but that is also more cost intensive. Hence when data is produced using whole-genome sequencing it has until now mostly resulted in low-coverage sequencing data. As of 2018, genome-wide data is available for more than 1000 ancient human individuals, representing a mixture of SNP capture and genome-sequence data. However, only about one third of these individuals have more than 1× coverage across the genome or for the targeted SNPs (Marciniak and Perry, 2017), and only a handful of high-coverage, high-quality ancient genomes have been sequenced (e.g. Rasmussen *et al.*, 2010; Prüfer *et al.*, 2014; Meyer *et al.*, 2012; Lazaridis *et al.*, 2014; Gamba *et al.*, 2014; Günther *et al.*, 2018).

10.2.5 Effects of Limited Amounts of Data on Downstream Analysis

Downstream analysis is usually population genetic in nature (see below). In order to conduct these types of analysis with standard population genetic tools, genotype calls are required.

However, calling diploid genotypes is not feasible for average (per-base) sequencing depths less than about 1, and it can sometimes be difficult to obtain more data for many samples. It is therefore important to note that low-coverage sequencing is also a limitation faced in many studies of modern populations, which has led to the development of specific approaches for genotype calling and analysis of such data. One possibility is to work with genotype likelihoods instead of genotypes called based on fixed cut-offs, which directly incorporates uncertainty about genotypes into the analysis. These approaches aim to calculate $P(D|G)$ for all possible genotypes G and the sequence data at a particular site D (Nielsen *et al.*, 2011; Korneliussen *et al.*, 2014). This allows both sequencing coverage at the site and the probability of sequencing errors (based on the base quality scores) to be taken into account. More advanced methods can also incorporate additional information such as known allele frequencies assuming Hardy–Weinberg equilibrium at the site. Furthermore, population genomic analysis methods have been extended to work on genotype likelihoods directly (Skotte *et al.*, 2013; Korneliussen *et al.*, 2014; Korneliussen and Moltke, 2015; Jørsboe *et al.*, 2017). Another direction much investigated in statistical genetics is the possibility of imputing sites that are missing (Servin and Stephens, 2007), potentially due to low coverage or sites that are not included in a particular genotype array. Such approaches have been proven powerful for present-day genomes (1000 Genomes Project, 2015; Lawson *et al.*, 2012) as well as for genome sequence data from prehistoric humans (Gamba *et al.*, 2014).

Another common strategy is to restrict the analysis to bi-allelic sites known to be polymorphic in present-day populations and to sample a single read per site, which is then used to make a pseudo-haploid call. This approach has the advantage of reducing the effect of sequencing errors and post-mortem damage as the possible alleles are known, but it also introduces an effect, mimicking genetic drift, private to the ancient individual (if the individual is assumed to be homozygous for the randomly drawn allele), which has impacts on several analyses. One approach to alleviate this problem is to ‘haploidize’ all individuals in an analysis and treat the data as from ‘mosaic’ haploid individuals, which has limitations for methods assuming diploid data and Hardy–Weinberg equilibrium (e.g. ADMIXTURE; Alexander *et al.*, 2009). Restricting analysis to known variable sites also has the inherent risk of ascertainment bias, if the reference panel contains variants specific to one group of interest. Table 10.1 displays common analysis types and approximate guidelines on how much sequence data would be needed to conduct those types of analysis.

The resulting data sets usually contain much missing data for the ancient samples. Direct comparisons between ancient individuals are restricted to sites with overlapping information, which is often limited to high-coverage sequence data or high-coverage SNP capture data. Including additional individuals from the same population can help to fill the gaps of missing data for population-based comparisons. Furthermore, large reference data sets including a diverse set of present-day individuals can be used as a scaffold to build up the axis of genetic variation. Ancient individuals can then be added onto these scaffolds using Procrustes alignment (Wang *et al.*, 2010; Skoglund *et al.*, 2012) or by projection (Keller *et al.*, 2012). Note that the limited amounts of data can affect results. This is illustrated in Figure 10.2 where such a projection has been performed for two ancient samples as well as for a large number of random sub-samples with different amounts of missing data.

10.3 Opportunities of Ancient DNA

Evolutionary biology studies a temporal process and aDNA allows evolutionary geneticists to study populations across time and space. This opportunity allows novel insights by applying

Table 10.1 Common analysis types and data types together with coverage needed

Analysis type/data type	Genome-wide SNP capture	Whole-genome shotgun sequencing
Sex identification	$\geq 0.01\times^a$, also depending on inclusion of sex chromosomes	$\geq 0.01\times$
Continental scale relationship (e.g. PCA)	$\geq 0.01\times^a$	$\geq 0.01\times$
Relationship within continents (basic population structure and admixture analysis)	$\geq 0.1\times^a$	$\geq 0.1\times$
Mitochondrial haplogroup	$\geq 0.1\times^a$	$\geq 0.1\times$
Kin relationships between individuals	$\geq 0.1\times^a$	$\geq 0.1\times$
Y chromosome haplotype	$\geq 0.5\times^a$, also depending on inclusion of sex chromosomes	$\geq 0.5\times$
Genetic diversity (using two or more individuals)	$\geq 0.5\times^{a,b}$	$\geq 0.5\times$
Demographic inference	$\geq 1\times^{a,b}$	$\geq 1\times$
Detect unknown human/hominid ancestry	not possible	$\geq 1\times$
Heterozygosity (diversity based on single individual)	$\geq 5\times^{a,b}$	$\geq 5\times$
Rare variant analysis	not possible	$\geq 10\times$
Effective population size over time	not possible	$\geq 15\times$
Novel variant calls	not possible	$\geq 20\times$

^aDepending on the number of targeted sites, here assuming >500,000 SNPs for humans.

^bAffected by ascertainment bias, estimates are relative and depend on targeted SNP panel.

standard population genetic methodology to temporal data. The possibility to investigate temporal data also provides an active field of research for the development of new approaches specifically geared towards analyzing aDNA. In addition to the technical challenges associated with generating genomic sequence data from ancient remains, there are a number of challenges for analyzing temporal data, including small sample sizes, admixture over time causing partial population replacements, and diversity in archaeological attribution. Temporal genetic variation and differentiation have been studied from theoretical and methodological perspectives (Rodrigo and Felsenstein, 1999; Nordborg, 1998; Sjödin *et al.*, 2014; Skoglund *et al.*, 2014c), and some specific simulation approaches have been designed (Anderson *et al.*, 2005; Jakobsson, 2009; Excoffier *et al.*, 2013; Kelleher *et al.*, 2016), but much work still remains in our efforts to understand the population genetic properties of time-series data. We summarize some of the methods and approaches used for the analysis of aDNA to highlight the specific opportunities of having access to time-series data.

10.3.1 Population Differentiation in Time and Space

With aDNA data, we have access to the time dimension in addition to spatial dimension(s), and samples/populations can be differentiated over both time and space. There are many similarities to investigating population differentiation along a space dimension and the time dimension. Wright's F_{ST} (Wright, 1951) is a standard population genetic parameter to quantify

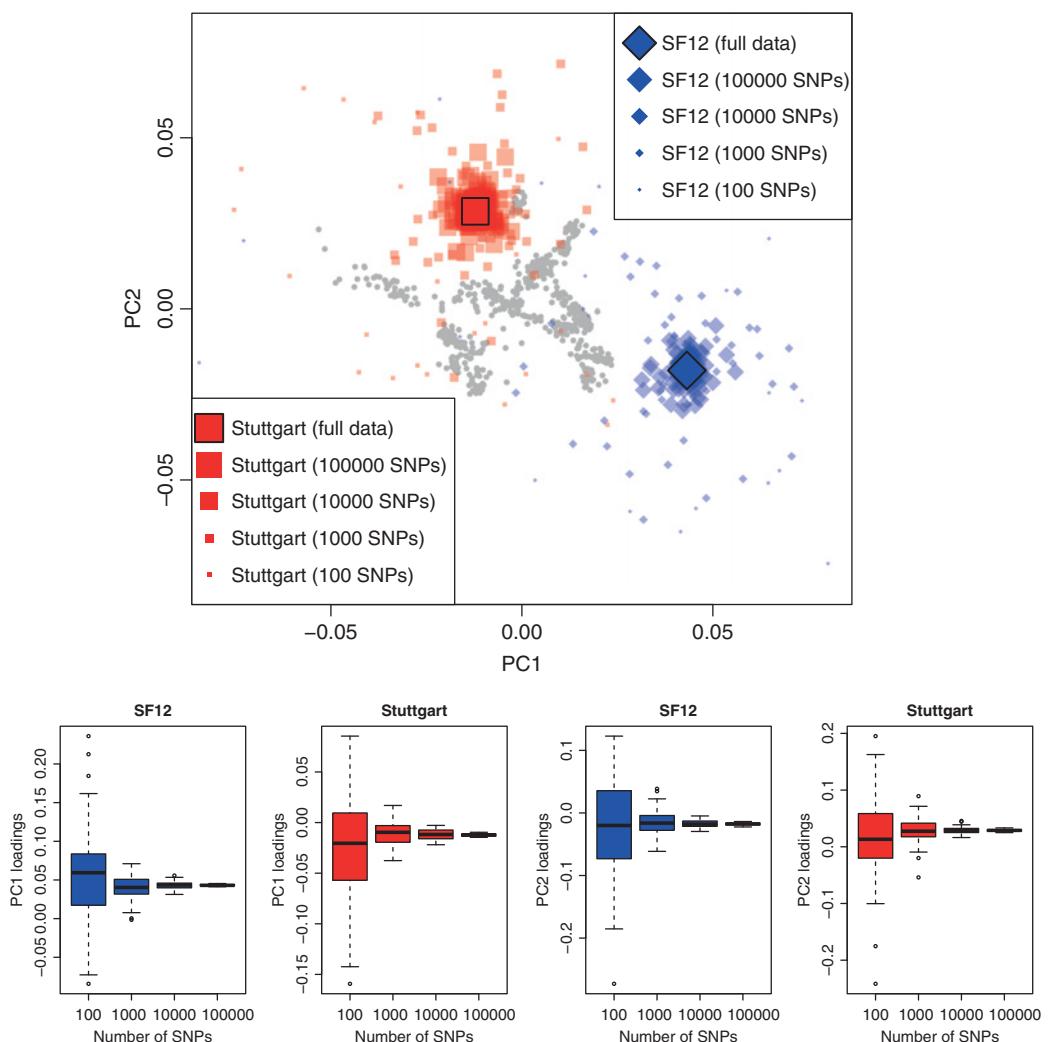


Figure 10.2 Principal components analysis to illustrate the uncertainty introduced by missing data: a Scandinavian hunter-gatherer ('SF12', ~9 kya; Günther *et al.*, 2018) and a early European farmer ('Stuttgart', ~7 kya; Lazaridis *et al.*, 2014) projected onto the PC1–PC2 space of present-day western Eurasians (gray dots) based on ~600,000 SNPs. The two ancient individuals were sub-sampled to show the noise generated by smaller data sets. The boxplots at the bottom show distributions of the PC1 and PC2 loadings for the sub-samples.

differentiation between/among (samples from) populations. Estimation of F_{ST} has become a standard tool to measure relations among populations (Nei, 1973), where the basic assumption is a model measuring genetic drift between populations. For instance, we can compare F_{ST} for a (two-population) split model and sampling from the two populations, and F_{ST} for a continuous model and sampling at two time-points (Figure 10.3(a),(b)). The Nei (1973) formulation of $F_{ST} = 1 - e^{-T_d}$, where T_d is the time since separation measured in $2N_e$ generations for two diverged populations that have both been sampled in the present, can be generalized to $F_{ST} = 1 - e^{-[(T_d + T_d - T_h)/2]}$, where the second sample is taken at a time T_h in the past instead of the present (Figure 10.3(b); see also Skoglund *et al.*, 2014c).

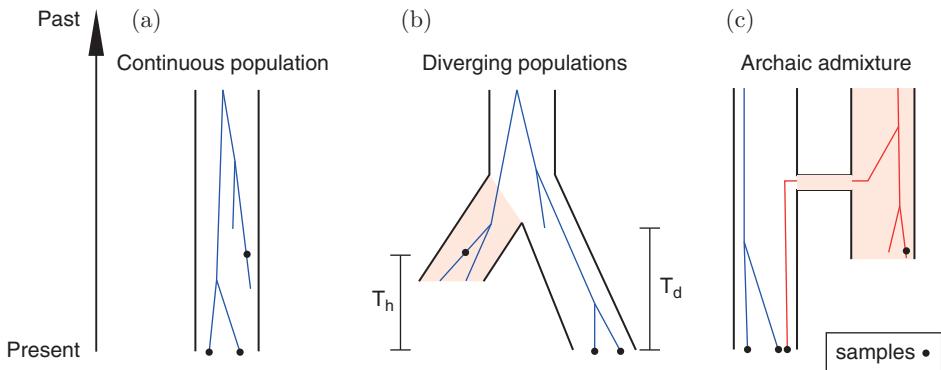


Figure 10.3 Illustration of different potential population models that can underlie a sample of present-day and ancient individuals. (a) A continuous population where the ancient sample is from a population that is directly ancestral to the present-day samples. (b) A population divergence model where the ancient sample comes from a sister population. The level of differentiation (F_{ST}) in this model can be related to split times (T_d) and sample times (T_h), measured both in time scaled with N_e (genetic drift; Wakeley, 2008; Skoglund *et al.*, 2011; Schlebusch *et al.*, 2012; Skoglund *et al.*, 2014c) and in number of generations (Schlebusch *et al.*, 2017). Similarly, private drift for an ancient sample (like the sample from the shaded population in (b)) can be used to test for continuity (Rasmussen *et al.*, 2014, 2015; Schraiber, 2018). (c) A model of two distinct populations, where one population becomes extinct (the shaded population) after having mixed with the other population. Several statistical approaches have been developed to test and quantify admixture under such a model (see Green *et al.*, 2010; Durand *et al.*, 2011; **Chapter 9**).

Estimating population divergence (based on samples from different populations) can be informative about the relationship among populations. Samples from ancient remains can represent populations that have few descendants today or populations that have been much affected by recent admixture. Inferring the population divergence between an ancient individual and present-day individuals can therefore be very informative about past events and the genetic make-up of past populations (Rasmussen *et al.*, 2014; Schlebusch *et al.*, 2017). The possibility of obtaining independent estimates for the population split-time (T_d in Figure 10.3(b)), one per branch (the shaded ancient population and the non-shaded present-day population in Figure 10.3(b)) can provide additional confidence when working with aDNA. Estimating split time for the ‘modern DNA’ sample and ‘ancient DNA’ sample separately also alleviates concerns for residual deaminations and other aDNA-specific properties. By setting up a general split model, and assuming two sampled gene copies (one diploid individual) from two populations, and assuming a constant ancestral population size, estimates for the population divergence time can be derived that is scaled in generations – and not in time scaled with N_e , which is a common approach (Schlebusch *et al.*, 2017). Using an estimate of the mutation rate, the divergence time (in generations) can be converted into years using a molecular clock and generation time assumption (Schlebusch *et al.*, 2017). Such estimates of population split times have been shown to be robust to low levels of migration and admixture (Schlebusch *et al.*, 2017).

Similar to F_{ST} , principal components analysis is commonly used to express spatial patterns of population differentiation, but it can also be used to express temporal patterns of differentiation. Figure 10.4 shows how PC1 captures the genetic differentiation over time with a monotonic cline from the oldest to the most recently sampled individuals (Skoglund *et al.*, 2014c). We can also see how the pattern changes for different demographic scenarios, making temporal sampling a potential approach to distinguish between population differentiation due to genetic drift in a continuous population or due to population subdivision (Skoglund *et al.*, 2014c).

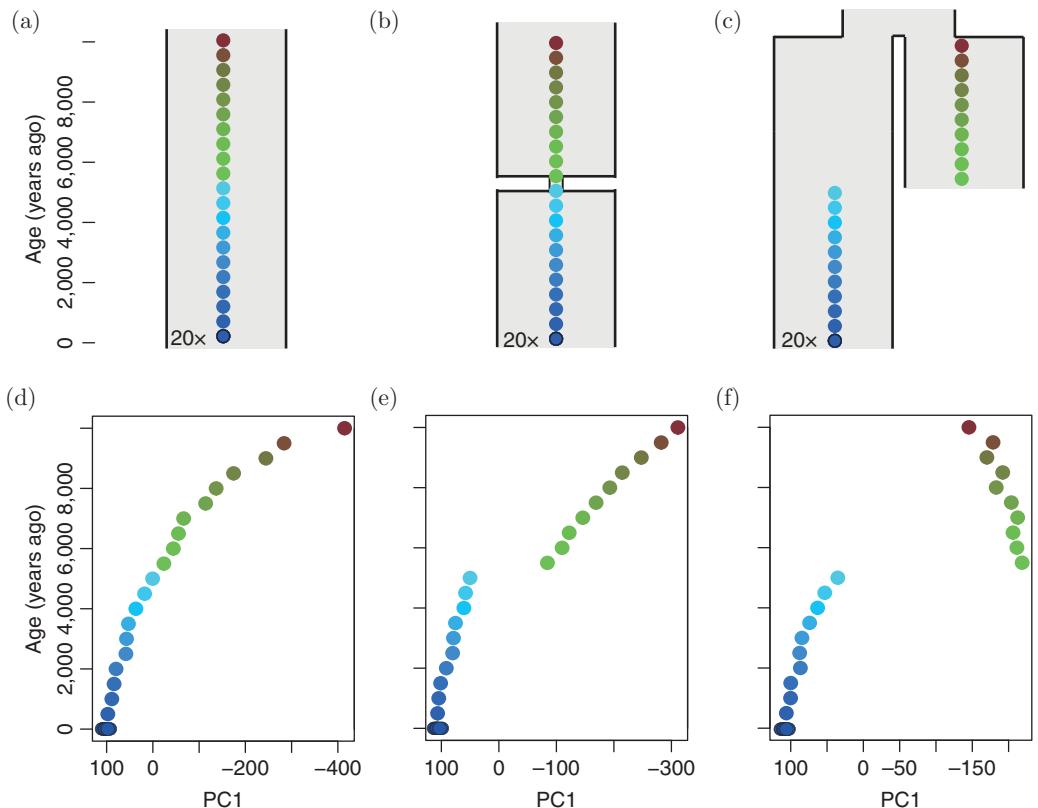


Figure 10.4 Temporal sampling distinguishes genetic drift from population structure. (a) Constant population size model. (b) Bottleneck model. (c) Replacement model. (d) Principal component 1 (PC1) stratified by sample time under the constant population size model. (e) PC1 stratified by sample time under the bottleneck model. (f) PC1 stratified by sample time under the replacement model. Each colored circle corresponds to a single sampled individual. There are 20 sampled individuals in (a)–(c) at time zero (the 20 individuals sampled at time zero end up on top of each other in (d)–(f)), and one sampled individual at each of the other time-points. F_{ST} between samples from before and after the bottleneck/replacement events at 5500 years ago fails to distinguish between the models ($F_{ST} = 0.0154 \pm 0.0003$ and 0.0153 ± 0.0003 , respectively). Illustration modified from Skoglund *et al.* (2014c).

10.3.2 Continuity

A question of particular interest in many aDNA studies is if an ancient population is the direct ancestor of a modern population (representing population continuity; Figure 10.3(a)), if the ancient sample represents a population on a related branch (Figure 10.3(b)), or if it represents another group (Figure 10.3(c)). Such knowledge is not only of interest for demographic inferences of population history, but also important for assumptions and interpretations of other analyses, for example changes in allele frequency trajectories (see below). Population genetic modeling is reasonably straightforward when two time-separated samples are assumed to be drawn from the same population with complete continuity between the two samples (Figure 10.3(a)), but such an assumption is perhaps naively unrealistic for most species, including humans (Nielsen *et al.*, 2017). Complete continuity models have often been evaluated using a simulation approach (e.g. Bramanti *et al.*, 2009; Malmström *et al.*, 2009; Sverrisdóttir *et al.*, 2014; Wilde *et al.*, 2014; Silva *et al.*, 2017). Rasmussen *et al.* (2014, 2015) used a framework

requiring diploid data from only two individuals to obtain maximum likelihood estimates of the coalescence rates private to each particular individual (e.g. to estimate the genetic drift private to the shaded population in Figure 10.3(b) from the sample in that population). Using specific assumptions about population sizes, Rasmussen *et al.* (2014, 2015) estimated ‘private drift’ from the probability of pairwise coalescence before divergence in each of the populations. If one individual represented a population ancestral to the other, the private drift in the ancient population would be ~ 0 . With a similar idea, Schraiber (2018) estimates the split time of two populations (T_d in Figure 10.3(b)) while taking post-mortem damage and sequencing errors from one (or more) low-coverage individuals into account. Again, the expected population split time (T_d in Figure 10.3(b)) for that ancient population would be ~ 0 if it represented a direct ancestor of the modern population and the samples rather represented draws from the model depicted in Figure 10.3(a). In most empirical applications, these methods have rejected complete population continuity as complete long-term isolation from external sources has been found extremely uncommon in humans (e.g. Nielsen *et al.*, 2017). To relax this unrealistically stringent assumption of no gene flow from known or unknown populations, Sjödin *et al.* (2014) developed a coalescent-based approach for investigating population continuity by allowing a certain level ($1 - C$) of gene flow from known or unknown populations. This opens up for a more nuanced treatment of questions regarding population continuity in terms of ‘level of contribution’ (C) from a particular population further back in time to a more recent population. Sjödin *et al.* (2014) used a hypothesis testing setup drawing information from single-locus data from two different points in time to assess contribution from an older to a younger population, and future approaches for multi-locus data, perhaps using estimation approaches (Chikhi *et al.*, 2001), will be important for understanding population continuity.

10.3.3 Migration and Admixture over Time

Ancient DNA makes it possible to observe how populations change over time and space, which makes the analysis of migration and admixture events one of the most common applications. Initially, exploratory analyses are typically conducted using PCA (e.g. Patterson *et al.*, 2006) or model-based clustering approaches (e.g. ADMIXTURE, Alexander *et al.*, 2009); see Chapter 8. These methods usually leverage a large set of modern-day individuals and populations to provide background variation. Such an approach solves the inherent aDNA limitations of low sample sizes and limited overlap of the genetic information in ancient samples, but also means that variation exclusive to ancient populations may be missed. Such exclusive variation will surely be missed when ancient individuals are projected onto modern genetic variation, which is commonly done for PCAs and occasionally for model-based clustering as well. From a logical perspective, the ideal situation would be to project later (more recent) samples/populations onto axes of variation built on older samples/populations, which will likely become the state of the art as the ancient data improves in both genome sequence coverage/quality and number of samples. In fact, very recent developments implement this idea of specifically modeling time between samples (Dystruct; Joseph and Pe'er, 2018) although their application to novel empirical data sets is still pending. Similarly, interpreting clustering analysis results based on a large number of modern samples combined with one or a few ancient samples can be tricky: what does it mean that an ancient sample is estimated to have shared ancestry with one or more ancestry clusters consisting of samples from present-day populations?

The exploratory analysis described above provides some insights into the relationships of ancient populations to each other and to modern groups, but they do not provide formal tests to assess population admixture or population continuity. Different types of statistics (often

denoted D and f) that measure allele sharing and shared drift between two, three or four populations (or individuals) (Patterson *et al.*, 2012; Durand *et al.*, 2011; see also Chapter 9) have proven to be powerful tools for testing specific hypotheses. f -statistics can be used to formally test admixture (f_3) and to estimate admixture proportions (f_4 ratio) and standard errors are usually estimated using a block-jackknife approach. One commonly used ‘spinoff’ application of f_3 -statistics in aDNA has been to assess genetic affinity among a large set of populations to a particular ancient sample as a heatmap (Raghavan *et al.*, 2014). These estimators are, however, sensitive to different error rates between samples, drift, and admixture from unobserved populations, which can complicate the interpretation of results (Rasmussen *et al.*, 2011; Rogers and Bohlender, 2015; Peter, 2016; Rodríguez-Varela *et al.*, 2017). Furthermore, more complex statistical frameworks based on f -statistics can be used to estimate population graphs and admixture proportions from multiple sources in both supervised and unsupervised ways (Patterson *et al.*, 2012; Pickrell and Pritchard, 2012).

With the number of individuals per time-period and archaeological site increasing, aDNA studies can also contribute to an understanding of the dynamics within prehistoric groups. Large sample sizes of sex chromosomes and autosomes allow sex-specific migration patterns to be studied (Goldberg *et al.*, 2017a), and combining information on diversity and runs of homozygosity has been used to illuminate social behaviors in past populations (Sikora *et al.*, 2017). Methods specifically designed to investigate genetic kinship in low-coverage data (Korneliussen and Moltke, 2015; Kennett *et al.*, 2017; Martin *et al.*, 2017; Kuhn *et al.*, 2018) will also contribute our understanding of the social dynamics within prehistoric groups.

10.3.4 Demographic Inference Based on High-Coverage Ancient Genomes

Recent years have seen a substantial increase in the total number of sequenced ancient individuals, but also in the genomic coverage per sample. This has opened the field of ancient population genomics to a more sophisticated set of methodology commonly used for genome data from present-day populations. These approaches exceed the allele frequency based methods discussed above by incorporating information on linkage disequilibrium and full genome sequences. For instance, methods utilizing linkage disequilibrium (Li and Durbin, 2011; Schiffels and Durbin, 2014; Terhorst *et al.*, 2017) can reveal population size changes as a function of time, resulting in better understanding of long-lost archaic (Meyer *et al.*, 2012; Prüfer *et al.*, 2014) and prehistoric humans (Lazaridis *et al.*, 2014; Schlebusch *et al.*, 2017; Günther *et al.*, 2018). Patterns of haplotype sharing among ancient populations have also been used to investigate admixture (Hofmanová *et al.*, 2016; Broushaki *et al.*, 2016; Martiniano *et al.*, 2017) and to date contacts between groups (Lipson *et al.*, 2017). High-coverage data also makes it possible to investigate sharing of rare alleles – a resource informative on recent population contact and population size exceeding the potential of standard analyses based on common variants (Schiffels *et al.*, 2016). These approaches promise new means of analyzing and testing more complex models of the relationship between modern and ancient samples (Gronau *et al.*, 2011; Excoffier *et al.*, 2013; Sikora *et al.*, 2017; Schlebusch *et al.*, 2017).

10.3.5 Allele Frequency Trajectories

Evolutionary and population geneticists have often utilized time series data to study selection and adaptation (Malaspinas, 2016). Most data sources have been experimental evolution trials and data collected from the wild. This was generally restricted to data from short time spans and species with relatively short generation times. The possibility of obtaining DNA from ancient remains has expanded the time range for this kind of analysis as it provides information

on allele frequency trajectories over (potentially) thousands of years. This realization has motivated a range of different methods that estimate different parameters (e.g. selection coefficient s , dominance coefficient, age of the selected allele) from the data or that scan genome-wide data for signals of selection. These approaches typically assume that all data was collected from a continuous population, an assumption that can be easily violated in practice (see above).

We first focus on methods for analyzing a single locus. Under neutrality ($s = 0$), the allele frequency p at a locus is not expected to change: $E(p_t) = p_0$, where p_0 would be the allele frequency at time $t_0 = 0$ and p_t is the allele frequency at time t . Adding selection ($s > 0$) and assuming an infinite population size (i.e. no genetic drift) as well as non-overlapping discrete populations (Malaspinas, 2016), we can approximate the change in allele frequency from one generation to the next as

$$p_{t+1} - p_t = \frac{1}{2}sp_t(1 - p_{t+1}),$$

assuming the selected allele enters the population at t_0 , which can also be seen as the age of the selected allele. We can approximate the allele frequency over time as a sigmoid curve of the shape

$$p_t = \frac{p_0}{p_0 + (1 - p_0)e^{-\frac{1}{2}s(t-t_0)}}.$$

Ancient DNA studies have the advantage of measuring p at different points in time, which makes it possible to use the above formulas to infer s . However, allele frequency changes in a real-life (finite) population over time are subject to oscillations due to the stochastic nature of genetic drift violating the ‘infinite size’ assumption (Figure 10.5). Combined with the often low sample sizes in aDNA studies (less than 20 chromosomes per temporal and geographic sampling point) inflating the variance around the estimates of population allele frequencies,

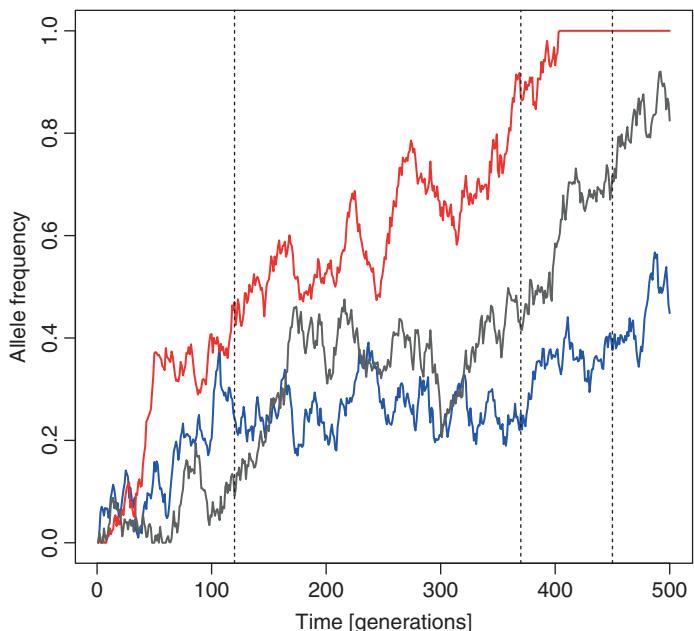


Figure 10.5 Simulated allele frequency trajectories with selection of three alleles with the same age, sampled at three points in time (dashed vertical lines).

these sources of noise represent major challenges for inference of different parameters from allele frequency trajectories based on aDNA making carefully developed statistical approaches essential. A number of different methods for estimating selection coefficient s and other parameters while taking the noise into account have been developed (e.g. Bollback *et al.*, 2008; Malaspinas *et al.*, 2012; Mathieson and McVean, 2013; Steinrücken *et al.*, 2014; Feder *et al.*, 2014; Sverrisdóttir *et al.*, 2014; Wilde *et al.*, 2014; Schraiber *et al.*, 2016; Loog *et al.*, 2017). In order to remain computationally feasible, these approaches rely on different approximations of the Wright–Fisher model. Depending on the expected strength of selection (weak versus strong), effective population size and other demographic effects (e.g. changing demographics over time or migration), some approaches can be more suitable in certain scenarios than others, making a sensible choice crucial for the interpretation of empirical data. In addition to these single-locus approaches, recent developments also perform genome-wide scans for loci under selection incorporating aDNA samples (Mathieson *et al.*, 2015; Racimo *et al.*, 2018). A review of currently used methods can be found in Malaspinas (2016).

10.4 Some Examples of How Genetic Studies of Ancient Remains Have Contributed to a New Understanding of the Human Past

In recent years, aDNA studies have led to some major breakthroughs in our understanding of the evolution and prehistory of our species. Some of these patterns had been erased by later demographic processes, implying that data from ancient samples was essential to detect these patterns. We describe the admixture between non-African anatomically modern humans and archaic hominins, as well as prehistoric migrations in western Eurasia.

10.4.1 Archaic Genomes and the Admixture with Modern Humans

One of the long-lasting debates in human evolution concerned the idea that modern humans either emerged from archaic forms of humans, *in situ*, in specific places across the globe, such as in Europe from Neanderthals, or emerged in Africa and subsequently replaced all archaic forms of humans outside Africa. The debate then nuanced into a discussion on whether there was any admixture with archaic humans as modern humans migrated out from Africa or if it was a complete replacement (Wolpoff *et al.*, 2000; Templeton, 2002; Garrigan and Hammer, 2006; Plagnol and Wall, 2006; Fagundes *et al.*, 2007; Wall *et al.*, 2009; Sánchez-Quijano and Lalueza-Fox, 2015). When the first (very limited) sequence data started appearing from Neanderthals (Krings *et al.*, 1997; Noonan *et al.*, 2006; Green *et al.*, 2006), the data was used to argue for different conclusions, including insufficient power and problems with contamination (Nordborg, 1998; Wall, 2000; Noonan *et al.*, 2006; Green *et al.*, 2006; Wall and Kim, 2007). With the draft sequence of a Neanderthal genome, Green *et al.* (2010) could conclusively demonstrate some level of admixture between non-Africans and Neanderthals using D - and f -statistics. The estimates of the proportion of Neanderthal ancestry that persists in the genomes of non-Africans is low (1–2% perhaps) (Sankararaman *et al.*, 2014; Prüfer *et al.*, 2014; Vernot and Akey, 2014) and points towards a complex history of interaction between Neanderthals and modern humans. For instance, East Asians have some 20% more archaic ancestry compared to Europeans (Wall *et al.*, 2013), which may reflect further admixture events in the ancestors of present-day East Asians after the population split from Europeans (Vernot and Akey, 2014) or possibly differences in selection against Neanderthal ancestry (Harris and Nielsen, 2016). Interestingly, SNP capture data in an early modern human from Romania who lived about 40 kya provided further evidence that introgression occurred

several times outside Africa (Fu *et al.*, 2015), although this individual/group did not contribute ancestry to present-day populations.

In addition to Neanderthals, at least one other form of archaic humans lived in Eurasia when the first modern humans started appearing on the continent. We know very little about their morphology and distribution, but the remarkable sequencing of the genome from one of an individual's finger phalanges, found in the Denisova Cave in Siberia, led to discovery of the 'Denisovans' (Reich *et al.*, 2010; Meyer *et al.*, 2012). The genetic data proved the existence of an unknown group of archaic humans, and only four individuals have been identified from this enigmatic group so far (Sawyer *et al.*, 2015; Slon *et al.*, 2017). The 'Denisovans' are most closely related to Neanderthals, on the same order of genetic differentiation as the deepest splits among modern humans (Prüfer *et al.*, 2014). Based on the limited data from these four 'Denisovans', genetic diversity appears to be lowest in Neanderthals, followed by 'Denisovans', with modern humans having greater levels of diversity (Sawyer *et al.*, 2015; Prüfer *et al.*, 2014). There are many peculiarities surrounding 'Denisovans'; for instance, it has been suggested that they carried genetic material from individuals of earlier forms of humans (via admixture), possibly *Homo erectus* (Prüfer *et al.*, 2014), and the scarcity of fossils lends credence to the idea that the group was short-lived or very low in numbers. Perhaps a more realistic scenario is to view 'Denisovans' as the eastern and/or southern end of a spectrum of archaic humans living in Eurasia (and possibly beyond), where Neanderthals represent the western end of the spectrum.

'Denisovans' have also (similarly to Neanderthals) interbred with modern humans. In particular, *D*- and *f*-statistics showed that some groups such as Melanesians in Oceania carry 3–6% genetic material tracing back to 'Denisovans' (Reich *et al.*, 2010; Meyer *et al.*, 2012; Prüfer *et al.*, 2014). Further analysis using *D*-statistics and PCA showed that mainland southeast Asians also carry genetic material from 'Denisovans' (Skoglund and Jakobsson, 2011), on the order of parts of a percent (Skoglund and Jakobsson, 2011; Prüfer *et al.*, 2014). Although this genetic material from Denisovan admixture is limited, it has played an important role in human adaptation to specific environments, such as high-altitude adaptation in Tibetans (Huerta-Sánchez *et al.*, 2014). The studies based on the first reliable genome data from archaic humans proposed two punctuated and very specific events for admixture between modern humans and archaic humans (Reich *et al.*, 2010; Green *et al.*, 2010), but since then we have learned that this admixture is much more common, with multiple events taking place between different modern human groups and different archaic groups (Skoglund and Jakobsson, 2011; Prüfer *et al.*, 2014; Huerta-Sánchez *et al.*, 2014; Vernot *et al.*, 2016; Browning *et al.*, 2018) and in both directions (Kuhlwilm *et al.*, 2016). Although a clearer picture of admixture between modern humans and archaic humans is emerging, we caution that our understanding of admixture models is not complete at present (Nielsen *et al.*, 2017).

10.4.2 Neolithic Revolution in Europe and the Bronze Age Migrations

10.4.2.1 The Neolithic Transition

The transition from a hunter-gatherer lifestyle to a sedentary farming lifestyle is one of the most important transitions in human history and forms the basis for the emergence of civilizations. This process – called the Neolithic transition – occurred independently in different parts of the world (Diamond and Bellwood, 2003). For western Eurasia, the first evidence of farming practices has been found in the fertile crescent dating to 11–12 kya. From there, farming spread into Anatolia and Europe, reaching the western and northern parts of the continent around 6 kya. A long-standing debate across different disciplines has been whether these farming practices were spread as an idea ('cultural diffusion', Whittle, 1996; Renfrew and Boyle, 2000) or via migration of a group of farmers across the continent ('demic diffusion', Ammerman and Cavalli-Sforza,

1984). Analysis of admixture proportions and PCAs conducted with genomic data from the first farmers from all over Europe clearly showed a strong differentiation between this group and different European hunter-gatherers (Skoglund *et al.*, 2012, 2014a; Sánchez-Quinto *et al.*, 2012; Lazaridis *et al.*, 2014; Gamba *et al.*, 2014; Olalde *et al.*, 2015; Günther *et al.*, 2015; Cassidy *et al.*, 2016). The genetic differentiation (measured by F_{ST}) among hunter-gatherers and farmers is comparable to the differentiation between modern-day populations from different continents (Skoglund *et al.*, 2014a). The genetic composition of the Mesolithic hunter-gatherers falls outside the variation of present-day populations, but they are genetically most similar to modern-day northern and northeastern European populations. Surprisingly, the early European farmers do not group with modern-day groups from the Near and Middle East, but they exhibit marked genetic similarities with modern-day southwestern Europeans, especially Sardinians (Skoglund *et al.*, 2012, 2014a; Lazaridis *et al.*, 2014; Keller *et al.*, 2012). This observation can be explained by local continuity in Sardinia, contrasting the many demographic turnovers in the Neolithic core area since the first farmers left that region (Omrank *et al.*, 2016; Lazaridis *et al.*, 2016). Direct studies on early farmers from Anatolia and the Levant confirmed this region as the source of the European Neolithic groups (Mathieson *et al.*, 2015; Omrank *et al.*, 2016; Lazaridis *et al.*, 2016; Kılınç *et al.*, 2016; Broushaki *et al.*, 2016), as predicted by archeology a long time ago (Childe, 1925). These early farming groups expanded first within Anatolia and the Near East and then started migrating into Europe (Kılınç *et al.*, 2016; Lazaridis *et al.*, 2016) in contrast to groups in the eastern part of the Fertile Crescent (Broushaki *et al.*, 2016; Lazaridis *et al.*, 2016; Gallego-Llorente *et al.*, 2016) who contributed limited genetic material to the early farmers of Europe.

Farming practices probably made more resources available to the populations, allowing for larger groups of people, as indicated by a substantially higher genetic diversity of farming groups (Skoglund *et al.*, 2014a; Gamba *et al.*, 2014). The two groups, hunter-gatherers and farmers, however, did not remain isolated from each other. Comparisons of the two groups, using PCA, D - and f -statistics and admixture modeling, have revealed some degree of admixture from hunter-gatherers into farmers (Skoglund *et al.*, 2014a; Haak *et al.*, 2015; Günther *et al.*, 2015). The hunter-gatherer lifestyle was eventually completely replaced by farming, but the farming groups genetically assimilated hunter-gatherers (Skoglund *et al.*, 2014a; Lazaridis *et al.*, 2014). Comparing groups across different time periods, farmers from the middle Neolithic and early Chalcolithic periods (6000 to 4500 years ago) display additional ancestry from Mesolithic hunter-gatherers compared to early Neolithic groups (Haak *et al.*, 2015; Günther *et al.*, 2015). The process of admixture must have continued for at least two millennia, which raises the question where people belonging to the hunter-gatherer gene-pool were living during this period. Generally speaking, the genetic result that hunter-gatherers and farmers mixed shows that neither the strict demic nor the strict cultural diffusion model accurately describes the events during the Neolithic transition (Pinhasi *et al.*, 2005; Fort, 2012).

10.4.2.2 Demographic Changes during the Late Neolithic and Bronze Age

The population turnover during the early Neolithic was not the last episode where a migrating group had a massive impact on peoples of Europe. The late Neolithic and early Bronze Age were also times of large-scale migrations into Europe. Northeastern Eurasia was likely populated, until some millennia ago, by a population that is lost today, and represented by a distinct gene pool (Lazaridis *et al.*, 2014; Raghavan *et al.*, 2014; Skoglund *et al.*, 2014a; Haak *et al.*, 2015). This group has likely contributed genetic material to Europe for a long time. For instance, 9000-year-old Scandinavian hunter-gatherers show affinities to this group (Günther *et al.*, 2018), but a dramatic shift of ancestry in central and western Europe occurs in the late Neolithic (Haak *et al.*, 2015; Allentoft *et al.*, 2015). This shift involved the Yamnaya culture herders from the Pontic-Caspian steppe who migrated to central Europe about 4500 years ago

(Haak *et al.*, 2015; Allentoft *et al.*, 2015). The steppe herders were themselves descendants from different hunter-gatherer groups from (modern-day) Russia (Haak *et al.*, 2015) and the Caucasus (Jones *et al.*, 2015). This migration resulted in the rise of the late Neolithic Corded Ware culture in central Europe and has been suggested to spread Indo-European languages (Allentoft *et al.*, 2015; Haak *et al.*, 2015). Interestingly, the migration of the Yamnaya herders was heavily male-dominated and lasted several generations (Goldberg *et al.*, 2017a,b), as evidenced by a depletion of 'Yamnaya-related' ancestry on the X chromosome compared to the autosomes in individuals excavated from central European Corded Ware sites. Taken together, the late Neolithic and early Bronze Age migrations brought a cultural shift to central Europe driven by male migration, possibly facilitated by new technology, way of life and conquest (Kristiansen *et al.*, 2017; Goldberg *et al.*, 2017a; Haak *et al.*, 2015; Allentoft *et al.*, 2015).

The late Neolithic and the Bronze Age were generally very dynamic times, which involved not only the Corded Ware groups but also people associated with the Bell Beaker culture and the later Unetice culture as well as several other groups. The dynamics of these groups eventually spread the genetic material over most of western and northern Europe, reaching the Atlantic coast and the British Isles only a few centuries later (Allentoft *et al.*, 2015; Haak *et al.*, 2015; Cassidy *et al.*, 2016; Martiniano *et al.*, 2017). These massive migrations homogenized European populations and reduced genetic differentiation to levels seen between modern Europeans (Lazaridis *et al.*, 2016). After the hunter-gatherers' recolonization of Europe following the Latest Glacial Maximum, and the migrations connected with the Neolithic transition, the late Neolithic/Bronze Age migrations from the east is likely the third most influential event for the composition and gradients of genomic variation among modern-day Europeans.

10.5 Summary and Perspective

The possibilities for studying individuals who lived long ago and even extinct species greatly improved in the last few decades. Much of this development has been driven by the great advances made in molecular genetics and specifically sequencing technologies. In many respects the coined terms 'paleogenetics/genomics', 'archeogenetics/genomics', and 'ancient DNA' are merely genetic/genomic investigations of individuals and populations who lived some time ago. As we have tried to illustrate in this chapter, virtually all population genetic approaches are perfectly applicable to this type of genetic information with a few additional prerequisites, such as handling contamination and DNA degradation. One promising direction is analysis methods specifically designed for aDNA that handle the specific properties of degraded DNA. A possibly grander challenge for the population genetic and statistical genetic community lies in expanding models and approaches to utilize the new dimension (time) of the data. While omitting to understand and handle the peculiar properties of DNA from prehistoric remains can cause serious biases, and this needs to be taken seriously, we foresee that genetic studies of ancient material will become common approaches in evolutionary biology, ecology, archeology, in addition to continuing to push the bounds for our understanding of human history and evolution.

Acknowledgements

We thank I. Moltke, L. Orlando, M. Sikora, P. Sjödin, F. Sánchez-Quinto, and S. Ramachandran for helpful comments and discussions. T.G. was supported by a starting grant from the Swedish Research Council and M.J. was supported by grants from the Swedish Research Council, European Research Council, Knut and Alice Wallenberg foundation, the Swedish National Bank's foundation for humanities and social sciences, and the Göran Gustafsson foundation.

References

- 1000 Genomes Project (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.
- Albrechtsen, A., Nielsen, F.C. and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* **27**, 2534–2547.
- Alexander, D.H., Novembre, J. and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**(9), 1655–1664.
- Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., Malaspina, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnnerup, N., Harvig, L., Baron, J., Della Casa, P., Dabrowski, P., Duffy, P.R., Ebel, A.V., Epimakhov, A., Frei, K., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolár, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Pálfi, G., Pokutta, D., Pospieszny, Ł., Price, T.D., Saag, L., Sablin, M., Shishlina, N., Smrčka, V., Soenov, V.I., Szeverényi, V., Tóth, G., Trifanova, S.V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando, L., Sicheritz-Pontén, T., Brunak, S., Nielsen, R., Kristiansen, K. and Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature* **522**(7555), 167–172.
- Ammerman, A.J. and Cavalli-Sforza, L.L. (1984). *The Neolithic Transition and the Genetics of Populations in Europe*. Princeton University Press, Princeton, NJ.
- Anderson, C.N.K., Ramakrishnan, U., Chan, Y.L. and Hadly, E.A. (2005). Serial SimCoal: A population genetics model for data from multiple populations and points in time. *Bioinformatics* **21**, 1733–1734.
- Bollback, J.P., York, T.L. and Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**(1), 497–502.
- Bramanti, B., Thomas, M.G., Haak, W., Unterlaender, M., Jores, P., Tambets, K., Antanaitis-Jacobs, I., Haidle, M.N., Jankauskas, R., Kind, C.-J., Lueth, F., Terberger, T., Hiller, J., Matsumura, S., Forster, P. and Burger, J. (2009). Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**(5949), 137–140.
- Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Pruefer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M. and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 14616–14621.
- Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M. and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, **38**(6), e87.
- Brotherton, P., Endicott, P., Sanchez, J.J., Beaumont, M., Barnett, R., Austin, J. and Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Research* **35**(17), 5717–5728.
- Broushaki, F., Thomas, M.G., Link, V., López, S., van Dorp, L., Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-del Molino, D., Kousathanas, A., Sell, C., Robson, H.K., Martiniano, R., Blöcher, J., Scheu, A., Kreutzer, S., Bollongino, R., Bobo, D., Davoudi, H., Munoz, O., Currat, M., Abdi, K., Biglari, F., Craig, O.E., Bradley, D.G., Shennan, S., Veeramah, K.R., Mashkour, M., Wegmann, D., Hellenthal, G. and Burger, J. (2016). Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**(6298), 499–503.
- Browning, S.R., Browning, B.L., Zhou, Y., Tucci, S. and Akey, J.M. (2018). Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**, 53–61.e9.

- Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., et al. (2013). Pulling out the 1%: Whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American Journal of Human Genetics* **93**(5), 852–864.
- Cassidy, L.M., Martiniano, R., Murphy, E.M., Teasdale, M.D., Mallory, J., Hartwell, B. and Bradley, D.G. (2016). Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proceedings of the National Academy of Sciences of the United States of America* **113**(2), 368–373.
- Chikhi, L., Bruford, M. and Beaumont, M. (2001). Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347–1362.
- Childe, V.G. (1925). *The Dawn of European Civilization*. Kegan Paul, Trench, Trubner & Co., London.
- Cooper, A. and Poinar, H. (2000). Ancient DNA: Do it right or not at ALL. *Science* **289**, 1139.
- Diamond, J. and Bellwood, P. (2003). Farmers and their languages: The first expansions. *Science* **300**(5619), 597–603.
- Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239–2252.
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C. and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10), e1003905.
- Fagundes, N.J.R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F.M., Bonatto, S.L. and Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 17614–17619.
- Feder, A.F., Kryazhimskiy, S. and Plotkin, J.B. (2014). Identifying signatures of selection in genetic time series. *Genetics* **196**(2), 509–522.
- Fort, J. (2012). Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proceedings of the National Academy of Sciences of the United States of America* **109**(46), 18669–18673.
- Fu, Q., Mittnik, A., Johnson, P.L., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., et al. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* **23**(7), 553–559.
- Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D. and Pääbo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**(7564), 216–219.
- Gallego-Llorente, M., Connell, S., Jones, E.R., Merrett, D.C., Jeon, Y., Eriksson, A., Siska, V., Gamba, C., Meiklejohn, C., Beyer, R., Jeon, S., Cho, Y.S., Hofreiter, M., Bhak, J., Manica, A. and Pinhasi, R. (2016). The genetics of an early Neolithic pastoralist from the Zagros, Iran. *Scientific Reports* **6**, 31326.
- Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kovári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczyk, P., Higham, T.F.G., Hofreiter, M., Bradley, D.G. and Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications* **5**.
- Garrigan, D. and Hammer, M.F. (2006). Reconstructing human origins in the genomic era. *Nature Reviews Genetics* **7**, 669–680.
- Ginolhac, A., Rasmussen, M., Gilbert, M.T.P., Willerslev, E. and Orlando, L. (2011). mapdamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**(15), 2153–2155.

- Goldberg, A., Gunther, T., Rosenberg, N.A. and Jakobsson, M. (2017a). Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 2657–2662.
- Goldberg, A., Gunther, T., Rosenberg, N.A. and Jakobsson, M. (2017b). Robust model-based inference of male-biased admixture during Bronze Age migration from the Pontic-Caspian Steppe. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E3875–E3877.
- Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. and Paabo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336.
- Green, R.E., Malaspinas, A.-S., Krause, J., Briggs, A.W., Johnson, P.L.F., Uhler, C., Meyer, M., Good, J.M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H.A., Ronan, M., Rothberg, J.M., Egholm, M., Rudan, P., Brajković, D., Kućan, Z., Gusić, I., Wikström, M., Laakkonen, L., Kelso, J., Slatkin, M. and Pääbo, S. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**(3), 416–426.
- Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science* **328**(5979), 710–722.
- Gronau, I., Hubisz, M.J., Galko, B., Danko, C.G. and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, **43**(10), 1031.
- Günther, T., Valdiosera, C., Malmström, H., Ureña, I., Rodriguez-Varela, R., Sverrisdóttir, Ó.O., Daskalaki, E.A., Skoglund, P., Naidoo, T., Svensson, E.M., Bermúdez de Castro, J.M., Carbonell, E., Dunn, M., Storå, J., Iriarte, E., Arsuaga, J.L., Carretero, J.-M., Götherström, A. and Jakobsson, M. (2015). Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proceedings of the National Academy of Sciences of the United States of America* **112**(38), 11917–11922.
- Günther, T., Malmström, H., Svensson, E.M., Omrak, A., Sánchez-Quinto, F., Kılınç, G.M., Krzewińska, M., Eriksson, G., Fraser, M., Edlund, H., et al. (2018). Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biology* **16**(1), e2003703.
- Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R.G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S.L., Risch, R., Rojo, M.A. Guerra, Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K.W. and Reich, D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**(7555), 207–211.
- Hanghøj, K., Seguin-Orlando, A., Schubert, M., Madsen, T., Pedersen, J.S., Willerslev, E. and Orlando, L. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epipaleomix. *Molecular Biology and Evolution* **33**(12), 3284–3298.
- Harris, K. and Nielsen, R. (2016). The genetic cost of Neanderthal introgression. *Genetics* **203**(2), 881–891.

- Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-del Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., et al. (2016). Early farmers from across Europe directly descended from neolithic aegeans. *Proceedings of the National Academy of Sciences of the United States of America* **113**(25), 6886–6891.
- Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, A.v. and Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* **29**(23), 4793–4799.
- Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**(7513), 194.
- Jakobsson, M. (2009). COMPASS: A program for generating serial samples under an infinite sites model. *Bioinformatics* **25**, 2845–2847.
- Jones, E.R., Gonzalez-Fortes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., Meshveliani, T., Bar-Yosef, O., Müller, W., Belfer-Cohen, A., Matskevich, Z., Jakeli, N., Higham, T.F.G., Currat, M., Lordkipanidze, D., Hofreiter, M., Manica, A., Pinhasi, R. and Bradley, D.G. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications* **6**, 8912.
- Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L. and Orlando, L. (2013). mapDamage2.0: Fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**(13), 1682–1684.
- Jørsboe, E., Hanghøj, K. and Albrechtsen, A. (2017). fastNGSAdmix: Admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33**(19), 3148–3150.
- Joseph, T.A. and Pe'er, I. (2018). Inference of population structure from ancient DNA. Preprint, bioRxiv 261131.
- Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M. and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics* **91**(5), 839–848.
- Kelleher, J., Etheridge, A.M. and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* **12**(5), e1004842.
- Keller, A., Graefen, A., Ball, M., Matzas, M., Boisguerin, V., Maixner, F., Leidinger, P., Backes, C., Khairat, R., Forster, M., Stade, B., Franke, A., Mayer, J., Spangler, J., McLaughlin, S., Shah, M., Lee, C., Harkins, T.T., Sartori, A., Moreno-Estrada, A., Henn, B., Sikora, M., Semino, O., Chiaroni, J., Roots, S., Myres, N.M., Cabrera, V.M., Underhill, P.A., Bustamante, C.D., Vigl, E.E., Samadelli, M., Cipollini, G., Haas, J., Katus, H., O'Connor, B.D., Carlson, M.R.J., Meder, B., Blin, N., Meese, E., Pusch, C.M. and Zink, A. (2012). New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications* **3**, 698.
- Kennett, D.J., Plog, S., George, R.J., Culleton, B.J., Watson, A.S., Skoglund, P., Rohland, N., Mallick, S., Stewardson, K., Kistler, L., et al. (2017). Archaeogenomic evidence reveals prehistoric matrilineal dynasty. *Nature Communications* **8**.
- Kılınç, G.M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A.M., Biçakçı, E., Baird, D., Dönertas, H.M., Ghalichi, A., Yaka, R., Koptekin, D., Acan, S.C., Parviz, P., Krzewinska, M., Daskalaki, E.A., Yüncü, E., Dagtas, N.D., Fairbairn, A., Pearson, J., Mustafaoglu, G., Erdal, Y.S., G. Çakan, Y., Togan, İ., Somel, M., Storå, J., Jakobsson, M. and Götherström, A. (2016). The demographic development of the first farmers in Anatolia. *Current Biology* **26**(19), 2659–2666.
- Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. In B., Shapiro and M., Hofreiter (eds.), *Ancient DNA: Methods in Molecular Biology*. Springer, New York, pp. 197–228.

- Korneliussen, T.S. and Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* **31**(24), 4009–4011.
- Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**.
- Krings, M., Stone, A., Schmitz, R.W., Krainitzki, H., Stoneking, M. and Pääbo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30.
- Kristiansen, K., Allentoft, M.E., Frei, K.M., Iversen, R., Johannsen, N.N., Kroonen, G., Pospieszny, Ł., Price, T.D., Rasmussen, S., Sjögren, K.-G., et al. (2017). Re-theorising mobility and the formation of culture and language among the Corded Ware Culture in Europe. *Antiquity* **91**(356), 334–347.
- Kuhlwilm, M., Gronau, I., Hubisz, M.J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H.A., Lalueza-Fox, C., de La Rasilla, M., et al. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**(7591), 429–433.
- Kuhn, J.M.M., Jakobsson, M. and Günther, T. (2018). Estimating genetic kin relationships in prehistoric populations. *PLoS ONE* **13**(4), e0195491.
- Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics* **8**(1), e1002453.
- Lazaridis, I., Patterson, N., Mitnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K.I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H.A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C.M., Brisighelli, F., Busby, G.B.J., Cali, F., Churnosov, M., Cole, D.E.C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S.A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B.M., Hervig, T., Hodoglugil, U., Jha, A.R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R.W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E.B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C.A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M.G., Ruiz-Linares, A., Tishkoff, S.A., Singh, L., Thangaraj, K., Villem, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E.E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D. and Krause, J. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**(7518), 409–413.
- Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E.R., Roodenberg, S.A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J.M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S.M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D.A., O'Reilly, S., Richards, M.B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J.F., Yengo, L., Hovhannisan, N.A., Patterson, N., Pinhasi, R. and Reich, D. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357), 493–496.
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature* **362**(6422), 709.
- Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A. and Wegmann, D. (2017). Atlas: Analysis tools for low-depth and ancient samples. Preprint, bioRxiv 105346.

- Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., *et al.* (2017). Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**(7680), 368.
- Loog, L., Thomas, M.G., Barnett, R., Allen, R., Sykes, N., Paxinos, P.D., Lebrasseur, O., Dobney, K., Peters, J., Manica, A., *et al.* (2017). Inferring allele frequency trajectories from ancient DNA indicates that selection on a chicken gene coincided with changes in medieval husbandry practices. *Molecular Biology and Evolution* **34**(8), 1981–1990.
- Malaspinas, A.-S. (2016). Methods to characterize selective sweeps using time serial samples: An ancient DNA perspective. *Molecular Ecology* **25**(1), 24–41.
- Malaspinas, A.-S., Malaspinas, O., Evans, S.N. and Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**(2), 599–607.
- Malaspinas, A.-S., Westaway, M.C., Muller, C., Sousa, V.C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J.Y., Crawford, J.E., *et al.* (2016). A genomic history of Aboriginal Australia. *Nature* **538**(7624), 207.
- Malmström, H., Gilbert, M.T.P., Thomas, M.G., Brandström, M., Storå, J., Molnar, P., Andersen, P.K., Bendixen, C., Holmlund, G., Götherström, A. and Willerslev, E. (2009). Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Current Biology* **19**(20), 1758–1762.
- Malmström, H., Storå, J., Dalen, L., Holmlund, G. and Götherström, A. (2005). Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution* **22**, 2040–2047.
- Marciniak, S. and Perry, G.H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics* **18**(11), 659–674.
- Martin, M.D., Jay, F., Castellano, S. and Slatkin, M. (2017). Determination of genetic relatedness from low-coverage human genome sequences using pedigree simulations. *Molecular Ecology* **26**, 4145–4157.
- Martiniano, R., Cassidy, L.M., Ó Maoldúin, R., McLaughlin, R., Silva, N.M., Manco, L., Fidalgo, D., Pereira, T., Coelho, M.J., Serra, M., *et al.* (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics* **13**(7), e1006852.
- Mathieson, I. and McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**(3), 973–984.
- Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E.R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J.L., de, J.M.B. Castro, Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M.A.R., Roodenberg, J., Vergès, J.M., Krause, J., Cooper, A., Alt, K.W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R. and Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**(7583), 499–503.
- Meyer, M., Arsuaga, J.-L., de Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., de Castro, J.M.B., Carbonell, E., *et al.* (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531**(7595), 504.
- Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Pruefer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson, N., Andres, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321–3323.

- Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**(6), 443.
- Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. and Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature* **541**(7637), 302.
- Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Platt, D., Paabo, S., Pritchard, J.K. and Rubin, E.M. (2006). Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**, 1113–1118.
- Nordborg, M. (1998). On the probability of Neanderthal ancestry. *American Journal of Human Genetics* **63**, 1237–1240.
- Ojalde, I., Schroeder, H., Sandoval-Velasco, M., Vinner, L., Lobón, I., Ramirez, O., Civit, S., García Borja, P., Salazar-García, D.C., Talamo, S., María Fullola, J., Xavier Oms, F., Pedro, M., Martínez, P., Sanz, M., Daura, J., Zilhão, J., Marquès-Bonet, T., Gilbert, M.T.P. and Lalueza-Fox, C. (2015). A common genetic origin for early farmers from Mediterranean Cardial and Central European LBK Cultures. *Molecular Biology and Evolution* **32**(12), 3132–3142.
- Omrak, A., Günther, T., Valdiosera, C., Svensson, E.M., Malmström, H., Kiesewetter, H., Aylward, W., Storå, J., Jakobsson, M. and Götherström, A. (2016). Genomic evidence establishes Anatolia as the source of the European Neolithic gene pool. *Current Biology* **26**(2), 270–275.
- Pääbo, S. (1985). Molecular-cloning of ancient Egyptian mummy DNA. *Nature* **314**, 644–645.
- Patterson, N., Price, A.L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**(12), e190.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. (2012). Ancient admixture in human history. *Genetics* **192**(3), 1065–1093.
- Pedersen, J.S., Valen, E., Velazquez, A.M.V., Parker, B.J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D.J., Kelly, T.K., Vang, S., et al. (2014). Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research* **24**(3), 454–466.
- Peter, B.M. (2016). Admixture, population structure, and F-statistics. *Genetics* **202**(4), 1485–1501.
- Pickrell, J.K. and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**(11), e1002967.
- Pinhasi, R., Fort, J. and Ammerman, A.J. (2005). Tracing the origin and spread of agriculture in Europe. *PLoS Biology* **3**(12), e410.
- Plagnol, V. and Wall, J.D. (2006). Possible ancestral structure in human populations. *PLoS Genetics* **2**, 972–979.
- Poinar, H., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R., Buigues, B., Tikhonov, A., Huson, D., Tomsho, L., Auch, A., Rampp, M., Miller, W. and Schuster, S. (2006). Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311**, 392–394.
- Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Onygerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J. and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**(7481), 43–49.
- Racimo, F. (2016). Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* **202**, 733–750.
- Racimo, F., Berg, J.J. and Pickrell, J.K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics* **208**, 1565–1584.
- Raghavan, M., Skoglund, P., Graf, K.E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T.W. Jr., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Roots, S., Maegi, R.,

- Campos, P.F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L.P., Fedorova, S.A., Voevoda, M.I., DeGiorgio, M., Sicheritz-Ponten, T., Brunak, S., Demeshchenko, S., Kivisild, T., Villems, R., Nielsen, R., Jakobsson, M. and Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**(7481), 87–91.
- Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., *et al.* (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**(7282), 757.
- Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., Jombart, T., Kivisild, T., Zhai, W., Eriksson, A., Manica, A., Orlando, L., De La Vega, F.M., Tridico, S., Metspalu, E., Nielsen, K., Ávila-Arcos, M.C., Moreno-Mayar, J.V., Muller, C., Dorch, J., Gilbert, M.T.P., Lund, O., Wesolowska, A., Karmin, M., Weinert, L.A., Wang, B., Li, J., Tai, S., Xiao, F., Hanihara, T., van Driem, G., Jha, A.R., Ricaut, F.-X., de Knijff, P., Migliano, A.B., Gallego Romero, I., Kristiansen, K., Lambert, D.M., Brunak, S., Forster, P., Brinkmann, B., Nehlich, O., Bunce, M., Richards, M., Gupta, R., Bustamante, C.D., Krogh, A., Foley, R.A., Lahr, M.M., Balloux, F., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Wang, J. and Willerslev, E. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**(6052), 94–98.
- Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W. Jr., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., Poznik, G.D., Gudmundsdottir, V., Yadav, R., Malaspina, A.-S., White, 5th, S.S., Allentoft, M.E., Cornejo, O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S., Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V.M., Lopes, M.C., Malhi, R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F., Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R. and Willerslev, E. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**(7487), 225–229.
- Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T.S., Moreno-Mayar, J.V., Poznik, G.D., Zollikofer, C.P., de León, M.S.P., Allentoft, M.E., Moltke, I., *et al.* (2015). The ancestry and affiliations of Kennewick Man. *Nature* **523**(7561), 455.
- Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P., Hublin, J.-J., Kelso, J., Slatkin, M. and Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060.
- Renaud, G., Slon, V., Duggan, A.T. and Kelso, J. (2015). Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology* **16**(1), 224.
- Renaud, G., Hanghøj, K., Willeslev, E. and Orlando, L. (2017). gargammel: A sequence simulator for ancient DNA. *Bioinformatics* **33**, 577–579.
- Renfrew, C. and Boyle, K.V. (2000). *Archaeogenetics: DNA and the Population Prehistory of Europe*. McDonald Institute for Archaeological Research, Cambridge.
- Rodrigo, A.G. and Felsenstein, J. (1999). Coalescent approaches to HIV population genetics. In K.A. Crandall (ed.), *The Evolution of HIV*. Johns Hopkins University Press, Baltimore, MD, pp. 233–272.
- Rodríguez-Varela, R., Günther, T., Krzewińska, M., Storå, J., Gillingwater, T.H., MacCallum, M., Arsuaga, J.L., Dobney, K., Valdiosera, C., Jakobsson, M., *et al.* (2017). Genomic analyses of pre-European conquest human remains from the Canary Islands reveal close affinity to modern North Africans. *Current Biology* **27**(21), 3396–3402.
- Rogers, A.R. and Bohlender, R.J. (2015). Bias in estimators of archaic admixture. *Theoretical Population Biology* **100**, 63–78.

- Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. and Reich, D. (2015). Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. *Philosophical Transactions of the Royal Society of London, Series B* **370**(1660), 20130624.
- Sánchez-Quijano, F. and Lalueza-Fox, C. (2015). Almost 20 years of Neanderthal palaeogenetics: Adaptation, admixture, diversity, demography and extinction. *Philosophical Transactions of the Royal Society of London, Series B* **370**(1660), 20130374.
- Sánchez-Quijano, F., Schroeder, H., Ramirez, O., Avila-Arcos, M.C., Pybus, M., Olalde, I., Velazquez, A.M.V., Marcos, M.E.P., Encinas, J.M.V., Bertranpetti, J., Orlando, L., Gilbert, M.T.P. and Lalueza-Fox, C. (2012). Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Current Biology* **22**(16), 1494–1499.
- Sankararaman, S., Mallick, S., Dannemann, M., Pruefer, K., Kelso, J., Pääbo, S., Patterson, N. and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**(7492), 354–357.
- Sawyer, S., Krause, J., Guschnski, K., Savolainen, V. and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7**, e34131.
- Sawyer, S., Renaud, G., Viola, B., Hublin, J.-J., Gansauge, M.-T., Shunkov, M.V., Derevianko, A.P., Pruefer, K., Kelso, J. and Pääbo, S. (2015). Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proceedings of the National Academy of Sciences of the United States of America* **112**(51), 15696–15700.
- Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics* **46**(8), 919–925.
- Schiffels, S., Haak, W., Paajanen, P., Llamas, B., Popescu, E., Loe, L., Clarke, R., Lyons, A., Mortimer, R., Sayer, D., Tyler-Smith, C., Cooper, A. and Durbin, R. (2016). Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nature Communications* **7**, 10408.
- Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Vicente, M., Steyn, M., Soodyall, H., et al. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**(6363), 652–655.
- Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B., Soodyall, H. and Jakobsson, M. (2012). Genomic variation in seven Khoi-San groups reveals adaptation and complex African history. *Science* **118**, 374–379.
- Schraiber, J.G. (2018). Assessing the relationship of ancient and modern populations. *Genetics* **208**(1), 383–398.
- Schraiber, J.G., Evans, S.N. and Slatkin, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics* **203**(1), 493–511.
- Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A., Willerslev, E., Krogh, A. and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**(1), 178.
- Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using paleoMix. *Nature Protocols* **9**(5), 1056.
- Seguin-Orlando, A., Schubert, M., Clary, J., Stagegaard, J., Alberdi, M.T., Prado, J.L., Prieto, A., Willerslev, E. and Orlando, L. (2013). Ligation bias in Illumina next-generation DNA libraries: Implications for sequencing ancient genomes. *PLoS ONE* **8**(10), e78575.
- Seguin-Orlando, A., Korneliussen, T.S., Sikora, M., Malaspina, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., Goebel, T., Westaway, M., Lambert, D., Khartanovich, V., Wall, J.D., Nigst, P.R., Foley, R.A., Lahr, M.M., Nielsen, R., Orlando, L. and Willerslev, E. (2014). Paleogenomics: Genomic structure in Europeans dating back at least 36,200 years. *Science* **346**(6213), 1113–1118.

- Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics* **3**, 1296–1308.
- Shapiro, B. and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: New insights from ancient DNA. *Science* **343**(6169), 1236573.
- Sikora, M., Seguin-Orlando, A., Sousa, V.C., Albrechtsen, A., Korneliussen, T., Ko, A., Rasmussen, S., Dupanloup, I., Nigst, P.R., Bosch, M.D., et al. (2017). Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science* **358**(6363), 659–662.
- Silva, N.M., Rio, J. and Currat, M. (2017). Investigating population continuity with ancient DNA under a spatially explicit simulation framework. *BMC Genetics* **18**(1), 114.
- Sjödin, P., Skoglund, P. and Jakobsson, M. (2014). Assessing the Maximum Contribution from Ancient Populations. *Molecular Biology and Evolution* **31**, 1248–1260.
- Skoglund, P. and Jakobsson, M. (2011). Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18301–18306.
- Skoglund, P., Götherström, A. and Jakobsson, M. (2011). Estimation of population divergence times from non-overlapping genomic sequences: Examples from dogs and wolves. *Molecular Biology and Evolution* **28**, 1505–1517.
- Skoglund, P., Malmström, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., Gilbert, M.T.P., Götherström, A. and Jakobsson, M. (2012). Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469.
- Skoglund, P., Malmstrom, H., Omrak, A., Raghavan, M., Valdiosera, C., Günther, T., Hall, P., Tambets, K., Parik, J., Sjögren, K.-G., Apel, J., Willerslev, E., Storå, J., Götherström, A. and Jakobsson, M. (2014a). Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* **344**, 747–750.
- Skoglund, P., Northoff, B.H., Shunkov, M.V., Derevianko, A.P., Paabo, S., Krause, J. and Jakobsson, M. (2014b). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2229–2234.
- Skoglund, P., Sjödin, P., Skoglund, T., Lascoux, M. and Jakobsson, M. (2014c). Investigating population history using temporal genetic differentiation. *Molecular Biology and Evolution* **31**, 2516–2527.
- Skotte, L., Korneliussen, T.S. and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**(3), 693–702.
- Slon, V., Viola, B., Renaud, G., Gansauge, M.-T., Benazzi, S., Sawyer, S., Hublin, J.-J., Shunkov, M.V., Derevianko, A.P., Kelso, J., et al. (2017). A fourth Denisovan individual. *Science Advances* **3**(7), e1700186.
- Steinrücken, M., Bhaskar, A. and Song, Y.S. (2014). A novel spectral method for inferring general diploid selection from time series genetic data. *Annals of Applied Statistics* **8**(4), 2203.
- Sverrisdóttir, O.O., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J.M., Arsuaga Ferreras, J.L., Götherström, A. and Thomas, M.G. (2014). Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Molecular Biology and Evolution* **31**(4), 975–983.
- Templeton, A. (2002). Out of Africa again and again. *Nature* **416**(6876), 45–51.
- Terhorst, J., Kamm, J.A. and Song, Y.S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**(2), 303.
- Vernot, B. and Akey, J.M. (2014). Resurrecting surviving neandertal lineages from modern human genomes. *Science* **343**(6174), 1017–1021.
- Vernot, B., Tucci, S., Kelso, J., Schraiber, J.G., Wolf, A.B., Gittelman, R.M., Dannemann, M., Grote, S., McCoy, R.C., Norton, H., et al. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**(6282), 235–239.

- Wagner, S., Lagane, F., Seguin-Orlando, A., Schubert, M., Leroy, T., Guichoux, E., Chancerel, E., Bech-Hebelstrup, I., Bernard, V., Billard, C., Billaud, Y., Bolliger, M., Croutsch, C., Čufar, K., Eynaud, F., Heussner, K.U., König, J., Langenegger, F., Leroy, F., Lima, C., Martinelli, N., Momber, G., Billamboz, A., Nelle, O., Palomo, A., Piqué, R., Ramstein, M., Schweichel, R., Stäuble, H., Tegel, W., Terradas, X., Verdin, F., Plomion, C., Kremer, A. and Orlando, L. (2018). High-throughput DNA sequencing of ancient wood. *Molecular Ecology* **27**, 1138–1154.
- Wakeley, J. (2008). *Coalescent Theory*. Roberts & Company, Greenwood Village, CO.
- Wall, J.D. (2000). Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154**, 1271–1279.
- Wall, J.D. and Kim, S.K. (2007). Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics* **3**(10), e175.
- Wall, J.D., Lohmueller, K.E. and Plagnol, V. (2009). Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular Biology and Evolution* **26**(8), 1823–1827.
- Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Steverson, L.S., Gignoux, C., Woerner, A., Hammer, M.F. and Slatkin, M. (2013). Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209.
- Wang, C., Szpiech, Z.A., Degnan, J.H., Jakobsson, M., Pemberton, T.J., Hardy, J.A., Singleton, A.B. and Rosenberg, N.A. (2010). Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical Applications in Genetics and Molecular Biology*, 9, e13.
- Wei, C.L.ß, Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E., Gould, B.A., Stinchcombe, J.R., Krause, J. and Burbano, H.A. (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *Royal Society Open Science* **3**(6), 160239.
- Whittle, A.W. (1996). *Europe in the Neolithic: The Creation of New Worlds*. Cambridge University Press, Cambridge.
- Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I.D., Schier, W., Thomas, M.G., et al. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences* **111**(13), 4832–4837.
- Wolpoff, M., Hawks, J. and Caspari, R. (2000). Multiregional, not multiple origins. *American Journal of Physical Anthropology* **112**, 129–136.
- Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

11

Sequence Covariation Analysis in Biological Polymers

William R. Taylor,¹ Shaun Kandathil,² and David T. Jones²

¹The Francis Crick Institute, London, UK

²University College London, London, UK

Abstract

Methods for the analysis of covariation between positions in large alignments of biological sequences have improved markedly over the last decade and, when combined with the large volumes of sequence data currently available, can now be used routinely to predict 3D macromolecular structure and interactions between molecules. A number of applications are reviewed, including: protein tertiary structure prediction for both globular and transmembrane proteins and disordered regions; protein–protein interactions; protein dynamics and RNA structure prediction. Recent developments in methodology are described, focusing on the application of machine learning approaches to the analysis of covariation signals, including deep learning (convolutional neural net) methods.

11.1 Introduction

Residues that make contact in a protein structure apply mutual selection pressure on the type of amino acids that can be substituted at each position. In its simplest form, the substitution of a positively charged residue at one position might increase the chance for a negatively charged residue to be substituted in an adjacent position. This idea of compensating substitutions has been current for as long as there have been two protein structures to compare. Observations by Kendrew and Perutz on their distantly related myoglobin and haemoglobin structures suggested that when an amino acid is substituted, an adjacent position might change to maintain a constant volume. The idea was developed by Klug and colleagues based on a virus coat protein for which there were seven homologous sequences (Altschuh *et al.*, 1987) and later was further generalised to allow for more general patterns of substitution (Taylor and Hatrick, 1994; Neher, 1994; Gobel *et al.*, 1994).

While these early studies established that the effect existed, the number of sequences at the time was too limited to allow the prediction of contacts that could be used to construct or provide much constraint on possible three-dimensional structures, even when taking account of phylogenetic effects (Pollock *et al.*, 1999). Despite some limited success predicting protein interactions (Pazos *et al.*, 1997) and tertiary structure (Bartlett and Taylor, 2008), the approach remained of limited use until around ten years ago when next generation sequencing (NGS) methods began to increase greatly the number of available sequences – where before

100 members would constitute a large protein sequence family, soon thousands of members became common.

This potential was augmented by new methods taking better account of significance level (Dunn *et al.*, 2008; Lee and Kim, 2009; Burger and van Nimwegen, 2008) and neighbour effects (Weigt *et al.*, 2009), which led to clearer contact predictions that had enough accuracy to help in the prediction and construction of tertiary structures. Applications, which will be reviewed briefly below, included globular proteins (Morcos *et al.*, 2012; Taylor *et al.*, 2012), membrane proteins (Hopf *et al.*, 2012; Nugent and Jones, 2012), inter-protein contacts (Hopf *et al.*, 2014; Ovchinnikov *et al.*, 2014) and even RNA (Eddy and Durbin, 1994; Taylor *et al.*, 2013; Weinreb *et al.*, 2016). A review of advances over this transition period can be found in Taylor *et al.* (2013) and de Juan *et al.* (2013), while Szurmant and Weigt (2017) provide a more recent review of protein–protein interactions.

The current work will focus mainly on recent methodological developments and emerging future directions with particular emphasis on machine learning approaches.

11.2 Methods

A number of methods to detect positional covariation or correlated substitution¹ are currently available as either servers or stand-alone programs that can be downloaded and run locally. Most servers will accept a single sequence and run standard search and alignment methods to generate the required multiple sequence alignment; however, an option is generally provided to upload a customised user-generated alignment.

While not an exhaustive list, the following sections give an overview of some of the more popular methods.

11.2.1 DCA Method

The ‘direct coupling analysis’ (DCA) method originally introduced by Weigt and co-workers allowed a theoretical maximum entropy approach (Lapedes *et al.*, 1999) to be put to practical use, with the solution of the contact weights achieved through a heuristic algorithm (Weigt *et al.*, 2009). Although the application in this paper was to protein–protein interactions, the intra-chain predictions were of sufficient quality to be useful in structure prediction (Sadowski *et al.*, 2011). Subsequent development of the method has involved the inclusion of filters on DCA (Cocco *et al.*, 2013). The original DCA method was provided as tools called EVcouplings and EVfold, with an additional tool dealing with contacts between proteins (described more below). The tools are provided as servers at <http://evfold.org/evfold-web/evfold.do>.

EVcouplings takes a single sequence or a user-supplied alignment and calculates couplings/contacts which can then be displayed on a structure, if known. For alpha-helical trans-membrane proteins, helices and topology can be additionally specified. Through an ‘advanced features’ menu, options to exercise additional control on the sequence search and other parameters and constraints, including predicted secondary structure, are provided. An option is also provided to compare the calculated contacts with known structures, but no algorithm is specified. Although the original DCA method (DI option) is retained, the current implementation uses, by default, the improved pseudo-likelihood method (PLM), called plmDCA (Ekeberg *et al.*, 2013). A stand-alone program called PLMC is also available (Hopf *et al.*, 2015).

¹ This approach is often referred to as *correlated mutation* analysis; however, as all the changes are accepted mutations, they are better referred to as *substitutions*.

The EVfold tool takes the calculated couplings/contacts and attempts to predict a three-dimensional structure from them following the approach described previously (Marks *et al.*, 2011).

11.2.2 PSICOV

To disentangle the direct from indirect couplings, the PSICOV method uses the sparse inverse covariance estimation method which is applied to the empirical covariance matrix between all amino acid types occurring at every alignment position (Jones *et al.*, 2012). Although this generates a large square symmetric matrix of rank $21L$ (where L is the length of the sequence and 21 refers to the 20 amino acids plus a gap symbol), the matrix is often singular due to missing data in the covariance matrix (i.e. unobserved amino acids in the alignment). PSICOV gets around this by trying to find an approximate inverse covariance or precision matrix. This is done by treating the problem as an optimisation problem, where an approximate precision matrix is found by attempting to solve the related linear equations approximately. Because an optimisation approach is used, other constraints can be applied, including the L_1 norm of the solution so as to emphasise a sparse solution. Because we know that contact maps are inherently sparse, by constraining the predicted maps to also be sparse, a great deal of noise reduction in the final output can be achieved. This method results in less loss of information compared to calculations that average over amino acid type and the predicted contacts were better than those achieved by other methods at the time, based on the prediction of contacts in proteins of known structure.

The current method called MetaPSICOV uses additional machine learning features and is more fully described below. The Jones group maintains a server at http://bioinf.cs.ucl.ac.uk/psipred_new/ which provides a simple input option to paste a single sequence with all subsequent database searches and alignment carried out automatically. Some basic help and documentation is provided.

11.2.3 plmDCA, GREMLIN and CCMpred

A more recent wave of covariation methods infer the direct residue interactions by estimating the couplings as parameters of a Markov random field by maximising their pseudo-likelihood (Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013). These PLMs are more accurate and run faster than some previous methods that require covariance matrix inversion. The approach underlies the plmDCA method (Ekeberg *et al.*, 2013), the GREMLIN method (Balakrishnan *et al.*, 2011; Ovchinnikov *et al.*, 2014, 2015) as well as the related CCMpred (Seemayer *et al.*, 2014) and PLMC (Hopf *et al.*, 2015) methods. The GREMLIN method is provided as a server at <http://gremlin.bakerlab.org/> giving access not only to methods to calculate contacts but also to extensive collections of pre-calculated models for both single proteins and complexes, including (where successful) models for the entire *E. coli* proteome and many of their interactions with each other.

The server takes a single sequence or an existing alignment (which can be uploaded), or, for calculating contacts between proteins, a pair of each. Options are provided to select the search method and vary two search parameters (E-value and the number of iterations), and recently the sequence databases have been extended to include metagenomic (or environmental sequencing) data (Ovchinnikov *et al.*, 2017). An additional pair of options can be used to control the degree of coverage of each hit and, for the remaining sequences, exclude regions that have too many gaps. Results are provided as lists of ranked contacts which are displayed as a contact map. This is also overlaid with the best matches found from a comparison with known structures (using methods described below). A unique and useful feature of this is to

display, in addition, contacts that arise from multimeric packing, which, according to a recent analysis, account for many of the longer predicted contacts between non-adjacent positions (Anishchenko *et al.*, 2017).

To find correlated pairs of positions between two different proteins requires the concatenation of two large multiple sequence alignments such that the sequences that are joined together are those that interact in their species of origin. When each protein is unique then the species name is sufficient to guide this match; however, given the vagaries of species nomenclature, a more direct option that is applicable to prokaryotes is to exploit the likelihood that interacting proteins will be co-located in an operon in the genome. The GREMLIN server uses a simple measure of ranked genome adjacency to infer sequences that should be concatenated. To control this, a parameter to limit the separation between genes (called Δ -gene) is provided.

In eukaryotes, which do not have operons, this simple principle cannot be applied and some recent attempts to address the sequence-pairing problem will be returned to below.

11.3 Applications

11.3.1 Globular Protein Fold Prediction

The main focus of applications of the correlated mutation approach has been in the prediction of the 3D (tertiary) structure of globular proteins. Although the approach does not address the ‘protein folding problem’ (as it does not model kinetics), it is the only practical method available to compute the tertiary structure given just protein sequence data.

Steric and geometric properties can be used to augment the prediction of structure, and some early methods relied heavily on these (Bartlett and Taylor, 2008), but as the methods improved and sequence data grew in volume, it became possible to calculate the structure in a more direct way, relying only on predicted contacts and stereo-chemistry, similar to the calculation of structure from nuclear magnetic resonance (NMR) data. However, even when the contact predictions are accurate, they remain relatively sparse and a unique solution to the restraints is not always, or easily, found – often requiring extensive sampling (Morcos *et al.*, 2012; Taylor *et al.*, 2012; Kamisetty *et al.*, 2013; Ovchinnikov *et al.*, 2015).

11.3.2 Transmembrane Protein Prediction

The approach is, of course, equally applicable to transmembrane structure prediction, and the longer secondary structure elements found in this class of protein, combined with their more limited packing arrangements (being confined in the membrane), have led to remarkably accurate predictions of quite large structures (Hopf *et al.*, 2012; Nugent and Jones, 2012).

11.3.3 RNA Structure Prediction

Covariation analysis of RNA sequences has been studied for almost as long as that of protein sequences, but with considerably better results (Eddy and Durbin, 1994). As the analysis of covariation between positions in a multiple sequence alignment depends only on the analysis of characters, it makes little difference to the calculation if these represent nucleotide bases rather than amino acids. This means that the plethora of current methods is equally applicable to RNA sequences or even a combination of RNA and protein (Weinreb *et al.*, 2016). Applications clearly show improved RNA secondary structure prediction (stemloops), but success with RNA tertiary structure prediction has been limited to simple structures with, typically, a few stemloops (De Leonardis *et al.*, 2015; Weinreb *et al.*, 2016).

The prediction of tertiary structure requires sequentially long-range base-pairing, and, with most bases ‘tied-up’ in local secondary structure pairings, there are relatively few, if any, found

between stemloops that can provide restraints on tertiary structure. However, when they occur, especially those in an extended interaction such as a pseudo-knot, the possible structures can be greatly restrained (Taylor and Hamilton, 2017; J. Wang *et al.*, 2017).

11.3.4 Protein Disordered Regions

One class of protein which is very hard to study by experimental structure determination methods is that of natively unfolded or disordered proteins (Van der Lee *et al.*, 2014). Perhaps the most interesting cases of disordered proteins are those proteins which adopt one or more ordered states upon binding to a cognate ligand or other protein chain, usually as part of a regulatory or signalling pathway in eukaryotes. This contextuality to the stable folding of these proteins makes them hard to study in the lab as the cognate ligands are usually unknown. This same contextuality also makes it hard to apply standard protein folding simulation techniques to this class of protein. Despite this complexity, it should be reasonable to assume that the ultimately stable states of a disordered protein should still leave their imprints in the covariation signals present in the multiple alignment for the family, assuming the family maintains reasonably similar function and binds the same set of cognate ligands. This is the basis behind the interesting study by Toth-Petroczy *et al.* (2016), who apply the PLMC method to alignments of disordered protein families and then attempt to determine one or more 3D structures based on the predicted contact data. One practical issue which the authors attempt to deal with is the difficulty of accurately aligning families of disordered proteins, where low-complexity sequences are found to be commonplace. Once this issue has been dealt with, the rest of the approach is more or less identical to the standard EVfold method, though in this case looking for more than one prominent conformation in the resulting ensemble of generated 3D models.

One difficulty in assessing the results is that there are only a very small number of cases where multiple binding conformations are either known or suspected from X-ray or NMR structure determination. This means that the authors were only able to validate their approach on 60 probable disordered protein binding interactions. Nevertheless their recovery of 79% of the binding residue interactions is very encouraging. Based on this, the authors went on to run a survey of 1000 predicted disordered regions in human proteins and find that, using their approach, at least half of the binding regions show signs of having one or more stable conformational states, that is, there are predicted contacts which suggest either stable tertiary or at least stable secondary structure, though somewhat disappointingly, they do not go as far as trying to build actual 3-D models of the regions. Of course, these observations remain highly speculative until experimental methods can be applied to verify some of the predictions, but nonetheless the results are extremely thought-provoking.

11.3.5 Protein–Protein Interactions

Some of the earliest applications of correlated mutation analysis were to the prediction of contacts between pairs of interacting proteins (Pazos *et al.*, 1997; Weigt *et al.*, 2009). The improved efficiency of recent methods, combined with increased numbers of sequences, has led to large collections of protein interaction models, many of which have been verified by known structures of the complex (Ovchinnikov *et al.*, 2014; Hopf *et al.*, 2014).

Using correlated substitution analysis to find residue contacts between interacting proteins requires that the sequences from each family are correctly paired up with each other, since only when the two proteins coexist and interact in the same organism will they be subject to the co-evolutionary pressures that give rise to the correlated substitution signal.

When each protein is unique, the species name is effectively all there is to guide this match; however, when there are multiple related sequences (paralogues) in each species, the pairing is

more difficult. As mentioned above, in bacteria a good guide can be gained from genome co-location as interacting proteins tend to be in a common operon but in eukaryotes this simple principle is not an option.

Some new methods that address the eukaryotic sequence matching problem will be returned to below.

11.3.6 Allostery and Dynamics

One of earlier methods of correlated mutation analysis, called ‘statistical coupling’, has been largely directed towards investigating how conformational changes are communicated through a protein structure (Lockless and Ranganathan, 1999; Süel *et al.*, 2003; Reynolds *et al.*, 2011). Despite much experimental support, for example, involving exhaustive site-specific mutagenesis of a small protein (McLaughlin *et al.*, 2012), the significance of using correlated mutation analysis to study allostery has been questioned (Livesay *et al.*, 2012). The authors of this review point out, among other things, that even close homologues can exhibit different allosteric mechanisms and, as all these sequences are combined in a multiple alignment, any effects should be lost through averaging.

While allosteric effects typically involve subtle changes of conformation, other protein mechanisms can involve very large movements. These may be apparent in the correlation signal depending on the degree to which the various conformational states are critical for function. For example, if a protein has two states in which a site is either open or closed, then if residue contacts are only formed in the closed state, that is all that will be seen in a correlation analysis (Figure 11.1). If the open state is also stabilised by contacts, then both sets will appear as predicted contacts that cannot be reconciled in a single molecular model. In both situations, molecular dynamics can be used to correlate the observed contacts with the conformational space explored by the protein structure (Morcos *et al.*, 2013) or used in a more general analysis of flexibility and folding (Sutto *et al.*, 2015). Large motions are also associated with the proteins that comprise molecular motors, and associated correlations of sequence have been seen in both the bacterial flagellum motor proteins (Pandini *et al.*, 2016) and the mitochondrial ATPase (Pandini *et al.*, 2015).

Long-range contact predictions, however, may not necessarily be explained only by motion within a single chain as contacts across subunits in a multimeric assembly can generate equivalent observations, with a recent survey indicating that many, if not most, can be accounted for in this way (Anishchenko *et al.*, 2017).

11.3.7 CASP

The community-wide experiment on the Critical Assessment of techniques for protein Structure Prediction (CASP) runs every two years. During the prediction season, groups interested in predicting protein structure attempt to predict the structure of target sequences in a blind manner. The targets are obtained from groups who have already obtained, or are about to obtain, a structure for these proteins by experiment. Crucially, many of these proteins have no structural relatives in publicly available databases, and so the experiment serves as a rigorous test of the ability of current methods to predict structures reasonably. Historically, this task of *free modelling* has been extremely challenging, with success being limited to small domains (shorter than around 120 residues).

Recently, the GREMLIN method was used by the Baker group to assist in the prediction of a structure (CASP target ID T0806; PDB ID 5CAJ – see Figure 11.2) for which no template structures were detectable by sequence similarity (Kinch *et al.*, 2016; Ovchinnikov *et al.*, 2016).

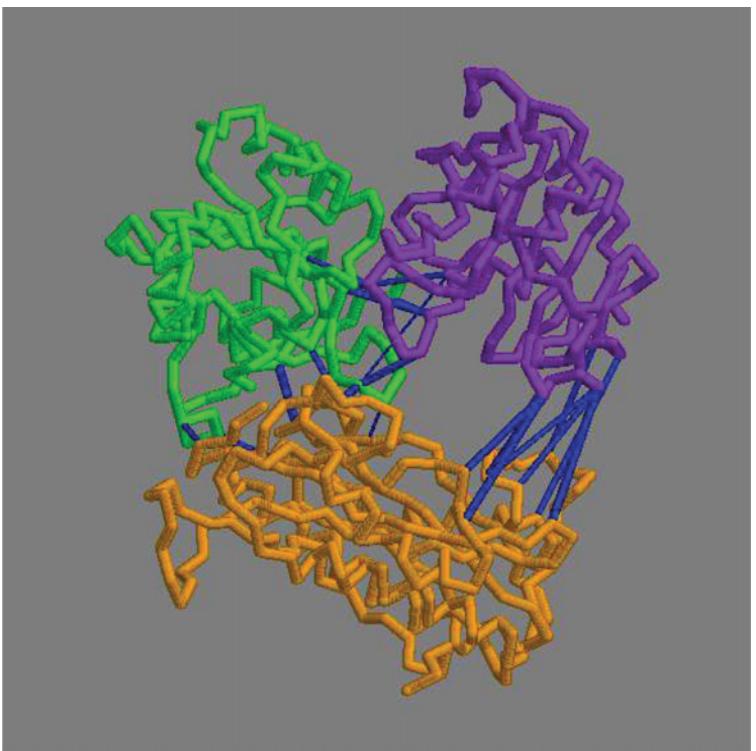


Figure 11.1 Long predicted contacts suggest domain motion. The structure of the cytosolic X-prolyl aminopeptidase, 3ctz, has three β/α domains (coloured purple–green–orange, sequentially), with the largest C-terminal domain terminating in two α -helices that do not contribute to the domain interfaces and were removed for clarity. The strongest contacts between the domains (predicted by the CCPred method) are shown as blue lines— with the thickest being strongest. The contacts between the first and last domains (purple/orange) are between residues that do not make close contact in the crystal structure and suggest that a hinge motion may bring them closer in their functional state.

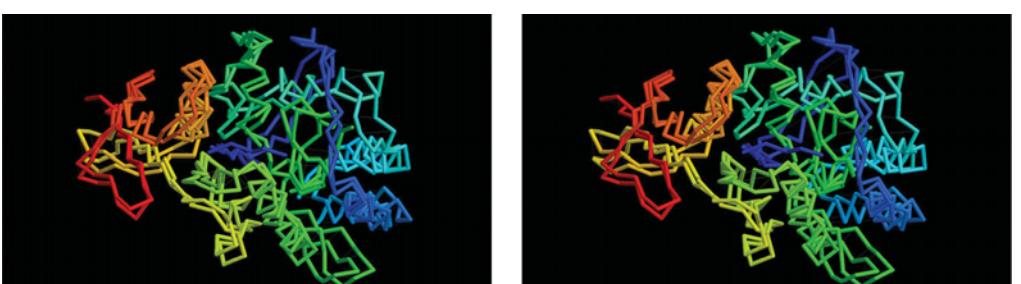


Figure 11.2 Predicted model for CASP target T0806. This target had no known template structures to provide a scaffold for prediction. The top model produced by the group of David Baker (University of Washington, USA) used an *ab initio* folding method (Rosetta) guided by pairwise residue distance restraints calculated by their GREMLIN method (described in the text). The top predicted model was a remarkably close match to the known structure (revealed after completion of the prediction exercise), despite the structure being a novel fold of considerable complexity. The predicted model even captured the unique amino terminal part of the chain (blue in the figure) that makes an unusual connection over the 'top' of an alpha helix. The model is shown superposed on the known structure as a (cross-eyed) stereo pair with both chains coloured from blue through the spectrum to red at the carboxy terminal. The only weak aspect of the prediction is that a pair of core beta-strands were modelled as an alpha-helix; nevertheless, the path of the chain (or 'topology') is correctly predicted with an RMS deviation under 5 Å over most of the chain. (See plot in Figure 11.3.)

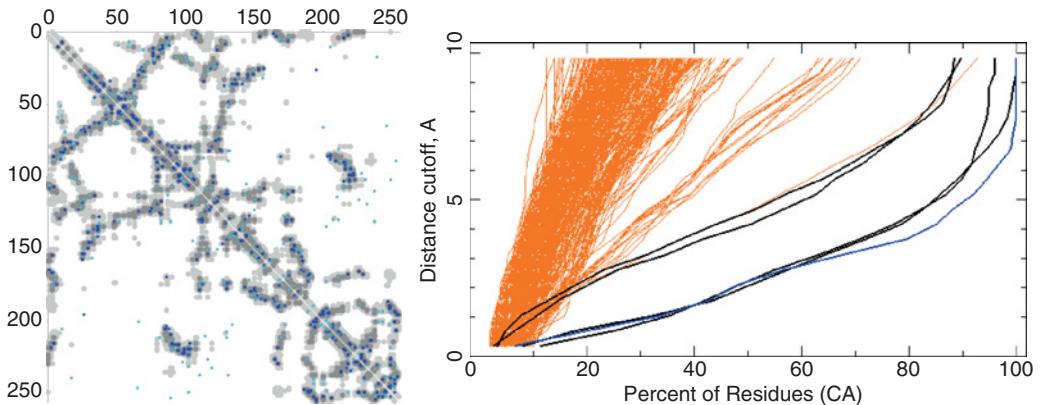


Figure 11.3 CASP target T0806 prediction data. *Left:* Shows the contact map for CASP target T0806 with the contacts from the (now) known structure plotted as grey dots and the contacts predicted by the GREMLIN method overlaid as blue dots of varying shade indicating strength of belief (dark = stronger). *Right:* The cumulative root mean square (RMS) deviation is plotted for the best predicted model (blue line), with the other models submitted by the Baker group plotted as black lines. In this plot the traces for the best models lie lower and further to the right (i.e. more positions giving a low RMS deviation). Models submitted by other groups participating in the CASP exercise are plotted as orange lines. Only the top model of the Jones group (University College London) approaches the Baker result. (Single orange line in the second-ranked cluster.)

The covariation signals were converted into residue–residue distance constraints and incorporated into their fragment-assembly program Rosetta Ab initio to produce structures. Among the CASP community, the prediction of T0806 is widely considered a breakthrough and shows that contacts predicted by residue covariation analyses can be used to successfully predict tertiary structure for reasonably large proteins (T0806 was 256 residues long). Of course, the prediction relied on the availability of sufficient sequence homologues of the target in order to predict contacts reasonably accurately – something that cannot always be obtained.

The ability of various groups to precisely predict residue–residue contacts has itself been assessed in a separate category in CASP. In recent years, some machine learning-based approaches (discussed below) have proven especially successful (Buchan and Jones, 2018; Wang *et al.*, 2018).

11.4 New Developments

11.4.1 Sequence Alignment

To collect the sequences required for analysis, the servers described above use standard sequence databank search tools such as PSI-BLAST (Altschul *et al.*, 1997), HMMER (Finn *et al.*, 2015) or HHblits (Remmert *et al.*, 2012). As well as searching, these methods compile a sequence alignment based on alignment to the probe sequence or profile – which generally evolves over a number of iterations. This ‘greedy’ approach to multiple sequence alignment can lead to errors, and in an iterated approach this can build to an error ‘catastrophe’ in which a misalignment allows unrelated sequences to become incorporated in the profile, leading to further corruption. This process, called *profile drift*, can easily degrade any correlation signal.

11.4.1.1 Family Membership Validation

While the current abundance of sequence data is of great benefit to the correlated substitution approach, it also creates a problem with checking that a sequence collection is valid and has not been corrupted by the inclusion of erroneous members. The correlation analysis not only

requires a large number of sequences (ideally, thousands) but also needs these to be well aligned and restricted to proteins that share a common structure. When dealing with many thousands of sequences, it is very difficult to verify these properties either 'by eye' or automatically.

To help with this alignment quality control problem, clustering methods, such as BLASTClust and CD-hit (Li *et al.*, 2001, 2012; Wei *et al.*, 2012), can be used to identify orphan sequences or possible false sub-families that might have become included through profile drift. However, whether these are true or false family members must still be decided subjectively. A statistical approach to aid in this decision is provided by the OD-seq analysis tool (Jehl *et al.*, 2015) which takes an alignment, distance matrix or set of unaligned sequences and finds outliers by examining the average distance of each sequence to the rest. Anomalous average distances are then found using the interquartile range of the distribution of average distances or by bootstrapping them. The algorithm has $O(N^2)$ complexity in the number of sequences, but this can be reduced using the mBed method in the CLUSTAL-Omega program to deal with large numbers of sequences – up to 100,000 (Sievers and Higgins, 2014).

Following similar lines, a method called MULSEL combines fast peptide-based pre-sorting of the sequence collection with a following cascade of mini-alignments, each of which are generated with the robust profile/profile method. From each mini-alignment, a representative sequence is selected, based on a variety of intrinsic and user-specified criteria that are combined to produce the sequence collection for the next cycle of alignment. For moderate sized sequence collections (tens of thousands) the method executes on a laptop computer within minutes. As similar sequences are progressively eliminated, those that remain are the most distantly related in the collection and can be identified either visually, using a variety of analysis methods combined with colouring schemes, or automatically by setting a cutoff on the lower limit to which sequences/profiles can be aligned. For example, the large rhodopsin family of 180,000 sequences was reduced to 14 representatives that could easily be verified as all being likely to have a common core structure (Taylor, 2016b).

11.4.1.2 Alignment Benchmarking and Improvement

Given a verified collection of sequences, the evaluation of whether these are all correctly aligned over their full length remains a difficult problem. Much effort has been invested in improving multiple sequence alignments over a long period of time, and one of the problems in this area is what the actual true reference alignment should consist of. The 'gold standard' that is generally taken is to use examples where the structure of the proteins is known, but even structure alignment can be ambiguous, especially in the loop regions, and, on the basis of a strict phylogenetic analysis, it has been questioned whether all such regions should even be considered as alignable (Loytynoja and Goldman, 2005).

In a meaningful alignment, especially one that is to be used to extract correlated positions, the only positions that should be aligned are those that are functionally equivalent and, at the risk of being circular, these should include those that exhibit covariation. This approach was used by Higgins and co-workers to formulate an alignment benchmark method called ConTest that uses a combination of the EVfold and PSICOV methods to identify positions that should be aligned. Using this criterion, they then evaluate a number of existing multiple sequence alignment methods (Fox *et al.*, 2016). However, the authors do not take the next step and use the degree of covariation signal as an evaluation function to refine the alignment itself. Given the circularity in such an iteration, the capacity for creating artefacts through such a scheme may be a danger.

11.4.2 Comparison to Known Structures

As mentioned above in the survey of methods, some of the web servers provide an option to match the predicted contacts against a database of known structure. Such an approach can

be used to detect so-called *analogous* folds, that is, proteins which share significant structural similarity but little or no sequence similarity, which can be used as templates for structural modelling. Two of these methods have been described in detail in the literature and both use different comparison methods.

The GREMLIN server automatically matches the predicted contacts using a variation of Taylor's iterated double-dynamic programming algorithm (Taylor, 1999). As the name suggests, this algorithm, called map_align, comprises two dynamic programming steps. In the first step, a score is computed for each row (corresponding to a set of specific residue contacts) of the first contact map with each row for the second contact map and dynamic programming is used to find the alignment of the contacts. These optimised sums of scores are then entered in a second matrix, and the optimal contact alignment is then found again by dynamic programming. At this point, however, the scores for individual row–row comparisons are overestimates since in the first step the alignments for each pair are independent. The second step matrix is then updated based on the current alignment. This process is repeated 20 times, after which the alignment has converged. The initial estimate of the similarity matrix is critical in getting at least part of the alignment correct as this serves as a nucleation point for aligning the rest of the contacts. To maximise the chance of success, a number of variations in the first step are tried. The full algorithm, along with pseudo-code and comparison to other methods, is described in the supplementary material accompanying the paper, and it was found to give better results than a number of alternative methods (Ovchinnikov, *et al.*, 2017). The results of matching are presented by the GREMLIN server as a ranked list of hits, each of which can be overlaid in a contact map with the predicted contacts.

The alternative comparison method of Di Lena *et al.* (2010) is used in the EigenTHREADER method of Buchan and Jones (2017). In this method, two contact maps can be aligned using the principle that most of the information in real symmetric matrices can be approximated using a subset of their eigenvectors corresponding to the largest eigenvalues. Then, using a Needleman–Wunsch-like procedure to align the truncated eigendecompositions, one obtains a fast and reasonably accurate way to search a database of contact maps. EigenTHREADER uses contact maps predicted by MetaPSICOV (Jones *et al.*, 2015) as queries for the search procedure.

11.4.3 Segment Parsing

The contact maps that are generated by the correlation analysis method are typically sparse as not all residues that make contact generate a correlation signal. To aid in their interpretation, knowledge of the local secondary structure state can be used, but as the protein structure will typically be unknown, the secondary structure needs to be predicted. Conversely, the distribution of predicted contacts can be employed to help in the prediction of secondary structure.

This interaction between these two sources was captured in a novel algorithm that simultaneously optimised the prediction of secondary structure segments to maximise the fit of the predicted contacts with those expected from interactions between the secondary structure elements (Taylor, 2016a). Packed secondary structures of like type (e.g. both alpha or both beta) appear in the contact map as a short stripe with close to unit slope (which is either positive for parallel interactions or negative for antiparallel). However, mixed alpha/beta packing has a slope determined by the ratio of the helical rise of the secondary structure type, and the method, called BOXSET, takes this into account when parsing the interactions.

An example application to an alpha-helical transmembrane protein is shown in Figure 11.4 using two sub-families with differing degrees of prediction quality. The algorithm was also applied to predict transmembrane helical interactions in proteins of unknown structure in the core of the bacterial flagellum motor (Taylor *et al.*, 2016).

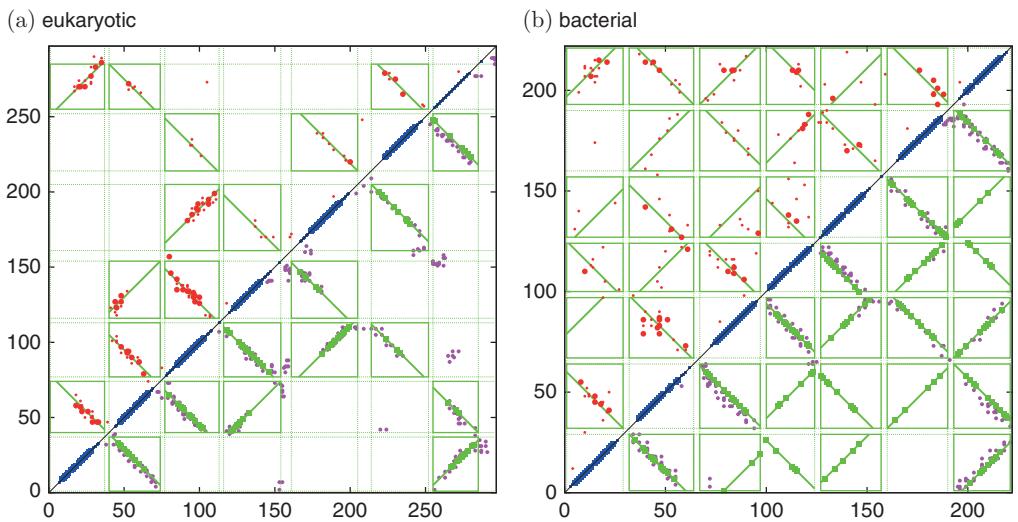


Figure 11.4 Rhodopsin parsed contact maps. The transmembrane helical segments predicted for rhodopsin are plotted in blue along the diagonal of each contact plot (with a larger mark where all methods agree). Each pair of segments has an associated interaction box (green), the boundaries of which have been iteratively refined to maximise a score based on the predicted contacts (red dots, top left). Those that fall within a box are analysed to determine a preferred packing orientation (green lines). In the lower right half of the plot, the lines are marked at the points corresponding to the orthogonal projection of the predicted contacts onto the line. For comparison, the contacts observed on the known structure are marked with magenta dots. The criss-cross pattern of these lines reflects the up-down alternate packing of the helices as they traverse the membrane. Panel (a) shows data for the large eukaryotic rhodopsin family (PDB code 1gzm) and (b) shows a similar plot for the bacteriorhodopsin protein (PDB code 2brd), with both sets of contacts predicted by the GREMLIN method (see text). The proteins have an equivalent structure but the prediction for the smaller bacterial family is much less accurate.

11.4.4 Machine Learning

Complex relationships between variables can be modelled using supervised machine learning methods, whereby the parameters of a model can be tuned in response to curated *training sets* of data. An advantage of such methods (in contrast to expert-imposed models such as DCA or PSICOV) is that they can learn properties of the data across distinct examples; in the case of contact prediction, this means that predictions for a new instance can be made using patterns learned from distinct protein families.

Another issue is that different contact predictors can often produce contact sets with only modest amounts of overlap (Jones *et al.*, 2012). Therefore, it seems reasonable that combining their predictions in some way may lead to increased precision. This can either be done as a simple average (Taylor, 2016a) or machine learning methods can be used to combine or aggregate predictions in a more sophisticated way.

11.4.4.1 PconsC

An example of such an approach, PconsC, uses the outputs from plmDCA (Hopf *et al.*, 2015) and PSICOV (Jones *et al.*, 2012) as input features, and correlates these to contacts using a random forests classifier (Skwark *et al.*, 2013). A total of eight sets of inputs were derived using the two contact prediction methods, with either PSI-BLAST (Altschul *et al.*, 1997) or jackHMMER (Finn *et al.*, 2015) being used to generate the input alignments, with four different E-value cutoffs.

A later version of the approach added input features such as sequence profiles, the sequence separation between the residue pair being evaluated, predicted secondary structure and solvent accessibility as input features (Skwark *et al.*, 2014). The method also used a stacked set of random forests classifiers (similar to stacked layers in a neural network), where a subset of each layer's output is added to the initial input feature set and used as the next layer's input.

The most recent version (Michel *et al.*, 2017) uses a single input alignment generated by HHblits (Remmert *et al.*, 2012) instead of the original eight, and eschews PSICOV in favour of two other contact predictors, namely GaussDCA (Baldassi *et al.*, 2014) and PhyCMAP (Wang and Xu, 2013).

11.4.4.2 MetaPSICOV

The MetaPSICOV method (Jones *et al.*, 2015) combines a large input feature set with two artificial neural network models to make predictions. The first neural network operates on the input features to produce an initial set of predictions, while the second network takes in a subset of the original input features, as well as the output of the first network to refine the predicted contact set. Contacts are predicted for one residue pair at a time.

The input feature set for the first neural network includes, among others: sequence composition statistics in windows of nine residues around each residue, as well as in a five-residue window at the midpoint between their two residues; predictions of secondary structure and solvent accessibility in each of these columns; raw and normalised mutual information, PSICOV (Jones *et al.*, 2012), mfDCA (Kaján *et al.*, 2014) and CCMpred (Seemayer *et al.*, 2014) scores for the residue pair in question; 16 features which encode sequence separation in various ranges; and a number of global features of the alignment, such as the logarithm of the number of effective sequences, and global fractions of each predicted secondary structure type.

The second neural network takes in the predicted scores from the first stage and considers windows of 11 residues centred on each residue when making the final predictions. Additionally, the second-stage network uses some of the original input features used by the first-stage network as well.

The method has proved to be very successful in recent CASP exercises (reviewed above).

11.4.5 Deep Learning Methods

A number of successful methods for contact prediction have taken advantage of recent advances in so-called *deep learning* (LeCun *et al.*, 2015), which broadly refers to the use of artificial neural network models employing more layers of hidden units than were traditionally feasible. The use of deeper models has been enabled by advances in methodology, the wide availability of easy-to-use programming frameworks, and advances in computing hardware as many state-of-the-art models can make efficient use of multiple graphical processing units to train on very large data sets, on practical timescales. A key advantage of using deeper network models is that hidden units in later layers can aggregate and compose the outputs of prior layers into higher-level features, which can model more complex relationships between the inputs and outputs of the network. This is sometimes termed *higher-level abstraction* in the literature. Although there are some additional factors to bear in mind when handling deeper networks with large numbers of tunable parameters and large data sets, the procedures for training deep networks remain broadly the same as for their shallow counterparts, namely an iterative procedure consisting of:

- a forward pass of some or all training examples through the network with its parameters (feeding the network inputs and getting the outputs);
- calculation of the cost or loss (deviation from the correct answer);

- back-propagation of the loss and calculation of its gradient with respect to each network parameter; and
- using the gradients to alter the parameters to define a new network state.

11.4.5.1 Convolutional Neural Networks

Some types of deep networks are particularly effective on data exhibiting various types of structure. An example of such structure is *spatial* structure, where nearby values of features in the input and/or output are strongly correlated such that the presence or absence of such local patterns in the data may be indicative of some property. Convolutional neural networks (LeCun *et al.*, 1998) are especially effective at detecting such local structure. They employ special layers of hidden units, called convolutional layers, in which individual units receive activations from only a subset of units in the previous layer. This is in contrast to ‘traditional’ neural network models where each hidden unit receives activations from every unit in the previous layer. The detection of local spatial patterns is achieved by tuning the values of a small 2D or 3D array of weights (referred to as a convolutional *kernel* or *filter*), with an additional bias parameter. The kernel or filter is usually smaller in spatial size than the input, and so is ‘swept’ across the input (as in a discrete convolution; hence the name ‘convolutional’ for this type of layer); see Figure 11.5.

The activation for each output unit in a convolutional layer is obtained by using the result of individual placements of the convolutional filter on the input. The weights in each filter are adapted to produce high activation values when the pattern in the weights of the filter matches a corresponding pattern in the input. This mechanism allows the convolutional layer to detect local spatial features in the input regardless of their precise location in the input (an example would be the recognition of a face regardless of whether it was in the left or right half of an image). The collection of units and activations from a convolutional layer is termed an *output feature map*. A single convolutional layer can be used to produce multiple output feature maps (by learning multiple convolutional filters) and so can be used to learn a number of different local patterns. Stacking convolutional layers allows complex models of spatial patterns to be built at increasing levels of abstraction, while using fewer free model parameters as compared to traditional, fully connected architectures.

RaptorX-contact (Wang *et al.*, 2017) is an effective example of a contact predictor employing deep convolutional networks. The model combines features that can be treated as one-dimensional (sequence profiles, predicted secondary structure and solvent accessibility) and two-dimensional (mutual information, CCMpred scores and contact potentials). This is achieved by the use of two deep neural networks, one operating on the 1D features, and one operating on 2D features. The output of the 1D network is transformed into a 2D set of features, and added to the other 2D input features before being used as input for the 2D network. The transformation from 1D to 2D features is performed by an operation similar to an outer product: the 1D vectors outputted for each residue from the first network are concatenated, giving a new 1D vector for each residue pair. This collection of vectors can be reshaped into a 2D array and added to the regular 2D features before being fed as input into the second deep net. The 1D and 2D networks in RaptorX-contact are composed of so-called *residual blocks* stacked one above the other. Each block consists of stacked activation² and convolutional layers with a *residual connection*, which adds the input to a given block of layers to its output. The 1D network uses a total of six 1D convolutional layers, each employing a filter size of 17, whereas the

2 An activation layer simply performs a predefined nonlinear transformation on its input, with no tunable parameters. In the case of RaptorX-contact, the rectified linear unit (ReLU) function is used.

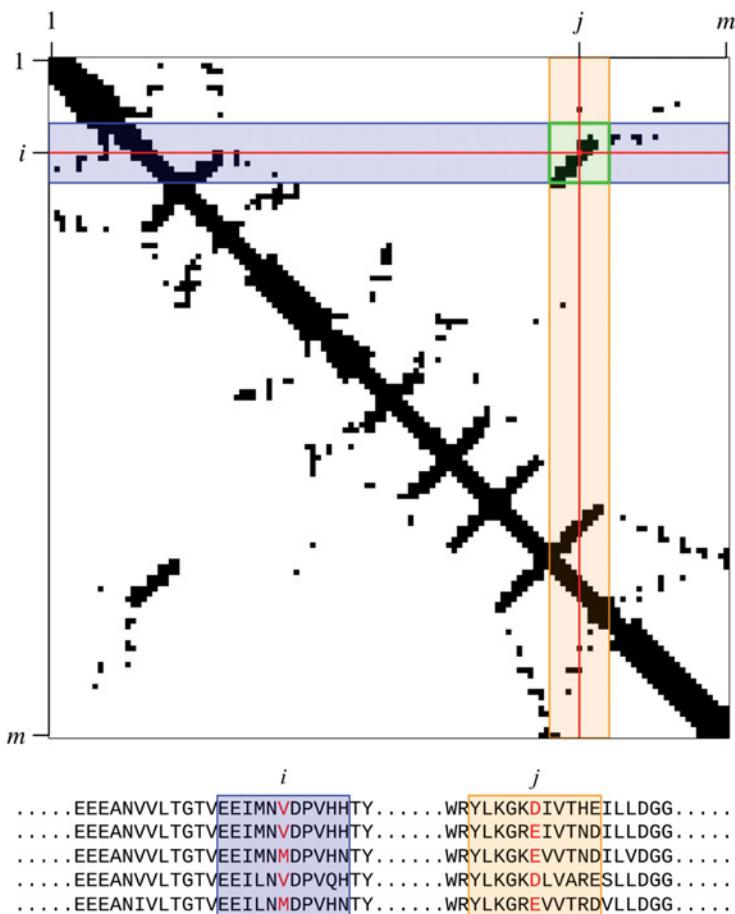


Figure 11.5 Spatial patterns in sequence-derived data and contact maps. The figure shows the relationship between pairwise (2D) sequence data and the columns of the alignment from which they are derived. For an alignment of m columns, pairwise data (and contact maps) form square matrices of dimension $m \times m$. When predicting a contact between a pair of residues (i, j), it can be desirable to look at the local pattern of contacts (and sequence statistics) in the vicinity of this residue pair (represented by the green box). This set of pairwise inputs or outputs corresponds to a pair of windows in the source alignment (blue and orange). A contact predictor can be trained by moving or sweeping the windows across the input and adjusting model parameters to match the corresponding output (contact maps). This approach can be implemented using convolutional neural networks, which efficiently implement this ‘sweeping’ operation.

2D network uses a variable number of 2D convolutional layers (the authors find the optimum number to be around 60) with filter sizes of 3×3 or 5×5 , with each layer generating 60 output feature maps (and hence learning 60 different convolutional filters each).

11.4.6 Sequence Pairing

Given the importance of identifying and modelling protein–protein interactions in eukaryotic cells, some new methods have been developed to address this problem. The most direct approach is to avoid the problem by mapping the eukaryotic proteins to a distantly related prokaryotic interaction (Rodriguez-Rivas *et al.*, 2016); however, this cannot be done for all interactions. As a step towards a more general solution, two similar methods try to directly

iterate the pairing of sequences using the predicted contacts as an evaluation function (Gueudré *et al.*, 2016; Bitbol *et al.*, 2016), while another uses just the phylogenetic structure of the families (Taylor, 2017).

11.4.6.1 Direct Contact Iteration

To calculate a set of predicted contacts while varying all possible pairings between paralogues in each species would be an enormous calculation, so both direct methods take an iterative refinement approach combined with a 'greedy' approach.

Bitbol *et al.* tested their method on the bacterial histidine-kinase/response-regulator system. These very large families were reduced to around 5000 sequences from each family and restricted to a fragment of the kinase comprising the alpha-helical hairpin that acts as a 'docking-platform' for the response-regulator. Despite these reductions, the run-time of the method remained considerable.

Gueudré *et al.* also tested their method on the histidine-kinase signalling system as well as proteins from the bacterial tryptophan synthesis operon, comprising the protein pairs TrpA/B and TrpE/G which interact. As with the previous method, the sequences used were fragments under 50 residues.

11.4.6.2 Phylogenetic Similarity

The phylogenetic approach (Taylor, 2017) was based on comparing sets of inter-sequence distances within one protein family of paralogues to the equivalent set of distances between the parologue sequences in another protein family. In theory, if evolution behaved in a regular manner and these distances could be trusted to a high degree of precision, then matching the two networks of distances would give a solution to the correct pairing of proteins. However, as the distances between sequences are not precisely replicated in the two families, the number of possible pairings between the two sets becomes very large since, with even a small degree of error, it becomes possible to match almost any sequence in the first set with any other in the second. This problem was overcome in a manner similar to the way in which phylogenetic trees are rooted by the inclusion of an outgroup.

In this problem, an 'outgroup' was added to each family from a species that had only one representative of each family. Since these sequences are unique (labelled), they can be immediately paired providing a reference point against which the other (unlabelled) sequences can be compared. Multiple different 'outgroups' were tried and a consensus pairing generated over the paralogues. In principle just one 'outgroup' species should be sufficient; however, it may have a poor (skewed) relationship with the other sequences, and multiple selections give a better overall coverage.

The final set of matchings was obtained using a bipartite graph-matching algorithm, similar to that used before in the equivalent problem of pairing secondary structure elements (Taylor, 2002).

11.4.7 Phylogeny Constraints

It had previously been realised that phylogenetic structure must contribute to the degree of observed covariance, but early attempts both to quantify (Pollock and Taylor, 1997) and exploit this effect (Pollock *et al.*, 1999) were of limited value due to the lack of sequence data available at the time. While such an effect must exist, later theoretical analysis has cast uncertainty over the underlying evolutionary mechanism (Talavera *et al.*, 2015).

The problem has recently been revisited, and with modern correlation methods and increased sequence data it has been possible to detect an underlying feature of the data that

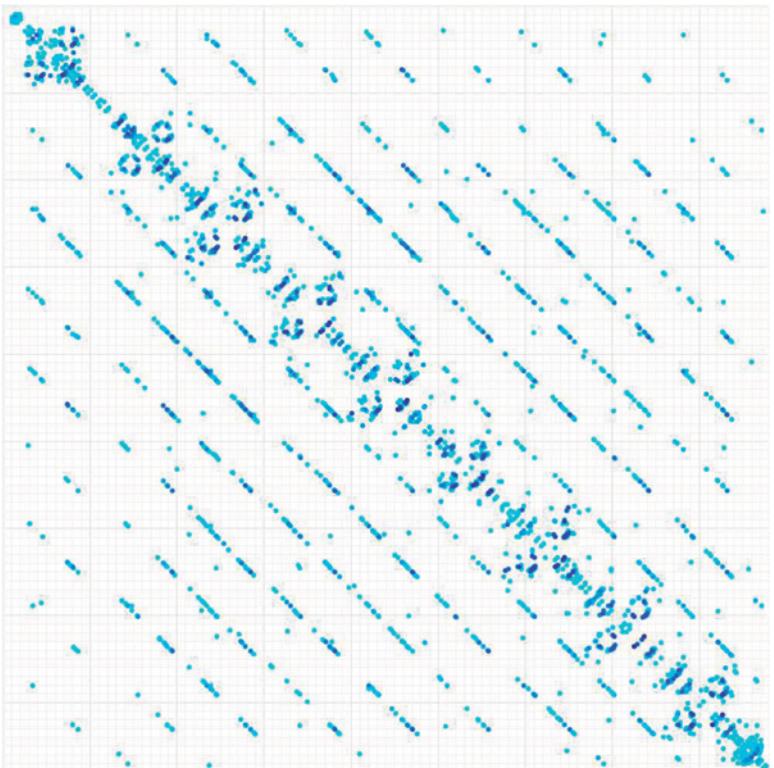


Figure 11.6 False predicted contacts from repeated domains: the contact map produced by GREMLIN for a segment of the sequence of fibronectin covering a number of type III repeats. A contribution to the pattern of domain-spaced parallel lines will also come from staggered misalignments of closely related or even identical sequences compiled by the search program to form the multiple alignment.

depends on both the number of sequences and the branch-lengths of their phylogenetic relationship (Qin and Colwell, 2018). Although at an early stage, the reconsideration of the phylogenetic structure of the data should lead to future improved methods.

While the studies mentioned above deal with phylogenetic structure among the sequences, a related problem concerns phylogenetic structure within the sequences. If a domain has recently duplicated and occurs in all members of the family across species, then equivalent positions in each domain will appear to be highly correlated even though they may form no contacts. The effect can be seen to dramatic effect in the multiple domains found in the large protein fibronectin in which parallel diagonal lines of covariation appear at increasing domain separations. (Figure 11.6). The degree to which this effect may contribute to less obvious duplications remains unquantified.

11.5 Outlook

Over the last decade, the method of covariance analysis of multiple aligned biological sequences has been transformed from a fringe activity of marginal interest and doubtful relevance into a mainstream tool providing accurate predictions of contacts within protein and RNA chains and interactions between molecules. Predicted contacts within a chain have shed light on both

structure and dynamics while predictions between proteins have informed on subunit assemblies, with both being utilised in combination with other data, such as low-resolution cryo-EM maps, to assist in model construction (Zhou *et al.*, 2015).

While improved methods have contributed to this revolution, the main driving force has been the enormous increase in available sequences. Even current methods perform badly if presented with only a few hundred sequences, which was typical ten years ago for a large family, but with ten- to a hundredfold more sequences often available in the current databanks, good contact predictions can be almost taken for granted. However, this is only generally true for proteins that have a representative in bacterial genomes. If the protein of interest is restricted to eukaryote species then the number of representatives is typically greatly reduced. While there are many thousand bacterial genomes covering a limited number of core proteins, there are still less than 1000 eukaryotic genomes covering a wide range of functions that are often quite restricted to subgroups. For example, if the protein of interest is an immune component, the contribution from plants will be minimal. In time, sequences from enough distinct species may be available for any protein, but in the meantime it will still be necessary to continue to improve basic methods.

The dichotomy between prokaryotic and eukaryotic sequences emerges again in predicting contacts between proteins. As mentioned above, the sequences in each multiple alignment need to be paired so correlation is only calculated along sequences from the same organism. For prokaryotic sequences, gene adjacency (operon co-location) provides a workable, if imperfect, solution but for eukaryotic sequences, despite some attempts at solving this problem, current methods remain limited. Finding a good solution is important as eukaryotes tend to have more paralogues (from past genome duplication events) and the correct pairing of these provides a way to boost their otherwise limited number of sequences. This is seen dramatically in the G-protein coupled receptor (rhodopsin) family mentioned above, where including paralogues results in 180,000 sequences.

On the method development front, after the introduction of Markov network pseudo-likelihood methods for resolving direct coupling interactions, little has changed in the underlying statistical model. Instead, the focus has shifted to post-processing these data and, in particular, using machine learning methods; either utilising multiple covariation sources (PconsC) or derived information, such as predicted structure (MetaPSICOV). The most recent logical progression of this trend has been towards deep learning methods. While these approaches undoubtedly generate a better overall prediction of contacts, it can be questioned whether they are actually providing any new information or simply just 'joining the dots' provided by the underlying correlation calculation. The matter may best be decided in future CASP exercises.

In summary, the approach of predicting protein (and RNA) contacts by position covariation has now developed into a powerful tool of general application to many structural problems. As the sequence databanks grow and new methodologies mature, in particular the introduction of deep learning, the power of the approach is clearly going to increase. Its one Achilles heel may be that it will always require enough sequences to be aligned, and even if fewer are required in the future, a protein that is restricted to, say, humans and our close relatives may never be tractable or an artificial sequence (with no homologues) will always be beyond reach.

Acknowledgements

This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK, the UK Medical Research Council, and the Wellcome Trust.

References

- Altschuh, D., Lesk, A., Bloomer, A.C. and Klug, A. (1987). Correlation of co-ordinated amino-acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology* **193**(4), 693–707.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.
- Anishchenko, I., Ovchinnikov, S., Kamisetty, H. and Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 9122–9127.
- Balakrishnan, S., Kamisetty, H., Carbonell, J., Lee, S.-I. and Langmead, C. (2011). Learning generative models for protein fold families. *Proteins* **79**, 1061–1078.
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M. and Pagnani, A. (2014). Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PLoS ONE* **9**, 1–12.
- Bartlett, G.J. and Taylor, W.R. (2008). Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in *de-novo* protein structure prediction. *Proteins: Structure, Function, and Bioinformatics* **71**, 950–959.
- Bitbol, A.-F., Dwyer, R., Colwell, L. and Wingreen, N. (2016). Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12180–12185.
- Buchan, D.W.A. and Jones, D.T. (2017). EigenTHREADER: Analogous protein fold recognition by efficient contact map threading. *Bioinformatics* **33**(17), 2684–2690.
- Buchan, D.W.A. and Jones, D.T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics* **86**, 78–83.
- Burger, L. and van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Molecular Systems Biology* **4**, 1–14.
- Cocco, S., Monasson, R. and Weight, M. (2013). From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. *PLoS Computational Biology* **9**, e1003176.
- de Juan, D., Pazos, F. and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics* **14**, 249–261.
- De Leonardis, E., Lutz, B., Ratz, S., Cocco, S., Monasson, R., Schug, A. and Weigt, M. (2015). Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Research* **43**, 10444–10455.
- Di Lena, P., Fariselli, P., Margara, L., Vassura, M. and Casadio, R. (2010). Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics* **26**(18), 2250–2258.
- Dunn, S., Wahl, L.M. and Gloor, G. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340.
- Eddy, S. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**, 2079–2088.
- Ekeberg, M., Lövkist, C., Lan, Y., Weigt, M. and Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* **87**, 012707.
- Finn, R., Clemmons, J., Arndt, W., Miller, B., Wheeler, T., Schreiber, F., Bateman, A. and Eddy, S. (2015). HMMER web server: 2015 update. *Nucleic Acids Research* **43**, W30–W38.
- Fox, G., Sievers, F. and Higgins, D. (2016). Using *de novo* protein structure predictions to measure the quality of very large multiple sequence alignments. *Bioinformatics* **32**, 814–820.

- Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics* **18**, 309–317.
- Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M. and Pagnani, A. (2016). Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 12186–12191.
- Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C. and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621.
- Hopf, T.A., Scharfe, C., Rodrigues, J., Green, A., Kohlbacher, O., Sander, C., Alexandre, M.J.J., Bonvin, A. and Marks, D. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430.
- Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Springer, M., Sander, C. and Marks, D.S. (2015). Quantification of the effect of mutations using a global probability model of natural sequence variation. Preprint, arXiv:1510.04612.
- Jehl, P., Sievers, F. and Higgins, D.G. (2015). OD-seq: Outlier detection in multiple sequence alignments. *BMC Bioinformatics* **16**, 269.
- Jones, D.T., Buchan, D.W.A., Cozzetto, D. and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190.
- Jones, D.T., Singh, T., Kosciolak, T. and Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**(7), 999–1006.
- Kaján, L., Hopf, T., Kalaš, M., Marks, D. and Rost, B. (2014). FreeContact: Fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85.
- Kamisetty, H., Ovchinnikov, S. and Baker, D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15674–15679.
- Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A. and Grishin, N.V. (2016). Evaluation of free modeling targets in CASP11 and ROLL. *Proteins: Structure, Function, and Bioinformatics* **84**, 51–66.
- Lapedes, A.S., Giraud, B., Liu, L. and Stormo, G.D. (1999). Correlated mutations in models of protein sequences: Phylogenetic and structural effects. In F. Seillier-Moiseiwitsch (ed.), *Statistics in Molecular Biology and Genetics*, IMS Lecture Notes Mongraph Series, Vol. 33. Institute of Mathematical Statistics, Hayward, CA, pp. 236–256.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* **521**, 436.
- LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324.
- Lee, B.-C. and Kim, D. (2009). A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics* **25**, 2506–2513.
- Li, W., Fu, L., Niu, B., Wu, S. and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics* **13**, 656–668.
- Li, W., Jaroszewski, L. and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* **17**, 282–283.
- Livesay, D., Kreth, K. and Fodor, A. (2012). A critical evaluation of correlated mutation algorithms and coevolution within allosteric mechanisms. *Methods in Molecular Biology* **796**, 385–398.
- Lockless, S.W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299.

- Loytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557–10562.
- Marks, D., Colwell, L., Sheridan, R., Hopf, T., Pagnani, A., Zecchina, R. and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766.
- McLaughlin Jr., R., Poelwijk, F. and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142.
- Michel, M., Skwark, M.J., Menéndez Hurtado, D., Ekeberg, M. and Elofsson, A. (2017). Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics* **33**(18), 2859–2866.
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D., Sander, C., Zecchina, R., Onuchic, J., Hwa, T. and Weigt, M. (2012). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E1293–E1301.
- Morcos, F., Jana, B. and Onuchic, J. (2013). Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20533–20538.
- Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America* **91**, 98–102.
- Nugent, T. and Jones, D.T. (2012). Accurate *de novo* structure prediction of large transmembrane protein domains using fragment-assemble and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1540–1547.
- Ovchinnikov, S., Kamisetty, H. and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030.
- Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D., Kamisetty, H., Grishini, N. and Baker, D. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248.
- Ovchinnikov, S., Kim, D.E., Wang, R.Y.-R., Liu, Y., DiMaio, F. and Baker, D. (2016). Improved *de novo* structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins: Structure, Function, and Bioinformatics* **84**, 67–75.
- Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G., Kim, D., Kamisetty, H., Kyripides, N. and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science* **355**, 294–298.
- Pandini, A., Kleinjung, J., Taylor, W., Junge, W. and Khan, S. (2015). The phylogenetic signature underlying ATP synthase c-ring compliance. *Biophysical Journal* **109**, 975–987.
- Pandini, A., Morcos, F. and Khan, S. (2016). The gearbox of the bacterial flagellar motor switch. *Structure* **24**, 1209–1220.
- Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of Molecular Biology* **271**, 511–523.
- Pollock, D.D. and Taylor, W.R. (1997). Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Engineering* **10**, 647–657.
- Pollock, D.D., Taylor, W.R. and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* **287**, 187–198.
- Qin, C. and Colwell, L. (2018). Power law tails in phylogenetic systems. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 690–695.
- Remmert, M., Biegert, A., Hauser, A. and Söding, J. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM–HMM alignment. *Nature Methods* **9**, 173–175.

- Reynolds, K.A., McLaughlin, R. and Ranganathan, R. (2011). Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575.
- Rodriguez-Rivas, J., Marsili, S.D.J. and Valencia, A. (2016). Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 15018–15023.
- Sadowski, M.I., Maksimiak, K. and Taylor, W.R. (2011). Direct correlation analysis improves fold recognition. *Computational Biology and Chemistry* **35**, 323–332.
- Seemayer, S., Gruber, M. and Söding, J. (2014). CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* **30**, 3128–3130.
- Sievers, F. and Higgins, D. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology* **1079**, 105–116.
- Skwark, M.J., Abdel-Rehim, A. and Elofsson, A. (2013). PconsC: Combination of direct information methods and alignments improves contact prediction. *Bioinformatics* **29**(14), 1815–1816.
- Skwark, M.J., Raimondi, D., Michel, M. and Elofsson, A. (2014). Improved contact predictions using the recognition of protein like contact patterns. *PLOS Computational Biology* **10**(11), 1–14.
- Süel, G.M., Lockless, S. and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* **10**, 59–69.
- Sutto, L., Marsili, S., Valencia, V. and Gervasio, F. (2015). From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 13567–13572.
- Szurmant, H. and Weigt, M. (2017). Inter-residue, inter-protein and inter-family coevolution: Bridging the scales. *Current Opinion in Structural Biology* **50**, 26–32.
- Talavera, D., Lovell, S. and Whelan, S. (2015). Covariation is a poor measure of molecular coevolution. *Molecular Biology and Evolution* **32**, 2456–2468.
- Taylor, W.R. (1999). Protein structure alignment using iterated double dynamic programming. *Protein Science* **8**, 654–665.
- Taylor, W.R. (2002). Protein structure comparison using bipartite graph matching. *Molecular & Cell Proteomics* **1**, 334–339.
- Taylor, W.R. (2016a). An algorithm to parse segment packing in predicted protein contact maps. *Algorithms for Molecular Biology* **11**, 17.
- Taylor, W.R. (2016b). Reduction, alignment and visualisation of large diverse sequence families. *BMC Bioinformatics* **17**(1), 300.
- Taylor, W.R. (2017). Algorithm for matching partially labelled sequence graphs. *Algorithms for Molecular Biology* **12**, 24.
- Taylor, W.R. and Hamilton, R.S. (2017). Exploring RNA conformational space under sparse distance restraints. *Scientific Reports* **7**, 44074.
- Taylor, W.R. and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering* **7**, 341–348.
- Taylor, W.R., Jones, D.T. and Sadowski, M.I. (2012). Protein topology from predicted residue contacts. *Protein Science* **21**, 299–305.
- Taylor, W.R., Hamilton, R.S. and Sadowski, M.I. (2013). Prediction of contacts from correlated sequence substitutions. *Current Opinion in Structural Biology* **23**, 473–479.
- Taylor, W.R., Matthews-Palmer, T.R. and Beeby, M. (2016). Molecular models for the core components of the flagellar type-III secretion complex. *PLoS ONE* **11**, e0164047.
- Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T., Berger, B., Sander, C. and Marks, D. (2016). Structured states of disordered proteins from genomic sequences. *Cell* **167**, 158–170.
- Van der Lee, R., Buljan, M., Lang, B., Weatheritt, R., Daughdrill, G., Dunker, A., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D., Kim, P., Kriwacki, R., Oldfield, C., Pappu, R., Tompa, P.,

- Uversky, V., Wright, P. and Babu, M. (2014). Classification of intrinsically disordered regions and proteins. *Chemical Reviews* **114**, 6589–6631.
- Wang, J., Mao, K., Zhao, Y., Zeng, C., Xiang, J., Zhang, Y. and Xiao, Y. (2017). Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Research* **45**, 6299–6309.
- Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Computational Biology* **13**(1), 1–34.
- Wang, S., Sun, S. and Xu, J. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics* **86**, 67–77.
- Wang, Z. and Xu, J. (2013). Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* **29**, i266–i273.
- Wei, D., Jiang, Q., Wei, Y. and Wang, S. (2012). A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics* **13**, 174.
- Weigt, M., White, R.A., Szurmantc, H., Hochc, J.A. and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 67–72.
- Weinreb, C., Riesselman, A., Ingraham, J., Gross, T., Sander, C. and Marks, D. (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975.
- Zhou, A., Rohou, A., Schep, D., Bason, J., Montgomery, M., Walker, J., Grigorieff, N. and Rubinstein, J. (2015). Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-em. *eLife* **4**, e10180.

12

Probabilistic Models for the Study of Protein Evolution

Umberto Perron,¹ Iain H. Moal,¹ Jeffrey L. Thorne,² and Nick Goldman¹

¹European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire, UK

²Department of Statistics & Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Abstract

Mathematical models are indispensable tools for characterizing the process of protein evolution, many aspects of which are not easily amenable to direct experimentation. Ideally, a model of protein evolution would provide a good description of the data and would simultaneously be parameterized in a manner that facilitates biological insight. Probabilistic descriptions of protein change are especially valuable because they engender a sound basis for likelihood-based statistical inference, and can provide the foundation for local and global alignment, phylogeny reconstruction, and prediction of protein structure and function. In this chapter, models of protein evolution are reviewed and their strengths and limitations are emphasized.

12.1 Introduction

The relationship between genotype and phenotype is central to biology. Proteins are at the heart of this relationship at its fundamental molecular level: the DNA coding for a protein is the genotype, whereas a protein's structure and its pattern of expression can be considered as a phenotype. Evidence that molecular biologists and biochemists have recognized the key role of proteins in this relationship can be found, for example, in the amount of resources allocated to characterize the proteome of numerous species and pathological conditions, the widespread use of RNA-Seq and now single-cell RNA-seq for measuring gene expression (see **Chapters 26 and 30**), and the considerable effort being invested in the development of ever more accurate methods for determining protein tertiary structure from protein sequence data (see **Chapter 11**).

The relationship between genotype and phenotype is surely as important to the study of evolution as it is to the rest of biology. Largely because of the complexity of the genotype–phenotype relationship, the process of protein evolution remains poorly understood. Many probabilistic models of protein evolution have been proposed and explored, but a comprehensive and satisfactory model is not yet on the horizon. Nevertheless, existing models provide a statistical foundation for beginning to characterize the process of evolution and for inferring evolutionary history. While these models still have obvious flaws, there have been substantial improvements over time.

A heightened interest in the ability of probabilistic models to extract biological information has accompanied the improvements, largely as a result of the emergence of the field of comparative genomics. In addition to the desire to understand protein evolution on a genomic scale, there are many tasks in comparative genomics for which understanding evolution may not be the goal, but for which evolution should be carefully treated because common ancestry is responsible for correlations among genomes. We cannot correctly consider data from related genomes to be independent samples; the extent and nature of the correlation structure are determined by the phylogeny that relates species and the processes by which genomes have evolved on these phylogenies.

Aside from being an inconvenience in some comparative analyses, correlation between proteins due to common ancestry provides the implicit underlying basis for established techniques of predicting protein function via sequence database searches, global sequence alignment, and homology modeling to predict protein structure. Common ancestry also plays a key role in predicting the phenotypic severity of new mutations and disease contributions of single nucleotide polymorphisms (e.g. Ng and Henikoff, 2001; Stone and Sidow, 2005; Cooper *et al.*, 2005; Mi *et al.*, 2007; Adzhubei *et al.*, 2013) and in emerging phylogenetic strategies for linking genes to phenotypic traits (e.g. Chikina *et al.*, 2016; Wu *et al.*, 2017; Collins and Didelot, 2018). All of these techniques have the potential to be improved via better treatments of protein evolution.

In this chapter, we describe some of the important features of existing models of protein evolution. We focus on models that have been employed for likelihood-based inference, and highlight promising research directions. Molecular evolution and comparative genomics are no longer fields where the ability to collect data is the main obstacle to progress. Instead, these fields are mainly limited by a lack of adequate tools for extracting information from existing data. We think that improved evolutionary models will inevitably lead to better tools.

12.2 Empirically Derived Models of Amino Acid Replacement

12.2.1 The Dayhoff and Eck Model

Dayhoff and collaborators (Dayhoff and Eck, 1968; Dayhoff *et al.*, 1972, 1978) introduced the first, and still most influential, probabilistic model of protein evolution. It describes the process of amino acid replacement in terms of a continuous-time Markov process on a phylogenetic tree, and is based upon the assumptions that all sites in a protein sequence evolve independently and identically, and that protein sequences evolve independently in different lineages of their common phylogeny. The independence assumptions are convenient because they allow the likelihood of an aligned set of protein sequences to be written as a product of likelihoods for individual alignment columns (i.e. site-likelihoods).

Site-likelihoods can be calculated via the pruning algorithm of Felsenstein (1981), mentioned in **Chapter 6**. Without repeating that algorithm here, we emphasize that it requires the calculation of transition probabilities: the probabilities of a possible state at the end of a branch on a phylogenetic tree, conditional upon the state at the beginning of the branch, the amount of evolution represented by the branch (i.e. the branch length), and the values of other model parameters. For the Dayhoff and Eck model there are 20 possible states, each representing one of the 20 amino acids that can be present at a protein sequence position at a given point in evolutionary time.

The parameters of the Dayhoff and Eck model are the instantaneous rates α_{ij} of replacement of amino acid type i by amino acid type j . Computation of transition probabilities from

underlying instantaneous replacement rates of a Markov process is a standard procedure, described for example by Liò and Goldman (1998). The rates in Dayhoff and Eck's formulation are constrained so that the model is time-reversible. In other words,

$$\pi_i \alpha_{ij} = \pi_j \alpha_{ji}, \quad \text{for all } i \text{ and } j, \quad (12.1)$$

where the frequency of each amino acid type j at equilibrium is denoted by π_j and is uniquely determined by the matrix of rates (α_{ij}). Informally, the time-reversibility property means that, without additional information, there is no way to determine which of two sequences is the ancestral protein and which is the descendant protein.

In fact, the Dayhoff and Eck model is the most general 20-state time-reversible homogeneous Markov model. Although four-state models of nucleotide substitution are computationally simpler, it is interesting to note that the most general four-state time-reversible model (Tavaré, 1986; Yang, 1994a), and even the simplest four-state model (Jukes and Cantor, 1969), were not studied until after the Dayhoff and Eck model was proposed.

For nucleotide sequence analysis using four-state general time-reversible (GTR) models, all parameters are typically estimated on a per-data-set basis. However, the number of parameters defining a 20-state GTR rate matrix is 209 (greatly exceeding the nine parameters needed to specify a four-state GTR rate matrix) and, as a consequence, it is generally considered that estimates of α_{ij} for a 20-state rate matrix are unlikely to be reliable unless the data being analyzed encompass thousands of amino acid substitutions between closely related sequences. Instead, estimates of relative rates of replacement for Dayhoff–Eck-type approaches are usually obtained prior to analysis of the aligned data set of interest. These estimates are generally based upon many data sets of aligned protein sequences and are fixed when subsequent data sets of interest are studied.

Dayhoff and Eck estimated the instantaneous rates α_{ij} implicitly, via a clever technique that relied upon comparing closely related protein sequences to derive transition probability matrices. When protein sequences are closely related, they will be identical at most positions. At the few positions where the sequences differ, the cause of the difference will usually be a single amino acid replacement event because the chance that two or more replacement events occurred at a position will be negligible. By comparing closely related sequences and neglecting the possibility of multiple amino acid replacements at the same positions, Dayhoff and Eck were able to infer implicitly the relative rates of the different kinds of amino acid replacements. Different methods for recovering the instantaneous rates α_{ij} from results such as those presented by Dayhoff and Eck (1968) were subsequently proposed, and are reviewed by Kosiol and Goldman (2005). Goldman *et al.* (1998) gave an alternative procedure to estimate rates α_{ij} directly from amino acid replacement counts.

Improved techniques for estimating the α_{ij} parameters are now available (e.g. Whelan and Goldman, 2001; Holmes and Rubin, 2002; Le and Gascuel, 2008; Dang *et al.*, 2014). The main improvements are the ability to deal with much larger training data sets, and making correct allowance for varying levels of sequence divergence and the consequent probabilities of multiple replacements at a position.

Subsequent to the Dayhoff and Eck (1968) publication, Dayhoff and collaborators published updated results as additional protein sequence data became available (Dayhoff *et al.*, 1972, 1978). References to the 'Dayhoff model' do not simply indicate the 20-state GTR model of amino acid replacement but instead have come to mean both the Dayhoff and Eck (1968) GTR model formulation and the specific estimates of the α_{ij} parameters obtained from the results of Dayhoff *et al.* (1978).

12.2.2 Descendants of the Dayhoff Model

The widely used Jones–Taylor–Thornton (JTT) model of amino acid replacement (Jones *et al.*, 1992) employs the same parameterization as Dayhoff and collaborators, but refers to estimates of α_{ij} that were derived from a still larger data set. Gonnet *et al.* (1992) also published an updated set of α_{ij} estimates. As methods were developed that made it computationally feasible to obtain maximum likelihood estimates of the parameters in a 20-state GTR model, Whelan and Goldman (2001) used this approach to generate the popular WAG model, based – like the models of Dayhoff and Eck (1968), Dayhoff *et al.* (1972, 1978) and Jones *et al.* (1992) – on data sets containing globular proteins encoded by nuclear DNA. An even larger and more varied data set, covering several protein families, was then used with a similar approach by Le and Gascuel (2008) in their LG model.

Amino acid replacements tend to involve chemically similar types of amino acids. There have been numerous attempts to classify amino acids on the basis of their physicochemical properties, and physicochemical distances among amino acid types have been proposed (e.g. Grantham, 1974; Taylor and Jones, 1993). However, these physicochemical distances between amino acid types are unlikely to directly reflect the propensity for amino acid replacements to occur in evolution between these types. The Dayhoff–Eck approach is appealing because estimates of relative amino acid replacement rates are empirically derived.

However, although the Dayhoff–Eck approach reflects the tendency for amino acid replacements to involve similar amino acid types, it describes evolution of the ‘average site’ in the ‘average protein’. In reality, there is variability of amino acid replacement patterns among proteins and among sites within proteins. A number of researchers have inferred models for specific types or families of proteins, aiming for greater accuracy through greater specificity. Adachi and Hasegawa (1996) derived amino acid replacement rate estimates from a data set consisting of all the mitochondrial proteins from each of 20 different vertebrate species. They detected some substantial differences between their estimates and those derived from nuclear-encoded proteins. Very similar findings were made by Yang *et al.* (1998), who studied mammalian mitochondrial proteins. Parameter estimates for protein evolution models derived from chloroplast genes, retroviral polymerase proteins and influenza proteins were obtained by Adachi *et al.* (2000), Dimmic *et al.* (2002) and Dang *et al.* (2010), respectively.

Dayhoff-type models can readily be modified to allow variation in composition of amino acids among proteins. Denoting by $\hat{\pi}_i$ and $\hat{\alpha}_{ij}$ the amino acid frequencies and replacement rates estimated with a Dayhoff-type approach (i.e. from a database of aligned protein sequences under the assumption that all families in the database have common composition and replacement rates), and by π'_i and α'_{ij} the corresponding values we will choose to use in the analysis of a particular protein, then the π'_i can be treated as parameters specific to each protein of interest to allow variation of amino acid composition among proteins. This is achieved by forcing the α'_{ij} for $i \neq j$ to obey

$$\alpha'_{ij} = \frac{\pi'_j}{\hat{\pi}_j} \hat{\alpha}_{ij} \quad (12.2)$$

(Cao *et al.*, 1994). The resulting replacement process defined by the α'_{ij} combines empirical information derived from databases (the $\hat{\alpha}_{ij}/\hat{\pi}_j$, also known as exchangeabilities: Whelan and Goldman, 2001) with the amino acid frequencies for the particular protein under study (π'_i), now estimated from that data set. Note that this approach maintains time reversibility (equation (12.1)). The benefit of this hybrid parameterization can be a sizeable improvement in model fit (Cao *et al.*, 1994). It is also attractive because it maintains the biologically realistic property of

allowing amino acid replacements to occur among chemically similar amino acids, and for the pragmatic reason that estimation of the 20 amino acid frequencies π'_j from individual protein alignments is computationally tractable and seems statistically robust. Goldman and Whelan (2002) obtained even greater improvements in model fit with a more general parameterization that makes the a'_{ij} a function of $\hat{\alpha}_{ij}$, the frequencies of the amino acid being replaced (π'_i and $\hat{\alpha}_i$), and the frequencies of the replacement amino acid (π'_j and $\hat{\alpha}_j$). The use of these parameterizations has thus been both to increase our understanding of the importance of particular amino acid residues in different proteins and to improve robustness of phylogenetic inferences.

12.3 Heterogeneity of Replacement Rates among Sites

There is variability not only of amino acid replacement among proteins, but also of amino acid replacement rates among sites within a single protein. Yang (1994b) introduced a practical method for incorporating heterogeneity of evolutionary rates among sites into models of nucleotide or amino acid substitution – see Chapter 6. Yang's innovation was to discretize a continuous gamma distribution of rates among sites into a relatively small number (C) of categories. Evaluation of a likelihood with Yang's discrete-gamma treatment of variation of rates among sites requires an amount of computation that is approximately a factor C more than would be needed if all sites were assumed to share identical rates. Yang (1994b) demonstrated that discretization of a gamma distribution into a relatively small number of rate categories (i.e. $C = 4\text{--}6$) generally fits the data well, with more categories giving no substantial change in parameter estimates or improvement in terms of likelihood.

It has been convincingly established that, for the majority of proteins, allowing heterogeneity of nucleotide substitution rates over sites represents a great improvement over models that assume homogeneity of evolutionary rate, a fact partly attributable to the triplet coding nature of DNA and to the structure of the genetic code. It is also now known that amino acid replacement models often fit data much better when allowance is made for rate variation among sites (e.g. Yang *et al.*, 1998; Goldman and Whelan, 2002; Le *et al.*, 2012; Dunn *et al.*, 2013). This has confirmed that rate variation over protein sites is a widespread evolutionary phenomenon that correlates with several biophysical traits (Echave *et al.*, 2016), and should be considered in all phylogenetic analyses of protein-coding sequences.

12.4 Protein Structural Environments

Although allowing empirically for particular proteins' distinct amino acid compositions and among-site rate variation has provided large improvements in fits of models of protein evolution, the models described so far shed little light on the cause of the variation, that is, on the interactions of mutation, selection and protein biochemistry and function that shape protein evolution. It is clear, from considering protein structure and function, that at least some of the variability will be associated with the structural environment of a site. For example, consider the solvent accessibility of a site. A site on the surface of a globular protein will be exposed to solvent, typically water. Sites with high solvent accessibility therefore tend to be occupied by hydrophilic amino acids. In contrast, sites buried in the interior of a protein are less accessible to solvent and are apt to be occupied by hydrophobic amino acids. Further, exposed sites tend to evolve about twice as fast on average as buried sites (Goldman *et al.*, 1998), most likely because residues at buried sites interact with many neighbouring residues; an amino acid replacement

at an exposed surface site is less likely to disrupt the position of other protein residues than is a residue at a buried site.

One of the most striking findings in the study of protein structure has been that proteins' secondary and tertiary structure will usually change very slowly during evolution. In cases where homologous proteins are known to perform the same biological functions, it is often the case that their structures are very similar even when the sequences that code for them are quite diverged (e.g. Chothia and Lesk, 1986; Russell *et al.*, 1997). In other words, protein sequence evolution tends in some sense to occur at a higher rate than the evolution of protein structure. This tendency can be exploited by models of amino acid replacement. Because structure changes more slowly than does sequence, a group of homologous protein sequences is likely to share a common underlying structure. Homologous amino acid residues (i.e. those that are in the same alignment column) are likely to be in a similar structural environment: for example, if a residue from one protein is part of an α -helix, then other sequences' residues in the same alignment column are also likely to be in an α -helix. If the tertiary structure of one protein sequence in an alignment has been experimentally determined, then each alignment column can be assigned a structural environment and thus a separate 20-state GTR model can be estimated for each environment (e.g. Overington *et al.*, 1990; Lüthy *et al.*, 1991; Topham *et al.*, 1993; Wako and Blundell, 1994a,b; Koshi and Goldstein, 1995, 1996; Thorne *et al.*, 1996; Goldman *et al.*, 1998; Le and Gascuel, 2010; Rios *et al.*, 2015). This strategy results in improved fit compared with standard models such as JTT or WAG.

Further, the structural environment at one site in a protein is not independent of the environment at other sites. Functional sites with lower rates of evolution may cluster together spatially (Huang and Golding, 2014). Secondary structure elements also induce clustering. For example, if a site is in an α -helix environment, then neighboring sites are also probably in the α -helix. It is possible to use this knowledge to further improve models of protein evolution. One approach has been the modeling of the organization of structural environments along a protein sequence as a first-order Markov chain. In protein sequence analysis, this approach was first used to predict secondary structure from single protein sequences (Asai *et al.*, 1993; Stultz *et al.*, 1993). The basis for the prediction was the tendency for certain kinds of amino acids to be found in certain secondary structure environments. Such models are referred to as hidden Markov models (HMMs) because the secondary structure underlying the sequence is not directly observed. Instead, it is 'hidden' but can be estimated on the basis of the amino acids that encode the protein sequence. A first-order Markov chain for the organization of protein structure along a sequence is not ideal because in reality protein structure is three-dimensional rather than linear, and the structural environment of a site may be strongly influenced by other sites that are nearby in the tertiary structure but are far separated along the linear protein sequence. However, a first-order Markov model for structural organization is computationally tractable and is clearly superior to assuming independence of structural environments among sites.

The HMM for organization of structural environment can be combined with the models of amino acid replacement for each structural environment to generate an integrated model of protein sequence evolution. This has been done for the study of globular proteins (Thorne *et al.*, 1996; Goldman *et al.*, 1998), transmembrane proteins (Liò and Goldman, 1999) and mitochondrial proteins (Liò and Goldman, 2002), using structural categories based on the secondary structure and solvent accessibility status of protein residues. With this approach, each category k of structural environment has its own equilibrium amino acid frequencies (π_i^k) and its own rates of amino acid replacement (α_{ij}^k). The algorithms used to calculate a likelihood with an HMM of protein structure were first applied to molecular sequence data by Churchill (1989) and are described in detail by Thorne *et al.* (1996). Felsenstein's pruning algorithm (Felsenstein, 1981) is used to compute likelihoods conditional on the unobserved states of the HMM, and

the HMM's transition probabilities are incorporated to deal with uncertainty about to which structure state each site belongs. The likelihood calculations now have a computational burden that is approximately a factor S more than without the HMM, where S is the number of distinct structure states considered; this remains fast enough to permit maximum likelihood phylogenetic inference. A by-product of this HMM approach is the ability to predict the underlying protein secondary structure (Goldman *et al.*, 1996).

Studies using these methods have shown the importance of structural environments to protein evolution. Different secondary structure elements do exhibit distinct patterns of evolution, presumably related to different residues' propensities to adopt the local structures and biochemical properties required of protein functions. It is also clear that taking a phylogenetic approach could be of advantage to other sequence analyses aimed at investigating protein function, most obviously the simple goal of protein structure prediction which remains difficult even after many years of intensive research and competition (e.g. Kryshtafovych *et al.*, 2005). In fact, recent progress in the ability to detect co-evolving pairs of protein sites is yielding substantial improvements in the ability to detect protein positions that are far apart along the sequence but nearby in tertiary structure, and this leads to better overall predictions of tertiary structure (Schaarschmidt *et al.*, 2018).

Although they change slowly, secondary structure and solvent accessibility do evolve. Kawabata and Nishikawa (2000) empirically estimated rates of change among secondary structure categories and solvent accessibility environments. This is a challenging endeavor, particularly because protein sequence alignment becomes difficult when proteins being compared are so diverged as to have moderately different structures. Kawabata and Nishikawa base a procedure for recognizing distantly related protein homologues on their model of secondary structure and solvent accessibility change, and they show that this procedure performs well. The future development of these or other ideas (see, for example, Grishin, 1997) into a complete model of protein sequence and structure evolution would be of great value.

12.5 Variation of Preferred Residues among Sites

Some, but clearly not all, variation of amino acid composition and replacement rates among sites can be explained by consideration of structural environments. The specific biochemical function of a protein, and even of its individual sites, means that the evolution of every site of every protein takes place under different constraints, potentially leading to different evolutionary patterns. Bruno (1996) made the first progress in this respect (see also Halpern and Bruno, 1998), devising a model that allowed each site in a protein to have its own equilibrium frequencies; we can denote the frequency of amino acid type i at site s by $\pi_{i,s}$. This highly flexible approach suffers from potential overparameterization problems because the 20 amino acid frequencies per site add 19 degrees of freedom per site to the model. To remedy the overparameterization problems, it is not sufficient to restrict analyses to data sets consisting of a large number of sequences because, if sequences in a data set are closely related, then the sequences will be highly correlated with one another and good estimates of $\pi_{i,s}$ may still not be confidently obtained. Therefore, Bruno's approach should work best when a data set contains a large number of highly diverged sequences.

Although this approach to allowing variation of 'preferred' residues among sites in a model of protein evolution is not computationally convenient, it is biologically attractive. The Dayhoff-type models allow the tendency of amino acid replacements to involve physicochemically similar amino acid types to be reflected in estimates of the replacement rates α_{ij} . On the other hand, this may instead be largely attributable to a tendency for the values of $\pi_{i,s}$ to be positively

correlated for physicochemically similar amino acids. These two explanations for the tendency of amino acid replacements to involve physicochemically similar amino acids are not mutually exclusive, but little is known about their relative importance.

A number of attempts to describe variation in amino acid frequencies over proteins' sites have followed Bruno's work. Wang *et al.* (2008) utilized principal components analysis to infer the existence of at least four major site-specific amino acid frequency classes in a data set of about 20 large protein alignments. Constructing a mixture model accounting for these four amino acid classes, plus a fifth class of global frequencies, they demonstrated significant improvements in model fit for real data sets.

Adopting a Bayesian perspective, Lartillot and Philippe (2004) largely overcome the overparameterization concerns associated with variation of 'preferred' amino acid types among sites. With their 'CAT' model, they assume each site belongs to some amino acid replacement process category, but that the category is not directly observed. Different replacement categories share a common set of amino acid exchangeability parameters taken from existing standard models, but have different equilibrium amino acid frequencies. The number of categories and their frequencies are jointly determined by a Dirichlet process prior distribution, and the set of 20 different amino acid frequencies for a given category has a Dirichlet prior. This Bayesian approach, with prior distributions for the number of replacement categories, the frequencies of the categories and the amino acid composition of each category, enables Lartillot and Philippe to greatly reduce the potential for overparameterization that occurs otherwise. They convincingly demonstrate that their Bayesian scheme is a substantial improvement over models that do not permit variation of preferred residues among sites. Although computationally intensive, implementations of the CAT approach exist that are practical for reasonably sized datasets.

Holmes and Rubin (2002) also assume each site of a protein belongs to one of a number of amino acid replacement categories that are not directly observed. Their innovative expectation-maximization (EM) algorithm yields maximum likelihood estimates of the amino acid replacement rates within each replacement category, along with estimates of the rates of category switches. Although their frequentist procedure may be subject to potential overparameterization problems and requires the number of replacement categories to be pre-specified, it has the advantage of not assuming that a particular protein site belongs to the same category for all of its evolutionary history, permitting the biologically appealing feature of sites' evolutionary dynamics potentially varying over evolutionary time.

Using similar models and building also upon a practical adaptation by Tuffley and Steel (1998) of ideas first described by Fitch and colleagues (e.g. Fitch and Markowitz, 1970; Fitch, 1971), several groups have shown that evolutionary patterns at sites or codons change over time and that these changes in pattern should not be ignored by evolutionary models (e.g. Galtier, 2001; Huelsenbeck, 2002; Guindon *et al.*, 2004; Inagaki *et al.*, 2004; Philippe *et al.*, 2005). This characteristic of protein evolution, somewhat related to Fitch's (1971) idea of concomitantly variable codons or 'covariants' and also known as heterotachy (Lopez *et al.*, 2002), may be attributable to changes in protein structural constraints over long evolutionary times. Absence of heterotachy in evolutionary models has been implicated in failures of phylogenetic inference methods, particularly in cases of phylogenies with long branches (e.g. Inagaki *et al.*, 2004; Delsuc *et al.*, 2005; Zhong *et al.*, 2011).

Building upon work by Yang and Roberts (1995), Groussin *et al.* (2013) suggested a model of protein evolution that can explain variation in amino acid composition among lineages. By doing this without adding a large number of parameters, the Groussin *et al.* (2013) approach is suitable for maximum likelihood analyses. A limitation of the Groussin *et al.* (2013) model

is that it does not permit process variation within a branch. Rate matrices are instead constant on each branch but can vary between branches.

Blanquart and Lartillot (2008) developed a highly flexible Bayesian approach that incorporates both heterogeneity of protein evolution among sites and among lineages. This flexibility is achieved by combining the CAT model (Lartillot and Philippe, 2004) with a stochastic process for breakpoints that represent locations on branches of the phylogenetic tree where the evolutionary process changes. Blanquart and Lartillot (2008) exploit Markov chain Monte Carlo (MCMC) techniques to perform inference with their highly flexible model.

12.6 Models with a Physicochemical Basis

An alternative approach to statistical models that implicitly (i.e. through empirical parameter estimation) reflect the tendency for amino acid replacements to involve physicochemically similar amino acid types is for this tendency to be reflected via the explicit inclusion of physicochemical properties into a model. Koshi *et al.* (1999) chose to quantify the physicochemical attributes of amino acid types according to hydrophobicity and bulk, properties expected to be of great importance to protein structure. Their approach assumes that each site in a protein belongs to exactly one of several classes, with each class k having its own set of equilibrium amino acid frequencies that are determined by hydrophobicity and bulk coefficients for that class. Although it can yield substantially better fits to data than Dayhoff-type models (Koshi *et al.*, 1999), this approach has not been widely pursued and seems to warrant additional investigation.

An alternative, semi-empirical, approach was developed by Zoller and Schneider (2013). Their method relies on principal components analysis to select the most relevant physicochemical parameters, creating an empirically determined parameterization for an amino acid substitution model. While similar to mixture models that exhibit heterogeneity among sites, Zoller and Schneider's model optimizes one substitution rate matrix per alignment as a combination of several 'universal' matrices and applies this single matrix to all sites. Zoller and Schneider (2013) report that their model compares positively with established amino acid substitution models, in particular for longer alignments, and is computationally more efficient than a mixture model approach.

12.7 Codon-Based Models

In reality, evolution occurs at the DNA level, with protein change being a consequence of phenomena such as sequence mutation, genetic drift and natural selection. It therefore makes sense to model protein evolution in terms of codons rather than amino acid types (Schöniger *et al.*, 1990; Goldman and Yang, 1994; Muse and Gaut, 1994; Ren *et al.*, 2005; Kosiol and Goldman, 2011).

Codon-based models are typically framed in terms of the 61 codons that specify amino acids in common genetic codes. Existing codon-based models follow largely the same approach as the Dayhoff model for amino acids. Codon replacements are modeled as a Markov process with α_{ij} , now describing the instantaneous rate of change from codon i to codon j , usually being defined as the product $\alpha_{ij} = \pi_j s_{ij}$ where π_j is the frequency of codon j and symmetry ($s_{ij} = s_{ji}$) of the exchangeabilities ensures reversibility (equation (12.1)).

Initially, limited data sets and computing power precluded the empirical estimation of exchangeabilities s_{ij} . Instead, mechanistic models were devised (Goldman and Yang, 1994; Muse and Gaut, 1994), with the s_{ij} themselves the products of parameters representing phenomena such as codon usage, transition-transversion bias and the strength of selection acting on each mutation, and further simplified by the assumption that multi-nucleotide substitutions do not occur instantaneously ($s_{ij} = 0$ if codons i and j differ by more than one nucleotide). This meant that the limited numbers of parameters could be reliably estimated on a per-data-set basis. Codon models have found widespread use in testing hypotheses regarding natural selection, through parameterizations that include factors describing synonymous and non-synonymous nucleotide substitution rates (Muse, 1996; Yang and Nielsen, 2000). Considerable attention has been devoted to their potential to detect both diversifying and purifying natural selection (e.g. Nielsen and Yang, 1998; Yang *et al.*, 2000; Kosakovsky Pond and Muse, 2005; Massingham and Goldman, 2005). These methods and applications are described in detail in **Chapter 13**. Scherrer *et al.* (2012) have investigated codon models that link non-synonymous rates and other parameters to the relative solvent accessibility of each position in the protein tertiary structure. Careful modeling choices can generate codon substitution models that exhibit relatively good fits to actual data but that are not parameter-rich (Miyazawa, 2013).

As a result of the growing amount of protein-coding DNA sequence data, empirical estimates of codon substitution rates are increasingly practical alternatives to less parameterized codon models (Schneider *et al.*, 2005; Doron-Faigenboim and Pupko, 2007; Kosiol *et al.*, 2007). Analysis of empirically derived codon models can lead to further insights into the actual patterns of mutation and selection in protein-coding sequences, for example regarding the prevalence of multi-nucleotide substitutions occurring instantaneously (Averof *et al.*, 2000; Whelan and Goldman, 2004; Kosiol *et al.*, 2007). Some recent work with codon-based models investigates whether substitution rates between amino acid pairs can be subdivided into a variable number of rate classes, dependent on the information content of the alignment (Delport *et al.*, 2010). In other cases, predefined partitions of the data (e.g. distinct protein domains) are analysed using different models in a multi-partition approach (Zoller *et al.*, 2015).

A mechanistic approach to modeling protein evolution has been to try to reconcile codon-based substitution rate matrices with molecular population genetics, explaining variation of preferred amino acid types among sites in terms of selective forces operating at the sites. The basic idea is that the rate of codon substitution is the product of mutation rate and the probability that a new mutation is eventually fixed in the population. Halpern and Bruno (1998) were the first to try this; their ideas were further developed by several other groups (e.g. Thorne *et al.*, 2007; Yang and Nielsen, 2008; Rodrigue *et al.*, 2010; Tamuri *et al.*, 2012). These approaches are beginning to provide results in accord with empirical observations from mutation experiments (Tamuri *et al.*, 2014), and take important steps toward narrowing the gulf between population genetics and the models employed in phylogenetics.

Bloom (2014) has developed an innovative approach that combines variation of preferred residues among sites, the mutation-selection ideas of Halpern and Bruno (1998), and experimental evolution. Bloom (2014) uses mutagenesis, functional selection, and deep sequencing to inform a codon substitution model for the influenza nucleoprotein. His high-throughput experimental data allow estimation of the relative fitnesses of each possible amino acid type at every nucleoprotein site. The relative fitnesses are used to parameterize the variation of preferred amino acids among protein positions. Bloom persuasively demonstrates that his approach provides a better fit than alternative models that are conventionally employed in phylogenetic studies. It is unclear whether a large proportion of protein families are amenable to this modeling strategy, but the strategy seems worthy of substantial additional effort.

12.8 Dependence among Positions

A limitation of widely used models of protein evolution is the assumption that the evolution of one position is independent of the sequence at other positions. Proteins are obviously three-dimensional, and evolution at one site in a protein is likely to be affected by the amino acids at sites that physically neighbour it in the tertiary structure of the protein. Covariation of residues at different sites may indicate, therefore, that these sites are nearby in the tertiary structure of the protein, and has been used to make inferences about protein tertiary structure and function (e.g. Pazos *et al.*, 1997; Pollock *et al.*, 1999; Tillier and Lui, 2003; Schaarschmidt *et al.*, 2018).

Parisi and Echave (2001; see also Bastolla *et al.*, 2003; Arenas *et al.*, 2013) introduced a noteworthy technique for simulating the evolution of protein-coding genes subject to constraints imposed by protein structure. They begin their simulations with a reference protein that has known tertiary structure, and assume that this structure does not change over time. They then propose a sequence-structure distance scoring function to assess how well protein sequences fold into the known structure. Such scoring functions have been actively developed and are well established for predicting protein folds from protein sequence data. The underlying idea behind the Parisi–Echave technique is that the sequence-structure compatibility function can be employed to help parameterize rates of non-synonymous substitutions.

Through simulations, Parisi and Echave (2001) convincingly demonstrated both that their model captured information that would be missed by other models of protein evolution and that this information capture could occur without requiring large numbers of free parameters to be added to the model. In a specific example, the simulation studies showed tendencies of certain amino acids to preferentially occupy certain sites in the left-handed β -helix domain of UDP-*N*-acetyl glucosamine acetyltransferases. When a group of actual acetyltransferases with this helix domain was examined, qualitatively similar tendencies were observed.

Following the Parisi and Echave work, simulated protein evolution with dependent change among positions has been widely employed to learn about the interplay between protein structure, expression, and evolution. Because this literature is both broad and rich, we do not attempt to overview it here. We instead concentrate on the much smaller number of probabilistic evolutionary models that have been employed for likelihood-based inference.

A model of change permitting dependence among sequence positions poses enormous challenges for conventional evolutionary inference procedures that use Felsenstein's (1981) pruning algorithm. The difficulty is that this algorithm relies upon converting instantaneous evolutionary rates to transition probabilities. With independently evolving nucleotides, amino acids or codons, the number of rows and columns in the rate and transition probability matrices equals 4, 20 or 61, respectively. But with a relatively general dependence structure, the number of rows and columns in the rate and transition probability matrices equals the number of possible sequences. Because this grows exponentially with sequence length, Felsenstein's pruning algorithm becomes computationally intractable with dependent change among sequence sites unless sequence length is extremely short.

Fornasari *et al.* (2002) introduced one way to reflect the amino acid replacement patterns observed in simulations that have dependent change among protein positions while maintaining the computational feasibility of likelihood-based inference with evolutionary independence among sites. They did this by performing large numbers of simulations with the aforementioned Parisi–Echave approach. Each simulation started with an actual protein sequence. Fornasari *et al.* (2002) recorded the number of times in these simulations that each type of amino acid was replaced by each other amino acid at each protein position. These replacement counts were used to derive position-specific replacement rate matrices that could be employed

by a model with independent change among protein sites. While this modeling approach is attractive and maintains the computational convenience of Felsenstein's pruning algorithm, the position-specific rate matrices are influenced by the amount of evolution that is simulated to obtain the replacement counts and the choice of this amount is somewhat arbitrary.

A promising alternative to Felsenstein's inference procedure was explored by Jensen and Pedersen (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001) as part of their efforts to understand the effects of overlapping reading frames and context-dependent mutation on molecular evolution (see also Siepel and Haussler, 2004; Hwang and Green, 2004; Christensen *et al.*, 2005). Jensen and Pedersen augmented the observed sequence data at the tips of evolutionary trees with 'sequence paths'. The sequence path on a particular branch of a phylogeny specifies all of the changes that occurred on the branch. For each change, the sequence path contains information about the time at which the change occurred, the sequence position at which the change occurred, and the nature of the change. Although the sequence path is not directly observed, MCMC techniques can be used to randomly sample possible sequence paths according to the appropriate probability density. The big advantage of the sequence path approach to inference is that it is often computationally tractable in cases where Felsenstein's pruning algorithm is difficult to apply because transition probabilities cannot be calculated or well approximated.

Combining the innovations of Parisi and Echave (2001) with the augmentation strategy of Jensen and Pedersen (Jensen and Pedersen, 2000; Pedersen and Jensen, 2001), Robinson *et al.* (2003) designed a model of protein-coding DNA sequence evolution that assumes a globular protein tertiary structure is known and is shared by the translated version of all protein-coding DNA sequences being analyzed. Codon substitution rates of the Robinson model are intended to reflect both the mutation processes and the influence of protein structure on non-synonymous change. With the Robinson model, this influence was governed by the effect of the non-synonymous change on the pairwise amino acid interactions and the change in hydrophobicity of the site that was modified. This influence could be assessed because the protein tertiary structure was assumed known and unchanging during evolution. The Robinson model yielded biologically plausible inferences regarding the evolutionary impact of hydrophobicity and solvent accessibility (Robinson *et al.*, 2003; Robinson, 2003; see also Choi *et al.*, 2007).

Rodrigue *et al.* (2006) carefully evaluated evolutionary models with rate parameterizations similar to that of Robinson (2003). Employing a new model comparison procedure (Lartillot and Philippe, 2004, 2006) as well as inference based upon sequence paths, Rodriguez *et al.* demonstrated that evolutionary models with dependence among sequence positions due to tertiary structure are statistically superior to those without dependence. However, these authors also concluded that available treatments of protein tertiary structure are not sufficient to produce satisfactory evolutionary models. They find that adding rate heterogeneity among sites via Yang's (1994b) discrete gamma technique generates model fit improvements beyond those realized solely by incorporating tertiary structure.

Subsequent efforts to incorporate site interdependencies brought about by structural constraints have employed more sophisticated treatments of protein structure. In these cases, an energy-like scoring system for sequence–structure compatibility (statistical potential) is used to evaluate the probability of fixation of a given mutation, assuming an underlying protein structure that remains constant. Terms related to pairwise distance interactions, torsion angles, solvent accessibility and flexibility of the residues are included in the potentials, so as to study the effects of the main factors known to influence protein structure. However, these structurally constrained models are often outperformed by some of the available site-independent models in terms of fit, possibly indicating that alternatives to coarse-grained statistical potentials should be explored in order to better model structural constraints (Kleinman *et al.*, 2010).

While the above approach uses a function to model the compatibility between sequence and tertiary structure, it may be adapted to include the effects of any function of the whole sequence. In an approach by Bordner and Mittelmann (2014), computational savings were generated by having the substitution rate at any given residue position be a function of only a subset of the other residues. By only considering interdependencies between sites that make side chain contacts in the tertiary structure of the protein, a model can be constructed in which the probability of observing two sequences separated by a short evolutionary distance can be factorised and represented as a factor graph. The short-distance models can then be concatenated to form a composite graph representing a specific phylogenetic tree and set of sequences, from which the likelihood can be estimated using a belief propagation algorithm. Preliminary tests using this approach show that it can capture relevant inter-site correlations on a short time-scale, and can be extended to correlations arising from quaternary protein structures.

Arenas *et al.* (2015) developed a new mean-field substitution model that generates independent site-specific amino acid distributions with constraints derived from knowledge about inferred stability of protein structures as they evolve. These in turn are based primarily on the number of other residues with which the residue at a particular site is in close contact. This model depends on a background distribution of amino acids and one selection parameter that maximizes the likelihood of the protein sequence. Here, the main determining factor of the site-specific distributions is the number of native contacts of that site and the most variable sites are those with an intermediate number of native contacts. The mean-field models yield larger likelihoods than models that only consider the native state, and produce stable sequences for most proteins, with more realistic average hydrophobicity due to taking into account misfolded conformations.

One reasonable interpretation of these findings is that sequence-structure compatibility measures yield models of protein evolution that are biologically meaningful and statistically valuable, extracting information that is otherwise not utilized, yet which are still incomplete summaries of natural selection. Although these constrained models are designed to reflect dependence due to protein tertiary structure, the same inference strategy could be applied to add evolutionary dependence due to other aspects of phenotype (e.g. protein expression or protein function).

12.9 Stochastic Models of Structural Evolution

Conventional probabilistic descriptions of protein evolution can be categorized as continuous-time, discrete-state Markov models, with the discrete states representing protein sequences or protein-coding DNA sequences. Incorporation of protein tertiary structure into these conventional models has usually been done by attempting to reflect how features of tertiary structure (e.g. solvent accessibility) affect rates of sequence change. This conventional approach is natural because DNA is the genetic material and it has discrete states. However, a limitation of this treatment is that some aspects of protein phenotype are usefully described with continuously distributed states and there is abundant uncertainty concerning how to map these to discrete (DNA-based) genotypic states. For these aspects of phenotype, there may sometimes be an advantage in describing how they change with continuous-state models. While a comprehensive review of how continuous-state models have been applied to evolution is beyond our intended scope, we summarize the promising line of research initiated by Challis and Schmidler (2012) because it represents an evolutionary treatment of protein tertiary structure that substantially differs from those already reviewed in this chapter.

Challis and Schmidler (2012) describe a Bayesian approach that considers data sets that consist of two protein sequences that are not aligned and that each have experimentally determined protein structures. We will not dwell on the details of their MCMC procedure, except for those that concern how protein tertiary structure changes over evolutionary time. Challis and Schmidler (2012) summarize the structural information for each protein via their α -carbon coordinates in three-dimensional Euclidean space, and employ a time-reversible continuous-time and continuous-state Markov model to describe how the α -carbon coordinates constituting one protein can be transformed during evolution into the α -carbon coordinates that constitute a related protein.

The Challis–Schmidler model of structure is quite simple, assuming independent change in spatial locations among the amino acids. For each of these amino acids, it further assumes that three separate Ornstein–Uhlenbeck (OU) processes are operating, with these three processes independently affecting the three coordinates that specify the location of an α -carbon in three-dimensional Euclidean space. This model of spatial evolution of amino acids is computationally tractable because OU processes are time-reversible and have both normally distributed transition probabilities and normally distributed stationary distributions.

Herman *et al.* (2014) generalized the inference procedure of Challis and Schmidler so that data sets consisting of more than two proteins could be jointly analyzed, as well as accounting for the fact that the spatial coordinates of experimentally determined protein structures have associated uncertainty. Although the Herman *et al.* (2014) treatment is too computationally demanding to handle data sets with large numbers of proteins, it can be employed to simultaneously reconstruct phylogenies and infer structural alignments between proteins. With their analyses of globins and cysteine proteases, Herman *et al.* (2014) provided persuasive evidence that adding structural information yielded biologically more plausible results concerning phylogeny and protein alignment than could be obtained with sequence data alone.

The Challis–Schmidler OU model of structural evolution has the notable limitation of independent change among spatial locations of different amino acids in a protein. At equilibrium, this means that the spatial locations of consecutive amino acids in a protein sequence would be independent of one another. A better evolutionary model would have the spatial locations of amino acids that are nearby in protein sequence change in a correlated fashion. Two recent works have improved models of structural evolution in this way. Golden *et al.* (2017) achieve this by describing the evolution of protein tertiary structure using a specialized stochastic process that operates in dihedral angle space. Larson *et al.* (2018) construct a different stochastic process, with correlated spatial drift of amino acids. Based on analyses of pairs of proteins, both of these recent efforts are promising. We expect that these approaches will soon be extended so that more than two proteins can be simultaneously analyzed for the purposes of phylogeny inference and structural alignment with these more biologically plausible models.

12.10 Conclusion

After more than half a century, progress continues to be made with probabilistic models that give increasingly good statistical descriptions of the patterns of protein evolution. But a useful model is not simply one that fits data well. For a model of protein evolution to help us to understand processes and pressures acting on evolving genomes, it is also important to establish a solid connection between the model parameters and the biological features that they represent. Although the best models of protein evolution are now far more realistic than the earliest ones, understanding of protein evolution is still at a primitive stage. Evidence for variation in evolutionary processes among sites, lineages, and protein families is strong. Unfortunately, the

extent to which this variation can be partitioned into components of interest (e.g. the structural environments of a site, the protein to which the site belongs, mutational tendencies at the site, interactions with other sites, protein function) remains unclear.

Ideally, protein evolution should be linked to the DNA that codes for the protein, the structure of the protein, the expression patterns of the protein, the function of the protein, and external influences on the protein that act via natural selection. Progress on modeling protein evolution will depend in part on the advances that are made in techniques for *in silico* prediction of phenotype from genotype. Protein-coding DNA that results in deleterious phenotypes is unlikely to be ancestral to extant DNA. Accurate *in silico* prediction of phenotype from genotype would lead to evolutionary models in which substitution rates are higher for selectively advantageous changes and lower for deleterious changes. Existing systems for *in silico* mapping of genotype to phenotype tend to be crude, and this is one reason why there is so much room for improvement in models of protein evolution.

Despite the primitive nature of models of protein evolution, software implementing those models for data analysis is in great demand. It must be emphasized that even the existing models of protein evolution, however unrealistic, are better than having no models at all. Explicit evolutionary models provide a basis upon which homologous sequences can be recognized, phylogenetic history can be inferred, evolutionary hypotheses can be evaluated, protein structure can be predicted, and our understanding of evolution can be quantified. The breadth of applications, data and understanding of how protein sequence leads to a phenotype that selection can act upon have, in recent years, led to many varied extensions and reformulations of these simple approaches. While some modeling directions are promising, a general framework within which the multiple and overlapping processes that shape the course of protein evolution has yet to be formulated. For these reasons, the development of model-based approaches to studying protein evolution is an attractive, active and valuable area of research.

Acknowledgements

J.L.T. was supported by National Institutes of Health grant GM118508. U.P. and N.G. were supported by the European Molecular Biology Laboratory. I.H.M. was supported by Biotechnology and Biological Sciences Research Council Future Leader Fellowship BB/N011600/1. We thank Greg Słodkowicz for his comments and suggestions.

References

- Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution* **42**, 459–468.
- Adachi, J., Waddell, P.J., Martin, W. and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**, 348–358.
- Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* **76**, 7.20.1–7.20.41.
- Arenas, M., Dos Santos, H.G., Posada, D. and Bastolla, U. (2013). Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* **29**, 3020–3028.
- Arenas, M., Sanchez-Cobos, A. and Bastolla, U. (2015). Maximum-likelihood phylogenetic inference with selection on protein folding stability. *Molecular Biology and Evolution* **32**, 2195–2207.

- Asai, K., Hayamizu, S. and Handa, K. (1993). Prediction of protein secondary structure by hidden Markov model. *Computer Applications in the Biosciences* **9**, 141–146.
- Averof, M., Rokas, A., Wolfe, K.H. and Sharp, P.M. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287**, 1283–1286.
- Bastolla, U., Porto, M., Roman, H.E. and Vendruscolo, M. (2003). Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *Journal of Molecular Evolution* **56**, 243–254.
- Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution* **25**, 842–858.
- Bloom, J.D. (2014). An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molecular Biology and Evolution* **31**, 1956–1978.
- Bordner, A.J. and Mittelmann, H.D. (2014). A new formulation of protein evolutionary models that account for structural constraints. *Molecular Biology and Evolution* **31**, 736–749.
- Bruno, W.J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution* **13**, 1368–1374.
- Cao, Y., Adachi, J., Janke, A., Pääbo, S. and Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution* **39**, 519–527.
- Challis, C.J. and Schmidler, S.C. (2012). A stochastic evolutionary model for protein structure alignment and phylogeny. *Molecular Biology and Evolution* **29**, 3575–3587.
- Chikina, M., Robinson, J.D. and Clark, N.L. (2016). Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular Biology and Evolution* **33**, 2182–2192.
- Choi, S.C., Hobolth, A., Robinson, D.M., Kishino, H. and Thorne, J.L. (2007). Quantifying the impact of protein tertiary structure on molecular evolution. *Molecular Biology and Evolution* **24**, 1769–1782.
- Chothia, C. and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**, 823–826.
- Christensen, O.F., Hobolth, A. and Jensen J.L. (2005). Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates. *Journal of Computational Biology* **12**, 1166–1182.
- Churchill, G.A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79–94.
- Collins, C. and Didelot, X. (2018). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology* **14**, 1–21.
- Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**, 901–913.
- Dang, C.C., Le, Q.S., Gascuel, O. and Le, V.S. (2010). FLU, an amino acid substitution model for influenza proteins. *BMC Evolutionary Biology* **10**, 99.
- Dang, C.C., Le, V.S., Gascuel, O., Hazes, B. and Le, Q.S. (2014). FastMG: A simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. *BMC Bioinformatics* **15**, 341.
- Dayhoff, M.O. and Eck, R.V. (1968). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington DC, pp. 33–41.
- Dayhoff, M.O., Eck, R.V. and Park, C.M. (1972). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5. National Biomedical Research Foundation, Washington DC, pp. 89–99.

- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington DC, pp. 345–352.
- Delport, W., Scheffler, K., Botha, G., Gravenor, M.B., Muse, S.V. and Kosakovsky Pond, S.L. (2010). CodonTest: modeling amino acid substitution preferences in coding sequences. *PLoS Computational Biology* **6**, e10000885.
- Delsuc, F., Brinkmann, H. and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**, 361–375.
- Dimmic, M.W., Rest, J.S., Mindell, D.P. and Goldstein, R.A. (2002). rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution* **55**, 65–73.
- Doron-Faigenboim, A. and Pupko, T. (2007). A combined empirical and mechanistic codon model. *Molecular Biology and Evolution* **24**, 388–397.
- Dunn, K.A., Jiang, W., Field, C. and Bielawski, J.P. (2013). Improving evolutionary models for mitochondrial protein data with site-class specific amino acid exchangeability matrices. *PLoS ONE* **8**, e55816.
- Echave, J., Spielman, S.J. and Wilke, C.O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics* **17**, 109–121.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Fitch, W.M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution* **1**, 84–96.
- Fitch, W.M. and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4**, 579–593.
- Fornasari, M.S., Parisi, G. and Echave, J. (2002). Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Molecular Biology and Evolution* **19**, 352–356.
- Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**, 866–873.
- Golden, M., García-Portugués, E., Søoslash, M., Mardia, K.V., Hamelryck, T. and Hein, J. (2017). A generative angular model of protein structure evolution. *Molecular Biology and Evolution* **34**, 2085–2100.
- Goldman, N. and Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution* **19**, 1821–1831.
- Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.
- Goldman, N., Thorne, J.L. and Jones, D.T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *Journal of Molecular Biology* **263**, 196–208.
- Goldman, N., Thorne, J.L. and Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.
- Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.
- Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864.
- Grishin, N.V. (1997). Estimation of evolutionary distances from protein spatial structures. *Journal of Molecular Evolution* **45**, 359–369.

- Groussin, M., Boussau, B. and Gouy, M. (2013). A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. *Systematic Biology* **62**, 523–538.
- Guindon, S., Rodrigo, A.G., Dyer, K.A. and Huelsenbeck, J.P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12957–12962.
- Halpern, A. and Bruno, W.J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution* **15**, 910–917.
- Herman, J.L., Challis, C.J., Novák, Á., Hein, J. and Schmidler, S.C. (2014). Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular Biology and Evolution* **31**, 2251–2266.
- Holmes, I. and Rubin, J.P. (2002). An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology* **317**, 753–764.
- Huang, Y.-F. and Golding, G.B. (2014). Phylogenetic Gaussian process model for the inference of functionally important regions in protein tertiary structures. *PLoS Computational Biology* **10**, e1003429.
- Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* **19**, 698–707.
- Hwang, D.G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13994–14001.
- Inagaki, Y., Susko, E., Fast, N.M. and Roger, A.J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaeabacteria in EF-1 α phylogenies. *Molecular Biology and Evolution* **21**, 1340–1349.
- Jensen, J.L. and Pedersen, A.-M.K. (2000). Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Advances in Applied Probability* **32**, 499–517.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–282.
- Jukes, T.H. and Cantor, C.R. (1969). Evolution of Protein Molecules. In H.N. Munro, (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–132.
- Kawabata, T. and Nishikawa, K. (2000). Protein structure comparison using the Markov transition model of evolution. *Proteins* **41**, 108–122.
- Kleinman, C.L., Rodrigue, N., Lartillot, N. and Philippe, H. (2010). Statistical potentials for improved structurally constrained evolutionary models. *Molecular Biology and Evolution* **27**, 1546–1560.
- Kosakovsky Pond, S. and Muse, S.V. (2005). Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution* **22**, 2375–2385.
- Koshi, J.M. and Goldstein, R.A. (1995). Context-dependent optimal substitution matrices. *Protein Engineering* **8**, 641–645.
- Koshi, J.M. and Goldstein, R.A. (1996). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* **27**, 336–344.
- Koshi, J.M., Mindell, D.P. and Goldstein, R.A. (1999). Using physical-chemistry based mutation models in phylogenetic analyses of HIV-1 subtypes. *Molecular Biology and Evolution* **16**, 173–179.
- Kosiol, C. and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution* **22**, 193–199.
- Kosiol, C. and Goldman, N. (2011). Markovian and non-Markovian protein sequence evolution: aggregated Markov process models. *Journal of Molecular Biology* **411**, 910–923.
- Kosiol, C., Holmes, I. and Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution* **24**, 1464–1479.

- Kryshtafovych, A., Venclovas, Č., Fidelis, K. and Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins* **61**, 225–236.
- Larson, G., Thorne, J.L. and Schmidler, S. (2018). Modeling dependence in evolutionary inference for proteins. In: Raphael B. (eds) *Research in Computational Molecular Biology. RECOMB 2018. Lecture Notes in Computer Science*, vol 10812, Cham. Springer.
- Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, 1095–1109.
- Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology* **55**, 195–207.
- Le, S.Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**, 1307–1320.
- Le, S.Q. and Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Systematic Biology* **59**, 277–287.
- Le, S.Q., Dang, C.C. and Gascuel, O. (2012). Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution* **29**, 2921–2936.
- Liò, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Research* **8**, 1233–1244.
- Liò, P. and Goldman, N. (1999). Using protein structural information in evolutionary inference: transmembrane proteins. *Molecular Biology and Evolution* **16**, 1696–1710.
- Liò, P. and Goldman, N. (2002). Modeling mitochondrial protein evolution using structural information. *Journal of Molecular Evolution* **54**, 519–529.
- Lopez, P., Casane, D. and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* **19**, 1–7.
- Lüthy, R., McLachlan, A.D. and Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* **10**, 229–239.
- Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive and purifying selection. *Genetics* **169**, 1753–1762.
- Mi, H., Guo, N., Kejariwal, A. and Thomas, P.D. (2007). PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research* **35**, D247–252.
- Miyazawa, S. (2013). Superiority of a mechanistic codon substitution model even for protein sequences in phylogenetic analysis. *BMC Evolutionary Biology* **13**, 257.
- Muse, S.V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Molecular Biology and Evolution* **13**, 105–114.
- Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.
- Ng, P.C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research* **11**, 863–874.
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.
- Overington, J., Johnson, M.S., Sali, A. and Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings of the Royal Society of London B* **241**, 132–145.
- Parisi, G. and Echave, J. (2001). Structural constrains and the emergence of sequence patterns in protein evolution. *Molecular Biology and Evolution* **18**, 750–756.
- Pazos, F., Helmer-Citterich, M., Ansieillo, G. and Valencia, A. (1997). Correlated mutations contain information about protein-protein interactions. *Journal of Molecular Biology* **271**, 511–523.

- Pedersen, A.-M.K. and Jensen, J.L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular Biology and Evolution* **18**, 763–776.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5**, 50.
- Pollock, D.D., Taylor, W.R. and Goldman, N. (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* **287**, 187–198.
- Ren, F., Tanaka, H. and Yang, Z. (2005). An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Systematic Biology* **54**, 808–818.
- Rios, S., Fernandez, M.F., Caltabiano, G., Campillo, M., Pardo, L. and Gonzalez, A. (2015). GPCRtm: an amino acid substitution matrix for the transmembrane region of class A G Protein-Coupled Receptors. *BMC Bioinformatics* **16**, 206.
- Robinson, D.M. (2003). *D.R.EVOL: Three Dimensional Realistic Evolution*. PhD thesis, North Carolina State University, Raleigh, NC.
- Robinson, D.M., Jones, D., Kishino, H., Goldman, N. and Thorne, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution* **20**, 1692–1704.
- Rodrigue, N., Philippe, H. and Lartillot, N. (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Molecular Biology and Evolution* **23**, 1762–1775.
- Rodrigue, N., Philippe, H. and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 4629–4634.
- Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *Journal of Molecular Biology* **269**, 423–439.
- Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. and Bonvin, A.M.J.J. (2018). Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins* **86**, 51–66.
- Scherrer, M.P., Meyer, A.G. and Wilke, C.O. (2012). Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology* **12**, 179.
- Schneider, A., Cannarozzi, G.M. and Gonnet, G.H. (2005). Empirical codon substitution matrix. *BMC Bioinformatics* **6**, 134.
- Schöniger, M., Hofacker, G.L. and Borstnik, B. (1990). Stochastic traits of molecular evolution – acceptance of point mutations in native actin genes. *Journal of Theoretical Biology* **143**, 287–306.
- Siepel, A. and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution* **21**, 468–488.
- Stone, E.A. and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* **15**, 978–986.
- Stultz, C.M., White, J.V. and Smith, T.F. (1993). Structural analysis based on state-space modeling. *Protein Science* **2**, 305–314.
- Tamuri, A.U., dos Reis, M. and Goldstein, R.A. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* **190**, 1101–1115.
- Tamuri, A.U., Goldman, N. and dos Reis, M. (2014). A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **197**, 257–271.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**, 57–86.
- Taylor, W.R. and Jones, D.T. (1993). Deriving an amino acid distance matrix. *Journal of Theoretical Biology* **164**, 65–83.

- Thorne, J.L., Goldman, N. and Jones, D.T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution* **13**, 666–673.
- Thorne, J.L., Choi, S.C., Yu, J., Higgs, P.G. and Kishino, H. (2007). Population genetics without intraspecific data. *Molecular Biology and Evolution* **24**, 1667–1677.
- Tillier, E.R.M. and Lui, T.W.H. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **19**, 750–755.
- Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. and Blundell, T.L. (1993). Fragment ranking in modelling of protein structure: conformationally constrained substitution tables. *Journal of Molecular Biology* **229**, 194–220.
- Tuffley, C. and Steel, M. (1998). Modeling the covariation hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.
- Wako, H. and Blundell, T.L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. *Journal of Molecular Biology* **238**, 682–692.
- Wako, H. and Blundell, T.L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. *Journal of Molecular Biology* **238**, 693–708.
- Wang, H.C., Li, K., Susko, E. and Roger, A.J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evolutionary Biology* **8**, 331.
- Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**, 691–699.
- Whelan, S. and Goldman, N. (2004). Estimating the frequency of events that cause multiple nucleotide changes. *Genetics* **167**, 2027–2043.
- Wu, J., Yonezawa, T. and Kishino, H. (2017). Rates of molecular evolution suggest natural history of life history traits and a post-K-Pg nocturnal bottleneck of placentals. *Current Biology* **27**, 3025–3033.
- Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.
- Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**, 306–314.
- Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.
- Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution* **25**, 568–579.
- Yang, Z. and Roberts, D. (1995). On the use of nucleic acid sequences to infer branchings in the tree of life. *Molecular Biology and Evolution* **12**, 451–458.
- Yang, Z., Nielsen, R. and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**, 1600–1611.
- Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000). Codon-substitution models for heterogeneous selection pressure. *Genetics* **155**, 431–449.
- Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V. and Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. *Genome Biology and Evolution* **3**, 1340–1348.
- Zoller, S. and Schneider, A. (2013). Improving phylogenetic inference with a semiempirical amino acid substitution model. *Molecular Biology and Evolution* **30**, 469–479.
- Zoller, S., Boskova, V. and Anisimova, M. (2015). Maximum-likelihood tree estimation using codon substitution models with multiple partitions. *Molecular Biology and Evolution* **32**, 2208–2216.

13

Adaptive Molecular Evolution

Ziheng Yang

Department of Genetics, Evolution and Environment, University College London, London, UK

Abstract

This chapter reviews statistical methods for detecting adaptive molecular evolution by comparing synonymous and non-synonymous substitution rates in protein-coding DNA sequences. A Markov process model of codon substitution is introduced first, which forms the basis for all later discussions. The simplest analysis under the model is the comparison of two gene sequences to estimate the number of synonymous substitutions per synonymous site (d_S) and the number of non-synonymous substitutions per non-synonymous site (d_N). Both maximum likelihood and Bayesian methods have been applied, as well as a number of heuristic methods. The rest of the chapter deals with joint analyses of multiple sequences on a phylogeny. I review Markov models of codon substitution that allow the non-synonymous/synonymous rate ratio ($\omega = d_N/d_S$) to vary among branches in a phylogeny or among amino acid sites in a protein. Those models can be used to construct likelihood ratio tests to identify evolutionary lineages under episodic Darwinian selection or to infer critical amino acids in a protein under diversifying selection. I use real-data examples to demonstrate the application of the methods. The chapter finishes with a discussion of the limitations of current methods, especially when used in analysis of genome-scale data sets.

13.1 Introduction

While it is generally accepted that natural selection is the driving force for the evolution of morphological traits (including behavioral, physiological and ecological traits), the importance of natural selection in molecular evolution has been a matter of debate. The neutral theory (Kimura, 1968; King and Jukes, 1969) maintains that most observed molecular variation (both diversity within species and divergence between species) is due to random fixation of mutations with fitness effects so small that random drift rather than natural selection dominates their fate. A number of tests of neutrality have been developed in population genetics (e.g. Kreitman and Akashi, 1995; Nielsen, 2005). Those tests have been applied to identify genes under positive selection from genome-wide analysis of within-species polymorphism (Fay *et al.*, 2001; Smith and Eyre-Walker, 2002). They infer selection by rejecting the null hypothesis of neutral selection, and it is often difficult to rule out a neutral model with demographic changes. See **Chapter 14**.

Another class of methods designed to detect adaptive molecular evolution relies on comparison of synonymous (silent) and non-synonymous (amino-acid-changing) substitution rates in

protein-coding genes. The synonymous and non-synonymous rates (d_S and d_N , or equivalently K_s and K_a) are defined as the numbers of synonymous and non-synonymous substitutions per site, respectively. The ratio of the two rates, $\omega = d_N/d_S$, then measures selective pressure at the protein level. If selection has no effect on fitness, non-synonymous mutations will be fixed at the same rate as synonymous mutations, so that $d_N = d_S$ and $\omega = 1$. If non-synonymous mutations are deleterious, purifying selection will reduce their fixation rate, so that $d_N < d_S$ and $\omega < 1$. If non-synonymous mutations are favored by Darwinian selection, they will be fixed at a higher rate than synonymous mutations, resulting in $d_N > d_S$ and $\omega > 1$. A significantly higher non-synonymous rate than the synonymous rate is thus evidence for adaptive evolution at the molecular level. This criterion was used to identify many cases of positive selection in the 1990s, including the human major histocompatibility complex (MHC) (Hughes and Nei, 1988), primate stomach lysozyme (Messier and Stewart, 1997), abalone sperm lysin (Lee *et al.*, 1995), vertebrate visual pigments (Miyamoto and Miyamoto, 1996), and HIV-1 *env* genes (Bonhoeffer *et al.*, 1995; Mindell, 1996; Yamaguchi and Gojobori, 1997). Development of powerful methods, such as those reviewed in this chapter, has led to identification of many more cases of molecular adaptation (Yang, 2014, Chapter 11), providing important insights into the mechanisms of molecular evolution.

The ω ratio has most often been calculated as an average over all codons (amino acids) in the gene of interest and over the entire evolutionary time that separates the gene sequences. The criterion that such an average ω is greater than 1 is a very stringent one for detecting positive selection (e.g. Kreitman and Akashi, 1995). Many amino acids in a protein must be under strong functional constraints, with very small ω . Many proteins also appear to be under purifying selection during most of the evolutionary history. Adaptive evolution most likely occurs at a few time points and affects only a few amino acids (Gillespie, 1991). In such cases, the ω ratio averaged over time and over sites will not be greater than 1 even if Darwinian selection has operated.

A remedy for this problem is to examine the ω ratio over a short evolutionary time period or in a short stretch of the gene such as functionally important domains. For example, Messier and Stewart (1997; see also Zhang *et al.*, 1997) used inferred ancestral genes to calculate d_N and d_S for each branch in the tree and identified two lineages in the lysozyme phylogeny for primates that went through positive selection. Similarly, Hughes and Nei (1988) found that $d_N > d_S$ at 57 amino acids in the MHC that constitute the antigen-recognition site, although $d_N < d_S$ in the whole gene. Those ideas have also been implemented as likelihood ratio tests. The *branch models* account for different ω ratios among branches in the tree (Yang, 1998; Yang and Nielsen, 1998) and are used to construct likelihood ratio tests of adaptive evolution along specific lineages, and have the advantage of not relying on inferred ancestral sequences. Similarly, the *site models* allow the ω ratio to vary among amino acid sites (Nielsen and Yang, 1998; Yang *et al.*, 2000; Kosakovsky Pond and Muse, 2005; Mayrose *et al.*, 2007). They do not require knowledge of functionally important domains and are used to test for the presence of critical amino acids under positive selection, and, when they exist, to identify them. Lastly, the *branch-site models* allow the ω ratio to vary both among branches on the tree and among amino acid sites in the gene, and are suitable for identifying episodic selection that has affected only a few codons in the gene (Yang and Nielsen, 2002; Yang *et al.*, 2005).

This chapter reviews statistical methods for phylogenetic analysis of protein-coding DNA sequences, with a focus on comparing synonymous and non-synonymous substitution rates to understand the mechanisms of protein sequence evolution. First, I will provide a brief introduction to Markov chain models of codon substitution, which is the basis for maximum likelihood (ML) estimation of d_N and d_S between two sequences as well as ML joint analysis of multiple sequences on a phylogeny. I will discuss different methods for comparing two sequences to estimate d_N and d_S . Besides ML (Goldman and Yang, 1994) and Bayesian methods (Angelis *et al.*, 2014), there are about a dozen heuristic methods (e.g. Miyata and Yasunaga, 1980; Nei

and Gojobori, 1986; Li *et al.*, 1985; Li, 1993; Ina, 1995; Yang and Nielsen, 2000). I will then discuss models that account for variable ω ratios among lineages and among sites, and use real data examples to illustrate their use in ML analysis. This chapter uses ML as the general framework. ML is known to have good statistical properties in large data sets and offers insights into the heuristic methods as well.

13.2 Markov Model of Codon Substitution

In molecular phylogenetics, we use a continuous-time discrete-state Markov process to describe the change between nucleotides, amino acids, or codons over evolutionary time (e.g. Whelan *et al.*, 2001). In this chapter, we focus on analysis of protein-coding DNA sequences, and the unit of evolution is a codon in the gene. Substitutions between the sense codons are described by a Markov process. Stop codons are excluded from the Markov process as they are usually not allowed in a protein. With the ‘universal’ genetic code, there are 61 sense codons and thus 61 states in the Markov process.

The Markov process is characterized by a rate (generator) matrix $Q = \{q_{ij}\}$, where q_{ij} is the substitution rate from sense codon i to sense codon j ($i \neq j$). Formally, $q_{ij}\Delta t$ is the probability that the process is in state j after an infinitesimal time Δt , given that it is currently in state i . The basic model we use in this chapter is similar to the models of Goldman and Yang (1994) and Muse and Gaut (1994), and accounts for the transition–transversion rate difference, unequal synonymous and non-synonymous substitution rates, and unequal base/codon frequencies. Mutations are assumed to occur independently among the three codon positions, and so only one position is allowed to change instantaneously. Since transitions ($T \leftrightarrow C$ and $A \leftrightarrow G$) tend to occur at higher rates than transversions ($T, C \leftrightarrow A, G$), we multiply the rate by κ if the change is a transition. The parameter κ is the transition/transversion rate ratio. Typical estimates of κ are 1.5–5 for nuclear genes and 3–30 for mitochondrial genes. To account for unequal codon frequencies, we let π_j be the equilibrium frequency of codon j and multiply substitution rates to codon j by π_j . We can either treat all π_j as parameters, with $60 (= 61 - 1)$ free parameters used, or calculate π_j from base frequencies at the three codon positions, with $9 = 3 \times (4 - 1)$ free parameters used. It is also possible to use the π_j to incorporate features of the mutation process, such as selection on codon usage (Yang and Nielsen, 2008). Similarly the discussion here uses the HKY mutation model, which accounts for different transition/transversion rates and unequal base compositions (Hasegawa *et al.*, 1985). More complex mutation models, such as the general time-reversible model, may be used as well (Yang and Nielsen, 2008).

To account for unequal synonymous and non-synonymous substitution rates, we multiply the rate by ω if the change is non-synonymous; ω is thus the non-synonymous/synonymous rate ratio, also termed the ‘acceptance rate’ by Miyata *et al.* (1979). In models considered here, the relationship holds that $\omega = d_N/d_S$. For most genes, estimates of ω are much less than 1. The parameters κ and π_j characterize processes at the DNA level, including natural selection, while selection at the protein level has the effect of altering ω . If natural selection operates on the DNA (such as selection on codon usage and constraints of RNA structure) as well as on the protein, the synonymous substitution rate will differ from the mutation rate.

Thus, the substitution rate from codon i to codon j ($i \neq j$) is

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition.} \end{cases} \quad (13.1)$$

For example, consider substitution rates to codon CTG (which encodes amino acid Leu). We have $q_{CTC, CTG} = \pi_{CTG}$ since the CTC (Leu) \rightarrow CTG (Leu) change is a synonymous transversion, $q_{TTG, CTG} = \kappa\pi_{CTG}$ since the TTG (Leu) \rightarrow CTG (Leu) change is a synonymous transition, $q_{GTG, CTG} = \omega\pi_{CTG}$ since the GTG (Val) \rightarrow CTG (Leu) change is a non-synonymous transversion, and $q_{CCG, CTG} = \kappa\omega\pi_{CTG}$ since the CCG (Pro) \rightarrow CTG (Leu) change is a non-synonymous transition. Also $q_{TTT, CTG} = 0$ since codons TTT and CTG differ at two positions.

While q_{ij} is the substitution rate to codon j (from codon i), the diagonal element q_{ii} of the rate matrix $Q = \{q_{ij}\}$ is given by the requirement that each row of the matrix sums to 0, with $q_{ii} = -\sum_{j \neq i} q_{ij}$ (e.g. Grimmett and Stirzaker, 1992, p. 241). Furthermore, without external information, molecular sequence data do not allow separate estimation of rate and time, and only their product is identifiable: for example, we cannot tell whether a sequence has evolved into another at a certain rate over a certain time or at twice the rate over half the time. It is then conventional to fix the average rate at 1 so that time t is measured by distance, the expected number of (nucleotide) substitutions per codon. This is achieved by multiplying matrix Q by a scale factor so that the expected number of nucleotide substitutions per codon (i.e. the rate) is 1:

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1. \quad (13.2)$$

The transition probability matrix over time t is

$$P(t) = \{p_{ij}(t)\} = e^{Qt}, \quad (13.3)$$

where $p_{ij}(t)$ is the probability that codon i will become codon j after time t . As long as the rate matrix Q can be constructed, $P(t)$ can be calculated for any t using, for example, matrix diagonalization. Note that over any time interval, there is a non-zero probability that any codon i will change to any other codon j , even if they are separated by two or three nucleotide differences; that is, for any $t > 0$, $p_{ij}(t) > 0$ for any codons i and j .

Finally, the model specified by equation (13.1) is time-reversible; that is, $\pi_i q_{ij} = \pi_j q_{ji}$ for any i and j . This means that

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \quad \text{for any } t, i \text{ and } j. \quad (13.4)$$

Note that $\pi_i p_{ij}(t)$ measures the amount of change from codons i to j over time t , while $\pi_j p_{ji}(t)$ measures the change in the opposite direction. Equation (13.4), known as the 'detailed balance', means that we expect to see equal numbers of changes from i to j and from j to i . I will mention some implications of reversibility in later sections.

13.3 Estimation of Synonymous and Non-synonymous Substitution Rates between Two Sequences and Test of Selection on the Protein

13.3.1 Heuristic Estimation Methods

About a dozen heuristic methods have been proposed to estimate d_S and d_N . Important basic concepts were developed in the early 1980s (Miyata and Yasunaga, 1980; Perler *et al.*, 1980; Gojobori, 1983; Li *et al.*, 1985), which we explain here with a hypothetical example. The critical question is whether natural selection has facilitated or hindered the fixation of non-synonymous mutations. Suppose we observe five synonymous and five non-synonymous differences (substitutions) between the gene sequences from two species. Can we conclude that synonymous and non-synonymous substitution rates are equal with $\omega = 1$? The answer is 'no'.

An inspection of the genetic code table suggests that all changes at the second codon position and most changes at the first position are non-synonymous, and only some changes at the third position are synonymous. As a result, we do not expect to see equal proportions of synonymous and non-synonymous mutations even if there is no selection at the protein level. Indeed, if the codons are equally frequent and mutations from any nucleotide to any other occur at the same rate, we expect 25.5% of mutations to be synonymous and 74.5% to be non-synonymous (Yang and Nielsen, 1998). If we use those proportions, it is clear that selection on the protein has decreased the fixation rate of non-synonymous mutations by about three times, and we can estimate $\omega = (5/5)/(74.5/25.5) = 0.34$. Similarly, if we can observe or estimate the numbers of synonymous (S) and non-synonymous sites (N) as well as the numbers of synonymous and non-synonymous substitutions we can estimate d_N and d_S and hence in turn ω . For example, suppose there are 300 codons in the gene and thus 900 nucleotide sites. We can then estimate the numbers of synonymous and non-synonymous sites to be $S = 900 \times 25.5\% = 229.5$ and $N = 900 \times 74.5\% = 670.5$, respectively. Since we observed five synonymous and five non-synonymous substitutions, we thus have $d_S = 5/229.5 = 0.0218$ and $d_N = 5/670.5 = 0.0075$, which in turn leads to the same estimate of $\omega = d_N/d_S = 0.0075/0.0218 = 0.34$.

All heuristic methods involve three steps and roughly follow the procedure above (for reviews, see Ina, 1996; Yang and Nielsen, 2000). First, we count the numbers of synonymous (S) and non-synonymous (N) sites in the two sequences; that is, the number of nucleotide sites in the sequence is partitioned into the synonymous and non-synonymous categories, measuring mutation/substitution opportunities before the operation of selection on the protein. This step is complicated by factors such as transition-transversion rate difference and unequal base/codon frequencies, both of which are ignored in our hypothetical example. Second, we count synonymous and non-synonymous differences between the two sequences; that is, the observed differences between the two sequences are classified into the synonymous and non-synonymous categories. This is straightforward if the two compared codons differ at one codon position only. When they differ at two or three codon positions, there exist four or six pathways from one codon to the other. The multiple pathways may involve different numbers of synonymous and non-synonymous differences and should ideally be weighted appropriately according to their likelihood of occurrence, although most heuristic methods use equal weighting. The third step is to apply a correction for multiple substitutions at the same site (also called multiple hits) since an observed difference may be the result of two or more substitutions. In our hypothetical example, we ignored the possibility of multiple hits and treated the observed differences as substitutions. All heuristic methods have used multiple-hit correction formulas based on nucleotide-substitution models, which assume that each nucleotide can change to one of three other nucleotides. When those formulas are applied to synonymous (or non-synonymous) sites only, this basic assumption of the Markov model is violated (Lewontin, 1989; Muse, 1996). Nevertheless, such corrections are usable when the sequence divergence is low.

The method of Miyata and Yasunaga (1980) and its simplified version (Nei and Gojobori, 1986) are based on the nucleotide-substitution model of Jukes and Cantor (1969), and ignore transition-transversion rate difference or unequal base/codon frequencies. As transitions are more likely to be synonymous at the third positions than transversions are, ignoring the transition-transversion rate difference leads to underestimation of S and overestimation of N . This effect is well known, and a number of attempts have been made to account for different transition and transversion rates in counting sites and differences (Li *et al.*, 1985; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995). Similarly unequal base/codon frequencies mean that substitution rates are not symmetrical (Moriyama and Powell, 1997). The YN00 method (Yang and Nielsen, 2000) takes into account both the transition-transversion rate difference and unequal base/codon frequencies.

13.3.2 Maximum Likelihood Estimation

Likelihood is a powerful and flexible methodology for estimating parameters and testing hypotheses (see **Chapter 1**). Since the data are observed, we view the probability of observing the data as a function (the likelihood function) of the unknown parameters. The likelihood or log-likelihood function is our inference tool and contains all information in the data about the parameters in the model. We estimate the unknown parameters by maximizing the likelihood function. Furthermore, the log-likelihood value under a model measures the fit of the model to data, and we can use the likelihood ratio test (LRT) to compare two nested models. The null distribution of the test is most often approximated by the χ^2 distribution. This approximation relies on large sample sizes (long sequences) and is found to perform well when the sequences include more than 50 or 100 codons (Yang and dos Reis, 2011). When the sequences are too short or when the two models are not nested, the null distribution can instead be generated using Monte Carlo simulation (Goldman, 1993). Below I describe the ML method for estimating d_N and d_S (Goldman and Yang, 1994).

13.3.2.1 Calculation of Likelihood

The data are two aligned protein-coding DNA sequences. As a numerical example, we will use the human and mouse acetylcholine receptor α genes. The first 15 codons of the gene are as follows:

Human GAG CCC TGG CCT CTC CTC CTG CTC TTT AGC CTT TGC TCA GCT GGC ...

Mouse GAG CTC TCG ACT GTT CTC CTG CTG CTA GGC CTC TGC TCC GCT GGC ...

We assume that different codons in the sequence are evolving independently according to the same Markov process. As a result, data at different sites are independently and identically distributed. Suppose there are n sites (codons) in the gene, and let the data at site h be $\mathbf{x}_h = \{x_1, x_2\}$, where x_1 and x_2 are the two codons in the two sequences at that site (see Figure 13.1(a)). In the above example, the data at site $h = 2$ are $x_1 = \text{CCC}$ and $x_2 = \text{CTC}$. The probability of observing data \mathbf{x}_h at site h is

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2). \quad (13.5)$$

The term in the sum is the probability that the ancestor has codon k and the two current species have codons x_1 and x_2 at the site. This probability is equal to the prior probability that the ancestor has codon k , given by the equilibrium frequency π_k , multiplied by the two transition probabilities along the two branches of the tree (Figure 13.1(a)). Since the ancestral codon k is unknown, we sum over all possibilities for k . Time reversibility of the Markov process implies that.

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_k p_{x_1 k}(t_1) p_{k x_2}(t_2) = \pi_{x_1} \sum_{k=1}^{61} p_{x_1 k}(t_1) p_{k x_2}(t_2) = \pi_{x_1} p_{x_1 x_2}(t_1 + t_2). \quad (13.6)$$

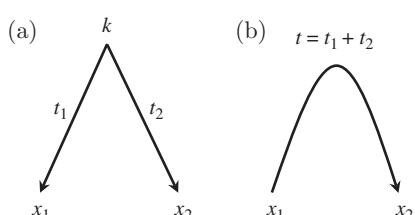


Figure 13.1 The tree for two sequences, with codons x_1 and x_2 for one codon site shown. Codon-substitution models considered in this chapter are all time-reversible and do not allow identification of the root. As a result, branch lengths t_1 and t_2 cannot be estimated separately (a), and only their sum $t = t_1 + t_2$ is estimable (b).

The last step follows from the Chapman–Kolmogorov theorem (e.g. Grimmett and Stirzaker, 1992, pp. 239–246). Thus the data are probabilistically identical whether we consider the two sequences to be descendants of a common ancestor (as in Figure 13.1(a)) or we consider one sequence to be ancestral to the other (as in Figure 13.1(b)). In other words, t_1 and t_2 cannot be estimated individually, or the root of the tree cannot be identified, and only the sequence distance $t = t_1 + t_2$ is estimable. Parameters in the model are thus the sequence distance t , the transition/transversion rate ratio κ , the non-synonymous/synonymous rate ratio ω , and the codon frequencies π_j . The log-likelihood function is then given by

$$\ell(t, \kappa, \omega) = \sum_{h=1}^n \log\{f(\mathbf{x}_h)\}. \quad (13.7)$$

An equivalent derivation of the likelihood function is to note that the data follow a multinomial distribution with 61×61 categories corresponding to the 61^2 possible site patterns (configurations), and with the multinomial probabilities given as functions of parameters in the model (t, κ, ω , etc.).

We usually estimate the codon frequencies (π_j) by the observed base/codon frequencies. To estimate parameters t, κ , and ω , we use a numerical optimization algorithm to maximize ℓ , since an analytical solution is intractable. Figure 13.2 shows a log-likelihood surface as a function of t and ω for the human and mouse acetylcholine receptor α genes. The model used in this example assumes equal transition and transversion rates and equal codon frequencies (with $\kappa = 1$ and $\pi_j = 1/61$ fixed), and thus involves two parameters. This is the model underlying the method of Miyata and Yasunaga (1980) and Nei and Gojobori (1986).

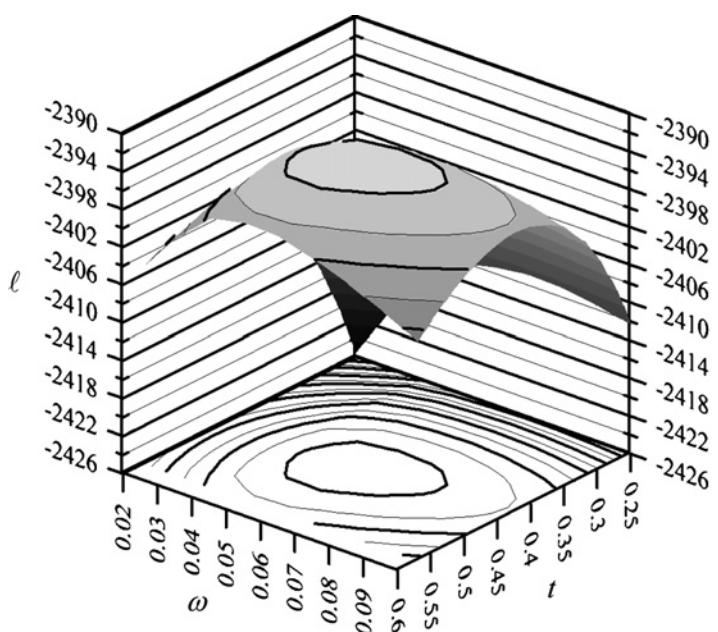


Figure 13.2 The log-likelihood surface contour as a function of parameters t and ω for the comparison of the human and mouse acetylcholine receptor α genes. The maximum likelihood method estimates parameters by maximizing the likelihood function. For these data, the estimates are $\hat{t} = 0.444$, $\hat{\omega} = 0.059$, with optimum log-likelihood $\ell = -2392.83$.

13.3.2.2 Maximum Likelihood Estimation of d_N and d_S

The distances d_N and d_S are defined as functions of parameters t , κ , ω , and π_j , and their ML estimates are simply functions of ML estimates of parameters t , κ , ω , and π_j . The description below thus gives both the definitions of d_N and d_S and also the ML method for their estimation. The basic idea is the same as explained in our hypothetical example before. Here we count sites and substitutions per codon rather than for the entire sequence. First, note that the sequence distance t is defined as the expected number of nucleotide substitutions per codon. We partition this number into the synonymous and non-synonymous categories. We note that

$$\begin{aligned}\rho_S^* &= \sum_{i \neq j, \text{aa}_i = \text{aa}_j} \pi_i q_{ij}, \\ \rho_N^* &= \sum_{i \neq j, \text{aa}_i \neq \text{aa}_j} \pi_i q_{ij}\end{aligned}\quad (13.8)$$

are the proportions of synonymous and non-synonymous substitutions, respectively, with $\rho_S^* + \rho_N^* = 1$ (equation (13.2)). The summation in ρ_S^* is taken over all codon pairs i and j ($i \neq j$) that code for the same amino acid, while the summation in ρ_N^* is taken over all codon pairs i and j ($i \neq j$) that code for different amino acids; aa_i is the amino acid encoded by codon i . The numbers of synonymous and non-synonymous substitutions per codon are then $t\rho_S^*$ and $t\rho_N^*$, respectively.

Next, we calculate the proportions of synonymous and non-synonymous *sites*. Let these be ρ_S^1 and ρ_N^1 . As noted before, these measure the substitution opportunities before the operation of selection at the protein level, that is, when $\omega = 1$ (Goldman and Yang, 1994; Ina, 1995). They are calculated similarly to equation (13.8), using the transition/transversion rate ratio κ and codon frequencies (π_j), except that $\omega = 1$ is fixed. We assume there are three nucleotide sites in a codon (see Yang and Nielsen, 1998, for a discussion of the effect of mutations to stop codons). The numbers of synonymous and non-synonymous sites per codon are then $3\rho_S^1$ and $3\rho_N^1$, respectively. The numbers of synonymous and non-synonymous substitutions per site are then $d_S = t\rho_S^*/(3\rho_S^1)$ and $d_N = t\rho_N^*/(3\rho_N^1)$, respectively. Note that $\omega = d_N/d_S = (\rho_N^*/\rho_S^*)/((\rho_N^1/\rho_S^1))$, where the numerator ρ_N^*/ρ_S^* is the ratio of the numbers of (observed) substitutions while the denominator ρ_N^1/ρ_S^1 is the ratio of the (expected) numbers of substitutions if $\omega = 1$.

Interpretation of d_N and d_S and definitions of a few other distances between two protein-coding genes were given by Yang (2014, Chapter 2). While the basic concepts discussed in the hypothetical example underlie both the ML and the heuristic methods for estimating d_N and d_S (and their ratio ω), differences exist between the two classes of methods. In the ML method, the probability theory (i.e. calculation of the transition probabilities by equation (13.3)) accomplishes several difficult tasks in one step: estimating mutational parameters such as κ , correcting for multiple hits, and weighting evolutionary pathways between codons. The Chapman–Kolmogorov theorem mentioned above ensures that the likelihood calculation accounts for all possible pathways of change between two codons, weighting them appropriately according to their relative probabilities of occurrence. When we partition the number of substitutions (t) into synonymous and non-synonymous categories, we only need to do it at the level of instantaneous rates (equation (13.8)), where there are no multiple changes, unlike comparison of two observed sequences, in which the two codons may differ at two or three codon positions.

13.3.2.3 Comparison with Heuristic Methods

In the heuristic methods, each of the three steps offers a challenge. For example, some methods ignore the transition–transversion rate difference. Others take it into account but it has been difficult to estimate κ reliably. Ina (1995) used the third codon positions and Yang and

Nielsen (2000) used so-called fourfold degenerate sites and non-degenerate sites to estimate κ , assuming that substitutions at those sites are either not affected or affected equally by selection at the protein level. Both methods use nucleotide-based correction formulas to estimate κ , which seem problematic. Use of a limited class of sites also leads to large sampling errors in the estimates. The steps of counting differences, weighting pathways, and correcting for multiple hits are complicated, when we want to incorporate major features of DNA sequence evolution such as the transition–transversion rate difference and unequal base/codon frequencies (Yang and Nielsen, 2000). Notably, the synonymous and non-synonymous status of a site changes over time and also with the nucleotides at other positions of the codon. As a result, nucleotide-substitution models used in heuristic methods may not be able to deal with the complexity of the codon-substitution process.

13.3.2.4 Testing for Selection on the Protein

Note that a significant difference between d_N and d_S means that natural selection has been operating on the protein and affected the fixation probability of non-synonymous mutations. After d_N and d_S are estimated, statistical tests can be used to test whether d_N is significantly higher than d_S . For the heuristic methods, a normal approximation is applied to the statistic ($d_N - d_S$). For ML, one can use an LRT, which compares the null hypothesis with ω fixed at 1 with the alternative hypothesis that does not place this constraint. The likelihood ratio statistic, which is defined as twice the log-likelihood difference between the two hypotheses ($2\Delta\ell$) is compared with a χ^2 distribution with one degree of freedom (df) to test whether ω differs from one. In practice, this test rarely detects positive selection (indicated by $\omega > 1$ or $d_N > d_S$), as d_N and d_S are calculated as averages over the entire sequence.

13.3.3 Bayesian Estimation

A Bayesian approach to estimating d_N and d_S between two sequences has been implemented by Angelis *et al.* (2014), by assigning gamma priors on parameters such as d and ω . This produces very similar estimates of ω to the ML estimate when the data are informative. However, when the compared sequences are very similar and non-synonymous or synonymous differences are absent, the ML estimate of ω may take the extreme values 0 or ∞ . In such cases, the Bayesian method, by applying a shrinkage through the prior, provides less extreme and biologically more reasonable estimates than the ML estimate. The method may have an advantage over ML estimation in comparative genomic studies, in which extreme ML estimates are quite common. In this regard, note that the ML estimate of ω does not have finite mean or variance, as the estimate is infinite in a proportion of data sets.

13.3.4 A Numerical Example

To see the differences among methods for estimating d_N and d_S , we compare the human and mouse acetylcholine receptor α genes (from Ohta, 1995), using ML as well as several heuristic methods (Table 13.1). The sequence has 456 codons (1368 nucleotides) after the start and stop codons are removed. With the ML method, we examine the effects of model assumptions. Some models ignore the transition/transversion rate ratio (with $\kappa = 1$ fixed) while others account for it (with κ estimated). Some ignore biased codon frequencies (Fequal) while others account for it to some extent (F1 × 4, F3 × 4, and F61; see legend to Table 13.1 for definitions of those models).

Most of these models are nested, and the χ^2 approximation can be used to perform LRTs. For example, we can compare models A and B in Table 13.1 to test whether the transition and transversion rates are equal. Model A is the null hypothesis and assumes that transition and transversion rates are equal ($\kappa = 1$). Model B does not impose this constraint and has one

Table 13.1 Estimation of d_N and d_S between the human and mouse acetylcholine receptor α genes

Model	p	$\hat{\kappa}$	\hat{S}	\hat{d}_N	\hat{d}_S	$\hat{d}_N/\hat{d}_S(\hat{\omega})$	ℓ
Heuristic methods							
Nei and Gojobori (1986)		1	321.2	0.030	0.523	0.058	
Li (1993)		N/A	N/A	0.029	0.419	0.069	
Ina (1995)		6.1	408.4	0.033	0.405	0.081	
YN00 (Yang and Nielsen 2000)		2.1	311.2	0.029	0.643	0.045	
ML methods							
(A) Fequal, $\kappa = 1$	2	1	348.5	0.029	0.496	0.059	-2392.83
(B) Fequal, κ estimated	3	2.8	396.7	0.031	0.421	0.073	-2379.60
(C) F1 \times 4, $\kappa = 1$ fixed	5	1	361.0	0.029	0.513	0.057	-2390.35
(D) F1 \times 4, κ estimated	6	2.9	406.5	0.031	0.436	0.071	-2376.12
(E) F3 \times 4, $\kappa = 1$ fixed	11	1	281.4	0.029	0.650	0.044	-2317.72
(F) F3 \times 4, κ estimated	12	3.0	328.1	0.030	0.545	0.055	-2303.33
(G) F61, $\kappa = 1$ fixed	62	1	261.5	0.028	0.736	0.038	-2251.92
(H) F61, κ estimated	63	3.0	319.5	0.030	0.613	0.048	-2239.33

Note: For ML analysis, p is the number of parameters and ℓ the log-likelihood value. Fequal: equal codon frequencies ($=1/61$) are assumed. F1 \times 4: four nucleotide frequencies are used to calculate codon frequencies (3 free parameters). F3 \times 4: nucleotide frequencies at three codon positions are used to calculate codon frequencies (9 free parameters). F61: all codon frequencies are used as free parameters (60 free parameters). Data are from Ohta (1995) and Yang and Nielsen (1998).

more free parameter (κ) than model A. The likelihood ratio statistic, $2\Delta\ell = 2 \times (-2379.60 - (-2392.83)) = 2 \times 13.23 = 26.46$, should be compared with the χ^2 distribution with $df = 1$, giving a p -value of 0.27×10^{-6} . So there is significant difference between the transition and transversion rates.

For these data, both the transition–transversion rate difference and unequal codon frequencies are clearly important. ML results under the most-complex model (F61 with κ estimated), which accounts for both factors, are expected to be the most reliable and will be used to evaluate other methods/models. The F3 \times 4 model is commonly used as it produces similar results to, and has far fewer parameters than, the F61 model. We note that heuristic methods give similar results to ML under similar models; for example, Ina's method give similar estimates to ML accounting for the transition–transversion rate difference and ignoring biased base/codon frequencies (Table 13.1 ML, Fequal, with κ estimated).

It is well known that ignoring the transition–transversion rate difference leads to underestimation of the number of synonymous sites (S), overestimation of d_S , and underestimation of the ω ratio. This effect is obvious in Table 13.1 when ML estimates with κ estimated are compared with ML estimates when κ is fixed at 1, or when the method of Nei and Gojobori (1986) is compared with those of Li (1993) and Ina (1995). Unequal codon frequencies often have opposite effects to the transition–transversion rate difference and lead to reduced numbers of synonymous sites. This is the pattern we see in Table 13.1, as estimates of S under the F3 \times 4 and F61 models are much smaller than under the Fequal model. The gene is GC-rich at the third codon position, with base frequencies to be 16% for T, 43% for C, 14% for A, and 27% for G. As a result, most substitutions at the third codon position are transversions between C and G, and there are more non-synonymous sites than expected under equal base/codon

frequencies. In this data set, the effect of unequal base frequencies is opposite to and outweighs the effect of the transition–transversion rate difference. As a result, the method of Nei and Gojobori (1986) *overestimates* rather than *underestimates* S and ω . The method of Ina (1995) accounts for the transition–transversion rate difference but ignores the codon frequency bias, and performs more poorly than the method of Nei and Gojobori (1986). The method of Yang and Nielsen (2000) accounts for both biases, and seems to produce estimates close to ML estimates under realistic models.

Whether different models and methods produce similar estimates of ω depends on a number of factors. With very little transition–transversion rate difference and little codon usage bias, different methods tend to produce similar results. For other data sets, estimates from different methods can be three or ten times different (Yang and Nielsen, 2000; Dunn *et al.*, 2001). Such large differences can occur even with highly similar sequences, as extreme transition–transversion rate difference or codon usage bias can drastically affect the counting of sites. One feature of the estimation is that when a method overestimates d_S , it tends to underestimate d_N at the same time, resulting in large errors in the ω ratio. This is because the total number of sites (or differences) is fixed, and if the method underestimates the number of synonymous sites (or differences) it will overestimate the number of non-synonymous sites as well, and vice versa. A worrying result is that the method of Nei and Gojobori (1986) can both underestimate and overestimate the ω ratio. In general, heuristic methods may be used for exploratory data analysis, and the ML method accounting for both the transition–transversion rate difference and the codon usage bias should be preferred.

13.4 Likelihood Calculation on a Phylogeny

Likelihood calculation for multiple sequences on a phylogeny may be viewed as an extension of the calculation for two sequences. The calculation is also similar to that under a nucleotide-substitution model (Felsenstein, 1981), although we now consider a codon rather than a nucleotide as the unit of evolution. We assume in this section that the same rate matrix Q (equation (13.1)) applies to all lineages and all amino acid sites. The data are multiple aligned sequences. We assume independent substitutions among sites (codons), so that data at different codon sites are independently and identically distributed. The likelihood is given by the multinomial distribution with 61^s categories (site patterns) for s sequences. Let n be the number of sites (codons) in the sequence and the data at site h be \mathbf{x}_h ($h = 1, 2, \dots, n$); x_h is a vector of observed codons in different sequences at site h . An example tree of four species is shown in Figure 13.3(a). As in the case of two sequences, the root cannot be identified, and is arbitrarily fixed at the node ancestral to sequences 1 and 2. The data \mathbf{x}_h can be generated by any codons j and k for the two ancestral nodes in the tree, and thus the probability of observing the data is a sum over all such possibilities,

$$f(\mathbf{x}_h) = \sum_j \sum_k [\pi_j p_{jx_1}(t_1) p_{jx_2}(t_2) p_{jk}(t_0) p_{kx_3}(t_3) p_{kx_4}(t_4)]. \quad (13.9)$$

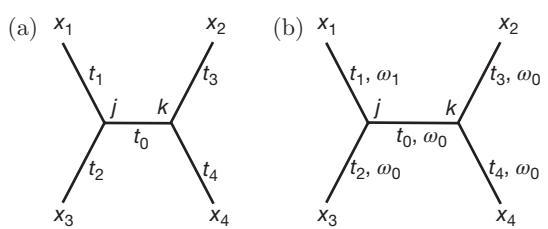


Figure 13.3 A tree of four sequences with codons at one site for nodes in the tree to illustrate (a) the basic codon model (M0) and (b) the branch model which assigns different ω parameters to different branches on the tree. Branch lengths t_0, t_1, \dots, t_4 are parameters. As the model is reversible and we do not assume the molecular clock, the root of the tree is unidentifiable and an unrooted tree is used. Adapted from Yang (1998).

The quantity in the square bracket is the contribution to $f(\mathbf{x}_h)$ from ancestral codons j and k , and is equal to the probability that the codon at the root is j (which is given by the equilibrium frequency π_j), multiplied by the five transition probabilities along the five branches of the phylogeny (Figure 13.3(a)). For a tree of s species with $s - 2$ ancestral nodes, the data at each site will be a sum over 61^{s-2} possible combinations of ancestral codons. In computer programs, we use the ‘pruning’ algorithm of Felsenstein (1981) to achieve efficient computation.

The log-likelihood is a sum over all sites in the sequence,

$$\ell = \sum_{h=1}^n \log\{f(\mathbf{x}_h)\}. \quad (13.10)$$

Compared with the case of two sequences, we now have the same parameters in the substitution model (κ , ω , and the π_j), but many more branch length parameters (e.g. t_0, t_1, \dots, t_4 in Figure 13.3(a) instead of the single t in Figure 13.1(b)). Again, numerical optimization algorithms have to be used to maximize the likelihood function.

It may be mentioned that equation (13.9) also gives the empirical Bayes approach (also known as the likelihood approach) to reconstructing ancestral character states (Yang *et al.*, 1995; Koshi and Goldstein, 1996). As mentioned above, the quantity in the square bracket is the contribution to the probability of the data $f(\mathbf{x}_h)$ by ancestral codons j and k . This contribution varies greatly depending on j and k , and the codons j and k that make the greatest contribution are the most probable codons for the two ancestral nodes at the site. This Bayes approach has the advantage, over the parsimony algorithm (Fitch, 1971; Hartigan, 1973), of using uses branch lengths and relative substitution rates between character states. Ancestral sequence reconstruction is widely used to construct heuristic methods for detecting adaptive molecular evolution, as will be discussed later in comparison with the ML method.

13.5 Detecting Adaptive Evolution along Lineages

13.5.1 Likelihood Calculation under Models of Variable ω Ratios among Lineages

It is easy to modify the model of the previous section to allow for different ω ratios among branches on a tree. The likelihood calculation under such a model proceeds in a similar way, except that the transition probabilities for different branches need to be calculated from different rate matrices (Q) generated using different values of ω . Suppose we want to fit a model in which the branch for species 1 of Figure 13.3(b) has a different ω ratio (ω_1), while all other branches have the same ‘background’ ratio ω_0 . Let $p_{ij}(t; \omega)$ denote the transition probability calculated using the ratio ω . Under this model, the probability of observing data \mathbf{x}_h is

$$f(\mathbf{x}_h) = \sum_j \sum_k \pi_j p_{jx_1}(t_1; \omega_1) p_{jx_2}(t_2; \omega_0) p_{jk}(t_0; \omega_0) p_{kx_3}(t_3; \omega_0) p_{kx_4}(t_4; \omega_0). \quad (13.11)$$

(compare with equation (13.9)).

Yang (1998) implemented models that allow for different levels of heterogeneity in the ω ratio among lineages. The simplest model (the ‘one-ratio’ model) assumes the same ω ratio for all branches in the phylogeny. The most general model (the ‘free-ratio’ model) assumes an independent ω ratio for each branch in the phylogeny. Intermediate models such as two- or three-ratio models assume two or three different ω ratios for lineages in the tree. Those models can be compared using the LRT to examine interesting hypotheses. For example, the likelihood values under the one-ratio and free-ratio models can be compared to test whether the ω ratios are different among lineages. Also, we can allow the lineages of interest to have a different ω ratio

from the background ω ratio for all other lineages in the phylogeny (as in Figure 13.3(b)). Such a two-ratio model can be compared with the one-ratio model to examine whether the lineages of interest have a different ω ratio from other lineages. Furthermore, when the estimated ω ratio for the lineages of interest (say, ω_1 in Figure 13.3(b)) is greater than 1, models with and without the constraint that $\omega_1 = 1$ can be compared to test whether the ratio is different from (i.e. greater than) 1. This test directly examines the possibility of positive selection along specific lineages.

It should be pointed out that variation in the ω ratio among lineages may not be sufficient evidence for adaptive evolution. First, relaxed selective constraints along certain lineages can generate variable ω ratios. Second, if non-synonymous mutations are slightly deleterious but not lethal, their fixation probabilities will depend on factors such as the population size of the species. In large populations, deleterious mutations will have a smaller chance of getting fixed than in small populations. Under such a model of slightly deleterious mutations (Ohta, 1973), species with large population sizes are expected to have smaller ω ratios than species with small population sizes. At any rate, an ω ratio significantly greater than 1 provides convincing evidence for Darwinian selection.

13.5.2 Adaptive Evolution in the Primate Lysozyme

In the following, we use the example of the lysozyme *c* genes of primates (Figure 13.4) to demonstrate the use of codon substitution models of variable ω ratios among lineages (Yang, 1998). Lysozyme is found mainly in secretions like tears and saliva as well as in white blood cells, where its function is to fight invading bacteria. Leaf-eating colobine monkeys have a complex foregut where bacteria ferment plant material, followed by a true stomach that expresses high levels of lysozyme, where its new function is to digest these bacteria (Stewart *et al.*, 1987; Messier and Stewart, 1997). It has been suggested that the acquisition of a new function may have led to high selective pressure on the enzyme, resulting in high non-synonymous substitution rates. In an analysis of lysozyme *c* genes from 24 primate species, Messier and Stewart (1997) identified two lineages with elevated ω ratios, indicating episodes of adaptive evolution in the lysozyme. One lineage, expected from previous analysis (Stewart *et al.*, 1987), is ancestral to colobine monkeys, and another, unsuspected, lineage is ancestral to the hominoids. The two lineages are represented by branches *h* and *c* in Figure 13.4, for a subset of the data of Messier and Stewart (1997).

As branches *h* and *c* are the lineages of interest, we test assumptions concerning three ω ratio parameters: ω_h for branch *h*, ω_c for branch *c*, and ω_0 for all other (background) branches. Table 13.2 lists log-likelihood values and ML parameter estimates under different models. The simplest model assumes one ω ratio (Table 13.2A) while the most general model assumes three

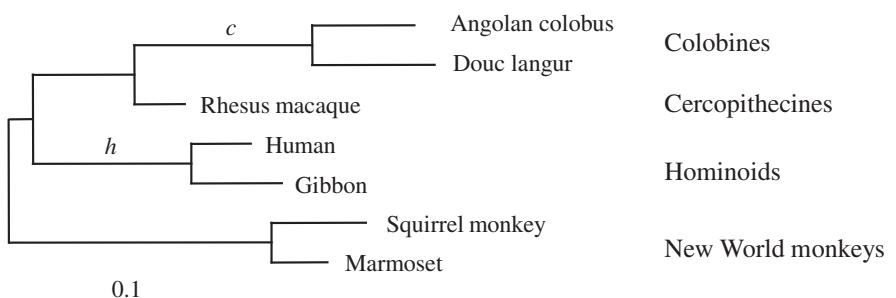


Figure 13.4 Phylogeny of seven primate species, for a subset of the lysozyme data of Messier and Stewart (1997), used to demonstrate the branch model with different ω ratios among branches. The analysis uses the unrooted tree, but the root is shown here for clarity. After Yang (1998).

Table 13.2 Log-likelihood values and parameter estimates under different models for the lysozyme *c* genes

Model	<i>p</i>	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_h$	$\hat{\omega}_c$
A. 1 ratio: $\omega_0 = \omega_h = \omega_c$	22	-906.02	4.5	0.81	= $\hat{\omega}_0$	= $\hat{\omega}_0$
B. 2 ratios: $\omega_0 = \omega_h, \omega_c$	23	-904.64	4.6	0.69	= $\hat{\omega}_0$	3.51
C. 2 ratios: $\omega_0 = \omega_c, \omega_h$	23	-903.08	4.6	0.68	∞	= $\hat{\omega}_0$
D. 2 ratios: $\omega_0, \omega_h = \omega_c$	23	-901.63	4.6	0.54	7.26	= $\hat{\omega}_H$
E. 3 ratios: $\omega_0, \omega_h, \omega_c$	24	-901.10	4.6	0.54	∞	3.65
F. 2 ratios: $\omega_0 = \omega_h, \omega_c = 1$	22	-905.48	4.4	0.69	= $\hat{\omega}_0$	1
G. 2 ratios: $\omega_0 = \omega_c, \omega_h = 1$	22	-905.38	4.4	0.68	1	= $\hat{\omega}_0$
H. 2 ratios: $\omega_0, \omega_h = \omega_c = 1$	22	-904.36	4.3	0.54	1	1
I. 3 ratios: $\omega_0, \omega_h, \omega_c = 1$	23	-902.02	4.5	0.54	∞	1
J. 3 ratios: $\omega_0, \omega_h = 1, \omega_c$	23	-903.48	4.4	0.54	1	3.56

Note: *p* is the number of parameters. All models include the following 21 common parameters: 11 branch lengths in the tree (Figure 13.4), 9 parameters for base frequencies at codon positions used to calculate codon frequencies, and the transition-transversion rate ratio κ . Source: Yang (1998).

ratios (E, I & J). The six possible two-ratio models (B-D and F-H) are used as well. In Models F-J, the ω ratio for the branch(es) of interest is fixed at 1.

The estimate of ω under the one-ratio model ($\omega_0 = \omega_h = \omega_c = \omega$) is 0.81, indicating that, on average, purifying selection dominates the evolution of the lysozyme. Estimates of ω_c for branch *c* range from 3.4 to 3.6 when ω_c is allowed free to vary (models B, E, and J in Table 13.2). Estimates of ω_h are always infinite when ω_h is assumed to be a free parameter (models C, E, and I), indicating the absence of synonymous substitutions along branch *h*. The estimate of the background ratio ω_0 is 0.54, when ω_h and ω_c are not constrained to be equal to ω_0 (models D, E, I, and J).

Results of LRTs are shown in Table 13.3. Tests A–E examine whether the ω ratio for the branch(es) of interest is different from (i.e. greater than) the background ratio, while tests A'–E' examine whether the ratio is greater than 1. For example, test E compares models B and E of Table 13.2 and examines the null hypothesis that $\omega_h = \omega_0$, with ω_c free to vary in both models; ω_h is significantly higher than ω_0 in this comparison. Such tests suggest that ω_h is significantly greater than the background ratio ω_0 ($P < 1\%$; Table 13.3, D and E) and also significantly greater than 1 ($P < 5\%$; Table 13.3, D' and E'). Similar tests suggest that ω_c is significantly greater than ω_0 ($P < 5\%$; Table 13.3, C), but not significantly greater than 1 (P ranges from 17% to 20%; Table 13.3, B' and C'). More detailed analyses of the dataset can be found in Yang (1998).

13.5.3 Comparison with Methods Based on Reconstructed Ancestral Sequences

Evolutionary biology has had a long tradition of reconstructing characters in extinct ancestral species and using them as observed data in all sorts of analyses. For molecular data, statistical methods (Yang *et al.*, 1995; Koshi and Goldstein, 1996) can be used to obtain more reliable ancestral reconstructions, taking into account branch lengths and relative substitution rates between characters (nucleotides, amino acids, or codons) (see the discussion below equation (13.10)). Overall, reconstructed molecular sequences appear much more reliable than reconstructed morphological characters (Yang *et al.*, 1995; Cunningham *et al.*, 1998).

Table 13.3 Likelihood ratio statistics ($2\Delta\ell$) for testing hypotheses concerning lysozyme evolution

Hypothesis tested	Assumption made	Models compared	$2\Delta\ell$
A. $(\omega_h = \omega_c) = \omega_0$	$\omega_h = \omega_c$	A & D	8.78**
B. $\omega_c = \omega_0$	$\omega_h = \omega_0$	A & B	2.76
C. $\omega_c = \omega_0$	ω_h free	C & E	3.96*
D. $\omega_h = \omega_0$	$\omega_c = \omega_0$	A & C	5.88*
E. $\omega_h = \omega_0$	ω_c free	B & E	7.08**
A'. $(\omega_h = \omega_c) \leq 1$	$\omega_h = \omega_c$	D & H	5.46*
B'. $\omega_c \leq 1$	$\omega_h = \omega_0$	B & F	1.68
C'. $\omega_c \leq 1$	ω_h free	E & I	1.84
D'. $\omega_h \leq 1$	$\omega_c = \omega_0$	C & G	4.60*
E'. $\omega_h \leq 1$	ω_c free	E & J	4.76*

* Significant at the 5% level, with $\chi^2_{1,5\%} = 3.84$.

** Significant at the 1% level, with $\chi^2_{1,1\%} = 6.63$.

Source: Yang (1998).

Messier and Stewart (1997) reconstructed ancestral sequences and used them to perform pairwise comparisons to calculate d_N and d_S along branches in the tree. Their analysis pinpointed two particular lineages in the primate phylogeny that may have gone through adaptive evolution. Crandall and Hillis (1997) took the same approach in an analysis of relaxed selective constraints in the rhodopsin genes of eyeless crayfishes living deep under the ground. Zhang *et al.* (1997) argue that the normal approximation to the statistic $d_N - d_S$ may not be reliable due to small sample sizes. Those authors instead applied Fisher's exact test to the counts of differences between the two sequences, ignoring multiple hits at the same site.

A major difference between the ML method discussed in this section and the heuristic approaches using ancestral reconstruction is that ML uses all possible ancestral characters (such as codons j and k for the two ancestral nodes in the tree of Figure 13.3), while the approach of ancestral reconstruction uses only the most likely codons and ignores the others. Ancestral sequences reconstructed by both parsimony and likelihood involve random errors, as indicated by the calculated posterior probabilities (Yang *et al.*, 1995). Using the optimal reconstruction and ignoring the suboptimal ones may cause a systematic bias. One kind of such bias is obvious if we use reconstructed ancestral sequences to estimate branch lengths, as both parsimony and likelihood tend to minimize the amount of evolution to select the most likely ancestral characters. Biases involved in estimation of d_N and d_S using reconstructed ancestral sequences are more complex. They can be reduced by considering sub-optimal as well as optimal reconstructions (Goldstein and Pollock, 2006; Matsumoto *et al.*, 2015), but this has not been pursued in the context of codon models. Furthermore, pairwise comparisons along branches of the phylogeny may not be as reliable as a simultaneous comparison of all sequences by ML.

It appears advisable that ancestral reconstruction be used for exploratory data analysis and ML be used for more rigorous tests. When the LRT suggests adaptive evolution along certain lineages, ancestral reconstruction may be very useful to pinpoint the responsible amino acid changes for experimental verification. Indeed, an interesting use of ancestral reconstruction is to provide ancestral proteins to be synthesized in the laboratory to examine its biochemical and physiological properties. Such studies of 'paleobiochemistry' were envisaged by Pauling

and Zuckerkandl (1963) several decades ago (Golding and Dean, 1998; Chang and Donoghue, 2000; Thornton, 2004).

13.6 Inferring Amino Acid Sites under Positive Selection

13.6.1 Likelihood Ratio Test under Models of Variable ω Ratios among Sites

Up to now, we have assumed that all amino acid sites in a protein are under the same selective pressure, with the same underlying non-synonymous/synonymous rate ratio (ω). While the synonymous rate may be homogeneous among sites, non-synonymous rates are well known to be highly variable. Most proteins have highly conserved amino acid positions at which the underlying ω ratio is close to zero. The requirement that the ω ratio, averaged over all sites in the protein, is greater than 1 is thus a very stringent criterion for detecting adaptive evolution. It would be much more realistic if we allowed the ω ratio to vary among sites.

We can envisage two scenarios, corresponding to fixed-effects and random-effects models in statistics. In the first scenario, we may know the different structural and functional domains of the protein, and can use such information to classify amino acid sites in the protein into several classes. The different site classes are assumed to have different ω ratios, which are parameters to be estimated by ML. Suppose we have K site classes, with the corresponding ω ratios $\omega_1, \omega_2, \dots, \omega_K$. The likelihood calculation under this model is rather similar to that under the model of one ω ratio for all sites (equations (13.9) and (13.10)), except that the correct ω ratio will be used to calculate the transition probabilities for data at each site. For example, if site h is from site class k ($k = 1, 2, \dots, K$) with the ratio ω_k , then $f(\mathbf{x}_h)$ of equation (13.9) will be calculated using ω_k . The likelihood is again given by equation (13.10). A few such models were implemented by Yang and Swanson (2002) and applied to MHC class I alleles, where structural information was used to identify amino acids at the antigen recognition site. The models were termed 'fixed-sites' models.

In the second scenario, we assume that there are several heterogeneous site classes with different ω ratios, but we do not know which class each amino acid site belongs to. Such models are termed 'random-sites' models by Yang and Swanson (2002) and will be the focus of discussion here. They use a statistical distribution to account for the random variation of ω among sites (Nielsen and Yang, 1998; Yang *et al.*, 2000). We assume that the synonymous rate is constant among sites, and allow only the non-synonymous rate to be variable, although the same approach can be taken to deal with synonymous rate variation (Kosakovsky Pond and Muse, 2005; Mayrose *et al.*, 2007). The branch length t is defined as the expected number of nucleotide substitutions per codon, averaged over sites. Suppose amino acid sites fall into K classes, with the proportions p_0, p_1, \dots, p_{K-1} , and ω ratios $\omega_0, \omega_1, \dots, \omega_{K-1}$ treated either as parameters or as functions of parameters in the ω distribution. To calculate the likelihood, we want to calculate the probability of observing data at each site, say data \mathbf{x}_h at site h . The conditional probability of the data given ω_k , $f(\mathbf{x}_h|\omega_k)$, can be calculated as described before (equation (13.9)). Since we do not know which class site h belongs to, we sum over all site classes (i.e. over the distribution of ω):

$$f(\mathbf{x}_h) = \sum_{k=1}^K p_k f(\mathbf{x}_h|\omega_k). \quad (13.12)$$

The log-likelihood is a sum over all n sites in the sequence,

$$\ell = \sum_{h=1}^n \log\{f(\mathbf{x}_h)\}. \quad (13.13)$$

Parameters in the model include branch lengths in the tree, κ , π_j , and parameters in the distribution of ω among sites. As before, we estimate the codon frequency parameters by the observed frequencies, and estimate the other parameters by numerical optimization of the likelihood.

A number of statistical distributions have been implemented by Nielsen and Yang (1998) and Yang *et al.* (2000). Positive selection is tested using an LRT comparing a null model that does not allow $\omega > 1$ with an alternative model that does. Computer simulations have highlighted two tests to be particularly effective (Anisimova *et al.*, 2001, 2002; Wong *et al.* 2004). The first compares the null model M1a (neutral), which assumes two site classes in proportions p_0 and $p_1 = 1 - p_0$ with $0 < \omega_0 < 1$ and $\omega_1 = 1$, and the alternative model M2a (selection), which adds a proportion p_2 of sites with $\omega_2 > 1$. M1a and M2a are slight modifications of models M1 and M2 in Nielsen and Yang (1998), which had $\omega_0 = 0$ fixed and which were found to be highly unrealistic for most data sets. As M2a has two more parameters than M1a, the χ^2 distribution may be used for the test. However, the regularity conditions for the asymptotic χ^2 approximation are not met, as M1a is equivalent to M2a by fixing $p_2 = 0$, which is at the boundary of the parameter space and as ω_2 is not identifiable when $p_2 = 0$. The use of χ^2 is expected to be conservative. The second test compares the null model M7 (beta), which assumes a beta distribution for ω , and the alternative model M8 (beta& ω), which adds an extra site class of positive selection with $\omega_s > 1$. The beta distribution $\text{beta}(p, q)$ can take a variety of shapes depending on its parameters p and q , such as L-, inverted L-, U-, and inverted U-shapes, but is restricted to the interval $(0, 1)$. It is thus a flexible null model. M8 has two more parameters than M7, so that χ^2 may be used to conduct the LRT. As in the comparison between M1a and M2a, use of χ^2 is expected to make the test conservative.

Another model, called M3 (discrete), is sometimes useful as well. This assumes a general discrete model, with the frequencies and the ω ratios (p_k and ω_k in equation (13.12)) for K site classes estimated as free parameters. All models discussed here may be considered special cases of this general mixture model. Model M3 may be compared with model M0 (one-ratio) to construct an LRT to test whether the selective pressure varies among sites.

After ML estimates of model parameters are obtained, we can use the empirical Bayes approach to infer the most likely site class (and thus the ω ratio) for any site. The marginal probability of the data $f(\mathbf{x})$ (equation (13.13)) is a sum of contributions from each site class k , and the site class that makes the greatest contribution is the most likely class for the site. That is, the posterior probability that a site with data \mathbf{x}_h is from site class k (with rate ratio ω_k) is

$$f(\omega_k | \mathbf{x}_h) = \frac{p_k f(\mathbf{x}_h | \omega_k)}{\sum_j p_j f(\mathbf{x}_h | \omega_j)} \quad (13.14)$$

(Nielsen and Yang, 1998). When the ω estimates for some site classes are greater than 1, this approach can be used to identify sites from such classes, which are potential targets of positive selection. The posterior probability provides a measure of accuracy. This is known as the naive empirical Bayes approach (NEB; Nielsen and Yang, 1998). A serious drawback with this approach is that it uses the MLEs of parameters as fixed constants in equation (13.14), ignoring their sampling errors. This may be a sensible approach in large or medium-sized data sets. However, in small datasets, the information content may be low and the parameter estimates may involve large sampling errors. This appears to be the major reason for the poor performance of the procedure in small data sets in several simulation studies (e.g. Anisimova *et al.*, 2002; Wong *et al.*, 2004; Massingham and Goldman, 2005; Scheffler and Seoighe, 2005). A more reliable approach is implemented by Yang *et al.* (2005), known as the Bayes empirical Bayes (BEB). BEB accommodates uncertainties in the MLEs of parameters in the ω distribution by

integrating numerically over a prior for the parameters. Other parameters such as branch lengths are fixed at their MLEs, as these are expected to have much less effect on inference concerning ω . A hierarchical (full) Bayesian approach is implemented by Huelsenbeck and Dyer (2004), using Markov chain Monte Carlo to average over tree topologies, branch lengths, as well as other substitution parameters in the model. This approach involves more computation but may produce more reliable inference in small uninformative data sets, where the MLEs of branch lengths may involve large sampling errors (Scheffler and Seoighe, 2005).

13.6.2 Methods that Test One Site at a Time

A heuristic approach to examining selective pressure indicated by the ω ratio at individual sites is to reconstruct ancestral sequences and then count synonymous and non-synonymous changes at each site. Comparison of the observed counts with a ‘neutral’ expectation may then allow us to decide whether the site is evolving under purifying selection or positive selection. On a large phylogeny, many changes may have accumulated at a single site for this approach to be feasible. Fitch *et al.* (1997) performed such an analysis of the hemagglutinin (HA) gene of human influenza virus type A and considered a site to be under positive selection if it has more non-synonymous substitutions than the average over the gene. Suzuki and Gojobori (1999) compared the counts with the neutral expectation that $d_S = d_N$ at the site, using the method of Nei and Gojobori (1986) to estimate d_S and d_N . Both the approaches of Fitch *et al.* (1997) and Suzuki and Gojobori (1999) used the parsimony algorithm to infer ancestral sequences. A large number of sequences are needed for the test to have any power (Suzuki and Gojobori, 1999; Wong *et al.*, 2004).

The use of reconstructed ancestral sequences may be a source of concern since the inferred sequences are not real observed data. In particular, positively selected sites are often the most variable sites in the alignment, at which ancestral reconstruction is the least reliable. This problem may be avoided by taking a likelihood approach, averaging over all possible ancestral states. Indeed, Suzuki (2004), Massingham and Goldman (2005), and Kosakovsky Pond and Frost (2005) implemented methods to estimate one ω parameter for each site using ML. Then at every site, an LRT is used to test the null hypothesis $\omega = 1$. This is called the *site-wise likelihood ratio* (SLR) test by Massingham and Goldman (2005) and the *fixed-effects likelihood* (FEL) model by Kosakovsky Pond and Frost (2005). A potential problem with those methods is that the number of ω ratios estimated in the model increases without bound with the increase of the sequence length, and ML is known to misbehave in such infinitely-many-parameter models (Stein, 1956; Felsenstein, 2001). Nevertheless, computer simulations have found good performance of those methods in large data sets with many sequences (Massingham and Goldman, 2005; Spielman and Wilke, 2016), and they have been used to estimate selective coefficients at individual sites in the protein sequence (Tamuri *et al.*, 2014). Note that one can achieve increased power for estimating ω for individual sites by including more sequences in the alignment.

The methods that test every site for positive selection are similar to the BEB procedure for identifying individual amino acid sites under positive selection (Yang *et al.*, 2005). To test whether the gene is under positive selection (i.e. whether the gene has any codon with $\omega > 1$), a correction for multiple testing should be applied (Wong *et al.*, 2004).

13.6.3 Positive Selection in the HIV-1 *vif* Genes

An example data set of HIV-1 *vif* genes from 29 subtype-B isolates is used here to demonstrate the likelihood models of variable ω ratios among sites. The data set was analyzed by Yang *et al.* (2000). The sequence has 192 codons. Several models are used in ML estimation, with the results shown in Table 13.4. Only parameters involved in the ω distribution are listed, as other

Table 13.4 Likelihood values and parameter estimates under models of variable ω ratios among sites for HIV-1 vif genes

Model code	ℓ	d_N/d_S	Estimates of parameters
M0. one-ratio (1)	-3499.60	0.644	$\hat{\omega} = 0.644$
M1a. neutral (2)	-3393.83	0.438	$\hat{p}_0 = 0.611, \hat{\omega}_0 = 0.080, (\hat{p}_1 = 0.389), (\hat{\omega}_1 = 1)$
M2a. selection (4)	-3367.86	0.689	$\hat{p}_0 = 0.573, \hat{\omega}_0 = 0.090, \hat{p}_1 = 0.346, (\hat{\omega}_1 = 1), (\hat{p}_2 = 0.081), \hat{\omega}_2 = 3.585$
M3. discrete (5)	-3367.16	0.742	$\hat{p}_0 = 0.605, \hat{\omega}_0 = 0.108, \hat{p}_1 = 0.325, \hat{\omega}_1 = 1.211, (\hat{p}_2 = 0.070), \hat{\omega}_2 = 4.024$
M7. beta (2)	-3400.44	0.440	$\hat{p} = 0.176, \hat{q} = 0.223$
M8. beta& ω (4)	-3370.66	0.687	$\hat{p}_0 = 0.909, \hat{p} = 0.222, \hat{q} = 0.312, (\hat{p}_1 = 0.091), \hat{\omega}_s = 3.385$

Note: The number of parameters in the ω distribution is given in parentheses after the model code. d_N/d_S is the average ω over all sites in the gene. Parameters in parentheses are given to ease interpretation but are not free parameters. Estimates of the transition–transversion rate ratio κ range from 3.6 to 4.1 among models. The data are from Yang *et al.* (2000).

parameters (branch lengths in the phylogeny, the transition–transversion rate ratio κ , and the base frequencies at the three codon positions) are common to all models. The model codes are those used in Yang *et al.* (2000) and in the CODEML program in the PAML package (Yang, 1997).

The one-ratio model (M0) assumes one ω ratio for all sites and gives an average ω ratio of 0.644, indicating that on average purifying selection is the dominating force during the evolution of the gene. The selection model (M2a) suggests about $\hat{p}_1 = 8.1\%$ of sites are under positive selection with $\hat{\omega}_2 = 3.58$, and has a significantly higher log-likelihood value than model M1a (neutral), which does not allow for sites under positive selection. The LRT statistic for comparing M1a and M2a is $2\Delta\ell = 2 \times 25.97 = 51.94$, much greater than $\chi^2_{2,1\%} = 9.21$. Similarly estimates under model M8 (beta& ω) suggest that about $\hat{p}_1 = 9.1\%$ of sites are under positive selection with $\hat{\omega}_s = 3.385$. M8 fits the data much better than M7 (beta), which does not allow for sites with $\omega > 1$. Thus the two LRTs, which compare M1a with M2a, and M7 with M8, provide significant evidence for the presence of sites in the gene under positive selection.

The discrete model (M3) suggests that about $\hat{p}_2 = 7.0\%$ of sites are under strong positive selection with $\hat{\omega}_2 = 4.0$, while a large proportion ($\hat{p}_1 = 33\%$) of sites are under weak positive selection or are nearly neutral with $\hat{\omega}_1 = 1.2$.

Figure 13.5 plots the posterior probabilities for site classes at each site under model M2a (selection), calculated using the NEB approach (equation (13.14)). Parameter estimates under M2a suggest that the three site classes have the ω ratios 0.090, 1 and 3.585 and are in proportions 57.3%, 34.6% and 8.1% (Table 13.4). Those proportions are the prior probabilities for site classes at every site. The observed data at the site alter those prior probabilities considerably, so that the posterior probabilities are very different from the prior. For example, the posterior probabilities for site 1 are 0.986, 0.014, and 0.000, and this site is almost certainly under purifying selection. In contrast, the probabilities at site 31 are 0.000, 0.012, and 0.988, and this site is most likely to be under strong diversifying selection (Figure 13.5).

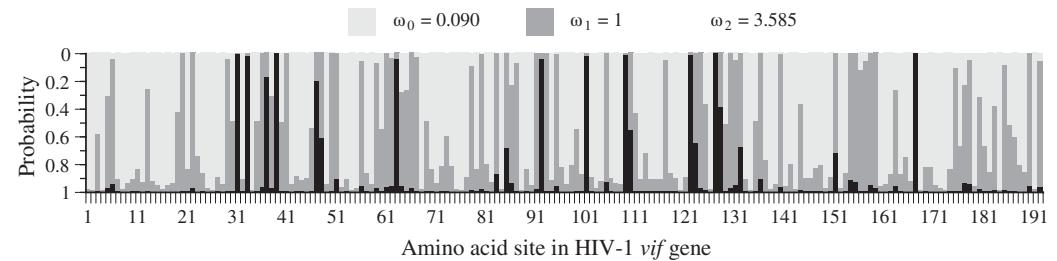


Figure 13.5 Posterior probabilities of site classes along the gene for the HIV-1 *vif* genes under the site model M2a (selection).

The above calculation used the NEB procedure (equation (13.14)), which treats the estimates of the proportions and ω ratios as true values. NEB is used here as it is simple to explain. In real data analysis, the BEB procedure should be used instead, which takes into account estimation errors in those parameters. The HIV *vif* data set is large enough for reliable estimation of the parameters in the ω distribution so that the two procedures produced very similar results. For example, at the 99% cutoff, both methods identified three sites to be under positive selection: 39F, 127Y, and 167K (the amino acids are from the reference sequence B_SF2). At the 95% cutoff, BEB identified five additional sites: 31I, 33K, 101G, 109L, and 122N, while the NEB list also included 63K and 92K.

13.7 Testing Positive Selection Affecting Particular Sites and Lineages

13.7.1 Branch-Site Test of Positive Selection

A natural extension to the branch and site models discussed above is the *branch-site* models, which allow the ω ratio to vary both among sites and among lineages (Yang and Nielsen, 2002; Yang *et al.*, 2005; Zhang *et al.*, 2005). Such models attempt to detect signals of local episodic natural selection (Gillespie, 1991), which affects only a few sites along particular lineages. Branches in the tree are divided *a priori* into foreground and background categories, and an LRT is constructed to compare an alternative hypothesis that allows for some sites under positive selection on the foreground branches with a null hypothesis that does not. Table 13.5 shows branch–site model A (Yang *et al.*, 2005; Zhang *et al.*, 2005). Along the background lineages, there are two classes of sites: the conserved sites with $0 < \omega_0 < 1$ and the neutral sites with $\omega_1 = 1$. Along the foreground lineages, a proportion $(1 - p_0 - p_1)$ of sites become under positive selection with $\omega_2 \geq 1$. Likelihood calculation under this model is very similar to that under the site models (equation (13.12)). As we do not know *a priori* which site class each site is from, the probability

Table 13.5 The ω ratios assumed in branch-site model A

Site class	Proportion	Background ω	Foreground ω
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1 - p_0 - p_1)p_0 / (p_0 + p_1)$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$(1 - p_0 - p_1)p_1 / (p_0 + p_1)$	$\omega_1 = 1$	$\omega_2 > 1$

Note: The model involves four parameters: p_0 , p_1 , ω_0 , ω_2 .

of data at a site is an average over the four site classes. Let $I_h = 0, 1, 2a, 2b$ be the site class that site h is from. We have

$$f(\mathbf{x}_h) = \sum_{I_h} p_k f(\mathbf{x}_h | I_h). \quad (13.15)$$

The conditional probability $f(\mathbf{x}_h | I_h)$ of observing data \mathbf{x}_h at site h , given that the site comes from site class I_h , is easy to calculate, because the site evolves under the one-ratio model if $I_h = 0$ or 1, and under the branch model if $I_h = 2a$ or $2b$.

To construct an LRT, we use model A as the alternative hypothesis, while the null hypothesis is the same model A but with $\omega_2 = 1$ fixed (Table 13.5). This is known as the branch–site test of positive selection. The null hypothesis has one parameter fewer, but since $\omega_2 = 1$ is fixed at the boundary of the parameter space of the alternative hypothesis, the null distribution should be a 50:50 mixture of point mass 0 and χ^2_1 (Self and Liang, 1987). The critical values are 2.71 and 5.41 at the 5% and 1% levels, respectively. One may also use χ^2 (with critical values 3.84 at 5% and 5.99 at 1%) to guide against violations of model assumptions.

As in the site-based analysis, the BEB approach can be used to calculate the posterior probability that a site is from site classes 2a and 2b, allowing identification of amino acid sites potentially under positive selection along the foreground lineages (Yang *et al.*, 2005).

Similar to the branch test, the branch–site test requires the foreground branches to be specified *a priori*. This may be easy if a well-formulated biological hypothesis exists, for example, if we want to test adaptive evolution driving functional divergences after gene duplication. The test may be difficult to apply if no *a priori* hypothesis is available. To apply the test to several or all branches on the tree, one has to correct for multiple testing (Anisimova and Yang, 2007).

13.7.2 Clade Models and Other Variants

Several other models of codon substitution also allow the ω ratio to vary both among lineages and among sites. Forsberg and Christiansen (2003) and Bielawski and Yang (2004) implemented the *clade* models. Branches on the phylogeny are *a priori* divided into two or more clades, and an LRT is used to test for divergences in selective pressure between the two clades indicated by different ω ratios. The clade models follow the ideas of Gu (2001) and Knudsen and Miyamoto (2001), who used different amino acid substitution rates as a proxy for protein functional divergence. There may not be any sites under positive selection with $\omega > 1$. Clade model C, implemented by Bielawski and Yang (2004), is summarized in Table 13.6. This assumes three site classes. Class 0 includes conserved sites with $0 < \omega_0 < 1$, while class 1 include neutral sites with $\omega_1 = 1$; both apply to all lineages. Class 2 includes sites that are under different selective pressures in the different clades, with ω_2 for clade 1, ω_3 for clade 2, and so on. With two clades, the model involves five parameters in the ω distribution: $p_0, p_1, \omega_0, \omega_2$, and ω_3 . An LRT can be

Table 13.6 The ω ratios assumed in clade models C and D

Site class	Proportion	Clade 1	Clade 2	Clade 3	Clade 4
0	p_0	ω_0	ω_0	ω_0	ω_0
1	p_1	ω_1	ω_1	ω_1	ω_1
2	$p_2 = 1 - p_0 - p_1$	ω_2	ω_3	ω_4	ω_5

Note: In clade model C, $\omega_1 = 1$ is fixed, while in clade model D, ω_1 ($0 < \omega_1 < \infty$) is estimated as a free parameter. In both models, ω_0 is estimated under the constraint $0 < \omega_0 < 1$.

constructed by comparing model C with the site model M1a (neutral), which assumes two site classes with two free parameters: p_0 and ω_0 (see Section 13.6). The χ^2 distribution may be used for the test. A more appropriate null model for the LRT was suggested by Weadick and Chang (2012), which is a variant of M2a and has three site classes with $\omega_0 < 1$, $\omega_1 = 1$, and $0 < \omega_2 < \infty$. The test then directly examines whether the ω ratios for the third site class differ among clades.

A *switching model* is implemented by Guindon *et al.* (2004), which allows the ω ratio at any site to switch among three different values: $\omega_1 < \omega_2 < \omega_3$. Besides the Markov process to describe substitutions between codons, a hidden Markov chain runs over time and describes the switches of any site between different selective regimes (i.e. the three ω values). The model has the same structure as the covarion model of Tuffley and Steel (1998; see also Galtier, 2001; Huelsenbeck, 2002), which allows the substitution rate for any site to switch between high and low values. The switching model is an extension of the site model M3 (discrete) discussed above, under which the ω ratio is fixed at every site. An LRT can thus be used to compare them. Guindon *et al.* (2004) found that the switching model fitted a data set of the HIV-1 *env* genes much better than the site models. An empirical Bayes procedure can be used to identify lineages and sites with high ω ratios, and it appears necessary to apply a correction for multiple testing.

Kosakovsky Pond *et al.* (2011) developed the random-effects branch–site model, which does not require the *a priori* specification of the foreground branches, but instead specifies the ω value for any branch–site combination as a random draw from a discrete distribution. The model may be useful for hypothesis generation in an exploratory analysis, when no biological knowledge is available to formulate the hypothesis to be tested.

13.8 Limitations of Current Methods

Both the test of positive selection along lineages and the test of positive selection at amino acid sites discussed may be expected to be conservative. The test for lineages under positive selection detects positive selection along a lineage only if the ω ratio averaged over all sites is significantly greater than 1. Since many or most sites in a protein are under purifying selection with the underlying ω ratios close to 0, this procedure constitutes a very conservative test. Similarly, the test for positively selected sites detects positive selection only if the underlying ω ratio averaged over all lineages is greater than 1. This assumption appears unrealistic except for genes under recurrent diversifying selection; for most genes, positive selection probably affects only a few lineages and only a few sites. In this regard, the branch–site and switching models may be more powerful as they allow the ω ratio to vary both among lineages and among sites.

The models discussed in this chapter assume the same ω ratio for all possible amino acid changes; at a positively selected site, changes to any amino acids are assumed to be advantageous. This assumption is unrealistic and appears to make the test conservative. It is well known that amino acids with similar physicochemical properties tend to exchange with each other at higher rates than dissimilar amino acids (Dayhoff *et al.*, 1965; Yang *et al.*, 1998). Some authors distinguish between radical and conservative amino acid replacements, and suggest that a higher radical than conservative rate is evidence for positive selection (Hughes *et al.*, 1990; Rand *et al.*, 2000; Zhang, 2000). However, this criterion is less convincing than the simple ω ratio and is found to be sensitive to assumptions about transition and transversion rate differences and unequal amino acid compositions (Dagan *et al.*, 2002).

The tests discussed here identify positive selection only if it causes excessive non-synonymous substitutions, and may have little power in detecting other types of selection such as balancing selection (Yang *et al.*, 2000). Furthermore, they require a moderate amount of

sequence divergence to operate and tend to lack power in population data. They may be useful for species data or fast-evolving viral genes only.

13.9 Computer Software

Heuristic methods for estimating d_N and d_S between two sequences, such as those of Nei and Gojobori (1986), Li *et al.* (1985), Li (1993) and Pamilo and Bianchi (1993), are implemented in MEGA (Kumar *et al.*, 2016) and in the CODEML program in the PAML package (Yang, 2007). Likelihood methods both for estimating d_N and d_S between two sequences (Goldman and Yang, 1994) and for joint sequence analysis on a tree are implemented in CODEML. The HyPhy package (Kosakovsky Pond *et al.*, 2005) also implements a number of the likelihood models discussed in this chapter.

References

- Angelis, K., dos Reis, M. and Yang, Z. (2014). Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. *Molecular Biology and Evolution* **31**, 1902–1913.
- Anisimova, M. and Yang, Z. (2007). Multiple hypothesis testing to detect adaptive protein evolution affecting individual branches and sites. *Molecular Biology and Evolution* **24**, 1219–1228.
- Anisimova, M., Bielawski, J.P. and Yang, Z. (2001). The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. *Molecular Biology and Evolution* **18**, 1585–1592.
- Anisimova, M., Bielawski, J.P. and Yang, Z. (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19**, 950–958.
- Bielawski, J.P. and Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *Journal of Molecular Evolution* **59**, 121–132.
- Bonhoeffer, S., Holmes, E.C. and Nowak, M.A. (1995). Causes of HIV diversity. *Nature* **376**, 125.
- Chang, B.S. and Donoghue, M.J. (2000). Recreating ancestral proteins. *Trends in Ecology and Evolution* **15**, 109–114.
- Comeron, J.M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159.
- Crandall, K.A. and Hillis, D.M. (1997). Rhodopsin evolution in the dark. *Nature* **387**, 667–668.
- Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution* **13**, 361–366.
- Dagan, T., Talmor, Y. and Graur, D. (2002). Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Molecular Biology and Evolution* **19**, 1022–1025.
- Dayhoff, M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (1965). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
- Dunn, K.A., Bielawski, J.P. and Yang, Z. (2001). Substitution rates in *Drosophila* nuclear genes: implications for translational selection. *Genetics* **157**, 295–305.
- Fay, J.C., Wyckoff, G.J. and Wu, C.-I. (2001). Positive and negative selection on the human genome. *Genetics* **158**, 1227–1234.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Felsenstein, J. (2001). Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution* **53**, 447–455.

- Fitch, W.M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.
- Fitch, W.M., Bush, R.M., Bender, C.A. and Cox, N.J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 7712–7718.
- Forsberg, R. and Christiansen, F.B. (2003). A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Molecular Biology and Evolution* **20**, 1252–1259.
- Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution* **18**, 866–873.
- Gillespie, J.H. (1991). *The Causes of Molecular Evolution*. Oxford University Press, Oxford.
- Gojobori, T. (1983). Codon substitution in evolution and the ‘saturation’ of synonymous changes. *Genetics* **105**, 1011–1027.
- Golding, G.B. and Dean, A.M. (1998). The structural basis of molecular adaptation. *Molecular Biology and Evolution* **15**, 355–369.
- Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182–198.
- Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.
- Goldstein, R.A. and Pollock, D.D. (2006). Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Molecular Biology and Evolution* **23**, 1444–1449.
- Grimmett, G.R. and Stirzaker, D.R. (1992). *Probability and Random Processes*. Clarendon Press, Oxford.
- Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution* **18**, 453–464.
- Guindon, S., Rodrigo, A.G., Dyer, K.A. and Huelsenbeck, J.P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12957–12962.
- Hartigan, J.A. (1973). Minimum evolution fits to a given tree. *Biometrics* **29**, 53–65.
- Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160–174.
- Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Molecular Biology and Evolution* **19**, 698–707.
- Huelsenbeck, J.P. and Dyer, K.A. (2004). Bayesian estimation of positively selected sites. *Journal of Molecular Evolution* **58**, 661–672.
- Hughes, A.L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170.
- Hughes, A.L., Ota, T. and Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Molecular Biology and Evolution* **7**, 515–524.
- Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution* **40**, 190–226.
- Ina, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: An indicator of mechanisms of molecular evolution. *Journal of Genetics* **75**, 91–115.
- Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In H. N. Munro (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–123.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.
- King, C.E. and Jukes, T.H. (1969). Non-Darwinian evolution. *Science* **164**, 788–798.

- Knudsen, B. and Miyamoto, M.M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 14512–14517.
- Kosakovsky Pond, S.L. and Frost, S.D.W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**, 1208–1222.
- Kosakovsky Pond, S.L., Frost, S.D.W. and Muse, S.V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679.
- Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W. and Scheffler, K. (2011). A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution* **28**, 3033–3043.
- Kosakovsky Pond, S.L. and Muse, S.V. (2005). Site-to-site variation of synonymous substitution rates. *Molecular Biology and Evolution* **22**, 2375–2385.
- Koshi, J.M. and Goldstein, R.A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution* **42**, 313–320.
- Kreitman, M. and Akashi, H. (1995). Molecular evidence for natural selection. *Annual Review of Ecology and Systematics* **26**, 403–422.
- Kumar, S., Stecher, G. and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**, 1870–1874.
- Lee, Y.-H., Ota, T. and Vacquier, V.D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution* **12**, 231–238.
- Lewontin, R. (1989). Inferring the number of evolutionary events from DNA coding sequence differences. *Molecular Biology and Evolution* **6**, 15–32.
- Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.
- Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.
- Massingham, T. and Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–1762.
- Matsumoto, T., Akashi, H. and Yang, Z. (2015). Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. *Genetics* **200**, 873–890.
- Mayrose, I., Doron-Faigenboim, A., Bacharach, E. and Pupko, T. (2007). Towards realistic codon models: Among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* **23**, i319–327.
- Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154.
- Mindell, D.P. (1996). Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 3284–3288.
- Miyamoto, S. and Miyamoto, R. (1996). Adaptive evolution of photoreceptors and visual pigments in vertebrates. *Annual Review of Ecology and Systematics* **27**, 543–567.
- Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution* **16**, 23–36.
- Miyata, T., Miyazawa, S. and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution* **12**, 219–236.
- Moriyama, E.N. and Powell, J.R. (1997). Synonymous substitution rates in *Drosophila*: Mitochondrial versus nuclear genes. *Journal of Molecular Evolution* **45**, 378–391.

- Muse, S.V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Molecular Biology and Evolution* **13**, 105–114.
- Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.
- Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.
- Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197–218.
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.
- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98.
- Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* **40**, 56–63.
- Pamilo, P. and Bianchi, N.O. (1993). Evolution of the *Zfx* and *Zfy* genes – rates and interdependence between the genes. *Molecular Biology and Evolution* **10**, 271–281.
- Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular ‘restoration studies’ of extinct forms of life. *Acta Chemica Scandinavica* **17**, S9–S16.
- Perler, F., Efstratiadis, A., Lomedica, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980). The evolution of genes: The chicken preproinsulin gene. *Cell* **20**, 555–566.
- Rand, D.M., Weinreich, D.M. and Cezairliyan, B.O. (2000). Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrial-encoded proteins. *Gene* **261**, 115–125.
- Scheffler, K. and Seoighe, C. (2005). A Bayesian model comparison approach to inferring positive selection. *Molecular Biology and Evolution* **22**, 2531–2540.
- Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.
- Smith, N.G. and Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024.
- Spielman, S.J. and Wilke, C.O. (2016). Extensively parameterized mutation-selection models reliably capture site-specific selective constraint. *Molecular Biology and Evolution* **33**, 2990–3002.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In J. Neyman (ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, pp. 197–206.
- Stewart, C.-B., Schilling, J.W. and Wilson, A.C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401–404.
- Suzuki, Y. (2004). New methods for detecting positive selection at single amino acid sites. *Journal of Molecular Evolution* **59**, 11–19.
- Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16**, 1315–1328.
- Tamuri, A.U., Goldman, N. and dos Reis, M. (2014). A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **197**, 257–271.
- Thornton, J. (2004). Resurrecting ancient genes: Experimental analysis of extinct molecules. *Nature Reviews Genetics* **5**, 366–375.
- Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences* **147**, 63–91.
- Weadick, C.J. and Chang, B.S. (2012). An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. *Molecular Biology and Evolution* **29**, 1297–1300.

- Whelan, S., Liò, P. and Goldman, N. (2001). Molecular phylogenetics: State of the art methods for looking into the past. *Trends in Genetics* **17**, 262–272.
- Wong, W.S.W., Yang, Z., Goldman, N. and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**, 1041–1051.
- Yamaguchi, Y. and Gojobori, T. (1997). Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 1264–1269.
- Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Computer Applications in BioSciences* **13**, 555–556.
- Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586–1591.
- Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford.
- Yang, Z. and dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution* **28**, 1217–1228.
- Yang, Z. and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**, 409–418.
- Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.
- Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**, 908–917.
- Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution* **25**, 568–579.
- Yang, Z. and Swanson, W.J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* **19**, 49–57.
- Yang, Z., Kumar, S. and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650.
- Yang, Z., Nielsen, R. and Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**, 1600–1611.
- Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-M.K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.
- Yang, Z., Wong, W.S.W. and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107–1118.
- Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution* **50**, 56–68.
- Zhang, J., Kumar, S. and Nei, M. (1997). Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes. *Molecular Biology and Evolution* **14**, 1335–1338.
- Zhang, J., Nielsen, R. and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**, 2472–2479.

14

Detecting Natural Selection

Aaron J. Stern¹ and Rasmus Nielsen²

¹Graduate Group in Computational Biology, University of California, Berkeley, Berkeley, CA

²Department of Integrative Biology, University of California; Department of Statistics, University of California, Berkeley, Berkeley, CA

Abstract

Understanding natural selection is at the core of many evolutionary and population genetic investigations. However, it is typically difficult to directly detect natural selection. Instead, it has to be inferred from observations of DNA sequence data. In this chapter, we will briefly introduce some standard models of natural selection used in population genetics. We will then review some of the main signatures of selection that can be identified by analyses of DNA sequence data, and finally provide an overview of some of the many different statistical methods that have been developed to identify natural selection. We will argue that the lack of tractable likelihood approaches has spurred a large literature on more *ad hoc* statistical approaches based on summary statistics.

14.1 Introduction

Natural selection arises when individuals differ in fitness due to genetic factors – that is, have heritable differences in survival probability (viability) or reproductive success (fertility). Other factors, such as mutation and genetic drift,¹ are also important evolutionary factors. However, selection plays a special role in driving adaptation, the evolutionary changes in response to environmental stimuli. Hence, understanding selection is the key to understanding how populations adapt to environments. Furthermore, at the molecular level, determining which genetic variants affect fitness provides information about which variants are important in interactions with the environment, including the response and susceptibility to disease.

In population genetics, the effects of natural selection are modeled as changes in allele frequencies. However, changes in allele frequencies can only rarely be directly observed, at least at this point in time (although with the increasing number of ancient DNA samples, this may become increasingly viable; see **Chapter 10**). Most studies aimed at detecting selection, therefore, focus on inferring selection indirectly from contemporary samples of DNA sequences. In this chapter we will discuss some of the statistical methods commonly used to infer selection (Section 14.4). Before doing so, we will briefly review some of the population genetic theory on natural selection (Section 14.2), and the general signatures associated with natural selection

1 The random sampling of gametes via which a new generation is formed by the previous generation.

(Section 14.3). Although we will review some basics of negative and balancing selection, we will focus much of our review on methods for detecting positive selection.

14.2 Types of Selection

We begin by reviewing the most common models of selection in population genetics theory, acquainting the reader with terminology and properties of these models.

For the sake of simplicity, we will initially consider selection acting on a single di-allelic locus, with alleles A and a , in a diploid population. The changes in the frequency of A from generation to generation – the *trajectory* of the allele – are in part determined by the relative fitnesses of the three possible genotypes, w_{AA} , w_{Aa} , and w_{aa} . We can write these relative fitnesses in terms of the *selection coefficients*, s_{AA} , s_{Aa} , and s_{aa} ,² and if we assume that the frequency of A at generation t is $X_t \in [0, 1]$, we expect the frequency in the subsequent generation X_{t+1} to be

$$E[X_{t+1}|X_t = p] = p \frac{1 + s_{Aa}q + s_{AA}p}{1 + 2s_{Aa}pq + s_{AA}p^2},$$

where $q = 1 - p$. We can then use standard techniques for recurrence relations to describe the expected trajectory through time. However, since genetic drift is acting on the population at the same time, the recurrence relation based on the expectations will only be a rough approximation that tends to work well when the effect of selection is strong relative to genetic drift. Adding genetic drift to the model results in discrete-time Markov chain models, such as the familiar Wright-Fisher model, which describes the trajectory in discrete generations forward in time assuming binomial sampling of alleles between generations (Fisher, 1999; Wright, 1931). For more details about this model, see Chapter 4, this volume.

These discrete generation models can be approximated in continuous time using diffusion equations (Kimura, 1955a,b) which have revealed many interesting mathematical results regarding natural selection, such as the probability of fixation of an allele (the probability that it reaches a frequency of 100%) or the expected time it will take for the allele to reach fixation or loss (Kimura, 1962; Ewens, 2012). One important insight gained from theoretical population genetics is the fact that the effect of genetic drift is stronger in populations with smaller effective population sizes (N_e). Population geneticists, therefore, often see genetic drift and selection as two different forces acting at the same time, where N_e and the selection coefficients determine which of these two forces have the strongest effect on the dynamics of the allele frequency trajectory.

14.2.1 Directional Selection

Models of selection on a di-allelic locus are sometimes classified into directional positive selection, directional negative selection, or balancing selection based on the values of the selection coefficients. Directional positive selection is the case where $w_{AA} \geq w_{Aa} \geq w_{aa}$, excluding the case where $w_{AA} = w_{Aa} = w_{aa}$, if A is the derived (mutant) allele. In this case we expect the derived allele frequency to increase through time; that is, $E[X_{t+1}|X_t = p] > p$ if $0 < p < 1$, and in the absence of genetic drift the selected allele will eventually go to fixation (reach a frequency of 100%).

Similarly, directional negative selection is the case where $w_{AA} \leq w_{Aa} \leq w_{aa}$, excluding the case where $w_{AA} = w_{Aa} = w_{aa}$. Here we expect the frequency of A to decrease on average, and in

² There is a one-to-one mapping between relative fitnesses and selection coefficients. For example, here we choose to define the selection coefficients by $w_{AA} = 1 + s_{AA}$, $w_{Aa} = 1 + s_{Aa}$, $w_{aa} = 1$, $s_{aa} = 0$. Then, given w_{AA} , we can obtain s_{AA} , and given w_{Aa} we can obtain s_{Aa} , and vice versa.

the absence of genetic drift A will approach a frequency of 0%. The special case of $w_{AA} = w_{Aa} = w_{aa}$ is the neutral case in which no selection is acting, and $E[X_{t+1}|X_t = p] = p$. In this case, fluctuations from p are only due to genetic drift, rather than both genetic drift and selection. Throughout this chapter, we often refer to a single selection coefficient s , rather than s_{AA} and s_{Aa} ; unless stated otherwise, we assume that $s_{Aa} = s$ and $s_{AA} = 2s$, that is, selection on A/a is additive, as well as positive directional.

14.2.2 Balancing Selection

Unlike directional selection, in which alleles under selection tend to be lost or fixed in the long term, balancing selection refers to selection schemes that maintain multiple alleles in the population. One situation that produces this effect is heterozygote advantage (overdominance), where $w_{Aa} > w_{AA}$ and $w_{Aa} > w_{aa}$. In this case, if we ignore the effects of drift, X_t converges over time to an intermediate frequency; if $w_{Aa} = 1$, $w_{AA} = 1 - s_{AA}$, and $w_{aa} = 1 - s_{aa}$, then the equilibrium frequency $x^* = \lim_{t \rightarrow \infty} X_t = s_{aa}/(s_{AA} + s_{aa})$, rather than the boundaries at 0 or 1. By contrast, under heterozygote disadvantage, where $w_{Aa} < w_{AA}$ and $w_{Aa} < w_{aa}$, there may exist an equilibrium frequency $x^* \in (0, 1)$ in the absence of genetic drift; however, if genetic drift causes the frequency to fluctuate away from x^* , then we expect X_t to be fixed or lost in the long term. For a deeper discussion of these models as well as the genomic signatures of balancing selection, we direct the reader to Charlesworth (2006).

In addition to heterozygote advantage, other selection schemes that can maintain variation include time- and space-varying selection. In these scenarios, directional selection can maintain variation, so long as the sign of the selection coefficient changes with sufficient frequency, either over time or space. Additionally, when the absolute fitness of an allele depends negatively on its own frequency – so-called negative frequency-dependent selection – alleles can also stabilize at intermediate frequencies.

14.2.3 Polygenic Selection

The simple di-allelic models described above are probably not realistic for much of the selection acting on the genomes of humans or most other organisms. Typically, a trait will be affected by multiple mutations and selection will, therefore, be polygenic (see, for example, Shi *et al.*, 2016). While much of the early literature in population genetics focused on selection affecting a single locus for reasons of mathematical simplicity, there has recently been a resurgence of interest in polygenic selection. This interest stems in part from the realization due to genome-wide association studies (GWASs) that most human traits of interest are highly polygenic (see, for example, Boyle *et al.*, 2017). However, methods for detecting polygenic selection from DNA sequence data are still in their infancy, and this chapter will focus primarily on methods aimed at detecting selection affecting a single locus. However, we note that a promising and very active research area is the development of methods for detecting and analyzing polygenic selection, and we later review several advances in this regard.

14.3 The Signature of Selection in the Genome

In the previous section we reviewed the basic behavior of alleles under selection assuming a simple di-allelic model of selection acting on a single locus. The trajectory of allele frequency changes can be estimated directly from experiments of viral or bacterial evolution (Bollback *et al.*, 2008; Lang *et al.*, 2013; Good *et al.*, 2017) or from analyses of ancient DNA (see, for example, Lazaridis *et al.*, 2014; Mathieson *et al.*, 2015; **Chapter 10**, this volume), and can be used

for quantifying and detecting selection (Malaspinas *et al.*, 2012; Feder *et al.*, 2014; Schraiber *et al.*, 2016). However, most of the time, such direct inference of the trajectory of allele frequency change is not possible. Instead, inference regarding past selection has to be made solely from observations of modern DNA. In this section we discuss some of the patterns that can be observed in DNA sequences that have been subject to selection. We then review statistical methods for detecting and quantifying these patterns.

14.3.1 The Signature of Positive Directional Selection

We begin by reviewing the signatures that arise due to positive directional selection (see Section 14.2.1). In this section we review signatures such as increased rates of substitution, changes in the allele frequency distribution around selected alleles, and the hitchhiking effect.

14.3.1.1 Rates of Substitution

An obvious consequence of positive selection is that favored alleles will have increased rates of substitution – that is, the rate at which these alleles fix at a frequency of 100% is greater than the rate at which neutral alleles fix. Many methods for detecting selection take advantage of this insight. However, there are factors other than selection that can increase the rate of substitution, such as increased mutation rate. Therefore, methods aimed at detecting positive selection by identifying increased rates of substitution must employ some standard of comparison to control for these factors. A common way of doing this is to compare mutations that *a priori* are, or are not, expected to be more likely to be under selection than other mutations. The most common comparison is of the number of non-synonymous and synonymous mutations that have fixed in protein coding regions (e.g. Hughes and Nei, 1988). Due to the redundancy of the genetic code, mutations in protein coding regions come in two flavors: those that change the amino acid sequence (non-synonymous changes) and those that do not (synonymous changes). We expect that most selection in protein coding regions acts at the amino acid level and that non-synonymous mutations, therefore, are more likely to experience selection than synonymous mutations. Hence, a signature of positive direction selection would be an increased number of fixed non-synonymous mutations compared to the number of fixed synonymous mutations. However, we note that selection may also act on synonymous mutations due factors such as codon usage preferences or maintenance of splice sites (Hershberg and Petrov, 2008; Sorek and Ast, 2003; Hockenberry *et al.*, 2018).

14.3.1.2 Frequencies of Selected Alleles

Wright (1938) showed that under equilibrium conditions, mutations are more likely to segregate at higher frequencies if they are under selection than if they are not. Thus, the frequency of an allele at a single point in time in itself provides information on whether a particular allele is under selection. However, if we make the assumption that the vast majority of mutations entering the population are more or less selectively neutral, then even high-frequency derived alleles must be primarily neutral (Eyre-Walker and Keightley, 2007). Allele frequency alone is therefore not a reliable indicator of selection, and we must look to additional genomic signatures to improve our power to discriminate between selection and neutrality.

Nonetheless, on aggregate, selected mutations will tend to have different frequencies than neutral mutations. Comparisons of the distribution of allele frequencies in different categories of mutations, such as non-synonymous and synonymous mutations, can therefore be used to infer selection acting on sets of mutations. The distribution of allele frequencies – the so-called *site frequency spectrum* (SFS) – in models of selection is typically modeled using Poisson

random field models pioneered by Sawyer and Hartl (1992). In these models, mutations enter the population according to a Poisson process, and selection and drift then act on the mutations to modify allele frequencies. Comparisons of the SFS stratified by different categories of mutations are an important tool in analyses of genomic data (Boyko *et al.*, 2008). In particular, selection acting on a specific category of sites causes the SFS for that category to differ from that of a category of sites assumed to be neutral, or the expected SFS under selective neutrality. The latter has a particularly simple expression: the expected proportion of mutations with allele frequency i , in a sample of size n , is given by $1/(ia_n)$, where $i \in \{1, 2, \dots, n-1\}$ and $a_n = \sum_{j=1}^{n-1} j^{-1}$ is a normalizing factor (Fu, 1995).

Hitherto, we have mostly considered polymorphisms in a panmictic population. Another common signature of selection is differentiated allele frequencies among populations. Natural selection can increase the level of genetic differentiation among populations if selection acts differently in different populations or geographic regions due to differences in environmental factors (Lewontin and Krakauer, 1973). Similarly, increased genetic differentiation among populations could also happen due to selection if selection acts on a recently arisen mutation that has not yet spread to other populations (Santiago and Caballero, 2005). In fact, increased genetic differentiation among populations is one of the most characteristic signatures of natural selection. However, many highly differentiated allele frequencies may be driven by a combination of genetic drift and restricted gene flow, rather than selection.

14.3.1.3 Hitchhiking

So far we have discussed the direct effect of selection on the selected allele itself. However, selection also affects variation at linked neutral sites in the genome. When a selected allele increases in frequency, linked neutral alleles will also increase in frequency. This is the so-called 'hitchhiking' effect (Smith and Haigh, 1974; Kaplan *et al.*, 1989). The consequence is a *selective sweep* (Figure 14.1), which in the genomic region surrounding the favored allele will lead to decreased variability (e.g. the number of segregating sites), increased identity by descent (IBD; i.e. DNA sequence identity due to recent common ancestry), and increased haplotype homozygosity (Braverman *et al.*, 1995). Importantly, these patterns differ at different times during the sweep. During the earlier phase in which the favored allele has reached intermediate frequencies in the population (an incomplete sweep), haplotypes carrying the selected allele are highly uniform, since these haplotypes have increased in frequency so fast that recombination and mutation have not had much time to act. Consequently, haplotype homozygosity and IBD at this locus will be high (Sabeti *et al.*, 2002; Albrechtsen *et al.*, 2010). Furthermore, as we demonstrate in Figure 14.2, neutral linked alleles will hitchhike along with the selected allele, resulting in an excess of alleles at frequency greater than or equal to that of the selected allele. As the sweep completes, haplotype homozygosity increases, and the frequencies of linked alleles shift towards 100% along with the selected allele; thus, the SFS in a region that has recently undergone a sweep often is bimodal, with peaks at the frequencies 1 and $n-1$. However, this signature of bimodality is transient, as with time many of these high-frequency hitchhiking alleles will fix. Thus, a longer-term signature of a completed sweep is an SFS with an excess of low-frequency alleles due to recent mutation in the region around the selected allele. With passing time and accruing mutation and recombination, so too the patterns of increased IBD and haplotype homozygosity in this region become less pronounced.

Another effect of hitchhiking is a change in linkage disequilibrium (LD) around the selected site. Kim and Nielsen (2004) showed that while LD between neutral linked alleles on either side of the selected allele increases on average, LD between the neutral linked alleles from opposite sides will be erased by the selective sweep.

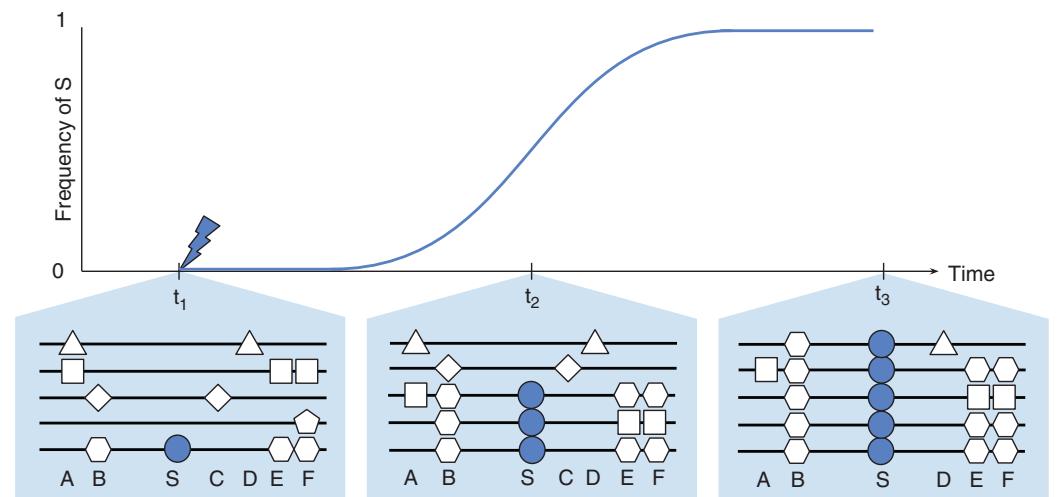


Figure 14.1 Genetic variation changes as a sweep progresses due to hitch-hiking. Individual haplotypes ($n = 5$) are denoted using a different shape for each haplotype in the sample at t_1 , to keep track of recombination events during the sweep. At t_1 , the selected allele S mutates into the population on the \square background. At this stage, there are six neutral polymorphisms ($A-F$) in the sample. At t_2 , the sweep is ongoing and the frequency of the selected allele is intermediate (incomplete selective sweep). Relative to a neutral allele, fewer recombination events have occurred around the selected allele by the time it reaches this frequency, due to its rapid increase in frequency driven by selection. As a result, diversity within haplotypes carrying S is much reduced compared to haplotypes carrying the disfavored ancestral allele, that is, there has been an increase in haplotype homozygosity within the allelic class carrying S . Note that at this stage, two recombinations have occurred, both between \square and \circlearrowleft . At t_3 , S has swept to fixation, along with the B allele. Another recombination event has occurred between \triangle and \circlearrowleft . Note the increase in high-frequency derived alleles and overall reduced levels of variability at t_3 relative to t_1 and t_2 , exemplified by the loss of diversity at three sites (B, C, S). Furthermore, the sample is IBD for the entire tract $B-S$.

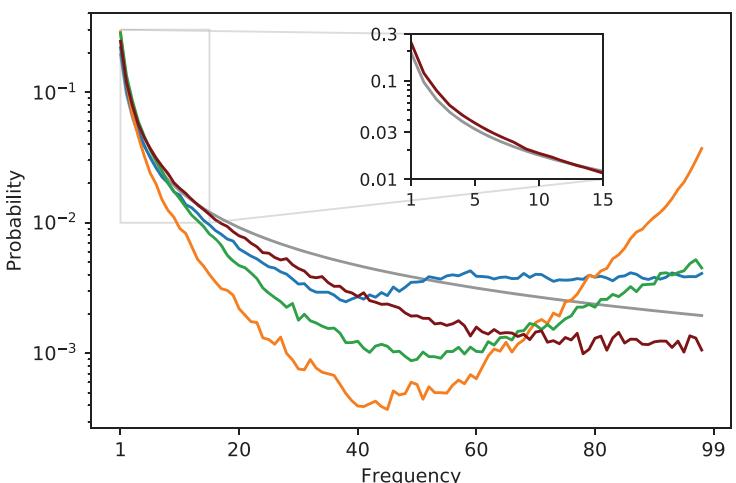


Figure 14.2 The SFS ($n = 100$) of a genomic region undergoing a selective sweep, in an equilibrium population of $N = 10,000$ sampled during different points in time in a selective sweep with $s = 0.1$. We let $\theta = \rho = 100$ (where θ and ρ are the population-scaled mutation and recombination rates, respectively) and average the SFS over 100 independent simulations. Gray, the null expected SFS; blue, the favored allele is at 50% in the population (incomplete sweep); orange, the favored allele just fixed; green, 4000 generations after fixation; maroon, 12,000 generations after fixation. Inset: Depicting the excess of low-frequency alleles lasting long after fixation.

So far, we have discussed models of the effect of a new, favored mutation rapidly increasing in frequency in the population. This model is known as a *hard sweep*. An alternative model involves *soft sweeps* (Hermisson and Pennings, 2005), in which selection is acting on either recurrent mutations or on standing variation, that is, on alleles that were already segregating in the population by the time that selection started to act. The signature of a soft sweep can differ substantially from that of a hard sweep. Przeworski *et al.* (2005) showed that while a sweep from standing variation (SSV) from a low frequency (less than $(2N_e s)^{-1}$) results in the same reduction in diversity as a hard sweep, SSVs affecting alleles of sufficiently high initial frequency (greater than $(2N_e s)^{-1}$) do not result in this decrease in diversity. The increased diversity resulting from SSVs relative to hard sweeps is due to accumulation of recombination near the selected site during the allele's neutral phase, before selection started acting.

Soft sweeps do share some signatures with hard sweeps, including decreased variability and increased haplotype homozygosity as well as increased IBD (Garud *et al.*, 2015; Albrechtsen *et al.*, 2010). However, these signatures are less pronounced for soft sweeps, because in this type of sweep several distinct haplotypes will increase in frequency, rather than a single haplotype under a hard sweep.

14.3.2 Balancing Selection

The effect of balancing selection is in many ways opposite to that of a selective sweep. Mutations are maintained in the population for a prolonged period of time, leading to an increase in variability in the region around the selected variant. The SFS in regions surrounding the selected allele will contain many more alleles of intermediate frequency than expected for neutral regions. For a hard selective sweep, as selection becomes stronger, the width of the genomic region affected by the sweep becomes larger. However, for balancing selection, the width of the region in which linked neutral variants are affected by selection is narrow and of the order of $(2N_e r)^{-1}$, where r is the recombination rate per site (Hudson and Kaplan, 1988).

14.3.3 Polygenic Selection

Like hard sweeps, polygenic selection has been proposed as a mechanism for rapid adaptation. But polygenic selection produces a very different genomic signature than classical sweeps. Polygenic adaptations can occur without any particular selected allele fixing or rising in frequency as quickly as a hard sweep. When many polymorphisms control fitness, it is possible for a population to adapt with subtle allele frequency changes spread across many sites, rather than a classical sweep at any one of these sites; these interactions across loci create a different genomic signature in surrounding regions, which we illustrate using IBD tracts and allele frequency trajectories (Figure 14.3). Polygenic selection may act on *de novo* mutations, standing variants, or recurrent mutation. While some younger alleles under selection are likely to carry some classical signatures of selection, such as elevated IBD, standing variants under selection are less likely to possess such a drastic excess of IBD. Hence, all in all the signatures of polygenic selection can be very subtle and therefore difficult to detect.

Pritchard *et al.* (2010) argued that in humans, polygenic selection has served as the major mode of recent adaptation. This is because human population-specific traits such as height and skin color exhibit high heritability and correlation to environment, and yet we observe a relative dearth of large allele frequency differences between human populations. These putative adaptations could be explained more feasibly as polygenic adaptations, rather than classical sweeps.

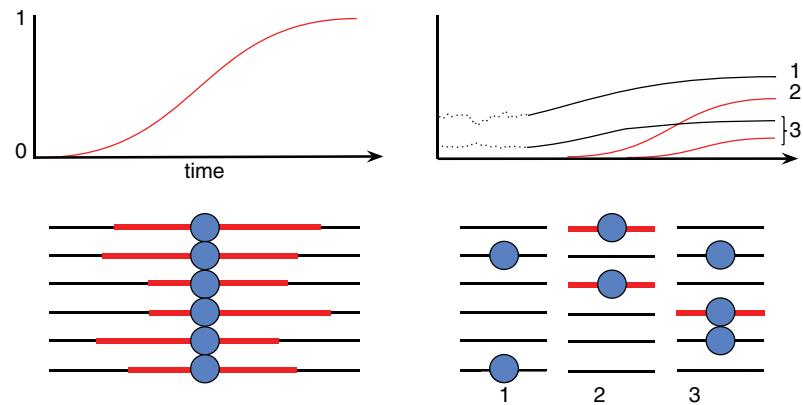


Figure 14.3 The genomic signatures of hard sweeps versus polygenic adaptations. In the bottom part of the figure, each row represents the chromosome/genome of an individual ($n = 6$), and each column represents a genomic region surrounding a favored allele, colored in blue. In the top part of the figure, allele frequency trajectories of selected alleles are shown, with *de novo* variants colored in red and standing variants in black. Left: In a hard sweep, selected alleles rise to high frequency or fixation with an excess of IBD surrounding the driving allele. The red tracts around the selected allele represent IBD to the ancestral haplotype carrying the selected allele. Right: Polygenic adaptation acting on both standing and *de novo* variation at three different loci. Here, fitness is determined by two standing variants (loci 1 and 3) and two *de novo* variants that arise after selection begins (loci 2 and 3). At locus 3 we demonstrate the interference of a recent recurrent mutation (the red tract signifies IBD between the present-day sample and the original haplotype carrying the mutation). Trajectories of alleles undergoing a sweep from the time of mutation are drawn in red, to signify the increased levels of IBD that tend to surround these alleles. This figure was adapted from Pritchard *et al.* (2010).

14.3.4 Confounders

A number of extraneous factors are frequently confounded with selection because they create a similar genomic signature to that left behind by selection. Several of these factors, such as genetic drift and increased mutation rate, have already been mentioned. However, there are a few more confounders of which to be wary. Most notably, selection scans are frequently confounded by unspecified non-equilibrium demography (Nielsen *et al.*, 2005; Teshima *et al.*, 2006). For example, a hard sweep will in the long-term cause an excess of low-frequency derived alleles, just as an expansion in population size will cause the same signature, even in the absence of selection (Figure 14.4). Galtier *et al.* (2000) showed that population size bottlenecks have a local effect on variation that is indistinguishable from that of a selective sweep. Similarly, balancing selection and recent selective sweeps can result in an excess of intermediate- and high-frequency derived alleles, respectively; these signatures can be mimicked under selective neutrality when sampled individuals hail from two different populations, unknown to the geneticist (Städler *et al.*, 2009). A common strategy for dealing with these demographic confounders is to control tests for selection using empirical distributions calculated genome-wide; this approach takes advantage of the tendency for demography to affect the entire genome, whereas signatures of selection tend to affect smaller genomic regions (Galtier *et al.*, 2000; Nielsen *et al.*, 2005).

Different modes of selection can also produce similar genomic signatures. Schrider *et al.* (2015) demonstrated that soft sweeps can be erroneously detected at the 'shoulders' of hard sweeps, far enough from the selected mutation to host increased diversity, yet close enough to have an aberrant level of diversity relative to the background. They also show that these shoulders can be erroneously identified as ongoing sweeps (Schrider *et al.*, 2015).

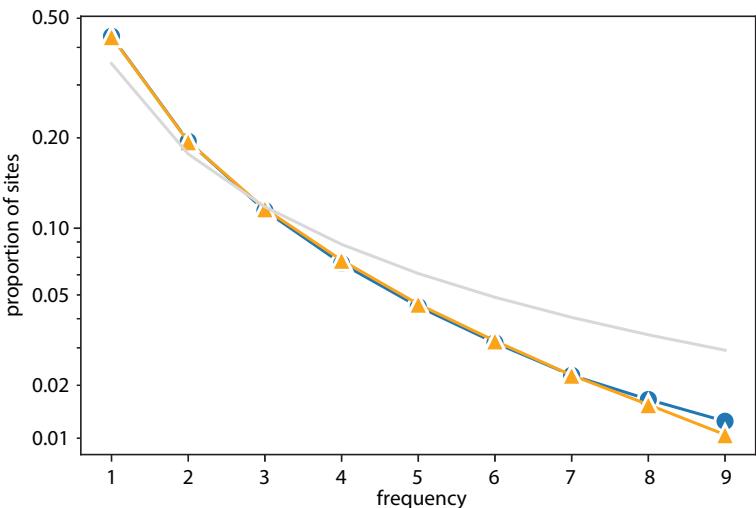


Figure 14.4 The expected site frequency spectra of $n = 10$ individuals under equilibrium demography (i.e. constant N_e in a panmictic population) with a selective sweep (\circ), fluctuating N_e with no selection (\triangle), and equilibrium demography with no selection (straight line). We obtain these SFSs by simulating 5000 unlinked loci with $\theta = \rho = 100$ (these parameters are the population-scaled mutation and recombination rates, respectively). The selective sweep has $N_e = 10^4$ and $s = 0.1$, conditioned on fixing 1.6×10^4 generations ago. The fluctuating N_e model has $N_e = 10^4$ for $0 \leq t < 1.2 \times 10^4$, $N_e = 8 \times 10^3$ for $1.2 \times 10^4 \leq t < 2 \times 10^4$, and $N_e = 2 \times 10^3$ for $t \geq 2 \times 10^4$.

14.4 Methods for Detecting Selection

In this section we give an overview of statistical methods that make use of the signatures of selection that we discussed in the previous section. We begin with a discussion of methods to detect selection based on summary statistics, such as substitution rates, the site frequency spectrum, and haplotype homozygosity. Common to most of the described methods is that they aim to detect selection by looking for deviations from the genetic patterns expected under neutrality (and compatible with selection). Thus, these methods are often called neutrality tests. Another commonality is that often these methods are applied in the context of a whole-genome scans, that is, the summary statistic is calculated at a set of sites throughout the genome to identify regions with extreme deviations. Later we discuss how to combine information across these various statistics; we discuss the challenges associated with likelihood-based inference, and review alternative approaches in this regard, such as composite likelihood, approximate Bayesian computation, and machine learning techniques.

14.4.1 Substitution-Based Methods

The pattern of an increased rate of substitution in a locus under positive selection (Section 14.3.1.1) has been extensively exploited to identify natural selection. Perhaps most famous are the tests based on the d_N/d_S ratio comparing two or more DNA sequences, typically from different species. The d_N/d_S ratio is the ratio of non-synonymous mutations per non-synonymous site to the number of synonymous mutations per synonymous sites. When comparing mutations among different species, the mutations largely reflect fixations between species (i.e. substitutions). The basic idea is that if no selection is acting on the mutations, then $d_N/d_S = 1$ in expectation. However, if positive selection is acting, then $d_N/d_S > 1$ in expectation. In its

original formulation (Li *et al.*, 1985; Nei and Gojobori, 1986), non-synonymous and synonymous sites were considered to be physical entities. However, because of the structure of the genetic code, both synonymous and non-synonymous mutations can occur in the same physical sites. A solution to this problem was proposed by Muse and Gaut (1994) and Goldman and Yang (1994) who developed Markov chain models of molecular evolution with a state space on the set of the 61 possible sense codons. Using such models, parameterized in terms of the rate of non-synonymous and synonymous rates of evolution, likelihood ratio tests of $H_0 : d_N/d_S = 1$ against alternatives of $H_A : d_N/d_S > 1$ can be established. Furthermore, these processes can be superimposed along the edges of a phylogeny to allow joint analysis of d_N/d_S ratios in multiple species. Popular computer programs implementing such tests include PAML (Yang, 2007) and HyPhy (Kosakovsky Pond and Muse, 2005). These tests have since been extended in various ways to detect selection acting along the edges of a phylogeny (Yang, 1998), acting in subsets of sites (Nielsen and Yang, 1998), or a combination of both (Zhang *et al.*, 2005). Methods for detecting $H_A : d_N/d_S > 1$ allowing d_N/d_S to vary among sites according to some distribution have, in particular, been useful, as the intensity of selection is likely to vary greatly among sites in real proteins. Furthermore, even in proteins experiencing substantial amounts of positive selection, we would expect $d_N/d_S < 1$ for most sites, as selection acts mostly to preserve function on most protein coding genes.

While d_N/d_S tests originally were intended mostly for comparative data (data from different species), they can also be similarly applied to data from within a species, although recombination then poses a challenge (Wong *et al.*, 2004). Analyses of d_N/d_S ratios are usually carried out assuming a fixed gene tree topology, which mostly is not a problem when only one sequence is included from each of a set of divergent species. However, when multiple sequences from the same species are included, different recombinant sites will have different gene tree topologies and the assumption of a single shared gene tree topology is no longer satisfied by the data. For more details about these types of tests, see Chapter 13, this volume.

14.4.2 Methods Comparing Substitutions and Diversity

Some of the most popular methods for detecting selection compare divergence between species with the amount of diversity within species. For example, the famous McDonald–Kreitman (MK) test establishes a 2×2 contingency table of the number of non-synonymous and synonymous mutations – NS and S , respectively – within and between species, estimated from multiple aligned sequences (McDonald and Kreitman, 1991). The MK test is then performed as a simple test of homogeneity. As the same underlying (set of) gene trees are shared by synonymous and non-synonymous mutations, $NS_{\text{within}} \sim \text{Bin}(\lambda, NS_{\text{within}} + S_{\text{within}})$ and $NS_{\text{between}} \sim \text{Bin}(\lambda, NS_{\text{between}} + S_{\text{between}})$, where $\text{Bin}(\cdot, \cdot)$ denotes the binomial distribution and λ is the ratio of the rate of new neutral non-synonymous to synonymous mutations. If no selection is acting, except to immediately eliminate strongly deleterious mutations, λ should be the same within and between species. Significant deviations from the null hypothesis can be caused by either positive selection, resulting in a decrease in

$$NI = \frac{NS_{\text{within}}/NS_{\text{between}}}{S_{\text{within}}/S_{\text{between}}},$$

or negative selection causing a similar increase in NI , where NI is the so-called ‘neutrality index’ (Meiklejohn *et al.*, 2007). However, some models of negative selection combined with fluctuations in population size may also cause decreases in the neutrality index (McDonald and Kreitman, 1991; Eyre-Walker, 2002). A related test, and the first test for detecting selection aimed at DNA sequencing data, is the HKA test (Hudson *et al.*, 1987). It is similar to the MK test in that it establishes a 2×2 contingency table comparing data within and between species.

However, instead of comparing non-synonymous mutations and synonymous mutations, it compares variability in different regions of the genome. The test can, therefore, be extended to an arbitrary number of loci, k , in a $2 \times k$ table. Unfortunately, the rationale for the use of a simple test of homogeneity used for the MK test does not hold for the HKA test and simulations are needed to test significance. As for many of the tests discussed in this review, these simulations must necessarily assume a specific demographic model and there is no reason to assume that the results are particularly robust to the assumptions regarding demography.

14.4.3 Methods Using the Frequency Spectrum

As previously mentioned in Section 14.3.1.2, the SFS describes the distribution of allele frequencies in multiple sites. Tests for selection acting on some category of mutations, such as non-synonymous substitutions, relative to synonymous substitutions, can also be carried out at the level of allele frequencies. A particular advantage of this approach is that parametric models of selection can be used to estimate distributions of selection coefficients, using the comparisons of the SFS in different categories of sites, if one of the categories can be assumed to be neutral (Williamson *et al.*, 2005; Eyre-Walker and Keightley, 2007; Boyko *et al.*, 2008). Some of our best estimates of the distributions of selection coefficients in humans and other organisms come from such comparisons (Eyre-Walker and Keightley, 2007; Boyko *et al.*, 2008).

However, some of the most popular methods for detecting selection are based on a comparison of the SFS, not to a presumed neutral category of mutations, but rather to the expected SFS under models of a standard neutrally evolving population. As we discussed in Section 14.3.1.3, certain deviations from this null expected SFS can be indicative of selection. The most commonly used methods in this regard are based on simple summary statistics of the frequency spectrum, the most famous of which is Tajima's D (Tajima, 1989):

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{z(S)}.$$

where S is the number of segregating sites in the sample, $\hat{\theta}_\pi$ is the average number of pairwise differences between individuals, and $\hat{\theta}_W$ is Watterson's estimator of θ , that is, $\hat{\theta}_W = S/a_n$ where $a_n = \sum_{j=1}^{n-1} j^{-1}$. (The term $z(S)$ standardizes the variance of D .)

Under the null model of a population with constant effective size and selective neutrality, $E_0[\hat{\theta}_\pi] = E_0[\hat{\theta}_W] = \theta \equiv 4N_e\mu$, where μ is the per-generation mutation rate of the locus. Thus, $E_0[D] = 0$, and deviations from the underlying neutral model can be tested as deviations from $D = 0$. When applying Tajima's D to genome-wide data, D is typically calculated in sliding or non-overlapping windows to obtain a genome-wide distribution to which each local value can be compared, which makes it possible to control for confounding factors such as non-equilibrium demography (see Section 14.3.4). Tajima's D detects selection primarily because the estimator $\hat{\theta}_\pi$ places a heavy weight on intermediate-frequency alleles, which are depleted after a selective sweep. Under these conditions, $E[D] < 0$, whereas under balancing selection, $E[D] > 0$.

Other SFS-based statistics related to Tajima's D have been proposed. One choice is Fu and Li's D (which we call D_{FL} to avoid ambiguity), defined as

$$D_{FL} = \frac{\hat{\theta}_W - \xi_1}{z(\xi)},$$

where ξ is the SFS, ξ_1 denotes the number of sites at which only one individual carries the derived allele (i.e. the number of singletons), and z is again a scaling factor to standardize the variance of D_{FL} (Fu and Li, 1993). Similarly to Tajima's D , D_{FL} has the property that $E_0[D_{FL}] = 0$.

Additionally, as previously mentioned, selective sweeps cause an excess of singletons after fixation; thus, like Tajima's D , we expect D_{FL} to take negative values after a sweep.

Another SFS-based statistic of this type is Fay and Wu's H , defined as

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{z(\xi)}$$

where

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

and ξ_i is the number of i -tons in the sample (Fay and Wu, 2000). A key property of this statistic is that it places high weight on high-frequency derived alleles, and is thus sensitive to very recent selection rather than old sweeps; this is because the excess of high-frequency derived alleles immediately following a sweep is extremely transient. Thus, like D and D_{FL} , we expect H to take negative values if a selective sweep occurred recently.

This class of methods has also been adapted specifically for detecting balancing selection by Siewert and Voight (2017), who introduced a statistic β that detects balancing selection by weighting alleles at intermediate frequency. As the reader might have noticed, these classical methods share a simple mathematical quality: they are all linear combinations of the SFS and some weight vector ω (see, for example, Achaz, 2009, who shows how to choose ω to detect specific violations of the neutral model with optimal power). As we discussed regarding Tajima's D , we can assess significance by calculating the genome-wide statistics under a particular choice of ω , and picking the most deviant regions. Other methods that use information from the SFS to detect selection include machine learning methods and composite likelihood methods, which we will discuss in later sections.

The performance of the methods based on summary statistics of the SFS varies wildly depending on whether a sweep has gone to fixation or is currently segregating (a so-called *incomplete* sweep); furthermore, they are easily confounded by demography such as population size bottlenecks (Galtier *et al.*, 2000; Teshima *et al.*, 2006). Power is also lessened when the sweep fixed a long time ago because mutation and recombination will have diminished the characteristic patterns of a sweep (Teshima *et al.*, 2006).

14.4.4 Methods Using Genetic Differentiation

The very first test of neutrality (Lewontin and Krakauer, 1973) was based on detecting patterns of increased genetic differentiation among populations (see Section 14.3.1.2). While such methods were perhaps, for some years, viewed with skepticism by many researchers due to the strong assumptions they have to make regarding the history of the populations, they have had a resurgence after the emergence of genomic data as a convenient and simple tool to scan the genome for evidence of local selection. The most common measure of genetic differentiation among populations is F_{ST} , which can be defined in various ways, and confusingly can take on the properties of both a statistic and a parameter (see Holsinger and Weir, 2009, for a discussion). A common definition of F_{ST} for two populations, in a single di-allelic locus, is (Wright, 1949)

$$F_{ST} = \frac{c_1 p_1 (1-p_1) + c_2 p_2 (1-p_2)}{\bar{p}(1-\bar{p})}$$

where $c_1, c_2 > 0$, $c_1 + c_2 = 1$ (i.e. c_1, c_2 are the proportion of samples from each population), p_1 and p_2 are the sample allele frequencies in the first and second population, respectively, and $\bar{p} = c_1 p_1 + c_2 p_2$ is the mean allele frequency (i.e. the frequency in the combined sample). Notice that the value of F_{ST} falls in $[0, 1]$; for highly differentiated allele frequencies we expect a value closer to 1, and for roughly undifferentiated frequencies we expect a value close to 0.

In genomic data, so-called F_{ST} scans are often used to detect selection, as elevated among-population genetic differentiation (high F_{ST}) may be a consequence of selection (Lewontin and Krakauer, 1973a; see also Section 14.3.1.2 above). Whether an F_{ST} value is significantly elevated can be tested using parametric models or simulations. For example, Beaumont and Balding (2004) developed a hierarchical Bayesian method for identifying outlier loci assuming a multinomial-Dirichlet likelihood function. However, significance is often not tested directly, but rather a list of the most extreme loci are presented without claims regarding significance.

Methods based on F_{ST} can also be extended to identify selection in an individual population by comparisons to multiple other populations. One particularly simple method for doing this is the so-called population branch statistic (PBS), which is based on transforming F_{ST} estimates between pairs of populations to an approximately linear distance and then inferring the amount of genetic drift distance on each branch of a tree with three populations (Yi *et al.*, 2010). Extreme drift on a population's branch at a particular locus is compatible with selection specific to that population acting on that locus.

More parametric methods will likely provide more power than simple methods based on F_{ST} . Furthermore, they can be used to test more specific hypotheses about the factors driving selection. Of particular interest in this regard are methods such as that of Coop *et al.* (2010) which uses a Bayesian model based on a Gaussian likelihood function for allele frequencies, to identify correlations between allele frequency changes across populations and specific environmental variables.

Methods based on genetic differentiation are also used to study polygenic traits. For example, the measure Q_{ST} is used to quantify differentiation of quantitative traits among populations (Lande, 1992; Spitz, 1993). This quantity is calculated analogously to F_{ST} , but for a phenotypic trait instead of allele frequencies, and is often directly compared with F_{ST} to infer selection acting on traits. Indeed, under simple neutral conditions we expect genome-wide $F_{ST} = Q_{ST}$. Under negative selection for the same phenotype value across populations, we expect $F_{ST} > Q_{ST}$, and under directional selection on differing phenotype values among populations, we expect $F_{ST} < Q_{ST}$.

Unfortunately, for many realistic models of population structure, such as hierarchical population models, the assumptions of many of the standard tests are violated. Berg and Coop (2014) recently developed a method to test for selection on polygenic traits while controlling for hierarchical population structure, applying the same principles as in the aforementioned Gaussian approximation deployed by Coop *et al.* (2010). For a particular trait, they consider the quantity

$$\vec{z} = 2\mathbf{A}\vec{p},$$

where \mathbf{A} is an $M \times L$ matrix of population-specific additive effect sizes of alleles on trait values (typically an estimate obtained from GWASs), \vec{p} is a vector of allele frequencies, M is the number of sub-populations and L is the number of alleles genotyped. Correcting for hierarchical population structure using the inverse sample covariance matrix \mathbf{F}^{-1} , they obtain a measure of the deviance of \vec{z} , called Q_X . Under the selectively neutral model with hierarchical structure, $Q_X \sim \chi^2_{M-1}$, and thus the significance of Q_X can be evaluated using the χ^2 distribution. A significant value of Q_X suggests the trait of interest has undergone recent selection. More

recently, Racimo *et al.* (2018) developed an even more generalized approach for inferring polygenic selection in the presence of population structure, allowing selection to act along specific edges of an admixture graph.

14.4.5 Methods Using Haplotype Structure

The methods discussed previously focused on allele frequencies and allele frequency changes. However, other features of the data can be leveraged for detecting selection; in particular, haplotype structure, a signature we introduced in Section 14.3.1.3. Many methods aimed at detecting ongoing selection (incomplete selective sweeps, a concept we introduced also in Section 14.3.1.3) focus solely on haplotypes – specifically, the pattern of increased haplotype homozygosity on chromosomes carrying the advantageous mutation.

Sabeti *et al.* (2002) developed a statistic called extended haplotype homozygosity (EHH) that estimates the probability that two randomly chosen haplotypes are identical up to a distance x around a particular candidate single nucleotide polymorphism (SNP) called the core SNP. More precisely, we can define EHH as the number of pairs of identical haplotypes in a window of length x divided by total number of pairs:

$$\text{EHH}(x) = \sum_{h \in \mathcal{H}(x)} \frac{\binom{n_h}{2}}{\binom{n}{2}},$$

where $\mathcal{H}(x)$ is the set of distinct haplotypes in the sample only considering sites within a distance x from the core SNP, n_h is the number of type- h chromosomes, and n is sample size.³ Notice that if we calculate $\text{EHH}(x)$ right around a particular site so that the window size is 0, then $\text{EHH}(0) \approx p^2 + q^2$, where p is the frequency of the core SNP and $q = 1 - p$. For increasingly large window sizes, $\text{EHH}(x)$ converges to zero because all n haplotypes become distinct when considering a sufficiently large region. However, the rate at which EHH decays to 0 with respect to x reflects the age of the core SNP, which depends on the strength of selection. A slow decay of EHH is compatible with recent selection; as discussed in the section on hitchhiking (Section 14.3.1.3), the region surrounding the selected allele tends to be depleted of variation and recombination, and thus EHH decays more slowly in this case than under selective neutrality. Thus, an elevated value of EHH around a core SNP serves as a convenient feature for detecting loci under positive selection. However, one general challenge, in addition to the reliance on phased data and a genetic map, is that other processes, such as a local reduction in mutation rate or an increase in negative selection, also can lead to increased haplotype homozygosity. Therefore, the relative EHH (rEHH) between different classes of haplotypes (i.e. haplotypes grouped by the allelic state of the core SNP) was proposed as a more robust method for detecting selection (Sabeti *et al.*, 2005).

A further development of the EHH family of statistics was proposed by Voight *et al.* (2006), who developed the integrated haplotype score (iHS), a statistic designed to detect ongoing selection. The iHS partitions haplotypes based on the ancestral/derived states of the core SNP, and is based on integrating EHH from the core SNP until EHH reaches a certain fixed value (typically 0.05). These integrated EHH values are called $i\text{HH}_A$ and $i\text{HH}_D$, and the statistic $\log(i\text{HH}_A/i\text{HH}_D)$ is then calculated genome-wide and standardized by the empirical mean and variance of this statistic using other genomic SNPs at the same frequency. Since selected alleles tend to carry longer surrounding IBD tracts than neutral alleles at the same frequency as

³ Note that $\sum_h n_h = n$; thus, another approximately equivalent calculation of EHH (assuming n is large) is $\text{EHH}(x) = \sum_h p_h^2$ where $p_h = n_h/n$.

the selected allele, we expect the most negative iHS values to indicate strongly selected derived alleles. Importantly, the iHS is standardized by allele frequency because low-frequency alleles tend to be younger and thus carry high amounts of IBD, even if they are selectively neutral. Notice that to calculate the iHS, it is assumed that a genetic map is known to integrate EHH with respect to distance. Ferrer-Admetlla *et al.* (2014) proposed an alternative to the iHS, the number of segregating sites by length (nSL), which avoids relying on a genetic map by, for each pair of haplotypes, using the number of mutations within the other $n - 2$ haplotypes to measure a mutational distance, leading to increased robustness against recombination and mutation rate variation. One important note is that the expectation of iHH is infinite under a standard neutral model, making statistics based on the iHH statistic highly sensitive to the choice of maximal window size for calculations of iHH (Ferrer-Admetlla *et al.*, 2014).

Importantly, the aforementioned measures of haplotype homozygosity are underpowered to detect soft sweeps (Garud *et al.*, 2015; Field *et al.*, 2016). Garud *et al.* (2015) developed an alternative haplotype-based statistic specifically designed to detect both hard and soft sweeps. They defined

$$H_{12} = (p_{h_{(1)}} + p_{h_{(2)}})^2 + \sum_{j>2} p_{h_{(j)}}^2,$$

where $h_{(1)}$ and $h_{(2)}$ are the first and second most frequent haplotypes in the set of distinct haplotypes \mathcal{H} , and $p_h = n_h/n$. Under a hard sweep we expect $p_{h_{(1)}} \gg p_{h_{(2)}}$, whereas under a soft sweep the discrepancy tends to be less severe; nonetheless, under a soft sweep we expect the several most frequent haplotypes to still dominate the haplotype distribution, and thus H_{12} is sensitive to both cases. To distinguish between hard and soft sweeps, they propose

$$H_1/H_2 = \frac{p_{h_{(1)}}}{\sum_{h \neq h_{(1)}} p_h}.$$

By the same intuition for defining H_{12} , here a high value of H_1/H_2 implies a hard sweep, whereas a low value implies a soft sweep.

Recently, Field *et al.* (2016) developed a haplotype score called the singleton density score (SDS) designed to have especially high sensitivity to detect extremely recent signatures of selection, relative to comparable methods such as the iHS. Their approach is based on the intuition that for ongoing or recent sweeps, the haplotypes carrying the favored allele have a dearth of singletons. To compute the SDS at a particular site, the SDS iterates through each diploid individual. For each individual, the distance between the nearest singletons up- and downstream of the core allele is computed, and these n distances are binned based on whether the individual is homozygous for the derived allele, homozygous for the ancestral allele, or heterozygous. A likelihood model is used to infer the ‘mean tip length’ of ancestral and derived lineages; essentially, long singleton distances imply a short mean tip length. The inferred ancestral and derived mean tip lengths, called \hat{t}_A and \hat{t}_D , respectively, are standardized similarly to the iHS. The SDS exploits the fact that we expect $\hat{t}_A > \hat{t}_D$ when the derived allele has risen sharply in frequency in the immediate past. Thus, the haplotypes surrounding a positively selected allele are expected to be depleted of singletons relative to a neutral allele segregating at the same frequency. The authors also designed a score called the trait SDS (tSDS), where the sign of the SDS is flipped if the ancestral allele is associated with increasing the value of the trait (e.g. associated with a positive change in height). This measure can be used to demonstrate polygenic selection on a trait by showing an excess in tSDS across associated sites.

While haplotype-based methods are mostly designed for detecting ongoing sweeps, rather than completed sweeps, there is also a distinct pattern of linkage disequilibrium arising after

a sweep that can be exploited for detecting sweeps. As discussed in Section 14.3.1.3, right after a completed sweep there will be increased LD to either side of the selected sweep, but no LD between SNPs from opposite sides of the selected sites (Kim and Nielsen, 2004). Kim and Nielsen (2004) proposed using a statistic, ω , to detect this pattern.

14.4.6 Why Full-Likelihood Methods Are Intractable for Population Samples

So far we have discussed various statistics used in tests aimed at detecting natural selection. The statistically minded reader might appropriately wonder at this point why there exists such a plethora of more or less *ad hoc* statistics, and why there are no methods for detecting selection based on likelihood functions that incorporate all information regarding the selection, including allele frequencies and haplotypes. Unfortunately, full likelihood methods that incorporate selection are considered computationally intractable. Some progress was made on models of weak selection without recombination (Krone and Neuhauser, 1997). However, these methods never scaled up to genomic data. Several methods have been developed that use simulations to approximate the likelihood for a single non-recombining locus under various assumptions (Slatkin, 2000; Coop and Griffiths, 2004). In particular, Coop and Griffiths (2004) developed a likelihood method for detecting and estimating the strength of selection by first simulating an ancestral allele frequency trajectory and then simulating a coalescence tree conditionally on the allele frequency trajectory. Like Krone and Neuhauser (1997), it is computationally intensive and the assumed absence of recombination makes it inapplicable to most data, such as human nuclear DNA.

Because full-likelihood inference is not viable for even small sample sizes, most methods for detecting selection rely on summary statistics that capture particular signatures of selection. The major challenge is that calculating the likelihood requires integrating out many sources of stochasticity, including allele frequency trajectories, the latent ancestral recombination graph (ARG)⁴ conditional on this trajectory, and neutral mutations superimposed on the ARG. The vast combinatorial space of ARGs makes analytical calculations impracticable. However, in lieu of tractable full likelihood methods, a number of methods have been developed that attempt to detect and/or quantify selection using functions that approximate the likelihood function.

14.4.7 Composite Likelihood Methods

Using diffusion theory, Stephan *et al.* (1992) and Kim and Stephan (2002) developed expressions for the distribution of sample allele frequencies (i.e. the SFS) as a function of the genetic distance from a recently completed sweep. Based on these calculations, Kim and Stephan (2002) could define a composite likelihood formed as the product of the individual likelihood functions calculated for each site along the length of the sequence, as a function of the site's recombination distance to the selected SNP and the selection coefficient. (For more on composite likelihood functions, see Chapter 1, this volume.) They then proposed to use a likelihood ratio to test the null hypothesis of no selection ($s = 0$) and to estimate the strength and location of the sweep. The advantage of this method over previous methods was multiple. Firstly, it employed all of the information from the allele frequency by using a full-likelihood approach to the allele frequencies. Secondly, it used the spatial distribution of SNPs and their allele frequencies to gain power and to locate the most likely selected SNPs. Nielsen *et al.* (2005) extended the method using an approximation by Durrett and Schweinsberg (2005) which considered the probability that a particular lineage in the genealogy 'escaped' a sweep; that is, the probability that a

⁴ A complex graphical structure that represents all of the genealogical and recombinant events occurring in a sample; see, for example, Griffiths and Marjoram (1996).

neutral allele linked to the non-beneficial allele recombined onto a beneficial background prior to loss of the non-beneficial allele. Using this result, the composite likelihood function could be calculated faster and could incorporate any SFS as the ‘background’ neutral allele frequency distribution to be tested against. This method has since been modified in multiple ways, including extensions to incomplete sweeps (Vy and Kim, 2015), modeling of population structure (Chen *et al.*, 2010), and incorporation of negative selection in the genomic background (Zhu and Bustamante, 2005).

14.4.8 Approximate Bayesian Computation

The previous section on composite likelihood methods introduced approaches for estimating the selection coefficient and the location of the selective sweep. However, there are other approaches for addressing this problem – in particular, methods based on approximate Bayesian computation (ABC; Beaumont *et al.*, 2002). ABC works by repeatedly sampling parameters (such as s) from a prior distribution and then subsequently simulating a genomic data set for each sampled parameter value. Without loss of generality, assuming we wish to approximate the posterior of s , we use an acceptance/rejection scheme where sampled values of s are rejected if the distance between the resulting simulated data and the observed data is sufficiently large. To determine distance between the simulated and observed data, many approaches use classical SFS- or haplotype-based summary statistics and calculate a distance (e.g. Euclidean distance) between the two summary vectors calculated for the observed and simulated data. After sampling and simulations are completed, the estimate of the posterior can be used to derive a maximum *a posteriori* (MAP) estimate of s , as well as its Bayesian credible interval. Unlike other methods based on individual summary statistics, ABC takes full advantage of the correlation structure between different types of summary statistics. For more details on ABC methods as well as priors, posteriors, Bayesian point estimation and uncertainty quantification, see Chapter 1, this volume.

Peter *et al.* (2012) developed an ABC method to jointly estimate selection coefficients, the time selection started, and the frequency of the selected mutation at the time selection started. The method could also perform model selection, distinguishing soft sweeps from hard sweeps. Similar methods have been used by Ormond *et al.* (2016) and McManus *et al.* (2017).

A common pitfall of ABC methods is that they are computationally intensive, and can suffer from the curse of dimensionality. That is, when the sample space of the summary statistics increases in dimensionality, so does the instability of the likelihood estimate (Sheehan and Song, 2016). Recently, Sugden *et al.* (2018) showed that using an average one-dependence estimate assumption of the structure of the likelihood can ameliorate this instability problem, while retaining some of the informative correlation structure of the likelihood.

14.4.9 Machine Learning Methods

An alternative to composite likelihood and approximate Bayesian methods, for incorporating more information from the data, are standard *supervised* machine learning methods. Broadly speaking, supervised methods aim to train some model to use statistics extracted from the data, often called *features*, with the goal of making accurate predictions based on such features. By contrast, so-called *unsupervised* methods are used to find structure in data, rather than generate predictions. Principal components analysis is one example of an unsupervised method, often applied in statistical genetics to illustrate and control for population structure (Novembre *et al.*, 2008; Price *et al.*, 2010).

Mathematically, we can summarize the supervised learning problem as follows. Assume the availability of a training set of pairs $\{(\vec{x}_i, y_i)\}_{i=1}^n$, where each pair consists of a feature vector \vec{x}_i

and its *label* y_i for each sample $i = 1, 2, \dots, n$. The objective is then to select a function f^* that maps feature vectors to labels such that

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n l(f(\vec{x}_i), y_i) \right\},$$

where $l(\hat{y}, y)$ is a *loss function* that is minimized when $\hat{y} = y^5$ (i.e. when the estimated labels \hat{y} perfectly match the true labels), and \mathcal{F} is a pre-specified set of prediction functions. This framework can be used to develop methods for inferring selection: for example, a so-called *classifier* can be trained to detect selection, where the labels y are binary variables such that $y = 0$ signifies ‘neutrality’ and $y = 1$ signifies ‘sweep’. Additionally, techniques such as linear regression can be used to estimate the value of the selection coefficient s . Notice, however, that a training set needs to be available for which the correct labels are known. In population genetics such training sets are rarely available. Instead, simulated data are used to train the classifiers. Also, the feature vector has to be chosen, and is usually based on the same type of summary statistics as used in other simulation-based methods. Schrider and Kern (2018) reviewed supervised machine learning in the context of population genetics and detecting selection, defining the problem and illustrating practical concerns in greater detail.

In the previously mentioned classical SFS-based methods such as Tajima’s D , the SFS is summarized as a linear combination that has expectation 0 under neutral equilibrium conditions. By contrast, a method by Ronen *et al.* (2013) (SFSelect) uses the full SFS as the high-dimensional analog to these classical methods. They simulate data under specific sweep and neutral conditions and train a machine learning model called a support vector machine (SVM; see Cortes and Vapnik, 1995) to classify SFS vectors. Similar approaches have been taken to integrate various haplotype statistics along with the SFS; Pavlidis *et al.* (2010) trained an SVM using a combination of LD- and SFS-based statistics, and the method EvolBoosting (Lin *et al.*, 2011) integrates both SFS- and haplotype-based statistics to detect selection using boosting (Bühlmann and Hothorn, 2007). SHiC (Schrider and Kern, 2016) is a method that extracts a number of haplotype- and SFS-based statistics from simulated data to train an Extra-Trees classifier (Geurts *et al.*, 2006). Notably, the latter method has been shown to retain good power to detect strong selection despite model misspecification during training (i.e. trained under equilibrium demography and tested under fluctuating N_e) (Schrider and Kern, 2016).

14.5 Discussion

As evident from the previous sections, there is a very large set of different methods for detecting selection. In fact, in this review we have only covered a subset of the most commonly used methods. For example, recent methods for detecting adaptive introgression are not covered (see, for example, Rosenzweig *et al.*, 2016). A common theme for many of the methods is that there typically is a trade-off between power and robustness. Since the emergence of the first neutrality tests (Lewontin and Krakauer, 1973), there has been an awareness that many demographic models can mimic the signature of selection. With the emergence of genomic data, it was generally hoped that this problem would vanish as the signature of demographic processes affects the entire genome, while selection may only affect one or a few loci. While genomic data certainly have helped identify signatures of selection, we are still facing challenges when

⁵ We also note that one can consider *regularized* loss functions, where l is not necessarily minimized when $\hat{y} = y$; such functions are useful to prevent model overfitting.

assigning *p*-values, or other methods of statistical confidence for inferences of selection. In the end, almost all methods rely to some degree on the assumption of a demographic model. By the very nature of the data, the null hypothesis considered will always be a composite hypothesis that also includes features of the demography. To address this problem, most studies rely on one of two possible strategies. (1) They may give up on including measures of statistical confidence and instead simply produce a list of the best candidates for targets of selection. One variant of this approach is the use of so-called ‘empirical *p*-values’ (e.g. Voight *et al.*, 2006), which in this context are simply quantiles of the empirical distribution of the test statistic. They are, therefore, not *p*-values in the classical sense and should probably more appropriately simply be reported as quantiles. (2) The alternative approach is to make specific assumptions about the demography, typically based on estimates of demography obtained from the same or other data. Simulations are then used in one form or another to generate the distribution of the test statistic under the null hypothesis. Variants of this approach include the machine learning and ABC methods which include simulations as an integrated part of the inference framework.

As previously mentioned, another major challenge of inferences of selection is that full-likelihood methods are not available. However, they may eventually be practicable, or at least closely approximated, by building on advances in inferring ARGs. ARG inference methods have historically been impractical for even modest sample size and locus length (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001). To ease computational costs, Li and Stephens (2003) calculated a heuristic for the conditional sampling distribution (CSD) of the *n*th sequence given *n* – 1 other sequences, which allowed them to conduct approximate maximum likelihood inference of recombination rates without explicitly sampling the ARG. McVean and Cardin (2005) showed that the so-called sequentially Markov coalescent (SMC) is remarkably consistent with the coalescent with recombination; this approximation, along with the SMC', a similar approximation due to Marjoram and Wall (2006), allows extremely efficient approximate simulation of the ARG and maximum likelihood inference of population size history under the pairwise sequentially Markov coalescent (Li and Durbin, 2011). Recently, Rasmussen *et al.* (2014) developed a probabilistic method to approximate the posterior distribution on ARGs, based on both the CSD and SMC/SMC' approximations. Their ARGweaver method efficiently samples posterior ARGs, and scales well with genome length and sample size. Based on these advances, it may be possible in the future to develop methods for detecting selection that more closely approximate the full-likelihood function.

References

- Achaz, G. (2009). Frequency spectrum neutrality tests: One for all and all for one. *Genetics* **183**(1), 249–258.
- Albrechtse, A., Moltke, I. and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**(1), 295–308.
- Beaumont, M.A. and Balding, D.J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**(4), 969–980.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025–2035.
- Berg, J.J. and Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genetics* **10**(8), e1004412.
- Bollback, J.P., York, T.L. and Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**(1), 497–502.

- Boyko, A.R., Williamson, S.H., Indap, A.R., Degenhardt, J.D., Hernandez, R.D., Lohmueller, K.E., Adams, M.D., Schmidt, S., Sninsky, J.J., Sunyaev, S.R., *et al.* (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* **4**(5), e1000083.
- Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**(7), 1177–1186.
- Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H. and Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**(2), 783–796.
- Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* **22**(4), 477–505.
- Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**(4), e64.
- Chen, H., Patterson, N. and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Research* **20**(3), 393–402.
- Coop, G. and Griffiths, R.C. (2004). Ancestral inference on gene trees under selection. *Theoretical Population Biology* **66**(3), 219–232.
- Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J.K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185**(4), 1411–1423.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20**(3), 273–297.
- Durrett, R. and Schweinsberg, J. (2005). A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and Their Applications* **115**, 1628–1657.
- Ewens, W.J. (2012). *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer, New York.
- Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics* **162**(4), 2017–2024.
- Eyre-Walker, A. and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**(8), 610.
- Fay, J.C. and Wu, C.-i. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**(3), 1405–1413.
- Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**(3), 1299–1318.
- Feder, A.F., Kryazhimskiy, S. and Plotkin, J.B. (2014). Identifying signatures of selection in genetic time series. *Genetics* **196**(2), 509–522.
- Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution* **31**(5), 1275–1291.
- Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I. and K, P.J. (2016). Detection of human adaptation during the past 2000 years. *Science* **354**(6313), 760–764.
- Fisher, R.A. (1999). *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford University Press, Oxford.
- Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology* **48**(2), 172–197.
- Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**(3), 693–709.
- Galtier, N., Depaulis, F. and Barton, N.H. (2000). Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**(2), 981–987.

- Garud, N.R., Messer, P.W., Buzbas, E.O. and Petrov, D.A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics* **11**(2), e1005004.
- Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* **63**(1), 3–42.
- Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**(5), 725–736.
- Good, B.H., McDonald, M.J., Barrick, J.E., Lenski, R.E. and Desai, M.M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature* **551**(7678), 45.
- Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**(4), 479–502.
- Hermisson, J. and Pennings, P.S. (2005). Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**(4), 2335–2352.
- Hershberg, R. and Petrov, D.A. (2008). Selection on codon bias. *Annual Review of Genetics* **42**, 287–299.
- Hockenberry, A.J., Stern, A.J., Amaral, L.A. and Jewett, M.C. (2018). Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. *Molecular Biology and Evolution* **35**(3), 582–592.
- Holsinger, K.E. and Weir, B.S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting *f_{ST}*. *Nature Reviews Genetics* **10**(9), 639.
- Hudson, R.R. and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**(3), 831–840.
- Hudson, R.R., Kreitman, M. and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**(1), 153–159.
- Hughes, A.L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**(6186), 167.
- Kaplan, N.L., Hudson, R.R. and Langley, C.H. (1989). The ‘hitchhiking effect’ revisited. *Genetics* **123**(4), 887–899.
- Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**(3), 1513–1524.
- Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**(2), 765–777.
- Kimura, M. (1955a). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* **41**(3), 144–150.
- Kimura, M. (1955b). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* **20**, 33–53.
- Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics* **47**(6), 713–719.
- Kosakovsky Pond, S.L. and Muse, S.V. (2005). HyPhy: Hypothesis testing using phylogenies. In R., Nielsen (ed.), *Statistical Methods in Molecular Evolution*. Springer, New York, pp. 125–181.
- Krone, S.M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51**(3), 210–237.
- Lande, R. (1992). Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution* **46**(2), 381–389.
- Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D. and Desai, M.M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**(7464), 571.
- Lazaridis, I., Patterson, N. and Mitnik, Alissa, *et al.* (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**(7518), 409–413.

- Lewontin, R. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**(1), 175–195.
- Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**(7357), 493–496.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**(4), 2213–2233.
- Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**(2), 150–174.
- Lin, K., Li, H., Schlötterer, C. and Futschik, A. (2011). Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics* **187**(1), 229–244.
- Malaspinas, A.-S., Malaspinas, O., Evans, S.N. and Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**(2), 599–607.
- Marjoram, P. and Wall, J.D. (2006). Fast ‘coalescent’ simulation. *BMC genetics* **7**, 16.
- Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K. and Gamba, C. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**(7583), 499–503.
- McDonald, J.H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *drosophila*. *Nature* **351**(6328), 652–624.
- McManus, K.F., Taravella, A.M., Henn, B.M., Bustamante, C.D., Sikora, M. and Cornejo, O.E. (2017). Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS Genetics* **13**(3), e1006560.
- McVean, G.A.T. and Cardin, N.J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London, Series B* **360**(1459), 1387–1393.
- Meiklejohn, C.D., Montooth, K.L. and Rand, D.M. (2007). Positive and negative selection on the mitochondrial genome. *Trends in Genetics* **23**(6), 259–263.
- Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**(5), 715–724.
- Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**(5), 418–426.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research* **15**(11), 1566–1575.
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**(3), 929–936.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. *Nature* **456**(7218), 98.
- Ormond, L., Foll, M., Ewing, G.B., Pfeifer, S.P. and Jensen, J.D. (2016). Inferring the age of a fixed beneficial allele. *Molecular Ecology* **25**(1), 157–169.
- Pavlidis, P., Jensen, J.D. and Stephan, W. (2010). Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* **185**(3), 907–922.
- Peter, B.M., Huerta-Sánchez, E. and Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genetics* **8**(10), e1003011.
- Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**(7), 459.
- Pritchard, J.K., Pickrell, J.K. and Coop, G. (2010). The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology* **20**(4), R208–R215.

- Przeworski, M., Coop, G. and Wall, J.D. (2005). The signature of positive selection on standing genetic variation. *Evolution* **59**(11), 2312–2323.
- Racimo, F., Berg, J.J. and Pickrell, J.K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics* **208**, 1565–1684.
- Rasmussen, M.D., Hubisz, M.J., Gronau, I. and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics* **10**(5), e1004342.
- Ronen, R., Udpa, N., Halperin, E. and Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics* **195**(1), 181–193.
- Rosenzweig, B.K., Pease, J.B., Besansky, N.J. and Hahn, M.W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology* **25**(11), 2387–2397.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. and Lander, E.S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**(6909), 832–837.
- Sabeti, P.C., Walsh, E., Schaffner, S.F., Varilly, P., Fry, B., Hutcheson, H.B., Cullen, M., Mikkelsen, T.S., Roy, J., Patterson, N., et al. (2005). The case for selection at CCR5-Δ32. *PLoS Biology* **3**(11), e378.
- Santiago, E. and Caballero, A. (2005). Variation after a selective sweep in a subdivided population. *Genetics* **169**(1), 475–483.
- Sawyer, S.A. and Hartl, D.L. (1992). Population genetics of polymorphism and divergence. *Genetics* **132**(4), 1161–1176.
- Schraiber, J.G., Evans, S.N. and Slatkin, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics* **203**(1), 493–511.
- Schrider, D.R. and Kern, A.D. (2016). S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLoS Genetics* **12**(3), e1005928.
- Schrider, D.R. and Kern, A.D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics* **34**(4), 301–312.
- Schrider, D.R., Mendes, F.K., Hahn, M.W. and Kern, A.D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**(1), 267–284.
- Sheehan, S. and Song, Y.S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology* **12**(3), e1004845.
- Shi, H., Kichaev, G. and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *American Journal of Human Genetics* **99**(1), 139–153.
- Siewert, K.M. and Voight, B.F. (2017). Detecting long-term balancing selection using allele frequency correlation. Preprint, bioRxiv 112870.
- Slatkin, M. (2000). Simulating genealogies of selected alleles in a population of variable size. *Genetical Research* **78**, 49–57.
- Smith, J.M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research* **23**(1), 23–35.
- Sorek, R. and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research* **13**(7), 1631–1637.
- Spitze, K. (1993). Population structure in *Daphnia obtusa*: Quantitative genetic and allozymic variation. *Genetics* **135**(2), 367–374.
- Städler, T., Haubold, B., Merino, C., Stephan, W. and Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**(1), 205–216.

- Stephan, W., Wiehe, T.H. and Lenz, M.W. (1992). The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical Population Biology* **41**(2), 237–254.
- Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M. and Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications* **9**(1), 703.
- Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3), 585–595.
- Teshima, K.M., Coop, G. and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research* **16**(6), 702–712.
- Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. *PLoS Biology* **4**(3), e72.
- Vy, H.M.T. and Kim, Y. (2015). A composite-likelihood method for detecting incomplete selective sweep from population genomic data. *Genetics* **200**(2), 633–649.
- Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R. and Bustamante, C.D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences* **102**(22), 7882–7887.
- Wong, W.S.W., Yang, Z., Goldman, N. and Nielsen, R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**(2), 1041–1051.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**(2), 97–159.
- Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proceedings of the National Academy of Sciences* **24**(7), 253–259.
- Wright, S. (1949). The genetical structure of populations. *Annals of Human Genetics* **15**(1), 323–354.
- Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**(5), 568–573.
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**(8), 1586–1591.
- Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**(5987), 75–78.
- Zhang, J., Nielsen, R. and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**(12), 2472–2479.
- Zhu, L. and Bustamante, C.D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**(3), 1411–1421.

15

Evolutionary Quantitative Genetics

Bruce Walsh¹ and Michael B. Morrissey²

¹Department of Ecology and Evolutionary Biology, University of Arizona

²School of Biology, University of St Andrews

Abstract

This chapter reviews fundamental principles of evolutionary quantitative genetics. First, we cover the concept of additive genetic variance, which is the component of genetic variance that is most relevant to evolutionary responses to selection. We then consider the evolutionary response to selection (i.e. how the mean of a trait changes from one generation to the next) in response to within-generation changes in the mean. We first consider univariate selection and the evolutionary response, and then consider multivariate cases. Next, we consider the dynamics of traits controlled by many loci, and models of how both the trait mean and additive genetic variances respond to selection over multiple generations. Finally, we consider more practical aspects of evolutionary quantitative genetics, including how quantities such as additive genetic variances may be estimated, how fitness is defined in the wild, and how natural selection can be measured in nature.

15.1 Introduction

Evolutionary quantitative genetics is a vast field, ranging from population genetics at one extreme to development and functional ecology at the other. The goal of this field is a deeper understanding of not only the genetics and inheritance of complex traits in nature (traits whose variation is due to both genetic and environmental factors) but also the nature of the evolutionary forces that shape trait variation and change in natural populations. Given the limitations of space, this chapter focuses on a review of the mechanics of generation-to-generation evolutionary change, in particular, empirical estimation of the quantities describing selection and its relationship to evolutionary change. A detailed treatment of the inheritance of complex traits in nature can be found in Lynch and Walsh (1998). Roff (1997) provides a good overview of the entire field, while the reader seeking detailed treatments on specific issues should consult Bulmer (1980) and Walsh and Lynch (2018). Bürger (2000) provides an excellent (but, in some places, highly technical) review of the interface between population and quantitative genetics. Much of the material in this chapter is greatly elaborated on in Walsh and Lynch (2018).

15.2 Resemblances, Variances, and Additive Genetic Values

In this section we introduce some key terminology and basic principles that are necessary before discussing more advanced and current issues later in the chapter. A more extensive treatment of basic principles is given in Falconer and MacKay (1996), while a very detailed treatment is given by Lynch and Walsh (1998).

15.2.1 Fisher's Genetic Decomposition

Although Yule (1902) can be considered the first paper in quantitative genetics, the genesis of modern quantitative genetics is the classic paper by R. A. Fisher (1918). Fisher made a number of key insights, notably that (i) with sexual reproduction, only a specific component of an individual's genotypic value (the mean value of that genotype when averaged over the distribution of environments and genetic backgrounds in the population) is passed on to its offspring, and (ii) we can estimate the variances associated with this and other components from the phenotypic covariances between appropriate relatives.

Fisher decomposed the observed phenotypic value z of an individual into a genotypic G and environment E value,

$$z = G + E. \quad (15.1)$$

The genotypic value can be thought of as the average phenotype (measured value for some trait) if the individual was cloned and replicated over the universe of environments it is likely to experience. As mentioned, Fisher noted that a parent does not pass along its entire G value to its offspring, as for any given locus, a parent passes along only one of its two alleles to a particular offspring. Thus, the genotypic value can be further decomposed into a component passed on to the offspring A (the *additive genetic value*) and a non-additive component, which includes the dominance, D , and any epistatic effects. We ignore the effects of epistasis here, whose complex features are extensively examined by Lynch and Walsh (1998) and Walsh and Lynch (2018). Thus,

$$z = \mu + A + D + E, \quad (15.2)$$

where μ is the population mean (by construction, the mean values of A , D , and E equal zero, such that $G = \mu + A + D$ in equation (15.1)). The additive values A are often referred to as *breeding values*, as the average phenotype of an offspring is just the average breeding value of its parents. Hence, the expected phenotypic value of an offspring is $\mu + (A_f + A_m)/2$, where A_f is the paternal (or sire) and A_m the maternal (or dam) breeding values.

Additive genetic values A , and particularly the variance in additive genetic values, are central to much of evolutionary quantitative genetics; we will see that they appear directly in the statistical mechanics that predict evolutionary responses to selection. They are the sum of the *average effects* of alleles on phenotypes. These average effects are defined by the slope of the regression of phenotype on genotype. The mechanics of this regression are treated in detail in Falconer and Mackay (1996) and Walsh and Lynch (2018). An important point is that, being average effects derived from regression, additive genetic values are independent of all other effects *by definition*. Consequently, the variance of these effects that are independent, and additive at the level of individuals, as in (15.2), is additive at the level of the population,

$$\sigma_z^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2, \quad (15.3)$$

where σ_x^2 represents the variance of each component. Further decomposition into additional components of genetic and environmental variance is possible; see Lynch and Walsh (1998, especially Chapters 4 to 7), and Falconer and Mackay (1996).

15.2.2 Additive Genetic Variances and Covariances

The concept of additive genetic variance (Section 15.2.1) becomes useful when investigating phenotypic similarity among relatives. When the phenotype can be decomposed as in equation (15.2), Fisher showed that the contribution of additive genetic effects to the phenotypic covariance between parents and their offspring is half the population variance of breeding values, σ_A^2 :

$$\sigma_A(z_p, z_o) = \frac{\sigma_A^2}{2}. \quad (15.4)$$

The coefficient of $\frac{1}{2}$ arises because parents and offspring share half their genes. Thus, twice the parent–offspring phenotypic covariance provides an estimate of σ_A^2 , which is called the *additive genetic variance*. There are a number of potential complications in applying equation (15.4), such as shared environmental effects and maternal effects when considering the mother–offspring covariance. See Falconer and MacKay (1996) and Lynch and Walsh (1998) for details on handling these complications. Additionally, equation (15.4) can be generalized to describe the contribution of additive genetic effects to similarity among arbitrary classes of relative. Under random mating,

$$\sigma_A(z_a, z_b) = \sigma_A^2 \cdot 2\Theta_{ab}, \quad (15.5)$$

where Θ is the *coefficient of coancestry*. Θ_{ab} is the probability that two alleles, randomly drawn from each of two individuals a and b , are identical by descent. Details are given in Lynch and Walsh (1998, Chapter 7).

The genetic basis of covariances between traits determines how the multivariate phenotype responds to selection, and so the contribution of additive genetic effects to such covariance will prove to be important. The covariance in breeding values between two traits (say, x and y), $\sigma_A(x, y)$, is needed for considering evolution of multiple traits. These *additive genetic covariances* can also be estimated from parent–offspring regressions, as the phenotypic covariance between one trait in the parent (p) and the other trait in the offspring (o) estimates half the additive genetic covariance of the traits, for example,

$$\sigma(x_p, y_o) = \sigma(y_p, x_o) = \frac{\sigma_A(x, y)}{2}. \quad (15.6)$$

See Lynch and Walsh (1998, Chapter 21) for more detail on components of genetic covariances among traits, and Walsh and Lynch (2018, Chapter 23) for information on a number of complications that arise under inbreeding.

Notably, two different mechanisms can generate a (within-individual) covariance in the breeding values of two traits: linkage (or rather *linkage disequilibrium*, which does not require physical linkage; see Chapter 2) and *pleiotropy*. When linkage disequilibrium occurs among loci (which need not be physically linked), alleles are correlated such that genotypes at any two loci may not be treated as independent draws from the population. In this case, even if each linked locus affects only a single trait, the correlation between alleles (linkage disequilibrium) generates a short-term correlation between the breeding values of the traits. Over time these associations decay through recombination randomizing allelic associations across loci. Alternatively, with pleiotropy a single locus can affect multiple traits. Covariances due to pleiotropic loci are stable over time in the absence of new mutations.

15.3 Parent–Offspring Regressions and the Response to Selection

With basic concepts and terminology in place, we are now in a position to consider the response to selection. In this section we will simply consider selection to be any change in the distribution of phenotype within a generation. For example, this could be the difference in mean phenotype

between an unselected cohort of individuals and those that survive culling. More generally, it is the difference in the distribution of phenotype among breeders, weighted by fecundity, and the distribution among all individuals born in the generation from which those breeders were produced. In later sections, we will elaborate on the statistical mechanics that relate effects of traits on fitness to within-generation changes in phenotype.

15.3.1 Single-Trait Parent–Offspring Regressions

Most of the theory of selection response in quantitative genetics assumes linear parent–offspring regressions that have homoskedastic residuals (i.e. the variance about the expected value is independent of the values of the parents). The simplest version considers the phenotypic value of the midparent, $z_{mp} = (z_m + z_f)/2$, with the offspring value z_o predicted as

$$z_o = \mu + b(z_{mp} - \mu), \quad (15.7)$$

where μ is the population mean and b is a regression coefficient. Since the slope b of the best linear regression of y on x , $y = a + bx$, is given by

$$b = \frac{\sigma(x, y)}{\sigma_x^2},$$

the slope of the midparent–offspring regression becomes

$$b = \frac{\sigma(z_o, z_{mp})}{\sigma^2(z_{mp})} = \frac{\sigma(z_o, z_m)/2 + \sigma(z_o, z_f)/2}{\sigma^2(z_m/2 + z_f/2)} = \frac{2\sigma_A^2/4}{2\sigma_z^2/4} = \frac{\sigma_A^2}{\sigma_z^2}. \quad (15.8)$$

This ratio of the (autosomal) additive genetic variance to the phenotypic variance is usually denoted h^2 (following Wright 1921a) and is called the (narrow-sense) *heritability*. Thus, the midparent–offspring regression is given by

$$E[z_o] = \mu + h^2(z_{mp} - \mu) = \mu + h^2 \left(\frac{z_m + z_f}{2} - \mu \right). \quad (15.9)$$

Equation (15.9) gives the expected value for an offspring. The actual value for any particular offspring is given by

$$z_o = \mu + h^2(z_{mp} - \mu) + e, \quad (15.10)$$

where the residual e of the parent–offspring regression has mean value zero and variance

$$\sigma_e^2 = \left(1 - \frac{h^4}{2} \right) \sigma_z^2. \quad (15.11)$$

15.3.2 Selection Differentials and the Breeder’s Equation

The parent–offspring regression allows us to predict the response to selection. Suppose the mean of parents that reproduce,¹ μ^* , is different from the population mean before selection, μ . Define the *directional selection differential* as $S = \mu^* - \mu$ (S is often simply called the selection

¹ This mean is intuitive for viability selection. For fecundity or sexual selection, it should be thought of as the mean phenotype, weighted by reproductive fitness. The mechanics described here hold regardless of the type of selection.

differential). From equation (15.9), the difference, R , between the mean value in offspring of these parents that reproduce, and the original mean of the population before selection, is

$$R = h^2 S. \quad (15.12)$$

This is the *breeder's equation* (Lush, 1937), which transforms the within-generation change in the mean (S) into a between-generation change in the mean (the selection response, R). Notice that strong selection does not necessarily imply a large response. If the heritability of a trait is very low (as occurs with many life-history traits), then even very strong selection results in very little (if any) response. Hence, selection (non-zero S) does not necessarily imply evolution (non-zero R).

15.3.3 Multiple-Trait Parent–Offspring Regressions

The (single-trait) parent–offspring regression can be generalized to a (column) vector of n trait values, $\mathbf{z} = (z_1, \dots, z_n)^T$. Letting \mathbf{z}_o be the vector of trait values in the offspring, \mathbf{z}_{mp} the vector of midparent trait values (i.e. the i^{th} element is just $(z_{f,i} + z_{m,i})/2$), and $\boldsymbol{\mu}$ be the vector of population means, then the parent–offspring regression for multiple traits can be written as having expected value

$$E[\mathbf{z}_o] = \boldsymbol{\mu} + \mathbf{H}(\mathbf{z}_{mp} - \boldsymbol{\mu}), \quad (15.13)$$

where \mathbf{H} is a multivariate generalization of the heritability h^2 that will be further explained in the coming subsections.

15.3.4 The Genetic and Phenotypic Covariance Matrices

In order to further decompose \mathbf{H} into workable components, we need to define the phenotypic covariance matrix \mathbf{P} whose ij^{th} element is the phenotypic covariance between traits i and j . Note that \mathbf{P} is symmetric, with the diagonal elements corresponding to the phenotypic variances and off-diagonal elements corresponding to the phenotypic covariances. In a similar manner, we can define the symmetric matrix \mathbf{G} , whose ij^{th} element is the additive genetic covariance (covariance in breeding values) between traits i and j . Note that when $i = j$, the entry in \mathbf{G} is equivalent to the additive genetic variance for the trait. Using similar logic to that leading to (15.8) (the slope of the single-trait regression), it can be shown that $\mathbf{H} = \mathbf{GP}^{-1}$, giving the multi-trait parent–offspring regression as

$$E[\mathbf{z}_o] = \boldsymbol{\mu} + \mathbf{GP}^{-1}(\mathbf{z}_{mp} - \boldsymbol{\mu}). \quad (15.14)$$

15.3.5 The Multivariate Breeder's Equation

Letting \mathbf{R} denote the column vector of responses (so that the i^{th} element is the between-generation change in the mean of trait i), and \mathbf{S} the vector of selection differentials, equation (15.14) allows us to generalise the breeder's equation to multiple traits,

$$\mathbf{R} = \mathbf{HS} = \mathbf{GP}^{-1}\mathbf{S}. \quad (15.15)$$

Equation (15.15) forms the basis for discussions about selection on multiple traits (Section 15.11). From equation (15.15), we can see the role of the proportion of variation and covariation that is generated by additive genetic effects on the transmission of within-generation changes in the mean (selection) to between-generation changes (the evolutionary response to selection). Whereas this is controlled by $h^2 = \sigma_A^2 / \sigma_p^2$ in the univariate case, it is controlled by the direct multivariate analog $\mathbf{H} = \mathbf{GP}^{-1}$ in the multivariate response to selection.

15.4 The Infinitesimal Model

The breeder's equation predicts the change in the mean following a single generation of selection from an unselected base population. The major assumption required is that the parent–offspring regression is linear (see also Walsh and Lynch, 2018, Chapter 6), but the question remains as to when this linearity holds. Further, the breeder's equation focuses only on the change in mean, leaving open the question of what happens to the variance (both phenotypic and genetic). The latter concern is of special relevance in evolutionary quantitative genetics, as stabilizing selection (which reduces the variance without necessarily any change in the mean) is thought to be common in natural populations. The infinitesimal model, introduced by Fisher (1918; see a comprehensive treatment in Bulmer, 1980), provides a framework to address these issues.

15.4.1 Linearity of Parent–Offspring Regressions under the Infinitesimal Model

Under the infinitesimal model, we assume that many (effectively an infinite number of) loci contribute to genetic variation for a given trait. Thus, from the central limit theorem, genotypic values (G) and breeding values are normally (or multivariate normally) distributed before selection (Bulmer 1971, 1976). Assuming environmental values (E) are also normal, then so is the phenotype z (this follows from equations (15.1) and (15.2)) and the joint distribution of phenotypic and genotypic values is multivariate normal. In this case, the regression of offspring phenotypic value z_o on parental phenotypes (z_f and z_m for the father and mother's values) is linear and homoskedastic.

15.4.2 Allele Frequency Changes under the Infinitesimal Model

Under the infinitesimal model, a trait is determined by an infinite number of unlinked and non-epistatic loci, each with an infinitesimal effect. A key feature of the infinitesimal model is that while allele frequencies are essentially unchanged by selection, large changes in the mean can still occur by summing infinitesimal allele frequency changes over a large number of loci. To see this feature, consider a trait determined by n completely additive and interchangeable loci, each with two alleles, Q and q (at frequencies p and $1 - p$), where the genotypes QQ , Qq , and qq contribute $2a$, a , and 0 (respectively) to the genotypic value. The resulting mean phenotype is $2nap$ and the additive variance (ignoring the contribution from gametic-phase disequilibrium) is

$$\sigma_A^2 = 2na^2p(1-p).$$

Since σ_A^2 is a function of na^2 , a must be of order $n^{-1/2}$. The change in mean due to a single generation of selection is $\Delta\mu = 2na\Delta p$. Assuming the frequency of Q changes by the same amount at each locus, $\Delta p = \Delta\mu/(2na)$. Since a is of order $n^{-1/2}$, Δp is of order $1/(n \cdot n^{-1/2}) = n^{-1/2}$, approaching zero as the number of loci becomes infinite. Thus, very small allele frequency changes at many loci can be the basis for large changes in mean phenotype.

What effect does this very small amount of allele frequency change have on the variance? Letting $p' = p + \Delta p$ denote the frequency after selection, the change in the *genic* variance, the additive genetic variance that would occur at linkage equilibrium, is

$$\begin{aligned}\Delta\sigma_a^2 &= 2na^2p'(1-p') - 2na^2p(1-p) \\ &= 2na^2\Delta p(1-2p-\Delta p) \\ &\approx a(1-2p)\Delta\mu.\end{aligned}$$

Since a is of order $n^{-1/2}$, the change in variance due to changes in allele frequencies is roughly $n^{-1/2}$ the change in mean. Thus, with a large number of loci, very large changes in the mean

can occur without any significant change in the variance. In the limit of an infinite number of loci, there is no change in the genic variance ($\Delta\sigma_a^2 = 0$), while arbitrarily large changes in the mean can occur.

15.4.3 Changes in Variances

As mentioned, under the infinitesimal model (and in the absence of drift) there is essentially no change in the genetic variances caused by allele frequency change. Changes in allele frequencies, however, are not the only route by which selection can change the variance (and other moments) of the genotypic distribution. Selection also creates associations (covariances) between alleles at different loci through the generation of gametic-phase (or linkage) disequilibrium, and such covariances can have a significant effect on the genetic variance. Disequilibrium can also change higher-order moments of the genotypic distribution, driving it away from normality and hence potentially causing parent–offspring regressions to deviate from linearity. This section provides an abbreviated treatment of the effects of selection on the variance. An expanded treatment is given in Walsh and Lynch (2018, Chapter 16). Additionally, Chapter 17 of that work gives background on how the genetic control of the phenotypic manifestation of environmental effects may allow the environmental variance to evolve, and Chapter 24 deals with departures from normality, both issues that are beyond the present scope.

Here, we focus on the consequences of linkage disequilibrium for the additive genetic variance. The additive genetic variance at a given time, $\sigma_{A,t}^2$, in the presence of linkage disequilibrium can be written as

$$\sigma_{A,t}^2 = \sigma_a^2 + d_t, \quad (15.16)$$

where σ_a^2 , the *genic variance*, is the additive genetic variance in the absence of disequilibrium and d_t the disequilibrium contribution at time t .

Assuming unlinked loci, Bulmer (1971) showed that the change in disequilibrium is given by

$$d_{t+1} = \frac{d_t}{2} + \frac{h_t^4}{2} \left(\sigma_{z^*,t}^2 - \sigma_{z,t}^2 \right), \quad (15.17)$$

where $\sigma_{z^*,t}^2 - \sigma_{z,t}^2$ is the within-generation change in the phenotypic variance. The first term represents the removal of linkage disequilibrium by recombination, while the second is the generation of linkage disequilibrium by selection. Note that equation (15.17) is the variance analog of the breeder's equation, relating the between-generation ($d_{t+1} - d_t$) and within-generation ($\sigma_{z^*,t}^2 - \sigma_{z,t}^2$) changes in the variance.

As an example, one can conduct a series of calculations starting with an unselected base population (where $d_0 = 0$). Iterating equation (15.17) gives the linkage disequilibrium (and hence the heritability, phenotypic variance, and response to selection) in any desired generation. Under directional selection, most of the linkage disequilibrium is generated in the first three to five generations (Figure 15.1), after which time d is very close to its equilibrium value \tilde{d} . Once selection stops, the current linkage disequilibrium is halved each generation, with the additive variation rapidly returning to its value before selection. This occurs because under the infinitesimal model, all changes in variances are due to linkage disequilibrium, which decays via recombination in the absence of selection.

Another important point regarding equation (15.17) is that, in the absence of any change in the mean phenotype ($S = 0$), selection can still act on the variance. Stabilizing selection, which removes extreme individuals and reduces the variance, is generally thought to be widespread in nature. The within-generation reduction in variance generates negative linkage disequilibrium, which in turn reduces the additive variance. Under the infinitesimal model, once selection on

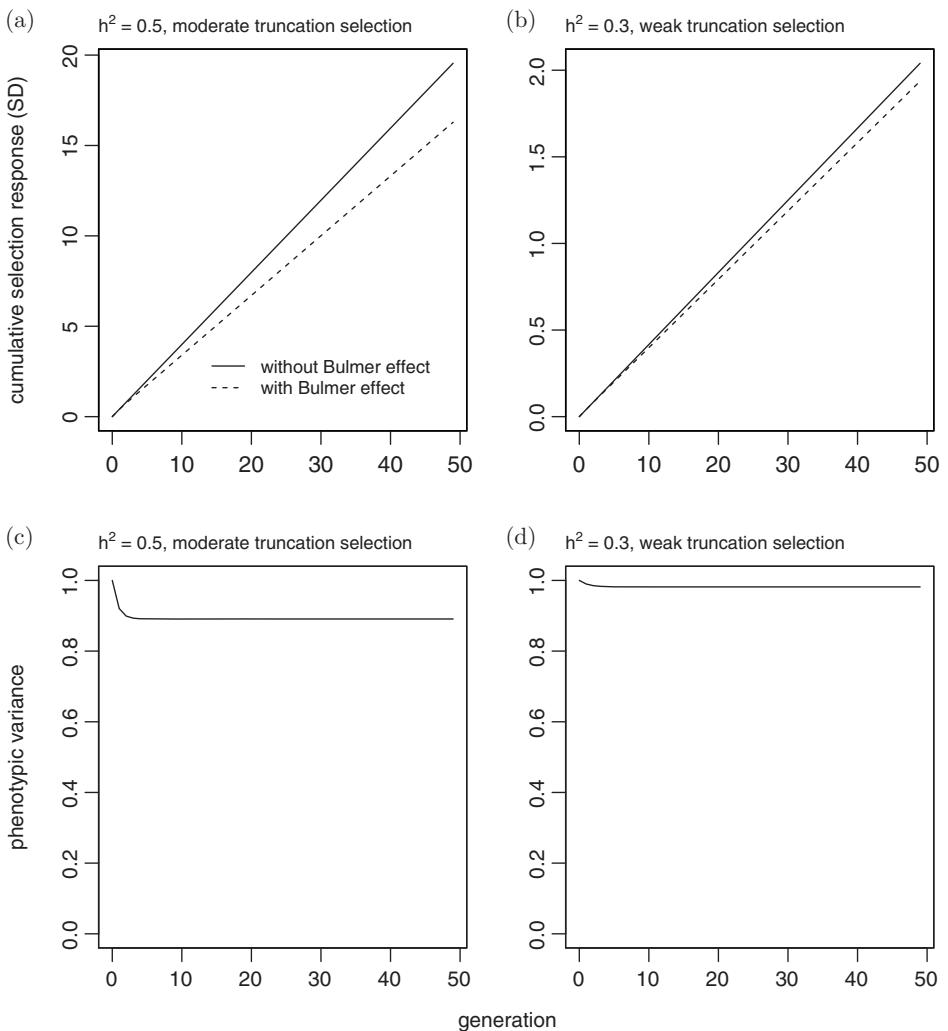


Figure 15.1 The influence of changes in the variance under directional selection on the response to selection. The Bulmer and Breeder's equations (equations (15.17) and (15.12), respectively) were iterated to obtain the trajectory of the mean ((a) and (b)) and the phenotypic variance ((c) and (d)), from a starting value of $d_0 = 0$. In (a) and (b) the solid lines represent the trajectory of mean phenotype if the genic variance, σ_a^2 , is assumed to be constant, and the dashed lines give the trajectories accounting for the Bulmer effect, which is the reduction in additive genetic and phenotypic variances due to the accumulation of linkage disequilibrium. (a) and (c) are obtained for moderate heritability and moderate selection (at least by the standards of artificial selection). Specifically, for ((a) and (c)), $h^2 = 0.5$ and truncation selection (breeding only from individuals with phenotypes above a critical value; see Walsh and Lynch 2018, Chapter 14) at the mean ($S \approx 0.8$ phenotypic standard deviations) was applied. Lower heritability and less intense selection were assumed for (b) and (d): $h^2 = 0.3$ and $S \approx 0.12$ standard deviations. These values approximate typical values observed in nature (Postma, 2014; Morrissey, 2016).

the variance stops, the variance returns to its initial value after a few generations. Likewise, under disruptive selection, individuals of intermediate value are selected against, increasing the variance. This generates positive d , increasing the additive variance. Again, when selection is stopped, the variance decays back to its initial (pre-selection) value.

Iteration of equation (15.17) allows any form of selection to be analyzed. For modeling purposes, it is often assumed that the within-generation change in the phenotypic variance can be written as

$$\sigma_{z^*,t}^2 = (1 - \kappa) \sigma_{z,t}^2 \quad (15.18)$$

where κ is a constant that does not change over time. Noting that $\sigma_{z,t}^2 = \sigma_{A,t}^2/h_t^2$ and substituting (15.18) into (15.17) recovers the result of Bulmer (1974),

$$d_{t+1} = \frac{d_t}{2} - \frac{\kappa}{2} h_t^2 \sigma_{A,t}^2 = \frac{d_t}{2} - \frac{\kappa}{2} \frac{[\sigma_a^2 + d_t]^2}{\sigma_z^2 + d_t}. \quad (15.19)$$

Again, simple iteration allows one to compute the variance in any generation.

15.4.4 The Equilibrium Additive Genetic Variance

The equilibrium variances, and hence the asymptotic rate of response under directional selection, or the equilibrium variance under stabilizing or disruptive selection, are easily obtained. At equilibrium, equation (15.19) implies

$$\tilde{d} = -\kappa \tilde{h}^2 \tilde{\sigma}_A^2 = -\kappa \frac{(\sigma_a^2 + \tilde{d})^2}{\sigma_z^2 + \tilde{d}}.$$

Solving for \tilde{d} gives the equilibrium additive genetic variance as

$$\tilde{\sigma}_A^2 = \sigma_z^2 \theta, \quad \text{where } \theta = \frac{2h^2 - 1 + \sqrt{1 + 4h^2(1 - h^2)\kappa}}{2(1 + \kappa)},$$

and the resulting heritability at equilibrium is

$$\tilde{h}^2 = \frac{\tilde{\sigma}_A^2}{\tilde{\sigma}_z^2} = \frac{\tilde{\sigma}_A^2}{\sigma_z^2 + (\tilde{\sigma}_A^2 - \sigma_A^2)} = \frac{\theta}{1 + \theta - h^2}. \quad (15.20)$$

The simple picture that emerges from directional selection under the infinitesimal model is that while the heritability may decrease slightly from its initial value (due to generation of linkage disequilibrium), the response to selection continues without limit. It is of note that recombination is typically very powerful for restoring σ_A^2 towards its equilibrium value. The medium-term effects of selection, acting through the effect of selection to reduce genetic variation via the generation of linkage disequilibrium, are probably typically modest in nature (Figure 15.1). Values of d_t and \tilde{d}_t are often assumed to be much larger than in fact they are, and are rarely calculated (but see Shaw *et al.*, 1992; Careau *et al.*, 2015).

Biological reality, of course, places limits on the phenotypic values that can be obtained by selection response. For example, directional selection may move the mean phenotype to a region on the fitness surface where stabilizing selection dominates (see Section 15.5). Likewise, selection on one trait may be opposed by selective constraints imposed by other traits also under selection, and a limit can be reached despite significant additive variance and a non-zero S in the trait being followed (Section 15.11). Additionally, drift and selection do cause changes in σ_A^2 via allele frequency change, and so finite population size and departures from the infinitesimal model do require consideration (see Walsh and Lynch, 2018, Chapters 16 and 24).

15.5 Inference of σ_A^2 and G

Since they are not directly observable, but are key determinants of evolution in response to selection, the empirical inference of genetic variances and covariances is an important topic. Methods for the inference of quantitative genetic parameters are extensively reviewed, particularly in Lynch and Walsh (1998; see especially Chapters 17 to 21). Around the time of that publication, the ‘animal model’ (Henderson, 1975), a mixed model-based system for using a pedigree-derived relationship matrix to allow direct inference of quantitative genetic parameters, became the predominant approach for inference of parameters such as σ_A^2 and G in nature and in many experiments. Introductions to variance component estimation with mixed models are given in Lynch and Walsh (1998, Chapters 26 and 27) and Walsh and Lynch (2018). General reviews aimed at evolutionary readers are given in Kruuk (2004) and Wilson *et al.* (2010). Walsh and Lynch (2018) Chapters 19 and 20 provide a somewhat more technical overview of the mixed model methodology associated with the animal model. Briefly, the principle benefits of the animal model arise from the fact that it can use relatedness information from all available classes of relatives simultaneously, and it can also be extended to control for confounding effects, such as maternal or other common environment effects.

It is beyond the scope of this chapter to recapitulate this background. However, it may be of use to provide an intuitive explanation of how a relationship matrix (typically pedigree-derived) is used in a mixed model in order to generate the analysis referred to as an ‘animal model’. The means by which the use of relatedness (see Lynch and Walsh, 1998, Chapter 7) information allows direct estimation of quantitative genetic parameters is not described in some of the introductions to the animal model that seek to be accessible, and the core concept is easily missed by many readers, when the animal model is described formally.

Consider an analysis of paternal half siblings (see Lynch and Walsh, 1998, Chapter 18). If an experiment generated groups of three offspring, with each group sharing a sire as a parent, and every offspring being born to a different and unrelated dam, each individual within a sib group would share one quarter of its genes with its sibs and no genes identical by descent with individuals from other sibships. The offspring of the first three sibling groups would thus have additive genetic relatedness matrix

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix}.$$

Imagine, now, that the genetic variance for a trait was $\sigma_A^2 = 8$ and the environmental variance was $\sigma_E^2 = 12$. The covariance of sibs would thus be $\frac{1}{4}\sigma_A^2 = 2$, since they are expected to share one quarter of their genes identical by descent. Providing that adequate randomization and environmental controls (and if not, statistical control of confounding variables) has been implemented, all individuals, regardless of whether they are sibs or not, would share no covariance via environmental effects. A model of all of the phenotypes simultaneously could be constructed according to

$$\begin{aligned}
 & \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \\ z_9 \end{bmatrix} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{A}\sigma_A^2 + \mathbf{I}\sigma_E^2) \\
 & \sim N \left(\mathbf{X}\boldsymbol{\beta}, \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & 1 & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 1 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 8 & & & & & & & & \\ & 12 & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \\ & & & & & & & & \end{bmatrix} \right) \\
 & \sim N \left(\mathbf{X}\boldsymbol{\beta}, \begin{bmatrix} 20 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 20 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 20 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 20 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 & 20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 20 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 20 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 20 \end{bmatrix} \right) \tag{15.21}
 \end{aligned}$$

where \mathbf{X} and $\boldsymbol{\beta}$ are a fixed effects design matrix and a vector of fixed effects, respectively, that may be fitted to account for confounding covariates. This is an animal model. In practice the values of the variance components ($\sigma_A^2 = 8$ and $\sigma_E^2 = 12$) are unknown. Working from the proportions to which observed covariances (as in equation (15.21)) reflect the covariances in the \mathbf{A}

matrix versus uncorrelated residuals (some multiple of an identity matrix), observed covariances are used to generate estimates of σ_A^2 and σ_E^2 . The power of the animal model comes from the fact that it can easily use information from different classes of relatives simultaneously, simply by using the A matrix that reflects the relationships among available phenotyped individuals.

15.6 Fitness

As we have seen, predicting the selection response under the infinitesimal model requires knowledge of the change in both mean and variance after an episode of selection. In an artificial selection program, the breeder or experimentalist can not only measure these components, but also largely set their values. In nature, the currency of selection is fitness, and the change in the phenotypic distribution is computed by first weighting individuals by their fitness values. Discussion of selection response in nature thus starts by trying to assign expected fitness values to particular trait values. This is the linking step that allows use of the above machinery for prediction of selection response. A greatly expanded treatment of measurement of individual fitness is given in Walsh and Lynch (2018, Chapter 29).

15.6.1 Individual Fitness

Loosely stated, the *lifetime* (or *total*) *fitness* of an individual is the number of descendants it leaves at the start of the next generation. When measuring the total fitness of an individual, care must be taken not to cross generations or to overlook any stage of the life cycle in which selection acts (Hadfield, 2012; Thompson and Hadfield, 2017). To accommodate these concerns, lifetime fitness is defined as the total number of zygotes (newly fertilized gametes) that an individual produces. Ideally an assessment of fitness should reflect the number of zygotes produced from individuals, starting when the focal individuals are themselves zygotes. Measuring total fitness from any other starting point in the life cycle (e.g. from adults in one generation to adults in the subsequent generation) can result in a very distorted picture of true fitness of particular phenotypes (Prout, 1965, 1969). If generations are crossed, measures of selection on a particular parental phenotype in reality are averages over both parental and offspring phenotypes, which may differ considerably.

Of course, an individual's demographic contribution to a population may be moderated by its own effects on its offspring's fitness. This fact has typically been handled in practice by counting fitness at some stage where offspring are essentially independent. Typically, the investigator makes some judgement as to some stage at which offspring are independent, and counts offspring surviving to this stage as *lifetime reproductive success*, as opposed to some operational system of counting the number of fertilized zygotes (which may be called *lifetime breeding success*). It should be noted that these terms are not necessarily used precisely in much of the literature on fitness and natural selection in the wild. It is typical to view lifetime reproductive success as an 'ultimate' measure of fitness (Clutton-Brock, 1988). However, this view is unfortunate: the practice of using numbers of offspring raised to independence in formal quantitative genetic studies of selection and evolution is likely to obscure, rather than illuminate, the roles played by parental performance in the evolutionary process (Hadfield, 2012; Thompson and Hadfield, 2017). Critically, evolutionary quantitative genetics is not blind to parental effects, and a variety of models exist to formally handle the evolutionary consequences of cross-generational effects (see especially Willham, 1963, 1972; Kirkpatrick and Lande, 1989). These models are reviewed in Walsh and Lynch (2018, Chapter 22) and in Hadfield (2012).

Systems for measuring lifetime fitness have been especially well developed for laboratory populations of *Drosophila* (reviewed by Sved, 1989). Measurements of lifetime fitness in field situations are more difficult and (not surprisingly) are rarely made. Attention instead is usually focused on particular episodes of selection or particular phases of the life cycle. *Fitness components* for each episode of selection are defined to be multiplicative. For example, lifetime fitness can be partitioned as a product of the probability of surviving to reproductive age, the number of mates and the number of zygotes per mating. The number of mates is a measure of *sexual selection*, while the viability and fertility components measure natural selection. A commonly measured fitness component is *breeding success*, the number of offspring per adult, which reflects both natural (fertility) and sexual selection (in males, the number of matings per adult). Clutton-Brock (1988) reviews estimates of reproductive success from natural populations.

Fitness components can themselves be further decomposed. For example, fertility in plants (the number of seeds per plant) might be partitioned as a product of the number of stems per plant, the number of inflorescences per stem, the average number of seed capsules per inflorescence, and the average number of seeds per capsule. Such a decomposition allows the investigator to ask questions of the form: do plants differ in number of seeds mainly because some plants have more stems, or more flowers per stem, or are there trade-offs between these?

Estimates of fitness can be obtained from either *longitudinal* or *cross-sectional* studies. A longitudinal study follows a cohort of individuals over time, while a cross-sectional study examines individuals at a single point in time. Cross-sectional studies typically generate only two fitness classes (e.g. dead versus living, mating versus unmated), and their analysis involves a considerable number of assumptions (Lande and Arnold, 1983; Arnold and Wade, 1984b). While longitudinal studies are preferred (Clutton-Brock and Sheldon, 2010), they usually require far more work and may be impossible to carry out in many field situations. Age-structured populations pose further complications in that proper fitness measures require knowledge of the population's demography – see Charlesworth (1994), Lande (1982), Lenski and Service (1982), and Travis and Henrich (1986) for details; recent progress has been made on defining fitness and measuring selection in age-structured populations in fluctuating environments (Engen *et al.*, 2009, 2012).

15.6.2 Episodes of Selection

As mentioned, individuals are often measured over more than one episode of selection. Imagine that a cohort of n individuals (indexed by $1 \leq r \leq n$) is followed through several episodes. Let $W_j(r)$ be the fitness measure for the j^{th} episode of selection for the r^{th} individual. For example, for viability $W_j(r)$ is either 0 (dead) or 1 (alive) at the census period. Note that coding survival numerically with 0s and 1s is not arbitrary (as is sometimes the case in statistical analysis of binary outcomes). Individuals that die are represented by zero individuals in the population after an episode of viability selection, and individuals that survive are represented by one copy of themselves. Let \bar{W}_j denote the mean fitness of individuals with a given phenotypic value following the j^{th} episode of selection. Relative fitness components $w_j(r) = W_j(r)/\bar{W}_j$ will turn out to be especially useful. At the start of the study, the frequency of each individual is $1/n$, giving for the first (observed) episode of selection

$$\bar{W}_1 = \frac{1}{n} \sum_{r=1}^n W_1(r).$$

Caution is in order at this point as *considerable selection may have already occurred prior to the life cycle stages being examined* – which can create severe complications (see Graffen, 1988; Hadfield, 2008). Following the first episode of selection, the new fitness-weighted frequency of the r^{th} individual is $w_1(r)/n$, implying

$$\overline{W}_2 = \sum_{r=1}^n W_2(r) \cdot w_1(r) \cdot \left(\frac{1}{n}\right).$$

In general, for the j^{th} episode of selection,

$$\overline{W}_j = \sum_{r=1}^n W_j(r) \cdot w_{j-1}(r) \cdot w_{j-2}(r) \cdots w_1(r) \cdot \left(\frac{1}{n}\right). \quad (15.22)$$

Note that if $W_j(r) = 0$ for any episode of viability selection, further fitness components for r are not realized.

Letting $p_j(r)$ be the fitness-weighted frequency of individual r after j episodes of selection, it follows that $p_0(r) = 1/n$ and

$$p_j(r) = w_j(r) \cdot p_{j-1}(r) = \frac{1}{n} \prod_{i=1}^j w_i(r). \quad (15.23)$$

Thus, equation (15.22) can also be expressed as $\overline{W}_j = \sum W_j(r) \cdot p_{j-1}(r)$. Using these weights allows fitness-weighted moments to be calculated; for example, the mean of a particular trait following the j^{th} episode is computed as

$$\mu_{z(j)} = \sum z(r) \cdot p_j(r), \quad (15.24)$$

where $z(r)$ is the value of the trait of individual r .

The directional selection differential S is computed as the difference between fitness-weighted means before and after an episode of selection, with the differential S_j for the j^{th} episode given by

$$S_j = \mu_{z(j)} - \mu_{z(j-1)}. \quad (15.25)$$

Selection differentials are additive over episodes, so that the total differential S following k episodes of selection is

$$S = \mu_{z(k)} - \mu_{z(0)} = (\mu_{z(k)} - \mu_{z(k-1)}) + \cdots + (\mu_{z(1)} - \mu_{z(0)}) = S_k + \cdots + S_1. \quad (15.26)$$

15.7 The Robertson–Price Identity, and Theorems of Selection

In this section we introduce several key results that are useful for relating different aspects of phenotypic selection (e.g. the covariance of trait and fitness, the regression of fitness on trait values, and the change in mean phenotype) to one another, and to the evolutionary response to selection. We attempt to clarify some aspects of how these results relate to different aspects of the relationship between genetic variation for traits and fitness, so as to address common misunderstandings. Finally, we discuss some ways in which these results may be applied in empirical studies.

15.7.1 Description of the Theorems

As first noted by Robertson (1966), and greatly elaborated on by Price (1970, 1972a), the directional selection differential equals the covariance of phenotype and relative fitness,

$$S = \sigma(w, z). \quad (15.27)$$

This identity is quite useful for obtaining the selection differential in complex selection schemes and (as is detailed below) forms the basis for a number of useful expressions relating selection and fitness. The Robertson–Price identity is applicable to selection over the entire lifetime, and also to analysis of selection during specific episodes of the life cycle or annual cycle.

To obtain the Robertson–Price identity, let μ_s be the fitness-weighted mean after selection and μ the mean before selection,

$$\begin{aligned} S &= \mu_s - \mu = \sum_{r=1}^n z(r) w(r) \frac{1}{n} - \mu \\ &= E[z w] - (1)E[z] \\ &= E[z w] - E[w]E[z] \\ &= \sigma(w, z), \end{aligned}$$

where we have used the fact that (by construction) $E[w] = 1$.

As presented in equation (15.27), the Robertson–Price identity describes *phenotypic selection*. If applied to breeding values (A ; see equation (15.2)), rather than phenotypes, the Robertson–Price identity describes the *within-generation change in breeding values due to selection*. This is the secondary theorem of selection (Robertson, 1966),

$$\Delta \bar{A}_z = \sigma(A_z, w), \quad (15.28)$$

where A_z are breeding values for focal trait z .

If applied to additive values for fitness, Fisher's (1930) fundamental theorem of selection is obtained from the Robertson–Price identity. Let the quantity considered in equation (15.27) be additive genetic values for relative fitness, A_w ,

$$\Delta \bar{A}_w = \sigma^2(A_w, w).$$

Some works on the fundamental theorem do not necessarily specify that the aspect of fitness in question is *relative fitness*. This reflects the fact that Fisher understood that fitness of an individual in an evolutionary sense is necessarily in relation to the fitness of others, and so the fact that the fundamental theorem pertains to relative fitness was treated as implicit. This has been one contributor to subsequent confusion about the fundamental theorem. Note that from the definition of additive genetic value, residuals of the regression of phenotype (in this case, fitness) on genetic value are independent of predictions from that regression (in this case, independent of additive genetic values for fitness), so

$$\Delta \bar{A}_w = \sigma(A_w, A_w) = \sigma_A(w), \quad (15.29)$$

which is the fundamental theorem of selection: *the partial change in (relative) fitness at any given time, attributable to selection, is given by the additive genetic variance in (relative) fitness in a population at that time*.

The fundamental and secondary theorems of selection are true theorems, in that they hold very generally, requiring only *definitions* of fitness (as differential demographic representation) and of additive genetic effects (as the sum of average effects of alleles on phenotype or fitness). However, fairly widespread misunderstandings persist as to how these theorems should be interpreted and, indeed, whether they are in fact theorems. It must be understood that the

theorems apply to the change in mean additive genetic values *within generations*. While evolutionary quantitative genetics generally treats selection as a phenotypic phenomenon (within-generation change in the distribution of phenotypes, weighted by fitness), the theorems pertain to a sort of genetical selection: the change in additive genetic values within generations. In practice, evolutionary quantitative geneticists are typically willing to *assume* that within-generation changes in breeding values are faithfully transmitted to the next generation. This *assumption* will not generally hold in the presence of non-additive genetic variation. Throughout this chapter, expressions given for evolutionary change represent the within-generation change in additive genetic values that are expected given different features of selection and genetic variation. They thus represent evolutionary change, under the assumption that changes in mean additive genetic values within generations are faithfully transmitted to subsequent generations.

Much confusion about the theorems of selection probably arises from evolutionary quantitative geneticists understanding the change in additive genetic values to refer to evolutionary change. Further elaboration and clarification is beyond the present scope, but the interested reader is referred to Ewens (1989), Price (1972b) and Walsh and Lynch (2018, Chapter 6).

15.7.2 Empirical Operationalization of the Theorems

The quantities on the right-hand sides of equations (15.28) and (15.29) are, in principle, estimable from data. In order to estimate these quantities, various assumptions must be introduced (e.g. an assumption of multivariate normality of breeding values and environmental effects for mixed model-based inference of these values), such that any analysis is not strictly an application of a theorem (i.e. it may go wrong!). However, there is potential value in estimating quantities such as the genetic variance in relative fitness, and genetic covariances of traits with relative fitness.²

Because evolution (changes in genetic value) requires covariance of genetic values with fitness, it requires genetic variation for fitness. Therefore, estimation of the genetic variance for relative fitness gives information about the maximum possible rate of evolution. It also provides a sanity check: if mechanics for predicting evolution based on phenotypic selection (given throughout this chapter) predict substantial evolutionary change, but there is no genetic variance in relative fitness, then this should indicate some inconsistency in predictions that should be further investigated. It is worth keeping in mind, however, that inference of genetic parameters for traits with high residual variation, especially fitness, is very challenging.

The secondary theorem has also been used to check the consistency of evolutionary predictions based on phenotypic selection analysis. Several formulations of such consistency checks are possible (Rausher, 1992; Morrissey *et al.*, 2010; reviewed in Walsh and Lynch 2018, Chapter 20). Briefly, if the phenotypic relationship between a trait and fitness is confounded, say by an environmental variable that affects both the trait and fitness, then trait–fitness covariance will occur that does not necessarily reflect natural selection. Any resulting covariance between trait and fitness will affect the selection differential S , in a way that does not contribute to evolutionary change (as in equation (15.12), with S given by the Robertson–Price identity, equation (15.27)), even if the trait is heritable. Predictions of the breeder’s equation (15.12) can thus, in

² Note that this quantity, $\sigma_A(z, w)$, is not necessarily the same quantity that appears in equation (15.28). The theorem pertains to the covariance of additive genetic values for traits with (phenotypic) fitness. Robertson used both quantities – compare Robertson (1966) and Robertson (1968). These quantities are the same if there is no gene–environment covariance for fitness. There is no mathematical reason why this covariance should be zero, but in practice, situations where gene–environment covariance may be strong within a population are probably limited, and will typically reflect a poor choice of scale on which to conduct an analysis.

principle, be checked by estimating the genetic covariance of traits with relative fitness, and comparing this quantity to R . Other checks, in particular comparing the genetic and phenotypic partial regressions of fitness on traits, are possible, and can serve as consistency checks to multivariate selection analyses discussed in sections 15.3.5, 15.11 and 15.12.

15.8 The Opportunity for Selection

How does one compare the amount of selection acting on different episodes and/or different populations? At first thought, one might consider using the standardized selection differential,

$$i = \frac{S}{\sigma_z}, \quad (15.30)$$

which is just the directional selection differential scaled in terms of trait standard deviations (i is often called the *selection intensity* and is widely used in breeding). The intensity of selection i is often calculated and referred to as a *variance-standardized selection differential*, which may be denoted S_σ . The drawback with i as a measure of *overall* selection on individuals is that it is *trait-specific*. Hence, i is appropriate if we are interested in comparing the strength of selection on a particular *trait*, but inappropriate if we wish to compare the overall strength of selection, for example as might occur due to simultaneous selection of multiple traits.

In a sense, a much cleaner measure (independent of the traits under selection) is I , the *opportunity for selection*, defined as the variance in *relative fitness*:

$$I = \sigma_w^2 = \frac{\sigma_W^2}{\bar{W}^2}. \quad (15.31)$$

This measure was introduced by Crow (1958, reviewed in 1989), who referred to it as the *index of total selection*. Crow noted that if fitness is perfectly heritable (e.g. $h^2(\text{fitness}) = 1$), then $I = \Delta\bar{w}$, the scaled change in fitness. Following Arnold and Wade (1984a,b), I is referred to as the opportunity for selection, as any change in the distribution of fitness caused by selection represents an opportunity for within-generation change. A key feature of I is that it bounds the maximal selection intensity i for *any* trait. From the Robertson–Price identity (15.27), the correlation between any trait z and relative fitness (which is bounded in absolute value by 1) is

$$|\rho_{z,w}| = \frac{|\sigma_{z,w}|}{\sigma_z \sigma_w} = \frac{|S|}{\sigma_z \sqrt{I}} = \frac{|i|}{\sqrt{I}} \leq 1,$$

implying

$$|i| \leq \sqrt{I}. \quad (15.32)$$

Thus, the most that the mean of any trait can be shifted within a generation is \sqrt{I} phenotypic standard deviations. I is thus a bound for i , but its practical utility depends on the correlation between relative fitness and the trait being considered. The opportunity for selection may be useful in some circumstances for distinguishing the intensity of either univariate or multivariate selection, particularly in experiments. However, in nature, the majority of fitness variation may be a result of stochasticity – different realizations of the number of offspring left by essentially identical individuals. In such cases, the intensity of selection may be disassociated both from selection on a given trait, and even from selection generally. This disassociation of I from i is demonstrated in Figure 15.2.

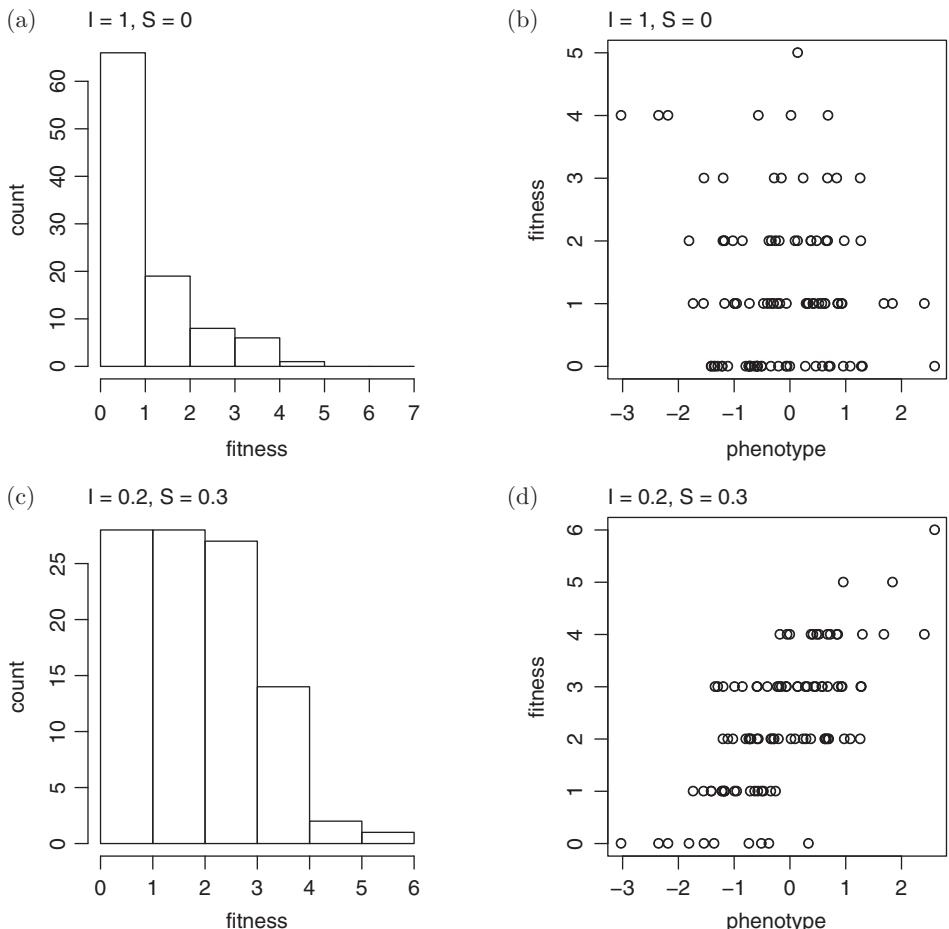


Figure 15.2 The potential for the opportunity for selection I to be disassociated from phenotypic selection. In nature, much variation in fitness may be random variation, such that large variation in fitness (a), or high opportunity for selection ((a) and (b)), need not be associated with systematic association of any particular trait, or indeed any traits at all, with fitness (b). Similarly, differences in the opportunity for selection ((a) and (c)), need not correspond to differences in the intensity of selection for any trait or traits ((b) and (d)). All plots show absolute fitness, but values I and S represent the variance in relative fitness, and the covariance of relative fitness with trait values, respectively.

15.9 Selection Coefficients

One of the strengths of the evolutionary quantitative genetic approach to studying selection and the adaptive response to selection is that it gives very specific meaning to different ways of expressing the association between traits and fitness. This meaning comes from the known formal relations between selection, genetic variation, and evolutionary change (e.g. equations (15.12), (15.15), and (15.30)). We can therefore say not only that natural selection is an association between trait and fitness, but also precisely what different aspects of this association, for example the covariance of traits and fitness, or the regressions of fitness on traits, mean. This reliance on the statistical mechanics of evolution is beneficial, not only for making quantitative predictions for specific traits and populations, but also for allowing comparison of selection across species, populations, traits, and so on. This section briefly reviews different measures of

the association between traits and fitness, and the quantitative relations of each to changes in the mean and variance of traits.

15.9.1 Measures of Selection on the Mean

A general way of detecting selection on a trait is to compare the (fitness-weighted) phenotypic distribution before and after some episode of selection (see equations (15.12) and (15.27)). Growth or other ontogenetic changes, immigration, and environmental changes can also change the phenotypic distribution, and great care must be taken to account for these factors. Another critical problem in detecting selection on a trait is that selection on phenotypically correlated traits can also change the distribution. Keeping this important caveat in mind, we first examine measures of selection on a single trait, as these form the basis for our discussion in Section 15.11 on measuring selection on multiple traits. Typically, selection on a trait is measured by changes in the mean and variance, rather than changes in the entire distribution. While often only the mean is examined, considerable selection can occur without any significant change in the mean (e.g. stabilizing selection). That said, we focus first on selection on the mean.

Two measures of within-generation change in phenotypic mean have been previously introduced: the directional selection differential S and the selection intensity i (the selection differential, standardized by the standard deviation of the trait, equation (15.30)). A third measure is the directional selection gradient,

$$\beta = \frac{S}{\sigma_z^2}. \quad (15.33)$$

As further detailed in Section 15.10.2, β is the slope of the linear regression of fitness on phenotype. These three measures (i , S , β) are interchangeable in their qualitative interpretation for selection acting on a single trait (they are quantitatively the same when the trait is expressed in units of its own standard deviation), but their multivariate extensions have very different behaviors (Section 15.11.1). A depiction of the scaling relationships between selection differentials S and selection gradients β is given in the comparison between different parts of Figure 15.3. In particular, the multivariate extension of β has generally been the measure of choice for characterizing natural selection, as it measures the amount of direct selection on a trait, while S (and hence i) accounts for both direct selection and indirect effects due to selection on phenotypically correlated traits.

Recalling that $h^2 = \sigma_A^2 / \sigma_z^2$, the response to selection is often written by breeders as

$$R = h^2 \sigma_z i = \sigma_A h i.$$

The response can also be expressed in terms of the directional selection gradient,

$$R = \sigma_A^2 \beta. \quad (15.34)$$

15.9.2 Measures of Selection on the Variance

Similar measures can be defined to quantify the change in the phenotypic variance. At first glance this change seems best described by $\sigma_{z^*}^2 - \sigma_z^2$, where $\sigma_{z^*}^2$ is the phenotypic variance following selection. The problem with this measure is that directional selection reduces the variance. Lande and Arnold (1983) showed that

$$\sigma_{z^*}^2 - \sigma_z^2 = \sigma (w, (z - \mu_z)^2) - S^2,$$

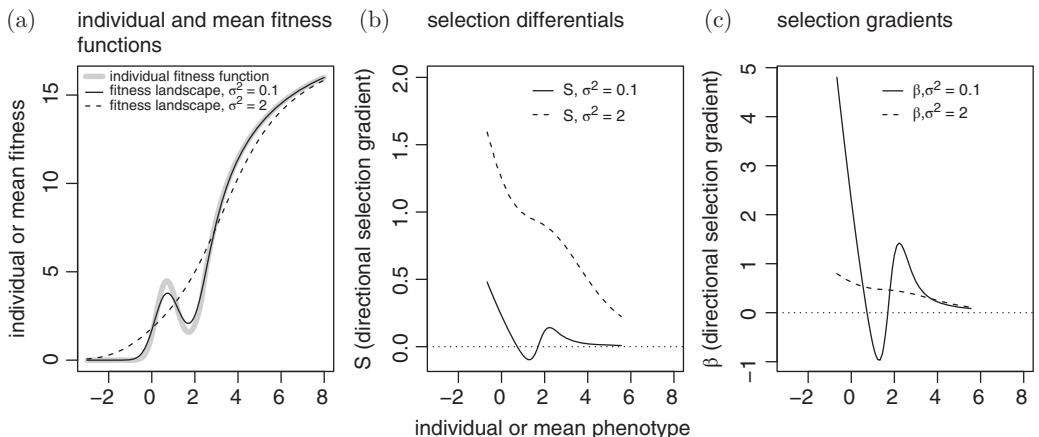


Figure 15.3 The relationship between fitness functions, fitness landscapes, and directional selection differentials and gradients. In (a), the thick gray line shows a hypothetical *fitness function*, specifying the relationship between *individual* phenotype and relative fitness. The black lines show the corresponding *fitness landscapes*, or relationships between *population mean* phenotype and fitness, which depend on both the individual fitness function and the distribution of phenotype. The solid line represents the fitness landscape for a phenotypic variance of 0.1 units², and the dashed line represents the corresponding landscape for a phenotypic variance of 2 units² (both assume a normal distribution of phenotype). In (b) and (c) directional selection differentials and gradients are depicted as a function of mean phenotype, again for σ_z^2 of 0.1 and 0.2 units². These are the slopes of the fitness landscapes, scaled by mean fitness.

implying that directional selection decreases the phenotypic variance by S^2 . With this in mind, Lande and Arnold suggest a corrected measure, the *stabilizing selection differential*

$$C = \sigma_{z^*}^2 - \sigma_z^2 + S^2, \quad (15.35)$$

which describes selection acting directly on the variance.

Analogous to S equaling the covariance between z and relative fitness, equation (15.35) implies that C is the covariance between relative fitness and the squared deviation of a trait from its mean,

$$C = \sigma(w, (z - \mu)^2).$$

As was the case with S , the opportunity for selection I bounds the maximum possible within-generation change in variance (Arnold, 1986). Expressing C as a covariance and using the standard definition of a correlation gives $C = \rho_{w,(z-\mu)^2} \sigma_w \sigma_{(z-\mu)^2}$. Since $\rho_{w,(z-\mu)^2}^2 \leq 1$, we have

$$C^2 \leq \sigma_w^2 \sigma_{(z-\mu)^2}^2 = I \cdot (\mu_{4,z} - \sigma_z^4),$$

where $\mu_{4,z}$ is the fourth central moment of the trait and σ_z^4 is the square of the phenotypic variance. Thus,

$$|C| \leq \sqrt{I(\mu_{4,z} - \sigma_z^4)}.$$

If z is normally distributed, the fourth central moment $\mu_{4,z} = 3\sigma_z^4$, giving

$$|C| \leq \sigma_z^2 \sqrt{2I}.$$

The quadratic analog of β , the *quadratic* (or *stabilizing*) *selection gradient* γ , was suggested by Lande and Arnold (1983),

$$\gamma = \frac{\sigma(w, (z - \mu)^2)}{\sigma_z^4} = \frac{C}{\sigma_z^4}. \quad (15.36)$$

As with β , γ is the measure of choice when dealing with multiple traits because it accounts for the effects of (measured) phenotypically correlated traits (Section 15.11).

While the distinction between differentials and gradients may seem trivial in the univariate case (only a scale difference), when considering multivariate selection, gradients have the extremely important feature of removing the effects of phenotypic correlations.

15.10 Fitness Functions and the Characterization of Selection

In this section we further explore the mechanics that link fitness functions, or the relationships between individual trait values and individual expected fitness, to the population-level selection coefficients discussed in Section 15.9. These mechanics will ultimately underlie the empirical inference of selection, which we will discuss in Section 15.12.

15.10.1 Individual and Mean Fitness Functions

$W(z)$, the expected fitness of an individual organism with phenotype z , describes a *fitness surface* (or *fitness function*), relating fitness and trait value. The *relative fitness* surface $w(z) = W(z)/\bar{W}$ is often more convenient than $W(z)$, and we use these two somewhat interchangeably. The *mode of selection* on a trait in a particular population is determined by the local geometry of the individual fitness surface over that part of the surface spanned by the population (for an example, see Figure 15.3). If fitness is increasing (or decreasing) over some range of trait values, a population having its mean value in this interval experiences *directional selection*. For instance, in Figure 15.3(a) a population with modest phenotypic variance, and a mean phenotype of, say, 4 units, would experience directional selection. If $W(z)$ contains a local maximum (i.e. if mean phenotype was about one unit in Figure 15.3), a population with members within that interval experiences *stabilizing selection*. If the population is distributed around a local minimum, *disruptive selection* occurs.

$W(z)$ may vary with genotypic and environmental backgrounds. In some situations (e.g. predators with search images, sexual selection, dominance hierarchies, truncation selection) the expected fitness of an individual with a given trait value depends on the distribution of traits in the population. In this case, fitness is said to be *frequency-dependent*.

Population mean fitness \bar{W} is also a function, or surface, describing the expected fitness of the population as a function of the distribution $p(z)$ of phenotypes in that population,

$$\bar{W} = \int W(z) p(z) dz.$$

Mean fitness can be thought of as a function of the parameters of the phenotypic distribution. For example, if z is normally distributed, mean fitness is a function of the mean μ_z and variance σ_z^2 for that population. Functions describing mean fitness as a function of properties of the population (i.e. mean fitness) are often referred to as *fitness landscapes*.

To stress the distinction between the $W(z)$ and \bar{W} fitness surfaces, the former is referred to as the *individual fitness surface* or function, the latter as the *mean fitness landscape*. Knowing the

individual fitness surface allows one to compute the mean fitness landscape for any specified $p(z)$, but the converse is not true.

Population mean fitness landscapes tend to be smoother than individual fitness functions. Essentially, the distribution of phenotype in a population ‘smooths’ out bumps in the individual fitness function. Thus, the distribution of phenotypes in a population can, in principle, greatly influence the evolutionary process. As we shall shortly see, selection favors evolution in the direction of increases in the population mean fitness landscape. Consider the two fitness landscapes plotted in Figure 15.3. Both of these represent how population mean fitness changes with mean phenotype, for the same individual fitness function. However, one (when $\sigma_z^2 = 2$) is monotonically increasing, and another has a local optimum (when $\sigma_z^2 = 0.1$).

Even though there is a maximum in the individual fitness function in Figure 15.3(a), whether or not this corresponds to the existence of stabilizing selection depends on the distribution of the trait. With a low phenotypic variance, there is selection for larger values below mean phenotypic values of about 0.7 units, and selection for larger values when the mean is above 0.7 units, and thus stabilizing selection, but only if the phenotypic variance is small (Figure 15.3(b),(c)). Similarly, there is disruptive selection around mean phenotype values of about 2 units (Figure 15.3(b),(c)), but again, only if the phenotypic variance is small. For large phenotypic variance, there are neither local minima nor maxima in the population mean fitness function (Figure 15.3(a)), and consequently, selection favors larger values on average, across all values of mean phenotype (Figure 15.3(b),(c)).

Just as the presence of a maximum (or minimum) in the individual fitness surface does not necessarily mean that stabilizing (or disruptive) selection occurs, selection coefficients describing non-directional selection (Section 15.9.2) cannot necessarily be related back to the mode of selection. For example, it is tempting to try to relate C and/or γ to the *mode of selection*. The terms stabilizing selection differential or gradient, which are often used to describe these selection coefficients may be slightly (or even highly) misleading, so Phillips and Arnold (1989) suggested C be referred to as the *quadratic selection differential*; *variance selection differential* (*covariance selection differential* for multivariate analysis) may be even more appropriate. Correction for the effects of directional selection is important, as claims of stabilizing selection based on a reduction in variance following selection can be due entirely to reduction in variance caused by directional selection. Similarly, disruptive selection can be masked by directional selection. Provided that selection does not act on traits phenotypically correlated with the one under study, C provides information on the nature of selection on the variance. Positive C indicates selection to increase the variance (as would occur with disruptive selection), while negative C indicates selection to reduce the variance (as would occur with stabilizing selection). As we discuss shortly, $C < 0$ ($C > 0$) is *consistent* with stabilizing (disruptive) selection, but not *sufficient*. A strictly monotonic fitness function that includes selection to reduce the variance, over and above the effect of directional selection to reduce the mean, is depicted in Figure 15.4.

15.10.2 Gradients and the Local Geometry of Fitness Surfaces

A conceptual advantage of β and γ is that they describe the average local geometry of the fitness surface. When z is normal and individual fitnesses are not frequency-dependent, β can be expressed in terms of the geometry of the *mean* fitness landscape,

$$\beta = \frac{\partial \ln \bar{W}}{\partial \mu_z} = \frac{1}{\bar{W}} \frac{\partial \bar{W}}{\partial \mu_z}, \quad (15.37)$$

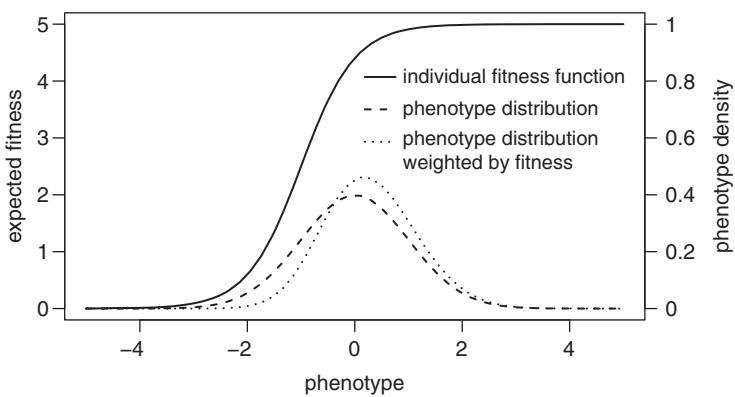


Figure 15.4 Selection affecting the variance is not necessarily indicative of stabilizing selection. The solid line depicts an individual fitness function that is strictly monotonic: fitness increases with increasing phenotype, and there is no maximum. However, there is selection on the variance: the variance of the distribution of phenotype, weighted by fitness, is lower than can be attributed to directional selection alone. This occurs because the individual fitness function is downwardly curved over the distribution of phenotype. In this example (with a trait with a mean of 0 and a variance of 1; the dashed line represents the distribution of phenotype), $C \approx -0.17$.

so that β is proportional to the slope of the \bar{W} surface with respect to population mean (Lande, 1979). β can also be expressed as a function of the *individual* fitness surface. Lande and Arnold (1983) showed, provided z is normally distributed, that

$$\beta = \int \frac{\partial w(z)}{\partial z} p(z) dz, \quad (15.38)$$

implying that β is the average slope of the individual fitness surface, the average being taken over the population being studied. Likewise, if z is normal,

$$\gamma = \int \frac{\partial^2 w(z)}{\partial z^2} p(z) dz, \quad (15.39)$$

which is the average curvature of the individual fitness surface (Lande and Arnold, 1983). Thus, β and γ provide a measure of the geometry of the individual fitness surface averaged over the population being considered.

A final advantage of β and γ is that they appear as the measures of phenotypic selection in equations describing selection response. Recall (equation (15.34)) that under the assumptions leading to the breeder's equation, $R = \Delta\mu = \sigma_A^2 \beta$. Similarly, for predicting changes in additive genetic variance under the infinitesimal model, the expected change in variance from a single generation of selection is

$$\Delta\sigma_A^2 = \frac{\sigma_A^4}{2} (\gamma - \beta^2),$$

which decomposes the change in variance into changes due to selection on the variance (γ) and changes due to directional selection (β^2). This is the contribution of selection to the disequilibrium component of the additive genetic variance (d_t , equation (15.16)). Written in terms of *selection gradients*, Bulmer's iteration (equation (15.17)) is thus

$$d_{t+1} = \frac{d_t}{2} + \frac{\sigma_A^4}{2} (\gamma - \beta^2).$$

15.11 Multivariate Selection

Having explored the theory about selection response on one trait only, we now move on to describe the consequences of selection acting on several traits simultaneously. To illuminate multivariate selection, we consider selection acting on one trait, and how evolutionary responses can occur in other, genetically correlated, traits.

15.11.1 Short-Term Changes in Means: The Multivariate Breeder's Equation

Using equation (15.15) and generalizing the regression definition of the univariate selection gradient $\beta = \frac{S}{\sigma_z^2}$ to its multiple regression analog $\beta = \mathbf{P}^{-1}\mathbf{S}$, we can express the multivariate breeders equation as

$$\begin{aligned}\mathbf{R} &= \mathbf{GP}^{-1}\mathbf{S} \\ &= \mathbf{G}\beta.\end{aligned}\tag{15.40}$$

Presented in this form, this nicely demonstrates the relationship between the two main complications with selection on multiple traits: the *within-generation* change due to *phenotypic* correlations and the *between-generation* change (response to selection) due to *additive genetic* correlations. Cheverud (1984) makes the important point that although it is often assumed a set of phenotypically correlated traits respond to selection in a coordinated fashion, this is not necessarily the case. Since β removes all the effects of phenotypic correlations and $\mathbf{R} = \mathbf{G}\beta$, phenotypic traits will only respond as a group if they are all under direct selection or if they are *genetically* correlated.

β is the vector of *partial regression coefficients* of relative fitness on traits. The regression of y on x is the quotient of covariance of x and y and the variance of x (Lynch and Walsh, 1998, Chapter 3, see especially p. 41).³ Multivariate selection gradients in the vector β thus represent the effects of each trait on fitness, holding all other traits constant. This property separates an important aspect of causation from correlation: if correlated traits have direct effects on fitness, these effects do not contribute to β for a given focal trait, so long as they are meaningfully measured and included in a multivariate selection analysis (Section 15.12). This is the divergence in meaning between β and \mathbf{S} in multivariate selection, which we noted earlier when we discussed the scaling differences between β , S , and i , in Section 15.9.1.

15.11.2 The Effects of Genetic Correlations: Direct and Correlated Responses

While the use of β removes any further evolutionary effect of phenotypic correlations, additive genetic correlations strongly influence the response to selection. If n traits are under selection, and g_{ij} is the additive genetic covariance between traits i and j , then the response in trait i to a single generation of selection is

$$R_i = \Delta\mu_i = \sum_{j=1}^n g_{ij} \beta_j = g_{ii} \beta_i + \sum_{j \neq i} g_{ij} \beta_j,\tag{15.41}$$

³ Those more familiar with ordinary least squares regression in practice will note that an estimated vector of partial regression coefficients is given by $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. This exactly reflects the definition of a vector of partial effects: $\mathbf{X}^T \mathbf{X}$ is the covariance matrix of the predictor variables, multiplied by the sample size, and $\mathbf{X}^T \mathbf{y}$ is the covariance vector of the predictors with the response, also multiplied by the sample size. In taking the quotient of these quantities, the sample size cancels out.

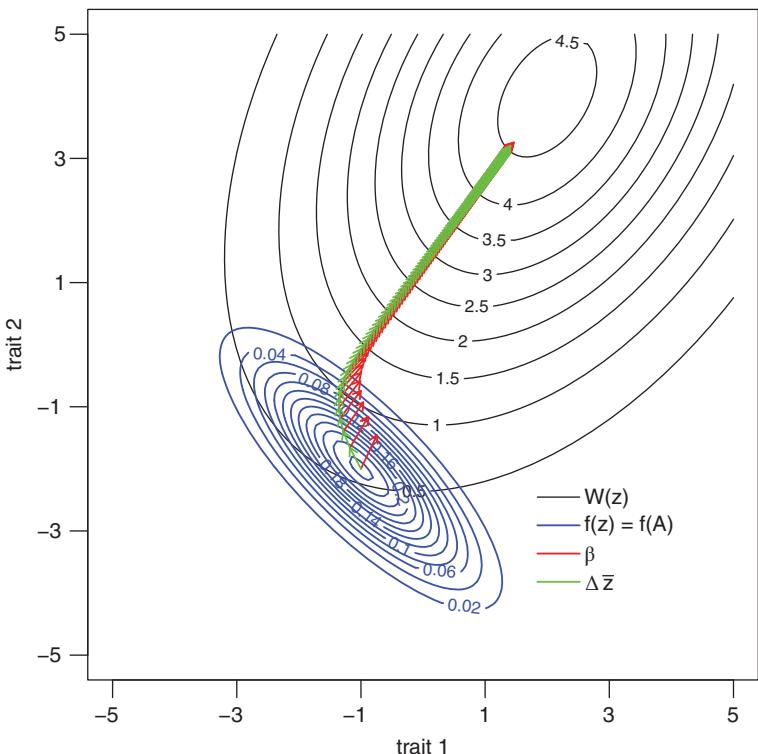


Figure 15.5 The effect of genetic correlations on the response to multivariate selection. Black contours represent a Gaussian fitness surface. Blue contours represent a bivariate normal distribution of trait values and additive genetic values (for simplicity, phenotype is assumed to be entirely determined by strictly additive genetic effects, and so the distributions of phenotype $f(z)$ and breeding values $f(A)$ are identical) in the first generation. Red arrows show the direct selection gradient β in each generation, and green arrows show the response to selection R . Initially, the effect of the genetic correlation is to deflect the response to selection away from the direction of maximally increasing population mean fitness.

so that the response has a component due to direct selection on that trait ($g_{ii} \beta_i$) and an additional component due to selection on all other genetically correlated traits.

If direct selection only occurs on trait i , a genetically correlated trait also shows a *correlated response to selection*. For example, for trait j , its response when only trait i is under selection is

$$R_j = g_{ij} \beta_i.$$

Figure 15.5 illustrates the response to multivariate selection. For simplicity, but without affecting the key results, two traits are assumed to be entirely genetically determined, and to have equal variances of 1 unit² (traits are often standardized this way in practice, following Lande and Arnold, 1983), and to have a (genetic) correlation of -0.8 . From a starting mean phenotype vector of $[-1, -2]^T$ the population evolves toward an optimal mean phenotype of $[2, 4]^T$. The fitness surface is modeled as a bivariate Gaussian function with stabilizing selection, and positive correlational selection (i.e. deviations from the optimum of one trait have less severe fitness consequences if the other trait diverges from the optimum in the same direction). The selection gradients in each generation point in the direction of most rapidly increasing population mean fitness (see Section 15.10). However, the response to selection in each generation is deflected away from β by the genetic correlation between the traits. Specifically, evolution

is initially much faster in the direction of decreasing values of trait 1 than increasing values of trait 2. If the genetic correlation were positive, then the response to direction would be biased toward simultaneously increasing or decreasing values (depending on the specific selection gradient vector) of the two traits.

15.11.3 Selection Gradients and Understanding which Traits Affect Fitness

In parallel to the literature of direct selection gradients (i.e. β in equation (15.40)), a largely parallel literature on path analysis of natural selection has developed in the past thirty years (Crespi and Brookstein, 1989; Kingsolver and Schemske, 1991; Conner, 1996; Scheiner *et al.*, 2000). It is not always clear how this literature is to be related the concept of selection gradients. In particular, proponents of path analysis have argued that partial effects of traits on fitness (direct selection gradients) cannot show what traits cause fitness variation. However, β are partial effects, that is, they show which traits directly affect fitness, accounting for correlations of traits and fitness resulting from confounding traits, and they should not be understood as reflecting all ways in which traits may materially affect fitness. It turns out that some key aspects of the two literatures have been essentially talking at cross-purposes.

Consider Figure 15.6. Assuming that all arrows represent non-zero effects of quantities on each other, all traits will covary with fitness. In other words, they will have non-zero selection differentials S (see equations (15.12) and (15.27)). Being *partial* effects, β (see Section 15.11.1) can be thought of as the (average) effects of each trait on relative fitness, holding all other traits constant. In other words, β does not represent the *causal* effect of a trait on fitness, but rather the direct component of its causal effect. Consequently, even though z_1 affects fitness, its direct selection gradient is zero. This is not an issue of estimation: this is the definition of the selection gradient for z_1 . Consequently, even though traits z_1 and z_2 are very different in the ecological interpretations we may wish to attach to them (one is materially relevant to fitness, the other is not), they are indistinguishable with respect to the main quantitative measures that are used to characterize natural selection.

The idea of path analysis is to distinguish between the types of ecological interpretations we might seek to attach to the relationships of z_1 and z_3 with fitness. According to Wright's (1921b, 1934) path rules, the ultimate effect of z_1 on fitness can be quantified (it is the product of the paths from z_1 to z_2 and from z_2 to expected fitness). Some have suggested that such a calculation could be used to generate an analog of β (Scheiner *et al.*, 2000). However, the

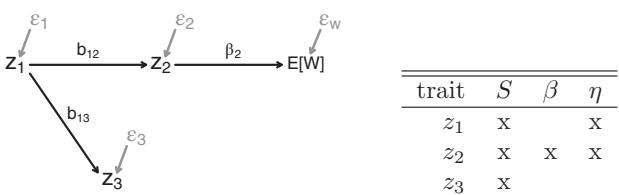


Figure 15.6 The distinction between *direct* effects on fitness, and *causal* effects on fitness, when traits have effects on one another. All traits either cause fitness variation (either directly, as in the case of z_2 , or as mediated by another trait, as for z_1), or are correlated with a trait that causes fitness variation (z_3 , which is correlated with z_1 , due to the effect of z_1 on z_3). Direct effects of traits on one another are denoted by b_{xy} , where x denotes causing and y the caused traits. β denotes the direct effect of a trait on fitness; in this example, only z_2 has a *direct* effect on fitness. The table on the right uses 'x' to denote selection coefficients that have non-zero values under the example scenario. S is the selection differential (see equations (15.33) and (15.40)), β denotes the direct selection gradient (see equations (15.12) and (15.27)), and η denotes an extended selection gradient (equation (15.44)).

directional selection gradient, as defined as an average partial effect (equation (15.38)), receives its meaning from the mechanics of the response to multivariate selection (equation (15.40)), and a path analysis-based gradient would not retain this meaning.

Gradients of a sort obtained by path analysis do have a relationship to the response to selection; it just is not via the same equation that relates direct selection gradients, β , to the statistical mechanics of evolution. If exogenous inputs to the system of traits (ϵ in Figure 15.6) can be decomposed into additive genetic (and residual) values analogously to equations (15.1) and (15.2), $\epsilon = A_\epsilon + e_\epsilon$, then the genetic covariance matrix among traits (as in Section 15.3.4) is given by

$$\mathbf{G} = \Phi \mathbf{G}_\epsilon \Phi^T, \quad (15.42)$$

where \mathbf{G}_ϵ will typically be a diagonal matrix with additive genetic exogenous variances on the diagonal, although correlation of exogenous genetic values can be accommodated. Φ is a lower triangular matrix containing the effects of each trait on all other traits, as defined by the rules of path analysis. Compactly,

$$\Phi = (\mathbf{I} - \mathbf{b})^{-1}, \quad (15.43)$$

where \mathbf{b} is a lower triangular matrix of direct effects of traits on one another. Given equations (15.43) and (15.42), an equation for evolutionary change based on the *total*, rather than *partial*, effects of traits on relative fitness can be derived by substituting equation (15.42) into the Lande equation (15.40), and noting that the total effects of traits on fitness (which we will denote by the vector η) are given by the product of their total effects on one another, and of their direct effects on relative fitness, $\eta = \Phi^T \beta$,

$$\begin{aligned} R &= \mathbf{G}\beta \\ &= \Phi \mathbf{G}_\epsilon \Phi^T \beta \\ &= \Phi \mathbf{G}_\epsilon \eta. \end{aligned} \quad (15.44)$$

This expression justifies the use of path analysis-based inferences of the *total* effects of traits on relative fitness as quantities that are similar to selection gradients. However, it also clarifies that η is a distinct quantity from β , as defined by Lande (1979; Lande and Arnold, 1983). More detailed discussion of benefits of and caveats to the use of the *extended* notion of the selection gradient encapsulated in η is given in Morrissey (2014a), which proposes the term *extended selection gradient* for η . In addition to being defined and calculated as the total direct effects of traits on relative fitness, η can be defined as the average partial derivatives of relative fitness with respect to exogenous inputs to a system of traits (i.e. to ϵ), analogously to the definition of β given in equation (15.38), which leads to a way of defining a quadratic extended selection gradient (Morrissey, 2015). An expanded illustration of the empirical calculation of extended selection gradients is given in Walsh and Lynch (2018, Chapter 30).

15.12 Inference of Selection Gradients

Inference of the form and strength of natural selection is one of the most important empirical tasks in evolutionary quantitative genetics. In this section we review some of the theory of selection gradients that is most pertinent to their empirical estimation. A greatly expanded treatment, including an up-to-date review of the literature, is given in Walsh and Lynch (2018, Chapters 29 and 30).

15.12.1 Ordinary Least Squares Analysis

Most of the empirical literature on natural selection reports selection gradients (see meta-analyses by, for example, Kingsolver *et al.*, 2001; Morrissey, 2016). The majority of selection gradient estimates have been generated using a remarkably simple and robust analysis developed by Lande and Arnold (1983). When phenotypes are multivariate normal, the selection gradient, β that appears in equations (15.34), (15.37) and (15.38) can be estimated by the ordinary least squares (OLS) regression of relative fitness on phenotype,

$$w_i = 1 + \beta(z_i - \bar{z}) + e_i, \quad (15.45)$$

where relative fitness w_i is individual fitness divided by population mean fitness, $w_i = W_i/\bar{W}$, z_i is individual phenotype and e_i are residuals. While fitness must be expressed as relative fitness (i.e. standardized to its mean), phenotype, z , may be expressed in arbitrary units, including original units (e.g. kilograms, millimeters, days), or may be standardized to the mean or standard deviation of phenotype, and it may be expressed on a logarithmic scale, as may be useful for any specific biologically motivated analysis.

The analysis is easily extended to multivariate phenotypes by applying a multiple regression version of equation (15.45),

$$w_i = 1 + \sum_{j=1}^k \beta_j(z_{ij} - \bar{z}_j) + e_i, \quad (15.46)$$

where j indexes k traits.

Lande and Arnold (1983) also developed theory relating the curvature of the relative fitness surface to changes in the covariance matrices of phenotype and breeding values. They also suggested an analysis to infer parameters describing the curvature of the fitness function that relate directly to the changes in \mathbf{P} and \mathbf{G} . *Quadratic selection gradients* γ , can be estimated by extension of equation (15.45) or (15.46),

$$w_i = \alpha + \sum_{j=1}^k \beta_j(z_{ij} - \bar{z}_j) + \sum_{j=1}^k \frac{1}{2}\gamma_j(z_{ij} - \bar{z}_j)^2 + \sum_{j=1}^k \sum_{l=j+1}^k \gamma_{jl}(z_{ij} - \bar{z}_j)(z_{il} - \bar{z}_l) + e_i. \quad (15.47)$$

γ_j and γ_{jl} are referred to jointly as *quadratic selection gradients*, but the latter may be referred to as *correlational selection gradients*. Together, they define that part of the nonlinear aspect of the fitness surface that can be explained by quadratic regression:

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 & \gamma_{12} & \gamma_{13} & \cdots \\ \gamma_{12} & \gamma_2 & \gamma_{23} & \cdots \\ \gamma_{13} & \gamma_{23} & \gamma_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

There is an extensive literature relating γ to evolutionary processes, starting with Lande and Arnold (1983). There has been substantial interest in using the geometric properties to characterize nonlinear selection, and there are extensive discussions of how such analyses may be pursued. There are many opportunities, and also many pitfalls, to the analysis of the geometry of nonlinear selection, and their discussion is beyond the scope of the present work. The reader is directed to Phillips and Arnold (1989), Blows (2007) and Morrissey (2014b).

One important caveat in the interpretation of γ must be addressed. The diagonal elements of γ , are often referred to as *stabilizing/disruptive selection gradients*, though the term *quadratic selection gradients* is also used, and is probably most appropriate. The idea is that positive

values of γ_j are indicative of disruptive selection, since they indicate that the fitness function is upwardly curved, and that negative values of γ_j are indicative of stabilizing selection, since they correspond to a downward curved fitness surface. This terminology is potentially confusing, for the same reason that the selection differential on the variance (C , see Section 15.9.2) is not directly related to the mode of selection. Positive (negative) values of γ_j for a given trait are *consistent* with the occurrence of disruptive (stabilizing) selection, but are not *sufficient*. In Figure 15.4, a purely monotonic function fitness function (with an asymptote) has negative values of γ , when the distribution of phenotype occupies the part of the fitness function that is downward curved.

15.12.1.1 A Note about the Factor of $\frac{1}{2}$ in Lande and Arnold's Analysis

The coefficient of $\frac{1}{2}$ in the $\dots + \sum_{j=1}^k \frac{1}{2} \gamma_j (z_{ij} - \bar{z}_j)^2 + \dots$ of equation (15.47) deserves special attention. If squared terms (i.e. $(z_{ij} - \bar{z}_j)^2$) are entered in a multiple regression, then the corresponding estimated regression coefficient will reflect only half of the average second derivative of relative fitness with respect to phenotype. This can be seen from taking the derivative of a term with a quadratic function $\frac{\partial^2}{\partial x^2} g \cdot x^2 = 2 \cdot g$. The least squares regression coefficient from the regression of w on $(z_{ij} - \bar{z}_j)^2$ is thus only half of the second partial derivative of the function relating relative fitness to phenotype. Dividing the covariate by 2 results in the corresponding regression coefficient being doubled, such that it reflects the (average) second partial derivative of the relative fitness function, recovering the correct definition of γ (equation (15.39)).

Most reports of γ in the literature probably represented only half of the true values, until the importance of including the factor of $\frac{1}{2}$ was clarified by Stinchcombe *et al.* (2008; note that Lande and Arnold, 1983, correctly included the factor of $\frac{1}{2}$ in their equation 16). In practice, one can correctly recover γ by dividing squared terms by $\frac{1}{2}$ before conducting multiple regression analyses. Alternatively, one can double regression coefficient estimates and their standard errors (but test statistics and p -values need not be modified), if non-halved covariates are used. Note that this correction is necessary only for quadratic gradients (in the strict sense of the diagonal elements of the γ matrix); it is not relevant to the inference of directional or correlational selection gradients.

15.12.2 Flexible Inference of Fitness Functions with Associated Selection Gradient Estimates

In light of the difficulties in relating curvature of relative fitness functions (as characterized by γ) to the mode of selection, it is desirable to consider the fitness function simultaneously with estimates of directional and quadratic selection gradients. Schlüter (1987) proposed that splines be used to characterize fitness functions. Splines allow the shape of (fitness) functions to be estimated without imposing any strong *a priori* constraints on the shape of the function. A comprehensive introduction to the use of splines (and related functions) to estimate functions in general is provided in Wood (2006).

It may be desirable to estimate the directional and quadratic selection gradients associated with any arbitrary estimated fitness function, such as that inferred by a spline-based analysis. This can be achieved by calculating the average derivative of the fitness function, where the average is taken over some estimate of the distribution of phenotype, and then scaling the average gradient to relative fitness. These calculations can thus be applied to absolute fitness

functions, as the expressions that follow include a rescaling to the relative fitness scale. Specifically, given an estimated fitness function $\hat{W}(z)$, and a sample of n individuals to characterize the distribution of phenotype, estimates of β and γ can be obtained by

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^n \frac{\partial \hat{W}(z_i)}{\partial z_j} \left(\frac{1}{n} \sum_{i=1}^n \hat{W}(z_i) \right)^{-1} \quad (15.48)$$

and

$$\hat{\gamma}_{jl} = \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \hat{W}(z_i)}{\partial z_j^2} \left(\frac{1}{n} \sum_{i=1}^n \hat{W}(z_i) \right)^{-1}, \quad (15.49)$$

where $\hat{W}'_j(z_i)$ and $\hat{W}''_{jl}(z_i)$ are the first and second partial derivatives of the estimated fitness function with respect to their subscripted quantities (the quadratic selection gradients is obtained when $j = l$), evaluated at each observed value of the mean phenotype (z_i).

This analysis is not limited to analyses with splines. It was first proposed by Janzen and Stern (1998) for obtaining directional selection gradients from logistic regression models (particularly of survival). Morrissey and Sakrejda (2013) suggested that the approach could be much more generally used, as the estimated fitness function $\hat{W}(z)$ can be very general. In addition to accommodating splines, $\hat{W}(z)$ can be characterized using *generalized* models, that is, models that can accommodate non-Gaussian error structures, which might in principle allow more efficient estimation of fitness functions and selection gradients. Morrissey and Sakrejda (2013) suggested that bootstrapping could be used to construct confidence intervals and to conduct hypothesis tests of selection gradients inferred from arbitrary fitness functions.

15.12.3 Normality and Selection Gradients

Analysis of selection gradients imposes unusual assumptions about normality on regression analyses. This issue has been a source of very substantial confusion. We conclude this chapter by addressing key sources of potential confusion in this area.

15.12.3.1 Normality of Residuals

Ordinary least squares regression, which has been the workhorse statistical procedure for inference of natural selection to date, is generally understood to assume that residuals are normally distributed. This understanding is mistaken. OLS mechanics can be derived from the Gauss–Markov theorem, and using that theorem, it can be shown that (a) OLS estimates are unbiased, assuming only that residuals have a true mean of zero (after accounting for the regression) for all values of the predictor variable, and (b) sampling variances (and thus standard errors) from OLS mechanics are correct, assuming that residuals are homoskedastic and independent (and have mean zero); these facts are illustrated by Rao (1973) and Fox (2009). Hypothesis tests and confidence intervals of OLS regression parameters require normality of residuals, but in practice this requirement will only hold for very small sample sizes – much smaller than the sample sizes used in the majority of studies of natural selection – due to the central limit theorem (Fox, 2009).

While the assumptions of residuals of fitness having a mean of zero for all values of phenotype, and being homoskedastic, are unlikely to hold in analyses of natural selection, it should not necessarily be assumed that deviation from these assumptions will be problematic for analyses of natural selection. In fact, OLS is expected to be remarkably robust to violation of these assumptions (Fox, 2009). More consideration of how influential outliers (i.e. individuals that have extreme value for both phenotype and fitness), which often occur on data sets for selection analysis, influence different approaches to selection analysis could be useful.

15.12.3.2 Normality of Phenotype and Additive Genetic Values

OLS does not typically make any assumptions about normality of predictor variables (i.e. of phenotype in the context of selection analysis). Phenotype is assumed to be multivariate normal in Lande and Arnold's (1983) OLS-based analysis of selection gradients. This should be seen as a property of selection gradient theory, not OLS mechanics. Normality of phenotype in selection gradient theory should probably be seen as a much more important requirement in selection analysis than assumptions about the distribution of residuals.

If the phenotype is skewed, then it is not possible to separate the directional and quadratic components in a quadratic regression analysis (equation (15.47); Lande and Arnold, 1983; Geyer and Shaw, 2008). Lande and Arnold (1983) provided a system for accounting for skew in separating aspects of selection pertaining to directionality versus curvature, but it has been little used.

A more important consequence of non-normality of phenotype is that it causes nonlinear components of selection to influence the evolution of mean phenotype. Since selection coefficients (differentials and gradients) derive their meaning from the places they occupy in the theory of generation-to-generation mechanics of natural selection, non-normality generates a particularly sticky situation. If the phenotype is non-normal, then the different expressions given here for selection gradients (see equation (15.33)) diverge, and some or all may not retain their relationships to the mechanics of evolution. Bonamour *et al.* (2017) has made substantial progress in clarifying the consequences of non-normality of both breeding values and phenotypes for the interpretation of selection gradients, using a quadratic fitness function model.

A general justification for the average derivative-based definition of the selection gradient (equations (15.38) and (15.39) and their operationalization via equations (15.48) and (15.49)) can be made when breeding values are (multivariate) normal, but not requiring normality of phenotype (i.e. environmental effects may be non-normal), assuming that breeding values and environmental effects are uncorrelated. Stein (1974) proved the proposition that if a random vector \mathbf{x} is multivariate normal, then the covariances of elements of \mathbf{x} with y (no assumption need be made about y) are given by

$$\Sigma(\mathbf{x}, y) = \Sigma(\mathbf{x})E\left[\frac{\partial y}{\partial \mathbf{x}}\right]. \quad (15.50)$$

Stein's lemma can give evolutionary change using the average derivative definition of the selection gradient. This will require that y represents relative fitness and that \mathbf{x} represents both additive genetic values (such that the left-hand side of equation (15.50) will represent the secondary theorem of selection equation (15.28)) and phenotype (such that $E[\partial y / \partial \mathbf{x}]$ in (15.50) will represent the selection gradient as in equation (15.38)). Note that if $z = \mu + A + E$, and A and E are uncorrelated (15.2), then $\partial z / \partial A = 1$, and so from the chain rule, $\partial w / \partial A = \partial w / \partial z$ for any fitness function. Thus, from equation (15.50), the average derivative of relative fitness, taken over any distribution of phenotype, correctly represents β , in the sense that equations (15.34) and (15.40) will correctly predict evolutionary change, if additive genetic values are normally distributed and independent of non-genetic effects.

15.13 Summary

The purpose of this chapter is twofold. First, it is intended to serve as an overview of the fundamental principles that govern the response to selection on generation-to-generation time-scales. Second, we have attempted to provide accessible guidance on some selected topics

relevant to current practice in evolutionary quantitative genetics, particularly in regard to the empirical estimation of natural selection. These mechanics, and practicalities involved in their application, are fundamental to all evolutionary quantitative genetics. However, this chapter inevitably fails to convey both the depth of understanding that exists of the fundamental principles, and also of the range of evolutionary quantitative theory that exists, for example, relating generation-to-generation processes to the fossil record and to phylogenies. This chapter is intended to serve primarily as an initial guide to, or perhaps as an executive summary of, some of the key topics contained in the much more comprehensive treatment given in Walsh and Lynch (2018).

References

- Arnold, S.J. (1986). Limits on stabilizing, disruptive, and correlational selection set by the opportunity for selection. *American Naturalist* **128**, 143–146.
- Arnold, S.J. and Wade, M.J. (1984a). On the measurement of natural and sexual selection: Theory. *Evolution* **38**, 709–719.
- Arnold, S.J. and Wade, M.J. (1984b). On the measurement of natural and sexual selection: Applications. *Evolution* **38**, 720–734.
- Blows, M.W. (2007). A tale of two matrices: Multivariate approaches in evolutionary biology. *Journal of Evolutionary Biology* **20**, 1–8.
- Bonamour, S., Teplicsky, C., Charmantier, A., Crochet, P.A. and Chevin, L.M. (2017). Selection on skewed characters and the paradox of stasis. *Evolution* **71**, 2703–2713.
- Bulmer, M.G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.
- Bulmer, M.G. (1974). Linkage disequilibrium and genetic variability. *Genetical Research* **23**, 281–289.
- Bulmer, M.G. (1976). Regressions between relatives. *Genetical Research* **28**, 199–203.
- Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, New York.
- Bürger, R. (2000). *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester.
- Careau, V., Wolak, M.E., Carter, P.A. and Garland, T. Jr. (2015). Evolution of the additive genetic variance-covariance matrix under continuous directional selection on a complex behavioural phenotype. *Proceedings of the Royal Society B* **282**, 20151119.
- Charlesworth, B. (1994). *Evolution in Age-Structured Populations* (2nd edition). Cambridge University Press, Cambridge.
- Cheverud, J.M. (1984). Quantitative genetics and developmental constraints on evolution by selection. *Journal of Theoretical Biology* **110**, 155–171.
- Clutton-Brock, T.H. (ed.) (1988). *Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems*. University of Chicago Press, Chicago.
- Clutton-Brock, T.H. and Sheldon, B.C. (2010). Individuals and populations: The role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends in Ecology and Evolution* **25**, 562–573.
- Conner, J.K. (1996). Understanding natural selection: An approach integrating selection gradients, multiplicative fitness components, and path analysis. *Ethology Ecology & Evolution* **8**, 387–397.
- Crespi, B.J. and Brookstein, F.L. (1989). A path-analytic model for the measurement of selection on morphology. *Evolution* **43**, 18–28.

- Crow, J.F. (1958). Some possibilities for measuring selection intensities in man. *Human Biology* **30**, 1–13.
- Crow, J.F. (1989). Fitness variation in natural populations. In W.G. Hill and T.F.C. Mackay (eds.), *Evolution and Animal Breeding*. CAB International, Wallingford, pp. 91–97.
- Engen, S., Lande, R. and Saether, B.-E. (2009). Reproductive value and fluctuating selection in an age-structured population. *Genetics* **183**, 629–637.
- Engen, S., Saether, B.-E., Kvalnes, T. and Jensen, H. (2012). Estimating fluctuating selection in age-structured populations. *Journal of Evolutionary Biology* **25**, 1487–1499.
- Ewens, W.J. (1989). An interpretation and proof of the Fundamental Theorem of Natural Selection. *Theoretical Population Biology* **36**, 167–180.
- Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th edition. Longman, Harlow.
- Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
- Fisher, R.A. (1930). *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford. Reprinted in 1958 by Dover Publications, New York.
- Fox, J. (2009). *A Mathematical Primer for Social Sciences*. Sage, Los Angeles.
- Geyer, C.J. and Shaw, R.G. (2008). Commentary on Lande-Arnold analysis. Technical Report No. 670. School of Statistics, University of Minnesota. <http://hdl.handle.net/11299/56218>.
- Graffen, A. (1988). On the uses of data on lifetime reproductive success. In T.H. Clutton-Brock (ed.), *Reproductive Success: Studies of Individual Variation in Contrasting Breeding Systems*. University of Chicago Press, Chicago, pp. 454–471.
- Hadfield, J.D. (2008). Estimating evolutionary parameters when viability selection is operating. *Proceedings of the Royal Society B* **275**, 723–734.
- Hadfield, J.D. (2012). The quantitative genetic theory of parental effects. In N.J. Royle, P.T. Smiseth and M. Kölliker (eds.), *The Evolution of Parental Care*. Oxford University Press, Oxford.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.
- Janzen, F.J. and Stern H.S. (1998). Logistic regression for empirical studies of multivariate selection. *Evolution* **52**, 1564–1571.
- Kingsolver, J.G. and Schemske, D.W. (1991). Analyzing selection with path analysis. *Trends in Ecology and Evolution* **6**, 276–280.
- Kingsolver, J.G., Hoekstra, H.E., Hoekstra, J.M., Berrigan, D., Vignieri, S.N., Hill, C.E., Hoang, A., Gilbert, P. and Beerli, P. (2001). The strength of phenotypic selection in natural populations. *American Naturalist* **157**, 245–261.
- Kirkpatrick, M. and Lande, R. (1989). The evolution of material characters. *Evolution* **43**, 485–503.
- Kruuk, L.E.B. (2004). Estimating genetic parameters in natural populations using the ‘animal model’. *Philosophical Transactions of the Royal Society of London, Series B* **359**(1446), 873–890.
- Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* **33**, 402–416.
- Lande, R. (1982). A quantitative genetic theory of life history evolution. *Ecology* **63**, 607–615.
- Lande, R. and Arnold, S.J. (1983). The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226.
- Lenski, E.E. and Service, P.M. (1982). The statistical analysis of population growth rates calculated from schedules for survivorship and fecundity. *Ecology* **63**, 655–662.
- Lush, J.L. (1943). *Animal Breeding Plans*, 2nd edition. Collegiate Press, Ames, IA.
- Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

- Morrissey, M.B. (2014a). Selection and evolution of causally covarying traits. *Evolution* **68**, 1748–1761.
- Morrissey, M.B. (2014b). In search of the best methods for multivariate selection analysis. *Methods in Ecology and Evolution* **5**, 1095–1109.
- Morrissey, M.B. (2015). Evolutionary quantitative genetics of nonlinear developmental systems. *Evolution* **69**, 2050–2066.
- Morrissey, M.B. (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology* **29**, 1882–1904.
- Morrissey, M.B. and Sakrejda, K. (2013). Unification of regression-based methods for the analysis of natural selection. *Evolution* **67**, 2094–2100.
- Morrissey, M.B., Kruuk, L.E.B. and Wilson, A.J. (2010). The danger of applying the breeder's equation in observational studies of natural selection. *Journal of Evolutionary Biology* **23**, 2277–2288.
- Phillips, P.C. and Arnold, S.J. (1989). Visualizing multivariate selection. *Evolution* **43**, 1209–1222.
- Postma, E. (2014). Four decades of estimating heritabilities in wild vertebrate populations: Improved methods, more data, better estimates? In A. Charmantier, D. Garant and L.E.B. Kruuk (eds.), *Quantitative Genetics in the Wild*. Oxford University Press, Oxford.
- Price, G.R. (1970). Selection and covariance. *Nature* **227**, 520–521.
- Price, G.R. (1972a). Extension of covariance selection mathematics. *Annals of Human Genetics* **35**, 485–490.
- Price, G.R. (1972b). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* **36**, 129–140.
- Prout, T. (1965). The estimation of fitness from genotypic frequencies. *Evolution* **19**, 546–551.
- Prout, T. (1969). The estimation of fitness from population data. *Genetics* **63**, 949–967.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.
- Rausher, M.D. (1992). The measurement of selection on quantitative traits: Biases due to environmental covariances between traits and fitness. *Evolution* **46**, 616–626.
- Robertson, A. (1966). A mathematical model of the culling process in dairy cattle. *Animal Production* **8**, 95–108.
- Robertson, A. (1968). The spectrum of genetic variation. In R.C. Lewontin (ed.), *Population Biology and Evolution*. Syracuse University Press, Syracuse, NY, pp. 5–16.
- Roff, D.A. (1997). *Evolutionary Quantitative Genetics*. Chapman & Hall, New York.
- Scheiner, S.M., Mitchell, R.J. and Callahan, H.S. (2000). Using path analysis to measure natural selection. *Journal of Evolutionary Biology* **13**, 423–433.
- Schluter, D. (1988). Estimating the form of natural selection on a quantitative trait. *Evolution* **42**, 849–861.
- Shaw, F.H., Shaw, R.G., Wilkinson, G.S. and Turelli, M. (1995). Changes in genetic variances and covariances: G whiz! *Evolution* **49**, 1260–1267.
- Stein, C.M. (1974). Estimation of the mean of a multivariate normal distribution. In J. Hájek (ed.), *Proceedings of the Prague Symposium on Asymptotic Statistics* 345–381. Univ. Karlova, Prague.
- Stinchcombe, J.R., Agrawal, A.F., Hohenlohe, P.A., Arnold, S.J. and Blows, M.W. (2008). Estimating nonlinear selection gradients using quadratic regression coefficients: Double or nothing? *Evolution* **62**, 2435–2440.
- Sved, J.A. (1989). The measurement of fitness in *Drosophila*. In W.G. Hill and T.F.C. Mackay (eds.), *Evolution and Animal Breeding*. CAB International, Wallingford, pp. 113–120.
- Thomson, C.E. and Hadfield, J.D. (2017). Measuring selection when parents and offspring interact. *Methods in Ecology and Evolution* **8**, 678–687.
- Travis, J. and Henrich, S. (1986). Some problems in estimating the intensity of selection through fertility differences in natural and experimental populations. *Evolution* **40**, 786–790.

- Walsh, B. and Lynch, M. (2018). *Selection and Evolution of Quantitative Traits*. Oxford University Press, Oxford.
- Willham, R.L. (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics* **19**, 18–27.
- Willham, R.L. (1972). The role of maternal effects in animal breeding: III. Biometrical aspects of maternal effects in animals. *Journal of Animal Science* **35**, 1288–1293.
- Wilson, A.J., Réale, D., Clements, M.N., Morrissey, M.B., Postma, E., Walling, C.A., Kruuk, L.E.B. and Nussey, D.H. (2010). An ecologist's guide to the animal model. *Journal of Animal Ecology* **79**, 13–26.
- Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL.
- Wright, S. (1921a). Systems of mating. I. The biometric relations between parent and offspring. *Genetics* **6**, 111–123.
- Wright, S. (1921b). Correlation and causation. *Journal of Agricultural Research* **20**, 557–585.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**, 161–215.
- Yule, G.U. (1902). Mendel's laws and their probable relation to intra-racial heredity. *New Phytologist* **1**, 193–207, 222–238.

16

Conservation Genetics

Mark Beaumont¹ and Jinliang Wang²

¹ School of Biological Sciences, Bristol University, Bristol, UK

² Institute of Zoology, Zoological Society of London, London, UK

Abstract

This chapter aims to give an overview of the statistical genetic methodology that has been motivated by and specifically developed for the conservation and management of populations. The initial focus is on methods that have been developed for the estimation of the effective population size. Statistical approaches based on samples taken at a single point in time, using genotypic and relatedness information, are first described, followed by methods based on temporally spaced samples. We then switch focus to consider statistical genetic approaches to estimate the census population size, an important complement to the estimation of effective size in population viability analyses. Next, methods for inferring aspects of population structure, gene flow, and admixture are described, from a conservation perspective. We also include a section on statistical aspects of genome-based ‘deintroduction’ strategies for ameliorating the effects of unwanted admixture. We conclude with a description of models for delimiting and discovering species using genetic information.

16.1 Introduction

The aim of conservation genetics is to uncover information about populations that can help conserve them. Statistical methods play a key role in reaching this aim: for example, in the estimation of the population size from genetic data (Wang, 2005), and helping to elucidate the connectivity of populations (Lowe and Allendorf, 2010). This information, often from small samples, can aid in monitoring a population and thus guide management, particularly to avoid extinction (Traill *et al.*, 2010). The field of conservation genetics has motivated the development of several statistical techniques, and this chapter aims to introduce some of the most important of these.

Some aspects of this chapter, those concerned with how to use genetic information to elucidate population history, may overlap with **Chapter 8**, but we have focused our description of the techniques primarily on the use of genetics to elucidate processes on recent time-scales relevant to conservation. Statistical aspects of the methods for identifying genes involved in adaptation, potentially informative from a conservation viewpoint (Pearse, 2016), are not covered in this chapter (see, for example, Hoban *et al.* 2016). Many of the statistical methods that are covered in this chapter have been developed for unlinked multi-allelic loci

(allozymes and microsatellites), but are also directly applicable to certain types of markers obtained by modern sequencing approaches (Baetscher *et al.*, 2018). In particular, although some of the examples are based on a relatively small number of (typically) unlinked microsatellite markers, most of the methods are directly applicable to micro-haplotypes obtained from restriction site associated DNA sequencing (assumptions of independence may become an issue with larger numbers of markers). We also describe some methods that are more appropriate for dense single nucleotide polymorphism (SNP) array and sequencing data when appropriate.

We first describe in some detail the methods that have been developed for estimating the effective population size, particularly important in predicting the deleterious effects of inbreeding and mutation accumulation. It is also possible, as we describe, to use genetic information to estimate the census population size, important for parameterizing stochastic population dynamic models (population viability analysis; Traill *et al.*, 2010). We then describe methods for elucidating population structure, and inferring recent gene flow and admixture. This information is useful for understanding the connectivity of populations, an important consideration in population viability analyses. Extinction through hybridization is also a concern (Wolf *et al.*, 2001), and we consider statistical aspects of the detection of admixture, and potential methods of 'deintrogession' to remove introgressed parts of the genome, thus allowing individuals from the original population to be recovered through selective breeding and conserved. Finally, we consider statistical aspects of genetic methods of species delimitation, potentially important for focusing conservation efforts.

16.2 Estimating Effective Population Size

Genetic variation (the varieties of genes and gene combinations) is the base for a population to adapt to ever-changing environments. It enables a population to evolve in response to new threats such as disease, parasites and predators, and to environmental changes such as global warming, ensuring its long-term survival and thriving. Fundamentally, genetic variation is generated by mutations and is lost by genetic drift (i.e. the random change in allele frequency in a finite population due to the fact that a generation is basically generated by sampling parental alleles from the current generation) at each generation. Other evolutionary forces, such as inbreeding, migration and selection, also impact on the loss or maintenance of genetic variation. The relative roles of these forces depend on the strength of genetic drift, which in turn depends to a large degree on the size of the population. In a large population, such that drift is weak, genetic variation is maintained at a dynamic equilibrium among evolutionary forces of mutation, drift, selection, inbreeding and migration. However, in a small population, genetic drift becomes the dominating force and as a result genetic variation is lost over time, endangering a population for survival in both the short term (due to inbreeding and inbreeding depression) and the long term (due to reduced adaptation potential).

The strength of genetic drift is determined by not only population size, but also many other factors, such as the number of breeding individuals, sex ratio, variation in reproductive success, and non-random mating (Crow and Kimura, 1970; Caballero, 1994). To summarize the impacts of all of these factors in measuring the strength of genetic drift, Wright (1931) proposed a parameter, effective population size (N_e). This is defined as the size of an idealized population (i.e. a population of monoecious diploid species with random mating including selfing, with constant population size over discrete generations, and with no selection, no migration and no mutation) which would have the same rate of genetic drift as actually observed in the population under consideration. Different real populations may have dramatically different census

sizes, mating systems (e.g. selfing and clonal reproduction, outbreeding, biparental inbreeding), inheritance modes (haploid, diploid,...) and variance in reproduction success. However, they could have the same N_e , and thus the same stochastic properties of genetic drift, leading to the same rate of loss of genetic variation. The variance effective size, N_{eV} , which measures the strength of drift, can be distinguished from the inbreeding effective size, N_{eI} , which measures the strength of the stochastic inbreeding process in a population (Crow and Kimura, 1970). However, N_{eV} and N_{eI} are identical in many cases such as a population of constant demography, and are similar in quantity in many other cases such as the average value over a certain period of time during which population size fluctuates. In this chapter, we do not distinguish N_{eI} and N_{eV} and refer to them collectively as N_e .

Because a small N_e means a high rate of loss of genetic variation, a high risk of inbreeding and inbreeding depression, a high threat to extinction due to low fitness and low adaptation potential, a population estimated to have a small N_e should be of conservation concern. Pedigree data or demographic parameters, such as census size and variance in reproductive success, can be used to calculate N_e (Caballero, 1994; Wang and Caballero, 1999). Unfortunately, except for well-studied populations such as humans, domestic animals and experimental animals, parameters describing the population pedigree and demography are difficult, if possible, to acquire. For wild populations, such data are likely to be more difficult to obtain than N_e . Fortunately, genetic drift and inbreeding leave traces of their impact on the patterns of genetic variation that can be revealed by genetic marker data. Various signals of genetic drift and inbreeding can be extracted from genetic marker data and translated into estimates of N_e (Wang, 2005).

Understandably, N_e can be defined over vastly different spatial and time-scales. It can be defined for a local population (or subpopulation), for all populations connected by migration in a region, or for the entire species. Similarly, it can be defined for the current or parental generation, for a few generations in the past, or for many generations into the past. We could call them current, recent and historical N_e , respectively. Of course, time-scales and spatial scales are closely interrelated. For example, historical N_e is highly likely to refer the species N_e because a sample of genotypes taken from a current local population may well have ancestries tracing back into ancestors many generations ago that scatter in the whole species range. For conservation purposes, usually it is the local N_e in the current time or the recent past that matters (Luikart *et al.*, 2010) and that permits the most accurate estimates in current practices (Wang *et al.*, 2016). Methods for estimating ancient or historical N_e have been reviewed elsewhere (e.g. Wang, 2005), and this section focuses on the estimation of current or recent N_e .

16.2.1 Methods Based on Heterozygosity Excess

The simplest genetic marker-based (e.g. codominant markers such as proteins and enzymes, microsatellites, SNPs) estimator of N_e relies on the estimation of heterozygosity excess. In a large random mating population at Hardy–Weinberg equilibrium, the frequency of a genotype is simply the product of the frequencies of the alleles in the genotype. This Hardy–Weinberg genotype frequency is disturbed in a small population; the observed heterozygotes are expected to be more than those calculated from observed allele frequencies by assuming Hardy–Weinberg equilibrium. The amount of proportional excess of heterozygotes is inversely proportional to the N_e of the population (Robertson, 1965; Crow and Kimura, 1970). Let us consider an idealized population in all aspects except that it consists of N_m breeding males and N_f breeding females. When N_m , N_f or both are small, males and females are expected to have slightly different allele frequencies due to genetic drift (i.e. random sampling errors in sampling $2N_m$ and $2N_f$ gametes to form the males and females, respectively). The smaller are the male and female numbers (N_m , N_f), the larger is the difference in allele frequency between sexes,

and the greater is the excess of heterozygotes in the offspring. Robertson (1965) derived the equation for the expected proportional excess of heterozygotes in offspring,

$$D = \frac{1}{8N_m} + \frac{1}{8N_f}. \quad (16.1)$$

Inserting the effective size of the population, $N_e = 4N_m N_f / (N_m + N_f)$ (see Caballero, 1994), into equation (16.1) yields

$$D = \frac{1}{2N_e}. \quad (16.2)$$

Equation (16.2) suggests that N_e can be estimated from the average heterozygosity excess observed at a number of genetic marker loci. In reality, the population D is unknown and must be estimated by using a sample of individuals drawn from it. Small sample size could also cause an excess of heterozygotes, and must be accounted for in obtaining an unbiased estimate of N_e . Such an estimator was proposed by Pudovkin *et al.* (1996),

$$\hat{N}_e = \frac{1}{2\hat{D}} + \frac{1}{2(\hat{D} + 1)}, \quad (16.3)$$

where $\hat{D} = (\hat{H}_o - \hat{H}_e)/\hat{H}_e$ estimates the proportional heterozygosity excess, with \hat{H}_o and $\hat{H}_e = 2\hat{p}(1 - \hat{p})$ the observed and expected (under Hardy–Weinberg equilibrium) heterozygosity, respectively. The estimated allele frequency \hat{p} , \hat{H}_o and \hat{H}_e are calculated from a sample of genotypes drawn from a population at a number of marker loci.

Estimator (16.3) was evaluated using simulated data (Pudovkin *et al.*, 1996), and was applied to the analyses of several empirical data sets (Luikart and Cornuet, 1999). From these and other studies, it is clear that the estimator is rather inaccurate, often giving rise to infinitely large N_e estimates for populations known to be small. This is because the signal of drift, in terms of heterozygosity excess, is rather weak, except when the actual N_e is very small. Even for a large sample taken from a population of small N_e , different markers may display highly variable degrees of heterozygosity excess. Many genetic markers (say, hundreds, depending on the actual N_e) are therefore needed to produce a reasonably precise estimate of N_e . One of the problems with this estimator is its strong assumption that the observed excess in heterozygosity comes solely from genetic drift. In real populations, however, many factors could affect heterozygosity and could easily overwhelm drift. Non-random mating, due to factors such as hidden population structure, biparental inbreeding or selfing, could cause deviation in genotype frequency from Hardy–Weinberg equilibrium globally for all markers. The naive estimator (16.3) without accounting for non-random mating will lead to grossly biased N_e estimates. Similarly, genotyping problems, such allelic dropouts and null alleles (Bonin *et al.*, 2004), could cause locus-specific heterozygosity deficit and thus an overestimate of N_e from (16.3). It is possible to quantify the effects on heterozygosity of non-random mating such as partial selfing or full-sib mating (Wang, 1996) and of genotyping problems, and to account for these effects in estimating N_e from heterozygosity excess. However, because these confounding effects can be greater than that of genetic drift and have to be estimated themselves, it is still difficult to obtain an unbiased and accurate N_e estimate from the heterozygosity excess approach.

16.2.2 Methods Based on Linkage Disequilibrium

16.2.2.1 Methods for Estimating the Present-Time N_e

Genetic drift causes association of alleles not only within a locus (i.e. heterozygosity excess) but also between loci, which is called linkage disequilibrium (LD) and which can also be used to

estimate N_e . For a detailed description of LD, see **Chapter 2**; here we will only describe the concept in enough detail to be able to explain LD-based methods for estimating N_e . Suppose the gamete possessing allele A at one locus and allele B at another locus has frequency p_{AB} in a population. When the population is large, randomly mating and not affected by selection and migration, then alleles at different loci are expected to be independent, and thus $p_{AB} = p_A p_B$, where p_A and p_B are the population frequencies of alleles A and B , respectively. In a small population with genetic drift, p_{AB} will deviate from $p_A p_B$, although on expectation we have $E[p_{AB}] = p_A p_B$. The deviation, $D_{AB} = p_{AB} - p_A p_B$, or its standardized form of the correlation between p_A and p_B , $r_{AB} = D_{AB}/(p_A(1-p_A)p_B(1-p_B))^{1/2}$, measures the extent of LD (Hill and Robertson, 1968). For two loci unaffected by selection in an isolated random mating population for a sufficiently long period of time, r_{AB} will attain an equilibrium distribution with mean $E[r_{AB}] = 0$ and variance approximately

$$V(r_{AB}) = \frac{(1-c)^2 + c^2}{2N_e c (2-c)} \quad (16.4)$$

(Weir and Hill, 1980), where c is the recombination rate between the two loci. Equation (16.4) shows that the distribution of r_{AB} becomes increasingly dispersed with decreasing values of both N_e and c . It suggests that, given a set of markers with known recombination rates among them, N_e can be estimated by employing the observed variance of pairwise r_{AB} values.

In practice, r_{AB} and $V(r_{AB})$ must be estimated from a sample of individuals drawn from the population. Sampling also contributes to $V(r_{AB})$. The contribution of a sample of n diploid individuals to $V(r_{AB})$ is roughly $1/n$ (Hill, 1981), thus

$$V(\hat{r}_{AB}) = E(\hat{r}_{AB}^2) = \frac{(1-c)^2 + c^2}{2N_e c (2-c)} + \frac{1}{n}. \quad (16.5)$$

Hill (1981) also derived the approximate variance–covariance matrix of \hat{r}_{AB}^2 among pairs of loci, which was used in a weighting scheme to derive a multilocus estimator of N_e . For a set of L markers, there are $k = L(L-1)/2$ pairs of loci, with the i th ($i = 1, 2, \dots, k$) pair having correlation \hat{r}_i , recombination fraction \hat{c}_i , and sample size n_i . The multilocus N_e estimator is

$$\frac{1}{\hat{N}_e} = \frac{\sum_{i=1}^k \gamma_i (\hat{r}_i^2 - 1/n_i) / (\gamma_i/\hat{N}_e + 1/n_i)^2}{\sum_{i=1}^k (1/\hat{N}_e + 1/(\gamma_i n_i))^{-2}}, \quad (16.6)$$

with approximate variance

$$V\left(\frac{1}{\hat{N}_e}\right) = \frac{2}{\sum_{i=1}^k \left(\frac{1}{\hat{N}_e} + \frac{1}{\gamma_i n_i}\right)^{-2}} \quad (16.7)$$

where $\gamma_i = ((1-c_i)^2 + c_i^2)/(2c_i(2-c_i))$. The variance (16.7) can be used to obtain uncertainties of estimator (16.6). Hill (1981) demonstrated, using a small *Drosophila* data set (genotypes of 198 flies at six enzyme loci) in an experimental population with roughly known N_e , that the multilocus estimator yielded reasonable N_e estimates.

Waples (2006) examined the accuracy of estimator (16.6) by simulations, for the case of a set of unlinked markers. He found that, despite being corrected for sample size n , the estimator still underestimates N_e substantially when n is small. Using simulations, he obtained empirical equations (functions of n) to re-correct sampling effects, and showed essentially unbiased N_e estimates can be obtained using (16.6) and his empirical re-correction equations. Waples and Do (2008) implemented these formulations in a computer program, LDNE, which has made

the LD methods popular among molecular ecologists and conservation biologists (Palstra and Ruzzante, 2008; Luikart *et al.*, 2010). This is because the LD method is simple, yet it is much more robust and accurate than heterozygosity excess methods. The data it requires, the individual genotypes at a small number of codominant genetic marker loci (e.g. microsatellites and SNPs) of a single sample of individuals taken at random from the population, are nowadays easily obtained from wild populations. Use of genomic SNP data could improve N_e estimation precision substantially (Waples *et al.*, 2016), but could also cause a bias due to marker linkage (i.e. limited genome size; see below).

It is worth mentioning some critical assumptions made in deriving the original LD estimators of N_e . Blindly applying the methods could lead to poor estimates when the assumptions are violated. One such assumption is an isolated random mating population. If the population is not isolated but receives immigrants regularly or irregularly, the observed LD would come from both drift and admixture (hybridization). If immigration is ignored, biased N_e estimates would be inevitable, as verified by simulations (Waples and England, 2011). When individual genotypes are sampled from a single subpopulation to estimate the local subpopulation N_e , an overestimate and an underestimate result when immigration into the subpopulation is high and low, respectively (Waples and England, 2011). In the former case, the estimated N_e approaches that of the entire meta-population as immigration increases. This is not surprising because, at a high migration rate, the meta-population structure dissolves and genetically it becomes homogenous. Therefore, individuals drawn locally (from within a subpopulation) still represent the entire population.

It is difficult to accommodate immigration in the LD estimator of N_e , even when the immigration rate is known. One may consider identifying and removing immigrants and their hybrid offspring before calculating the LD estimator. However, current methods available (e.g. Rannala and Mountain, 1997; Pritchard *et al.*, 2000; discussed in Section 16.4 below) for identifying immigrants and hybrids from multilocus genotype data have limited power. They are reliable when subpopulations are highly differentiated, genetic marker information is sufficient, and the hybrids are of recent hybrid ancestry (say, in the past two or three generations). It is increasingly difficult to identify a hybrid from an immigrant ancestor occurring at an increasing number of generations into the past. Unfortunately, however, these ancient hybridization events still have impacts on LD and thus the estimates of N_e .

The genetic structure generated by non-random mating (e.g. biparental inbreeding or selfing) or by age structure for populations with overlapping generations could also cause problems of the LD estimator. Using simulations, Waples *et al.* (2014) found that LD calculated from mixed-age adult samples is overestimated and thus N_e is underestimated in species with different simulated life tables. How to deal with the genetic structure of a population caused by overlapping generations, non-random mating, subdivision and migration in LD-based estimation of N_e is an urgent issue that deserves special attention in future studies.

16.2.2.2 Methods for Estimating N_e in the Recent Past

More closely linked loci (i.e. smaller recombination rate c) are expected to have a higher LD, and to provide more information about N_e (Hill, 1981). Although the formulations (16.5) and (16.6) apply to both linked ($c < 0.5$) and unlinked ($c = 0.5$) loci, they have been applied mainly to unlinked microsatellites in practice, thanks to work by Waples and colleagues. With the rapidly decreasing cost and increasing ease of DNA sequencing, SNPs in the thousands or more are being collected and analysed regularly by ecologists and conservation and evolutionary biologists. Such SNP loci are linked to different extents, depending on the locus density. Using the linkage and genotype information of the genomic loci could substantially improve LD

estimators. Linkage disequilibria of pairs of loci with different recombination rates shed light on the drift that occurred in a population at different time points in the past, allowing the inference of the N_e trajectory (i.e. N_e as a function of generation t back into the past) that the population has undergone from present-time samples. The observed LD between loci separated by large distances along the chromosome, or more appropriately by large genetic distances, reflects relatively recent N_e , whereas LD over short recombination distances depends on relatively ancient N_e (Hayes *et al.*, 2003). A number of methods (Hayes *et al.*, 2003; Burren *et al.*, 2014; Barbato *et al.*, 2015; Mezzavilla and Ghirotto, 2015; Saura *et al.*, 2015; Hollenbeck *et al.*, 2016) have been developed to exploit the rich LD information in data of densely spaced markers for inferring the N_e at different time points in the past. These estimators simply use the average of \hat{r}_{AB}^2 in a bin with average recombination rate c to estimate the effective population size $\sim 1/(2c)$ generations ago (Hayes *et al.*, 2003; Mezzavilla and Ghirotto, 2015; Barbato *et al.*, 2015). They are implemented in software packages, such as NeON (Mezzavilla and Ghirotto, 2015) and SNeP (Barbato *et al.*, 2015), to facilitate the applications. Analyses of some empirical data sets, such as sheep and cattle SNP data (Barbato *et al.*, 2015) and human genome-wide SNP data (Mezzavilla and Ghirotto, 2015), produced sensible results. However, several issues remain with genomic LD-based N_e estimation methods.

First, despite using the same approach, different implementations have been adopted in calculating $E(\hat{r}_{AB}^2) = V(\hat{r}_{AB})$, in accounting for the sampling effect, and in binning marker pairs. For the latter, a variety of more or less arbitrary binning schemes were employed, such as binning for generation classes in the past (Corbin *et al.*, 2012; Hollenbeck *et al.*, 2016), for distance classes with a constant range for each bin (Kijas *et al.*, 2012), and for distance classes in a linear fashion but with larger bins for pairs of loci with higher genetic distances (Burren *et al.*, 2014). It is unclear which binning scheme is the most appropriate, and whether the best scheme varies with sample (e.g. SNP density and sample size) and population (e.g. the actual N_e trajectory) properties. The sampling effect on $E(\hat{r}_{AB}^2)$ is quantified as $1/(\beta n)$ where n is the number of sampled individuals and $\beta = 1$ and $\beta = 2$ when the gametic phase is unknown and known, respectively (Hill, 1981; Barbato *et al.*, 2015). However, Sved *et al.* (2013) showed that the sampling contribution is $1/(2n - 1)$ rather than $1/(2n)$ for the case of known phase. The difference between the two corrections for sample size might be small for tightly linked markers and for large sample size n . However, it could lead to biased estimates of $E(\hat{r}_{AB}^2)$ in (16.5) and thus biased estimates of N_e for the case of unlinked or loosely linked loci and small sample size. In brief, it is unclear which of these different implementations of the same approach works best, and why.

Second, \hat{r}_{AB}^2 observed in a sample for loci with any c value is due to the cumulative results of genetic drift and recombination over potentially many generations before the sampling point, and thus has information for N_e at different time points in the past. Hence, since current estimators simply use the average of \hat{r}_{AB}^2 in a bin with average recombination rate c to estimate the effective population size $\sim 1/(2c)$ generations ago, estimation of N_e trajectories from linked markers can potentially be substantially improved by using this information.

16.2.3 Methods Based on Relatedness

A direct effect of a small effective population size is that individuals in the population become increasingly related and inbred. Indeed, the rate of increase in average coancestry between genes within (due to inbreeding within) and between (due to relatedness of) individuals is inversely proportional to N_e (Crow and Kimura, 1970). Thus, estimating this rate of increase in identity by descent of genes at genetic marker loci could lead to an estimate of the current effective population size.

It was shown (Crow and Denniston, 1988; Caballero, 1994) that a diploid population composed of N_m males and N_f females at each discrete generation has effective population size

$$\frac{1}{N_e} = \sum_{s=m,f} \frac{1}{16N_s} \left[\left(\frac{1}{\mu_{sm}} + \frac{1}{\mu_{sf}} \right) (1 - \alpha) + \left(\frac{\sigma_{sm}^2}{\mu_{sm}^2} + \frac{2\sigma_{sm,sf}}{\mu_{sm}\mu_{sf}} + \frac{\sigma_{sf}^2}{\mu_{sf}^2} \right) (1 + 3\alpha) \right], \quad (16.8)$$

where $\mu_{sr} = N_r/N_s$ is the mean and σ_{sr}^2 is the variance of the numbers of offspring of sex r for a parent of sex s ($s,r = m$ for males and f for females), respectively, $\sigma_{sm,sf}$ is the covariance between the number of male offspring and the number of female offspring from a parent of sex s , and α measures the deviation from Hardy–Weinberg equilibrium (equivalent to F_{IS} ; Wright, 1969). Wang (2009) showed that these variances and covariances of reproductive successes can be transformed into the probabilities that two offspring (both males, both females, one male and one female) come from the same parent (i.e. siblings). These probabilities can be further expressed in terms of the frequencies that two offspring taken at random from the population are full siblings, Q_{FS} , and half siblings, Q_{HS} . Equation (16.8) then becomes

$$\frac{1}{N_e} = \frac{1+3\alpha}{4}(Q_{HS} + 2Q_{FS}) - \frac{\alpha}{2} \left(\frac{1}{N_m} + \frac{1}{N_f} \right). \quad (16.9)$$

Equation (16.9) shows the functional relationship between sibship frequencies and the N_e of the parental population, and enables the estimation of N_e from a single sample of multilocus genotypes. All quantities in (16.9) are estimable from the genotypes of a sample of individuals taken at random from the population. Q_{HS} and Q_{FS} can be estimated from a maximum likelihood sibship assignment analysis (Wang and Santure, 2009; Jones and Wang, 2010), which essentially reconstructs a two-generation pedigree of the sampled individuals from their genetic marker data. α can be estimated separately from the same data with an F_{ST} -like approach (Wang, 2009), or is simply assumed to be 0 for an outbred population when marker genotype frequencies do not deviate significantly from Hardy–Weinberg equilibrium. The difficulty comes from the estimation of the numbers of breeding males, N_m , and females, N_f . Except for some special cases such as haplodiploid species or different male and female mating systems (e.g. male monogamy and female polygamy), paternal and maternal sibship cannot be distinguished in a sibship assignment analysis using autosomal marker data (Wang and Santure, 2009), and as a result N_m and N_f cannot be individually estimated reliably. However, as argued in Wang (2009), the analysis still yields a reasonably good estimate of the composite quantity $1/N_m + 1/N_f$. Furthermore, the quantity enters into the equation as a product with α , which is usually small. Therefore, the quantity $1/N_m + 1/N_f$ still has little effect on the estimated N_e even when it is not very well estimated.

The sibship frequency approach to N_e estimation is conceptually analogous to the capture–mark–recapture (CMR) approach (Pradel, 1996) to population census size estimation used widely in animal ecology. In the CMR approach, a random sample of individuals from the population is captured, marked, and then released to mix with uncaptured individuals. Later, another random sample of individuals is captured, and the proportion of marked individuals within the sample is the recapture rate. The total population size can be estimated by dividing the number of marked individuals (i.e. the size of the first sample) by the recapture rate. Despite the conceptual similarity, genetic CMR differs from traditional CMR in several important aspects. First, the capture or recapture unit is not an individual as in the traditional CMR approach, but a sib family. When one offspring from a sib family is included in the sample, the family is marked and ‘captured’. When another one or more offspring from the same family are included in the sample and are detected as such by a sibship assignment analysis, then the family is ‘recaptured’.

The frequency of sibling pairs in a sample is equivalent to recapture rate, and is inversely proportional to N_e . Second, the ‘mark’ used is not an artificial one, such as a ring on the leg or wing, but the natural, high-resolution, and permanent DNA markers. Hence, the capture or recapture does not necessarily involve physically handling the animals; DNA samples can be obtained non-invasively (e.g. from faeces, feathers or shed skins) even without seeing the animals, and analysed for N_e estimation (Luikart *et al.*, 2010). Third, a single sample of individuals (or more precisely multilocus genotypes) acquired at a single time point is sufficient to capture and recapture sib families and thus to yield an N_e estimate of the population.

Compared with other N_e estimation approaches, the sibship frequency approach explicitly accounts for non-random mating as quantified by α in equations (16.8) and (16.9). Close relative mating, such as partial selfing or biparental inbreeding, which results in a positive α , usually leads to a reduced N_e if unaccounted for. Using α to account for non-random mating, the sibship frequency approach can yield unbiased N_e estimates while other approaches which ignore non-random mating result in underestimated N_e . Similarly, immigration either has no effect or has some effect that can be quantified and corrected for in the sibship frequency approach. When a sample of individuals is drawn from a subpopulation before immigration, the sibship frequency approach is unaffected. When it is drawn after immigration such that it could contain first-generation immigrants, the multilocus genotypes can be used to identify and then remove the immigrants before conducting sibship assignment and N_e analysis (for how to identify these, see Section 16.4.2). Immigration and hybridization that happened in the more remote past, before the parental generation, has no effect on the sibship method for estimating the N_e of a subpopulation at the parental generation. Including individuals whose parents are immigrants or F1 hybrids in a sample, for example, has no effect on sibship-based estimates of N_e of the focal subpopulation at the parental generation. In contrast, it affects LD and thus the estimates of N_e based on LD.

A unique desirable property of the sibship estimator of N_e is that it provides not only an estimate of N_e , but also some information about the variances and covariance of reproductive successes. This is important for conservation (Wang *et al.*, 2016), because effective management methods aimed at maintaining the genetic diversity of an endangered species rely on detailed demographic information (such as variance in reproductive success, census size, sex ratio). The composite parameter N_e allows for the explanation of the current genetic variation, and for the prediction of future trends. However, without external information, it lends no insight into the demography of the population. A small N_e , for example, can be due to many possible causes, such as biased sex ratio, fluctuation or bottleneck of census size, inbreeding, high variance in reproductive success (due perhaps to social structure, reproductive dominance). Without knowing the exact cause or causes, it is impossible to devise an effective management plan to reverse the situation. The information about the variance in reproductive success provided by a sibship assignment analysis is invaluable for the informed conservation management of endangered species.

Formulations similar to (16.9) for an outbred dioecious diploid species have also been proposed for haplodiploid species (Wang, 2009) and monoecious species with partial selfing (Wang, 2016a). For a population with overlapping generations, marker-based parentage analysis of a sample of individuals of mixed sex and age classes can yield information about the life table (i.e. age-specific reproductive rates and age-specific survival rates) and thus the N_e and generation interval of a population (Wang *et al.*, 2010). It is also conceivable that this genetic CMR approach can be further extended to use more remote relatives to estimate N_e in generations further into the past. This is especially desirable for large populations from which recapturing a sib family is difficult. If no sib family is recaptured from a large random mating population ($\alpha = 0$), then $Q_{HS} + 2Q_{FS} = 0$ and equation (16.9) suggests an infinitely large N_e . For a

given realistic sample size (say, a few hundred), it becomes more difficult to recapture a sib family with an increasing N_e . However, such a sample has a much higher probability of containing more remote relatives, such as cousins, whose frequency indicates the N_e of the grandparent generation. Unfortunately, remote relatives are difficult to infer reliably and to use in N_e estimation. While there are only two sibling types (half and full siblings sharing a single parent and both parents, respectively), there are quite a few different cousin types sharing between two and seven grandparents. Differentiating these types and distinguishing them from closely competing relationships (e.g. unrelated) is a formidable task. With genomic data having thousands of SNPs, it is now realistic (Sun *et al.*, 2016) to infer remote relatives. How well the approach performs in comparison with other methods requires further work.

Densely spaced genomic data (SNPs and DNA sequences) can also be used to identify segments of a chromosome that are identical by descent (IBD) (e.g. Browning and Browning, 2015; **Chapter 20**, this volume). These IBD segments come from the same common ancestor in the recent past without mutations and recombination. The frequency and size distributions of IBD segments shed light on both time and extent of genetic drift (N_e) in the past, allowing for the estimation of N_e trajectories in the past few hundred generations. For example, the presence of large and small IBD segments signifies recent and historical drift, while the frequencies of large and small IBD segments signify the strength of drift (N_e) in recent and remote past generations. Essentially, this IBD segment approach can be viewed as a genetic CMR method as well, capturing and recapturing families in recent (represented by large IBD segments) and remote past (represented by small IBD segments) generations. Several models (Harris and Nielsen, 2013; Palamara *et al.*, 2012) identify and use IBD segments of length greater than 1 cM or 100 bp to estimate population demography in the past. They fit one or multiple periods of exponential growth or contraction models to the IBD segments data, and choose the N_e trajectory of the maximal likelihood as the best estimate. As pointed out by Browning and Browning (2015), these parametric approaches have difficulties in modelling complex or unanticipated features of population demography because the user must pre-specify the class of models to be considered, and computational and statistical constraints limit the number of parameters that can be considered. Browning and Browning (2015) proposed a nonparametric method for accurately estimating recent effective population size by using inferred long IBD segments. They showed that their method works well for simulated data and several empirical data sets.

16.2.4 Methods Based on Temporal Changes in Allele Frequency

A genetically large population (i.e. N_e large) will show little allele frequency changes at a neutral locus over the long term due to an equilibrium of various evolutionary forces (e.g. mutations), and over the very short term (e.g. a couple of generations) because drift is negligibly small. This is true regardless of the inheritance patterns (e.g. ploidy levels), mating system (e.g. partial inbreeding such as selfing), and population structure (e.g. subdivision and migration). A small population (i.e. N_e small), however, will show random allele frequency changes at a locus over both the short and long term. This stochastic process is dominated by genetic drift, and will lead to the fixation of one allele and the loss of the other alleles at a locus. Measuring the extent of temporal changes in allele frequencies of neutral genetic markers provides a method to quantify the strength of genetic drift and thus N_e . Based on this logic, Krimbas and Tsakas (1971) proposed to use genetic marker genotype data of two temporally separated samples of individuals in measuring allele frequency changes and in estimating the N_e of the population during the sampling interval. This so-called ‘temporal method’ was subsequently refined by many others, resulting in quite a few (allele frequency) moment estimators (e.g. Nei and Tajima, 1981; Pollak, 1983; Waples, 1989).

These moment estimators measure the standardized variance, F , in temporal changes of allele frequency at a number of marker loci, and link it to causal factors, N_e and sample size S in estimating N_e . Several measures of F are available, with some subtle differences resulting in differing sampling variances and sensitivities to rare alleles. Nei and Tajima's single-locus F estimator is

$$\hat{F} = \frac{1}{k} \sum_{i=1}^k \frac{(x_i - y_i)^2}{(x_i + y_i)/2 - x_i y_i}, \quad (16.10)$$

where x_i and y_i are the observed frequencies of allele i at a locus with k alleles in the first and second samples, respectively. Multiple markers provide independent \hat{F} estimates, which are averaged in estimating N_e .

Both the genetic drift that occurs in the sampling interval from generations 0 and t (when the first and second samples are taken) and the small sizes of samples taken at 0 and t are expected to contribute to \hat{F} . The sampling contribution is quantified by $1/(2S_0) + 1/(2S_t)$, where S_0 and S_t are the sample sizes (i.e. number of diploid individuals) at sampling points 0 and t , respectively. The drift contribution is a function of average N_e during the sampling interval, and the length of the sampling interval, proportional to t/N_e . The exact amount of drift contribution depends on the sampling scheme (Nei and Tajima, 1981; Waples, 1989), whether individuals are drawn effectively with replacement (e.g. when population census size is much larger than N_e , or sampled individuals are returned to the population) or not. In the former case, which was called sampling scheme II (Waples, 1989), the expected value of \hat{F} is

$$E[\hat{F}] = \frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N_e},$$

and thus N_e is estimated as (Nei and Tajima, 1981)

$$\hat{N}_e = \frac{t}{2\left(\hat{F} - \frac{1}{2S_0} - \frac{1}{2S_t}\right)}. \quad (16.11)$$

In the latter case, which was called sampling scheme I (Waples, 1989), estimator (16.11) with the numerator t replaced by $t - 2$ gives \hat{N}_e .

These simple moment estimators work well, as verified by simulations and application to empirical data sets (Nei and Tajima, 1981; Waples, 1989). One of their problems is the bias and low precision when sampling interval t is short, say $t = 1$ or 2 (Nei and Tajima, 1981). This is because drift signal is rather weak compared with sampling when t is small, and because some approximations made in deriving estimators like (16.11) do not work well. Another problem is caused by rare alleles, which could again lead to biased and low-precision estimates. This is understood by considering a simple example. Suppose a small population has a rare allele at the initial sampling point, and the allele is sampled. However, the allele drifts and is lost from the population before the second sample at time t is taken. Of course, the second sample would not include any copies of the allele. The moment estimator would give an overestimate of N_e from information of this allele, because the allele would be counted as lost at the second sampling generation, while actually it is lost earlier. The likelihood methods do not suffer from such a problem because they naturally weigh information from different loci and different alleles within a locus. A rare allele observed in only one sample gets a smaller weight than a common allele observed in both samples at times 0 and t . The moment estimators of Jorde and Ryman (2007) for \hat{F} and \hat{N}_e are less biased by rare alleles but have a higher variance than those of Nei and Tajima (1981) presented above, compromising their overall accuracy.

To overcome the difficulties of the moment estimators, likelihood or coalescent Bayesian methods (e.g. Williamson and Slatkin, 1999; Anderson *et al.*, 2000; Wang, 2001; Berthier *et al.*, 2002; Beaumont, 2003; Laval *et al.*, 2003) have been proposed to tackle the temporal data for N_e estimation. For a di-allelic locus in a diploid population of effective population size N (note, subscript e is dropped for simplicity and clarity in this section), the probability of obtaining the temporal genotype data is (Williamson and Slatkin, 1999)

$$P(\mathbf{n}_0, \mathbf{n}_t | N) = \sum_{\mathbf{p}_0, \mathbf{p}_t} P(\mathbf{n}_0 | \mathbf{p}_0) P(\mathbf{n}_t | \mathbf{p}_t) P(\mathbf{p}_t | \mathbf{p}_0, N) P(\mathbf{p}_0 | N), \quad (16.12)$$

where \mathbf{n}_0 and \mathbf{n}_t are vectors of counts of each of the different possible alleles in samples obtained at time 0 and t (in generations) respectively, $P(\mathbf{n}_0 | \mathbf{p}_0)$ is the probability of observing \mathbf{n}_0 given the vector of the (unknown) population allele frequencies \mathbf{p}_0 at time 0, $P(\mathbf{n}_t | \mathbf{p}_t)$ is the probability of observing \mathbf{n}_t given the vector of the (unknown) population allele frequencies \mathbf{p}_t at time t , $P(\mathbf{p}_t | \mathbf{p}_0, N)$ is the probability of \mathbf{p}_t conditional on \mathbf{p}_0 and N , and $P(\mathbf{p}_0 | N)$ is the probability of \mathbf{p}_0 given N . The probability of obtaining allele count data given N , $P(\mathbf{n}_0, \mathbf{n}_t | N)$, is the likelihood of N . Maximizing $P(\mathbf{n}_0, \mathbf{n}_t | N)$ leads to a maximum likelihood estimate of N .

In equation (16.12), population allele frequencies \mathbf{p}_0 and \mathbf{p}_t are unknown and of no interest. However, they are necessary for calculating the probability of a sample allele count configuration. Given \mathbf{p}_0 and \mathbf{p}_t , \mathbf{n}_0 and \mathbf{n}_t are independently multinomially distributed. $P(\mathbf{p}_t | \mathbf{p}_0, N)$ can be calculated by the standard Wright–Fisher transition matrix (see Ewens, 2004) whose element m_{ij} is the multinomial probability of moving (drifting) from state i to j , where $i, j = 0, 1, \dots, 2N$ are the count of an allele at the di-allelic locus in the population. $P(\mathbf{p}_0 | N)$ is unknown and Williamson and Slatkin (1999) assumed an equal probability for each of the $2N + 1$ configurations (i.e. $\mathbf{p}_0 = \{p_i = 1/(2N + 1)\}$ for $i = 0, 1, \dots, 2N$).

Williamson and Slatkin (1999) showed that equation (16.12) yields good estimates of N_e from simulated and real data. An advantage of this likelihood method compared with moment methods is that it can easily handle more than two temporal samples, either to get an average N_e estimate for the entire sampling period or to fit a simple growth model (e.g. exponential) for estimating an N_e trajectory. However, the method assumes di-allelic loci for computational feasibility, which means it cannot be applied to multi-allelic genetic markers such as microsatellites and some protein markers. The number of allele frequencies to be integrated out in calculating (16.12) is $(2N + k - 1)! / ((2N)!(k - 1)!)$ for a k -allele locus (Wang, 2001). It increases very rapidly with k , being $2N + 1$, $(N + 1)(2N + 1)$, and $(N + 1)(2N + 1)(2N + 3)/3$ when $k = 2, 3, 4$, respectively. The computation burden scales as N^2 , N^4 and N^6 for $k = 2, 3$ and 4. To overcome the computational difficulty for multi-allelic ($k > 2$) markers, Anderson *et al.* (2000) proposed an importance sampling approach to calculating the likelihood for multi-allelic loci. However, the computation load is still rather high and the Monte Carlo method introduces some sampling errors (for more details about importance sampling, see Chapter 1, this volume).

The likelihood computation for multi-allelic markers was made feasible and practical by the work of Wang (2001). He made several improvements on the calculation of (16.12), the most important ones being the conversion of a k -allele locus to k di-allelic loci and the approximation of the transition matrix for calculating $P(\mathbf{p}_t | \mathbf{p}_0, N)$. For the former, the i th converted di-allelic locus has allele i and a pooled allele \bar{i} , which consists of all of the k alleles except for allele i . To account for the dependency of alleles within a locus (i.e. $\sum_{i=1}^k p_i \equiv 1$), a weight of $(k - 1)/k$ is applied to each of the converted di-allelic loci from a k -allele locus (Wang, 2001). For the latter, Wang (2001) noticed that off-diagonal elements further away from the diagonal in the square $((2N + 1) \times (2N + 1))$ transitional matrix make smaller contributions in calculating $P(\mathbf{p}_t | \mathbf{p}_0, N)$. To a good approximation but with a great improvement in computational efficiency and a mass reduction in computer storage requirements, only diagonal and close to diagonal

elements are calculated, stored and used in computing $P(\mathbf{p}_t | \mathbf{p}_0, N)$. Wang (2001) showed that a tri-allelic locus has the highest dependency among allele frequencies. Yet his approximation works very well, yielding N_e estimates with precisions and accuracies almost indistinguishable from those obtained from the exact likelihood calculation method using a three-allele transition matrix.

These computational improvements made it possible to apply the maximum likelihood method to the analysis of large data sets (with many loci and many alleles per locus) drawn from populations of large N_e . The first extensive simulations were made by Wang (2001) to compare moment and likelihood estimators. It was shown, over many different scenarios of population (e.g. N_e) and sample (e.g. number of loci, number and frequency distribution of alleles) properties, that the likelihood method yields more accurate results. The improvements are substantial only when sampling interval (t) is short, or when rare alleles are abundant.

Alternative coalescent-based likelihood methods have also been developed to analyse data of multi-allelic markers (Berthier *et al.*, 2002; Beaumont, 2003; Anderson, 2005). These assume that the changes in gene frequencies can be accurately described by the diffusion approximation, or that no more than two coalescent events occur at any single generation. As a result, the methods could lead to biased N_e estimates for very small populations.

The above moment and maximum likelihood estimators assume the traditional model on temporal allele frequency changes: a single isolated population with discrete non-overlapping generations without selection. The model has been extended in several directions. First, many long-lived species have overlapping generations. The temporal approach is robust to the age structure in a population with overlapping generations when the sampling interval t is large (Nei and Tajima, 1981). When t is small, however, Waples and Yokota (2007) showed by simulations that various sampling regimes (sampling only newborns, only adults and all age classes in proportions) invariably result in biased N_e estimates. Jorde and Ryman (1995) developed a moment N_e estimator applicable to populations with overlapping generations. However, their method relies on numerous age-specific survival rates and age-specific reproduction rates of the focal population, which are usually unknown.

Second, the assumption of an isolated population may not be tenable in reality, especially over long sampling intervals. When immigration into the focal population is present during the sampling interval but is ignored, either an over- or underestimate of N_e results (Wang and Whitlock, 2003), depending on the sampling interval t and migration rate m . Wang and Whitlock (2003) developed both moment and maximum likelihood methods to estimate N_e and m jointly, using temporal data from the focal and source populations. They showed that both N_e and m can be estimated reasonably well given sufficient temporal genotype data. The limitation of their models is that they assume either a single large source population, or a single small source population. A more general model with $n > 2$ subpopulations providing and receiving migrants is yet to be developed. Such a model has n effective sizes and $n(n - 1)$ migration rates, with a total of n^2 parameters.

Third, a locus might be under direct or indirect (due to linkage with a selected locus) selection, such as adaptive or balancing selection. Such a locus would show faster or slower changes in allele frequency than neutral loci. Ignoring selection may lead to biased N_e estimates. Recently, the temporal methods have been extended to estimate N_e of a population and the selection coefficient, s , of a locus from temporal genotype data (Bollback *et al.*, 2008; Mathieson and McVean, 2013; Foll *et al.*, 2015). These new methods use hidden Markov models to make the best estimates of N_e and s that explain the observed allele frequency changes owing to drift and selection.

In brief, quite a few genetic marker-based methods are available for estimating the current or recent N_e , using a specific piece of information (such as heterozygosity excess, LD,

temporal changes in allele frequency, sibship frequency or more generally estimated genealogy). Gilbert and Whitlock (2015) provide a comparison of the performance of different methods. It is also possible to use multiple sources of information (such as expected heterozygosity, number of alleles, in addition to those listed above) in arriving at an estimate of N_e . The approximate Bayesian computation (ABC) methods (Beaumont *et al.*, 2002; Beaumont, 2010) were proposed to estimate N_e using many summary statistics calculated from genetic marker data, such as heterozygosity and LD (e.g. Cornuet *et al.*, 2008; Tallmon *et al.*, 2008). One difficulty with the ABC approach to N_e estimation lies with the interpretation of the estimate. Is it the N_e of the parental generation or earlier, or an average of some generations in the past? Of course, this is irrelevant when N_e is constant over generations, but in reality few natural populations maintain a constant size.

16.3 Estimating Census Size by the Genotype Capture–Recapture Approach

It seems a trivial matter to estimate the census size of a population (N_c), since one may naively think that it can be easily obtained by counting the number of individuals in a population. However, the counting approach is feasible only in some very simple cases such as a small population in which individuals are recognizable and countable. As mentioned by Hilborn (2002) (see also Luikart *et al.*, 2010), ‘counting fish is like counting trees, except they are invisible and they keep moving’. Even if they are visible, they could be still individually unrecognizable and thus uncountable. N_c is particularly difficult to estimate accurately for elusive animal species, ranging from fishes to other aquatic organisms and forest-dwelling mammals (Luikart *et al.*, 2010). Unfortunately, N_c is one of the basic and most important parameters necessary for understanding the demographic and genetic status of wildlife, and for the effective management and conservation to maintain population health and genetic diversity.

N_c is a simple concept, but in reality it is defined and used in different ways due perhaps to the difficulty in estimating it. N_c can be and has been defined as the total number of individuals, the number of individuals of a given age or of several different age classes. The most widely used definition is the number of adults in a population (Luikart *et al.*, 2010). Juveniles are not counted because they may not survive to become adults and thus do not contribute genetically. They may also contribute little demographically for many species, like fry and tadpoles in fish and amphibian species. We follow this definition but note that the same or similar estimation methods can be applied to N_c defined otherwise.

A widely used method for estimating N_c is capture–mark–recapture (White and Burnham, 1999), so called because the procedure involves randomly capturing animals, marking them physically (e.g. attaching rings to legs or wings, clipping fins from fish), releasing them to allow them to mix completely with uncaptured individuals in the population, and then performing a ‘recapture’ in which a new random sample of individuals is captured. The proportion of marked individuals (recapture rate) in the recapture signifies N_c : the higher the recapture rate is, the lower the value of N_c . The logic is simple, but the implementation (at least for some species) is difficult and the estimates may not be reliable. It assumes that individuals are both captured and recaptured at random, and captured and marked individuals mix with unmarked individuals completely before recapture occurs. Violation of these assumptions can lead to serious estimation bias. Furthermore, for high-quality estimates, substantial proportions of individuals in a population must be captured and recaptured. This can be extremely costly and may be impossible for many species (e.g. large marine mammals).

Genetic methods, equivalent to CMR, have been developed to estimate N_c , overcoming several difficulties of the traditional CMR. First, animals do not have to be captured and marked artificially, as nature has done this for us. Every individual is uniquely (except for the rare case of identical twins or clone mates) marked with their DNA, which is permanent (as compared with the temporal and easily lost physical markers) and is easily genotyped using a general protocol (e.g. polymerase chain reaction). Therefore, the marked proportion is 100%, regardless of the species or populations. The genetic markers also have much richer information than physical markers. For example, these markers can be used to decipher the genealogical relationships of the individuals. Second, by using DNA extracted from non-invasive samples (e.g. faeces, hair, shed skin, feathers, urine, menstrual blood, snail slime tracks), genetic CMR can be carried out to estimate N_c without capturing, handling, or even seeing the animals. Third, non-invasive DNA-based methods can yield an N_c estimate from samples collected in a single sampling session (rather than multiple sampling sessions for traditional CMR). The principle of the genetic methods is simple. Within a single sampling session, an individual is 'captured' if its multilocus genotype is observed once, and is 'recaptured' $n - 1$ times if its multilocus genotype is observed n times (i.e. n samples show the same multilocus genotype and thus come from the same individual). Then the recapture rate can be used in a way similar to the tradition CMR to estimate N_c . The genetic method's ability to estimate N_c from a single sampling session without handling live animals is extremely helpful in studying species that are costly or time-consuming to sample. Indeed, the availability and the ease of genetic CMR made it possible to estimate the population size of many species (e.g. Creel *et al.*, 2003; Eggert *et al.*, 2003; Schwartz *et al.*, 2007) that are difficult or impossible to study using traditional CMR.

However, genetic CMR does not come without its own problems. The usual assumptions, such as random sampling and homogeneity of capturing rate among individuals, apply to both traditional and genetic CMR. Violation of these assumptions could lead to highly biased estimates of N_c from either CMR method. An additional assumption made by genetic CMR is that genotype data are informative and reliable, so that different multilocus genotypes represent different individuals, and multilocus genotypes from the same individual are identical. Under this assumption, individuals are identified and recapture rate are simply calculated by counting unique and duplicated multilocus genotypes. If genetic marker data were error-free, then increasing the number of genetic marker loci would always increase genetic marker information, and individuals would have unique multilocus genotypes (except for the case of identical twins and clone mates) which can be replicated faithfully when a sufficient number of markers are genotyped. In reality, however, genotyping errors are inevitable, except for a very small sample size. Genotyping errors, such as allelic dropouts and false alleles, are especially abundant in data obtained from non-invasive samples (see Pompanon *et al.*, 2005), due to the low quantity and low quality of DNA extracted from non-invasive samples.

Genotyping errors generate a dilemma in individual (or duplicate) identification and N_c estimation from genetic marker data (especially from non-invasive sources). How many genetic markers should be genotyped and used in the analysis? Without genotyping errors, the more markers are genotyped, the more informative the data will be, and the more accurate the results obtained for individual identification and N_c estimation will be. However, this simple relationship is broken in the presence of genotyping errors. More genetic marker loci would contain more information which is unfortunately buried in more noise (genotyping errors). Let us consider the probability that a genotype at L loci has at least one error, E . If genotyping errors occur independently and equally at a rate e per locus, then $E = 1 - (1 - e)^L$. E increases monotonically and rapidly with L . Even if e is negligibly small, E can still be substantial when L is large. For example, a 10-, 50- and 250-locus genotype is expected to contain at least one genotyping error with a probability of $E = 1.0\%$, 4.9% and 22.1% , respectively, when $e = 0.001$, and of $E = 9.6\%$,

39.5% and 91.9%, respectively, when $e = 0.01$. Note that a single error in a multilocus genotype would suffice to generate a false or ‘ghost’ individual, and thus inflate N_c estimates. This result has prompted several researchers to suggest that individual identification should use the minimum number of genetic loci required to attain a low probability of identity among samples from different individuals (Waits *et al.*, 2001; Creel *et al.*, 2003). This suggestion is, however, difficult to implement in practice. How do we determine this ‘optimal’ number of loci? In principle, it is dependent on marker allele frequency distributions and marker error rates, and should be found by minimizing both types of errors (false and unidentified individuals). It is unclear how best to find this optimal set of loci in practice.

16.3.1 Methods Based on Multilocus Genotype Mismatches

Several approaches to handling genotyping errors have been proposed and applied to obtain unbiased and accurate estimates of N_c (Wang, 2016b). The simplest one is to use a threshold number, T_m , of one or two mismatches between two multilocus genotypes in determining whether they come from the same individual (thus duplicates) or from two different individuals. If the observed number of mismatches is below the threshold, then the two multilocus genotypes are regarded as duplicates that come from the same individual, and the mismatches are regarded as due to genotyping errors. Otherwise, the mismatches cannot be fully explained by genotyping errors and the two multilocus genotypes are regarded as coming from separate individuals. This approach has been implemented in several computer programs, such as GENECAP (Wilberg and Dreher, 2004). Allowing $T_m = 1$ or 2 mismatches could reduce ghost individuals and thus the overestimation of N_c substantially. However, this threshold is obviously arbitrary. The optimal T_m should depend on factors such as the mistyping rates and the number of loci. While $T_m = 1$ or 2 may suffice to reduce ghost individuals when both mistyping rates and number of loci are low (say, $e < 0.05$ and $L < 20$ for high-quality microsatellites), more mismatches should be allowed for when e is high (e.g. for genotypes from low-quality non-invasive samples), L is large (e.g. many SNPs), or both. Furthermore, with heterogeneity in missing data and mistyping rates among individuals and among loci, this mismatch threshold approach is even more problematic.

16.3.2 Methods Based on Pairwise Relatedness

A fundamental flaw of the above mismatch method is that it does not use the precious marker information efficiently and thus has low accuracy. In fact, by using an ‘optimal’ set of markers and the corresponding value of $T_m = 1$ or 2, data on more markers are simply discarded. A desirable method should use all data available and have increasing power and accuracy with an increasing number of markers. This is especially important now with more and more genomic data with many SNPs coming into use in conservation genetics. A more powerful approach to individual identification and N_c estimation from genotype data is via pairwise relatedness analysis. Relatedness between two multilocus genotypes can be estimated by a number of estimators, which are resilient to genotyping errors (Wang, 2007) and implemented in quite a few computer programs (e.g. Wang 2011; Hardy and Vekemans, 2002). Compared with the mismatch method, relatedness analysis also uses allele frequency as well as genotype information in identifying duplicated individuals from other competitive relationships such as full-sibs (Ringler *et al.*, 2015). For diploid species, two multilocus genotypes have an expected relatedness, r , of 1 and 0.5 if they come from the same individual and from two first-degree relatives (full-sibs and parent offspring), respectively (and less for less closely related individuals). Therefore, they can be inferred to represent duplicates of the same individual when their estimated relatedness

is closer to 1 than to 0.5. Otherwise, they are inferred to represent distinct individuals. Ringler *et al.* (2015) applied the relatedness method to track the survival of amphibian (dendrobatid frog (*Allobates femoralis*)) larvae through adulthood using 14 highly polymorphic microsatellites. They showed the method was more accurate than other genetic methods (e.g. genotype matching) and produced highly reliable results as checked by matching ventral patterns of juveniles and adult individuals.

16.3.3 Methods Based on Pairwise Relationships

The similarity between two multilocus genotypes is explained mainly by the underlying genealogical relationship of the individuals from which the genotypes come. A parent–offspring dyad, for example, shares one allele IBD at each locus, while identical twins or duplicates should have an identical genotype at each locus. Genotyping errors destroy the similarity patterns to an extent determined by their rate of occurrence. Most often they decrease the similarity, but occasionally they may increase it. It is possible to calculate the likelihood of two multilocus genotypes falling into each of a number candidate relationships, given the error rate at each of the genotyped loci. The maximum likelihood relationship is the one that has the maximal likelihood value among those of all candidate relationships. For individual identification in N_c estimation, the relevant candidate relationships are duplicates (DP) and the competing relationships of full sibs (FS), half sibs (HS), parent–offspring (PO), and unrelated (UR). The competing relationships are chosen because FS, HS and PO have relatedness values close to DP, while UR is usually the most abundant relationship in a sample of individuals. If DP has the highest likelihood, then the two multilocus genotypes are inferred to come from the same individual (i.e. duplicates); otherwise, they are inferred to come from distinctive individuals. One of the advantages of this dyadic relationship approach over the pairwise relatedness approach is that, as shown below, genotyping errors occurring at each locus can be easily accounted for.

These candidate relationships are characterized by three IBD coefficients Δ_i , where Δ_i is the probability that the two non-inbred individuals share exactly i ($i = 0, 1, 2$) pairs of gene copies IBD at a locus. In diploid species, $\Delta = \{\Delta_0, \Delta_1, \Delta_2\} = \{0, 0, 1\}$ for DP, $\{0.25, 0.5, 0.25\}$ for FS, $\{0.5, 0.5, 0\}$ for HS, $\{0, 1, 0\}$ for PO, and $\{1, 0, 0\}$ for UR, and the relatedness r above mentioned is a simple function of these coefficient: $r = \frac{1}{2}\Delta_1 + \Delta_2$. The probability of an observed genotype $G_X = \{a, b\}$ for individual X and an observed genotype $G_Y = \{c, d\}$ for individual Y at a locus with k codominant alleles is

$$P[\{a, b\}, \{c, d\} | \Delta] = \sum_{u=1}^k \sum_{v=1}^k \sum_{w=1}^k \sum_{x=1}^k P[\{u, v\}, \{w, x\} | \Delta] P[\{a, b\} | \{u, v\}] P[\{c, d\} | \{w, x\}], \quad (16.13)$$

(Wang, 2016b), where

$$\begin{aligned} P[\{u, v\}, \{w, x\} | \Delta] &= (2 - \delta_{uv}) p_u p_v (\Delta_0(2 - \delta_{wx}) p_w p_x + \frac{1}{4} \Delta_1 (2 - \delta_{wx}) ((\delta_{uw} + \delta_{vw}) p_x \\ &\quad + (\delta_{ux} + \delta_{vx}) p_w + \Delta_2 (\delta_{uw} \delta_{vx} + \delta_{ux} \delta_{vw} - \delta_{uw} \delta_{vx} \delta_{ux} \delta_{vw}))) \end{aligned} \quad (16.14)$$

is the probability that X and Y have (true underlying) genotypes $\{u, v\}$ and $\{w, x\}$, respectively, conditional on their relationship or IBD coefficients Δ , and δ_{uv} (and similarly for other δ variables) is the Kronecker delta variable with $\delta_{uv} = 1$ and 0 when $u = v$ and $u \neq v$, respectively. The probability that a genotype $\{u, v\}$ is observed as $\{a, b\}$ due to false alleles and allelic dropouts, $P[\{a, b\} | \{u, v\}]$, is derived by Wang (2004) using the rates of false alleles and allelic dropouts estimated for the locus. $P[\{c, d\} | \{w, x\}]$ is calculated similarly.

Equations (16.13) and (16.14) give the likelihood for a single locus. For multiple independent loci, the likelihood is simply the product of single-locus likelihood values. Using both simulated and empirical data sets, Wang (2016b) showed that the dyadic relationship approach described above gives much more accurate individual identifications than the genotype mismatch method. Overall, it is also more accurate than the pairwise relatedness approach.

It is worth noting that these results and the methods described for relatedness and relationship inference are mainly relevant when only a few microsatellites are available. We have focused on these here because this is the data that is typically available in the context of estimating population sizes. With genome-wide SNP data, other more powerful methods are available for estimating Δ and r (e.g. Purcell *et al.* 2007, Albrechtsen *et al.* 2009). There are even methods that allow for such estimation from sequencing data that is of such low depth that the genotype can only be called with high uncertainty (Korneliussen and Moltke, 2015).

16.3.4 Methods Based on Pedigree Reconstruction

Each of the above three approaches considers a pair of multilocus genotypes in determining whether they come from the same individual or different individuals, in isolation of other multilocus genotypes. This pairwise approach is simple to implement, fast in computation, but incurs two serious problems.

First, the pairwise inferences, when put together, may be incompatible. Among three multilocus genotypes X , Y and Z , for example, $\{X, Y\}$ and $\{X, Z\}$ might be inferred to be two pairs of duplicates, while Y and Z might be inferred to come from separate individuals. This is obviously incompatible. The frequency of incompatibilities increases rapidly with an increase in genotyping error rates and in individual duplication levels. Scat- or hair-based non-invasive samples often exhibit massive replications, with potentially hundreds of replicated samples per individual (e.g. Creel *et al.*, 2003). At this level of replications, even a small genotyping error rate could result in many incompatible pairwise inferences, which would translate to grossly biased estimates of the number of distinct individuals in a sample, recapture rate, and N_c .

Second, pairwise approaches could have low power due to their insufficient use of marker data (Wang, 2004). In parentage analyses, for example, one offspring possesses one of the two paternal alleles and thus provides only 50% information for the parent. The probability that both parental alleles are present in the genotypes of a number of n offspring is $1 - 2^{1-n}$, and thus the power of parentage assignment rises rapidly with an increasing n when all siblings are considered jointly for parentage assignment. The same argument applies to the inference of other pedigree relationships (Wang, 2007), including identical twins (or duplicates).

To solve these among other problems of the pairwise approaches, Wang (2016b) adapted his sibship reconstruction method to partition the entire sample of N multilocus genotypes into an unknown number (m) of individual clusters. Each cluster represents a distinct individual, containing one or more multilocus genotypes inferred to come from the same individual. The new method begins by reconstructing sibship (Wang, 2004), assuming either polygamy or monogamy and each multilocus genotype represents one distinct individual. It then further partitions each inferred full sibship into a number of individual clusters. For a full sibship containing M multilocus genotypes, there are B_M possible partitions (or configurations), where B_M is the Bell number. Three multilocus genotypes ($M = 3$) of X , Y and Z , for example, have $B_3 = 5$ different partitions, $\{X, Y, Z\}$, $\{XY, Z\}$, $\{XZ, Y\}$, $\{YZ, X\}$, $\{XYZ\}$, where different individuals in a partition are separated by a comma. The partition $\{XY, Z\}$ means, for example, two individual clusters, with cluster 1 containing genotypes X and Y coming from one individual and cluster 2 containing genotype Z coming from another individual. The challenge is that the possible number of partitions, B_M , increases combinatorically with M . Even for moderate values of $M = 5, 10$

and 15, for example, the corresponding B_M values are 52, 115,975 and 1,382,958,545, respectively. Apparently, it is impossible to consider all partitions even when $M = 15$. Wang (2016b) proposed a systematic method to construct and search through a small fraction of the B_M possible partitions for a full sibship of M multilocus genotypes. The partition with the maximal likelihood is reported as the best estimate.

The likelihood of a partition is calculated by integrating all possible individual genotypes and genotyping errors. As an example, consider a partition with m individual clusters, with cluster c ($= 1, 2, \dots, m$) containing n_c genotypes g_{cj} ($j = 1, 2, \dots, n_c$) at a locus with k alleles. The likelihood of this partition is

$$\sum_{u=1}^k p_u \sum_{v=1}^k p_v \sum_{w=1}^k p_w \sum_{x=1}^k p_x \prod_{c=1}^m \frac{1}{4} \left(\sum_{a=u,v} \sum_{b=w,x} \prod_{j=1}^{n_c} P[g_{cj}|a,b] \right), \quad (16.15)$$

where the probability of observing genotype g_{cj} given its underlying true genotype $\{a,b\}$ is calculated in the same way as $P[\{c,d\}|\{w,x\}]$ in equation (16.14). The computation of equation (16.15) can be greatly reduced by pooling unobserved alleles and by pooling identical parental genotypes (e.g. $\{u, v\}$ and $\{v, u\}$) and parental genotype combinations (e.g. $\{\{u, v\}, \{w, x\}\}$ and $\{\{w, x\}, \{u, v\}\}$), as in sibship likelihood calculations (Wang, 2004). For multiple loci in linkage equilibrium, the likelihood is simply the product of single locus values calculated by (16.15).

The analyses of extensive simulated data and an empirical data set (Ringler *et al.*, 2015) showed that the likelihood-based partition method is generally one or two orders more accurate for individual identification than the pairwise methods described in Sections 16.3.1–16.3.3. Its accuracy is especially high when the sampled multilocus genotypes have poor quality (i.e. teeming with genotyping errors and missing data) as is usually true with non-invasive samples, and when samples are highly replicated, a situation also typical of non-invasive sampling used in estimating population size (Creel *et al.*, 2003).

16.4 Inferring Genetic Structure

Conservation of wild populations needs information not only about N_c and N_e , but also about the genetic structure. For an endangered species, we may ask questions such as how many populations there are, how different (differentiated) they are, whether hybridization exists, and which population a focal individual (say, a poached animal) comes from. Statistical methods based on population genetics have been developed to address these issues using genetic marker data. We present some of these below.

16.4.1 Measuring Genetic Differentiation

The genetic differentiation among well-defined population segments (e.g. defined by geographic features such as locations, rivers and mountains) can be measured by Wright's fixation index, F_{ST} (Wright, 1943). This was proposed to measure the population subdivision or structure. It was defined originally as 'the correlation between random gametes within subdivisions, relative to gametes of the total population' (Wright, 1965). Equally, it can be understood as the inbreeding coefficient of a hypothetical individual formed by the union of gametes taken at random from within subdivisions, relative to the total population. These definitions pave the way for estimating F_{ST} from pedigree data.

Wright (1965) also showed that, for a neutral locus without mutations, F_{ST} is the amount of genetic variation between populations (V_B) as a proportion of the total variation (V_T), which

is composed of within-population (V_W) and between-population (V_B) variation. Suppose an allele has frequency p_i in subdivision i . The mean and variance of p_i are $\mu = 1/n \sum_{i=1}^n p_i$ and $V_B = (1/n) \sum_{i=1}^n (p_i - \mu)^2$, and

$$F_{ST} = V_B / V_T, \quad (16.16)$$

where $V_T = p(1-p)$. This formulation paves the way for estimating F_{ST} from genetic marker data (see below). Under the assumption of neutrality and no mutations, different genetic markers, irrespective of their properties such as number and frequencies of alleles, have the same expected value of F_{ST} . Each genetic marker can be used to make an independent estimate, and the average estimate from many loci gives an accurately estimated F_{ST} .

Several genetic marker-based estimators of F_{ST} were proposed and widely used in evolution, ecology and conservation literature. For example, Nei (1973) proposed his coefficient of gene differentiation, G_{ST} , to measure genetic differentiation using multi-allelic markers. G_{ST} is defined as

$$G_{ST} = \frac{H_T - H_S}{H_T}, \quad (16.17)$$

where H_T and H_S are the heterozygosity (or gene diversity) of the total population and the average heterozygosity of subpopulations expected under Hardy–Weinberg equilibrium, respectively. In the special case of a di-allelic locus, Nei (1973) showed that $H_T = 2p(1-p)$ and $H_T - H_S = 2V_B$, and thus G_{ST} in equation (16.17) is identical to F_{ST} in equation (16.16). For a locus with $k > 2$ alleles, G_{ST} is equal to the weighted average of F_{ST} for all alleles (Nei, 1973). To estimate G_{ST} from genetic marker data, a sample of individuals is taken from each of a number of populations, and each sampled individual is genotyped at a number of genetic marker loci. The data are then analysed for estimating \hat{H}_T and \hat{H}_S , using the nearly unbiased estimators of Nei and Chesser (1983).

Cockerham (1969, 1973) introduced an analogous measure of differentiation, coancestry θ , which does not rely on s , the number of subpopulations, which is usually unknown. Weir and Cockerham (1984) developed an estimator of θ using genetic marker data. Although conceptually different, Cockerham's θ and Wright's F_{ST} give very similar results in practice, except when s is very small.

Wright's F_{ST} was designed to measure population structure or the extent of genetic differentiation between subpopulations due to drift within (or small N_e of) subpopulations and migration between subpopulations. Its estimators (e.g. θ and G_{ST}) give unbiased and consistent estimates when the used genetic markers are neutral (i.e. not under direct or indirect selection) and have low mutation rates. Highly mutable markers, however, will show lower and variable differentiation different from the expected F_{ST} of the population (Whitlock, 2011). As a result, they should not be used in F_{ST} estimators such as θ and G_{ST} . In Wright's (1943) infinite island model, he showed (see also Takahata and Nei, 1984) that

$$F_{ST} \approx \frac{1}{4N_e(m+u)+1}, \quad (16.18)$$

where the marker mutation rate is u in the infinite allele model. Equation (16.18) shows that mutations in the infinite alleles model have exactly the same effect on differentiation as migration in the island model. Migration and mutations have a large impact on harmonizing the population, and a migration or mutation rate of $1/N_e$ per generation would constrain G_{ST} to a maximal value of 0.2. It also suggests that use of highly polymorphic markers of high u might lead to an underestimated F_{ST} .

In practice, many empirical studies (e.g. Balloux *et al.*, 2000; Carreras-Carbonell *et al.*, 2006) confirmed that microsatellites yielded differentiation estimates that were unexpectedly low among highly differentiated subspecies, as evidenced by morphological and other information. This has prompted some researchers to question F_{ST} as a proper measure of differentiation (e.g. Hedrick, 2005; Jost, 2008; Meirmans and Hedrick, 2011), and to propose either new differentiation statistics, such as D (Jost, 2008), or corrected or standardized F_{ST} , F'_{ST} (Hedrick, 2005), for replacing F_{ST} . However, while it is true that F_{ST} calculated from highly polymorphic markers (i.e. high mutation rate u) might underestimate differentiation as shown in equation (16.18), what is to be blamed is not F_{ST} *per se*, but the markers or the naive method used for its estimation. One should not use markers of high mutation rates naively as if they were unaffected by mutations in F_{ST} estimation. New differentiation statistics, such as D , are shown to have more serious problems than F_{ST} , such as marker polymorphism dependent, improper measurement of differentiation, no biological meaning and use (e.g. Ryman and Leimar, 2009; Whitlock, 2011; Wang, 2012).

Recently, Wang (2015) showed, using both analytical and simulation approaches, that marker-based F_{ST} or G_{ST} underestimates population differentiation caused solely by demographic factors (migration, population subdivision and subpopulation sizes) only when mutations occur at a higher rate than genetic drift (i.e. $u > 1/N_e$) and migration (i.e. $u > m$). Otherwise, all markers, regardless of their polymorphisms, should yield the same F_{ST} or G_{ST} value in expectation and thus should give replicated estimates of population differentiation. The study also pointed out that in cases where mutational effects are important, there is a negative, roughly linear correlation between single-locus estimates of G_{ST} and H_S . Motivated by this, Wang (2015) proposed to calculate the correlation coefficient between G_{ST} and H_S across loci, r_{GH} , for detecting mutational effects on G_{ST} . A highly negative and significant r_{GH} reveals that mutations have impacted the G_{ST} estimated from the markers, so that the mean G_{ST} value across markers may well underestimate population differentiation.

16.4.2 Population Assignment

Population assignment is the problem of inferring which population(s) a specific individual is from based on genetic data. In the simplest case, there has been no admixture between any individuals from different populations and thus the focal individual has ancestry from one population only. When this is the case the population assignment problem can be thought of as identifying and determining the origin of immigrants, which is the problem we will focus on in the first part of this section. One of the approaches to identifying immigrants is population assignment analysis, proposed by Paetkau *et al.* (1995). They sampled individuals from each of a number of populations, genotyped them at a number of genetic marker loci, and tried to find out whether an individual was a resident or an immigrant of the population from which it was drawn. When an individual was an immigrant, which source population did it come from? To address these questions, they made an estimate of the allele frequencies at each locus of each population, using the genotype data of individuals sampled from the population and assuming no immigrants were sampled. For each individual, they calculated the probability that it belonged to each of the sampled populations, using the multilocus genotype of the individual and the estimated allele frequencies at each locus of each population. The individual was then assigned to the population that makes the observed data the most probable. The rationale behind this approach is that, in the absence of immigration, the genotype of each individual is a random draw from its own population gene frequencies and therefore, providing gene frequencies vary sufficiently among populations, individuals should be assigned back to their own populations. Migrants can be detected because they have a higher probability of being assigned

to a population other than the one from which it is sampled. In the study of Paetkau *et al.* (1995), four populations of polar bears were sampled, using eight microsatellite loci. They found that 60% of individuals were assigned to the population in which they were sampled, 33% to the nearest population and 7% to more distant populations. Using the same loci in studying brown bears from seven populations in NW Canada, Paetkau *et al.* (1998) found that 92% individuals were assigned to the population from which they were sampled, and those assigned to other populations illustrated biologically plausible patterns of dispersal.

The above assignment method works well when marker information is sufficient and populations are well differentiated. However, population-specific allele frequencies were estimated by assuming no immigrants were included in a sample from the population. This is a good approximation when migration rate is low. Otherwise, a sample of individuals from a population may contain immigrants from different source populations. Ignoring immigrants in calculating allele frequencies blurs the differentiation among populations and incurs reduced power. Another issue in population assignment analysis is whether the genotype of the focal individual for assignment should be used or discarded in calculating allele frequencies. Including the focal individual in allele frequency calculations would lead to an overestimated probability that it is a resident of the population from which it is sampled, when it is actually an immigrant. However, discarding the focal individual in allele frequency calculations may lead to a zero-valued allele frequency which may derail the assignment analysis.

Some of these issues are addressed in a study by Rannala and Mountain (1997). Their approach differs from that of Paetkau *et al.* in that they include Bayesian estimation of the allele frequencies in each population, and a likelihood ratio test to compare different hypotheses. They consider a number of source populations with a number of marker loci sampled. For ease of exposition and illustration, only one locus and two diploid populations are considered here. The observed allele counts are given by the vectors \mathbf{a}_0 and \mathbf{a}_1 for the two populations. The posterior distribution for the unknown allele frequency vector \mathbf{x} is given by $P(\mathbf{x}|\mathbf{a}_0)$ and $P(\mathbf{x}|\mathbf{a}_1)$. To estimate this, a uniform Dirichlet prior, $D(1/k, \dots, 1/k)$, is assumed, where k is the number of alleles observed across all sampled individuals. Using this prior, they obtained a posterior $P(\mathbf{x}|\mathbf{a}_0) = D(\mathbf{a}_0 + 1/k)$ and $P(\mathbf{x}|\mathbf{a}_1) = D(\mathbf{a}_1 + 1/k)$. Rannala and Mountain describe their prior as assigning equal probability density to the frequencies of the alleles, although this only occurs with a Dirichlet $D(1, \dots, 1)$. Clearly the choice of prior may depend on the genetic model assumed.

Assuming an individual has no immigrant ancestry, its genotype would be a multinomial sample of size 2 taken, for example, from the posterior $P(\mathbf{x}|\mathbf{a}_0)$ defined above. Integrating out \mathbf{x} , the marginal probability of an individual having genotype $\mathbf{X} = (X_i, X_j)$ is then multinomial Dirichlet,

$$P(\mathbf{X}|\mathbf{a}_0) = \int P(\mathbf{X}|\mathbf{x})P(\mathbf{x}|\mathbf{a}_0)d\mathbf{x}.$$

For multiple loci, the probabilities for each locus can be multiplied together, under the assumption that they are independent.

If we allow for the focal individual to be admixed, and thus have ancestry from more than one population, the population assignment problem becomes a bit more complicated. In this case other hypotheses should be considered, for example, whether the individual has ancestry from one immigrant d generations earlier. The probability of observing genotype $\mathbf{X} = (X_i, X_j)$ in an individual where one allele is drawn at random from one population and one is drawn at random from the other population is $P(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1) = \frac{1}{2}(P(X_i|\mathbf{a}_0)P(X_j|\mathbf{a}_1) + P(X_j|\mathbf{a}_0)P(X_i|\mathbf{a}_1))$, where the probabilities are calculated as single draws from the multinomial Dirichlet outlined above.

Rannala and Mountain (1997) extend the argument above to consider the probability of observing the genotype, $P(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)$, given that an individual has one ancestor coming, d

generations ago, from a different population (with all the other ancestors coming from the same resident population). The probability that the individual has an immigrant allele is $1/2^{d-1}$, in which case the probability is as given above, and the probability that the individual has both alleles coming from the resident population (i.e. no immigrant allele) is $1 - 1/2^{d-1}$, which is calculated as a sample of size 2 from the multinomial Dirichlet, $P(\mathbf{X}|\mathbf{a}_0)$ or $P(\mathbf{X}|\mathbf{a}_1)$.

Rannala and Mountain suggest that for particular individuals, hypotheses can be tested using likelihood ratio tests of the form

$$\Lambda = \frac{P(\mathbf{X}|\mathbf{a}_0)}{P(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)}.$$

Critical regions of a given size can be estimated by parametric bootstrapping (i.e. Monte Carlo simulations of genotypes \mathbf{X} with probability $P(\mathbf{X}|\mathbf{a}_0)$ under the null hypothesis). The power of the tests can be estimated by additionally simulating under the alternative hypothesis genotypes \mathbf{X} with probability $P(\mathbf{X}|\mathbf{a}_0, \mathbf{a}_1, d)$.

A comparative analysis of the performance of the two approaches on test data sets has been carried out by Cornuet *et al.* (1999). They introduce the use of a distance-based method for assigning individuals to populations. The method is analogous to that of Paetkau *et al.*, but individuals are assigned to populations with the smallest genetic distance. Cornuet *et al.* study the most commonly used genetic distances for microsatellite or allozyme data, modified to take into account that individuals are compared with populations. In addition to the question of assigning individuals to particular populations, Cornuet *et al.* also consider the question whether an individual is likely to have come from the population in which it resides. In order to test the latter, which they term testing for exclusions, Cornuet *et al.* suggest simulating genotypes at random from an estimate of the sample population gene frequencies and comparing the likelihood (or genetic distance) of the individual's genotype with the distribution of likelihoods or distances from random sampling.

Cornuet *et al.* tested the methods using data simulated from a model of diverging populations with mutations according to either an infinite allele model or a stepwise mutation model. Their overall conclusion was that the Rannala and Mountain method outperformed all other methods, and the genetic distance methods, in general, performed less well than the methods using maximum likelihood estimation. Since Rannala and Mountain's calculation of likelihoods differs from that of Paetkau *et al.* only in the estimation of population frequencies \mathbf{x} , this implies that, taking a Bayesian approach, the choice of a suitable prior distribution for \mathbf{x} is an important consideration. Tests for exclusions also appear to work well using the Rannala and Mountain method to calculate likelihoods. However, Cornuet *et al.* noted that the method used for simulating null distributions was an important component of testing – with increasing numbers of loci there was a tendency to incorrectly exclude individuals from all populations.

16.4.3 Population Clustering and Inference of Ancestry Proportions

Several issues remain with the method of Rannala and Mountain discussed above. First, it is necessary to conduct a large number of hypothesis tests, requiring some care in specifying the critical values for assessing significance. Second, it would be better if the estimates of the population allele frequencies could take into account the possibility that some individuals are immigrants rather than residents. Third, hypothesis testing is limited to the populations that are actually included in the survey, whereas immigrants may come from other, unsurveyed, populations (Cornuet *et al.*, 1999).

More importantly, the above-described assignment and F_{ST} methods rely on predefined populations. In some cases, however, it is impossible to sort the sampled individuals into predefined

populations, and therefore to study population structure by the assignment and F_{ST} methods. An example is the mixed stock analysis, where individuals are sampled from a common breeding or feeding location representing different but unknown populations. Such cryptic population structure is common in conservation genetics studies of wildlife. For example, a batch of confiscated animals or animal products (e.g. ivory) may represent several different but undefined populations.

The focus of this section is to describe a method that overcomes several of these limitations. This method, described in more detail in **Chapter 8**, was proposed by Pritchard *et al.* (2000) and is a Bayesian clustering method that sorts individuals in a sample either probabilistically or proportionally into K unknown populations using multilocus genotype data of sampled individuals. We will refer to the model where the sorting is probabilistic as the ‘non-admixture model’ and the model where sorting is proportional as the ‘admixture model’. In both models it is assumed that each of the assumed populations is characterized by a set of allele frequencies at each marker locus, and these markers are in Hardy–Weinberg equilibrium and linkage equilibrium in each population. The non-admixture model assumes that each individual’s genome comes exclusively from one of the K assumed populations, and yields, for each individual, a vector of K numbers (summing to 1) with each specifying the estimated probability that the individual comes from the corresponding population. The admixture model assumes that each individual’s genome can come from one or multiple populations, and yields, for each individual, a vector of K numbers (summing to 1) with each specifying the estimated proportion of the individual’s genome (membership) from the corresponding population.

The non-admixture model tries to estimate the posterior distribution

$$P(\mathbf{Z}, \mathbf{p}|\mathbf{X}) \sim P(\mathbf{Z})P(\mathbf{p})P(\mathbf{X}|\mathbf{Z}, \mathbf{p}),$$

where \mathbf{X} denotes the genotypes of the sampled individuals, \mathbf{Z} denotes the (unknown) populations of origin of the individuals, and \mathbf{p} denotes the (unknown) allele frequencies in all populations (note that \mathbf{X} , \mathbf{Z} , \mathbf{p} are all multidimensional vectors). The admixture model is more complicated, with an added \mathbf{Q} matrix of admixture proportions for each individual, and \mathbf{Z} is expanded to include the membership (i.e. source population) of each allele copy at each locus of each individual. The data is \mathbf{X} , which is used to learn about \mathbf{Z} (the non-admixture model), \mathbf{Q} (admixture model), and \mathbf{p} . Pritchard *et al.* took a Bayesian approach to solve the problem. Initially, sampled individuals were randomly allocated to the K non-empty clusters (populations). Using Markov chain Monte Carlo (MCMC) over many iterations, they improve (update) individual membership assignments and population allele frequency estimates iteratively (for more details about MCMC, see **Chapter 1**, this volume). The results of an analysis of wildcats by the admixture model (Beaumont *et al.*, 2001) are shown in Figure 16.1 to illustrate the method’s capability in inferring admixture and individual ancestry proportions.

Pritchard *et al.*’s (2000) method was implemented in a computer program STRUCTURE, which has proven to be highly powerful and popular. Among many other uses, STRUCTURE can be employed to estimate the most likely number of populations (K) that the sampled individuals represent, the population membership assignments of sampled individuals, identifications of hybrids or admixture. Later, the work was improved in several respects to increase its application scope and power. In the original implementation of the model, allele frequencies at a locus are assumed uncorrelated among populations. Falush *et al.* (2003) implemented the correlated allele frequency model in STRUCTURE, and showed that the new model could detect subtle population subdivisions that were not detectable using the uncorrelated allele frequency model, and could also provide a F_{ST} estimate for each population. They also developed methods to handle linkage between loci. This ability makes it possible to detect admixture events

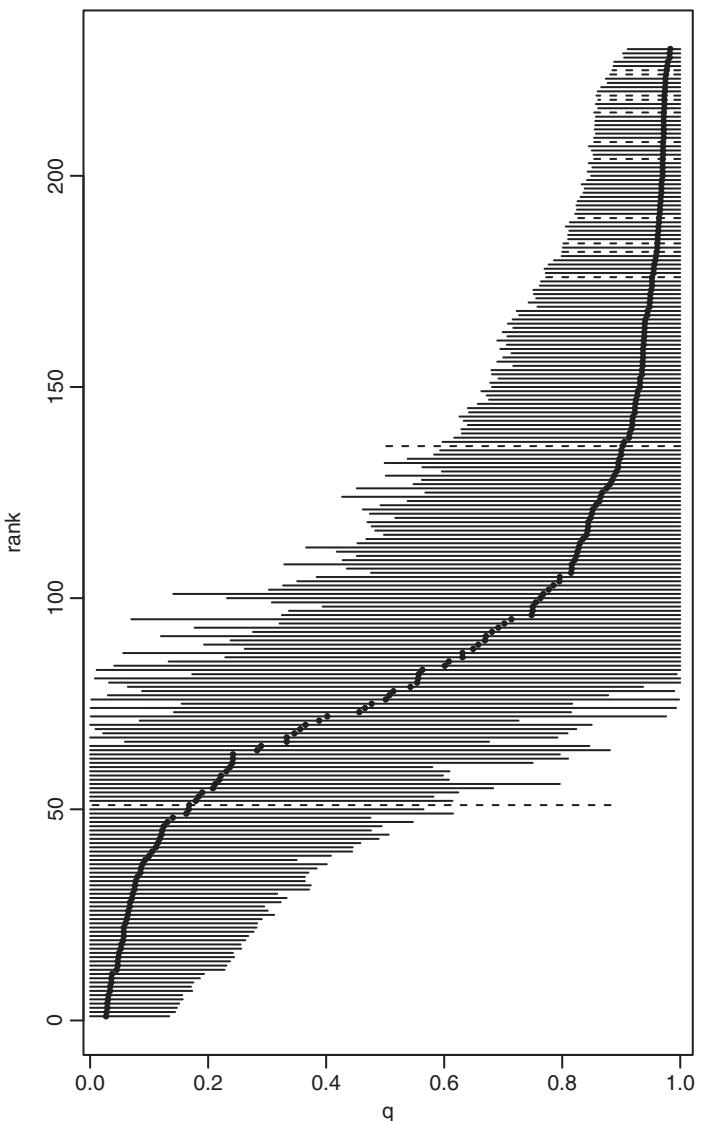


Figure 16.1 The means and 95% credible intervals for estimates of the probability (q) that individual wild-living cats have purely non-domestic ancestry. These are plotted against the rank of the mean estimate. Horizontal dashed lines refer to museum specimens (from Beaumont *et al.*, 2001).

farther back into the past and to infer the population of origin of chromosomal segments (i.e. not individual loci). Falush *et al.* (2007) also extended the methods to allow for uncertainties of genotypes so that STRUCTURE can use dominant markers (such as amplified fragment length polymorphisms) and codominant markers with null alleles. Hubisz *et al.* (2009) extended the methods by allowing the use of non-genetic information in assisting the inference of population structure. They showed that individual location information can be used to modify the prior distribution for each individual's population assignment, thus helping to achieve better inference results when genetic marker data have limited information or when population structure is weak.

With Pritchard *et al.*'s original work and co-workers' extensions, STRUCTURE has become a powerful, indispensable tool in studies of evolution, ecology, conservation and human medicine. However, nothing is perfect, and STRUCTURE is no exception. While it does an excellent job in assigning individual memberships (or ancestry proportions), it has difficulties in finding the number of populations, K . Pritchard *et al.* developed an *ad hoc* procedure for estimating $P(\mathbf{X}|K)$, the probability of data \mathbf{X} given K , using the sample mean and variance $P(\mathbf{X}|K)$ estimates from MCMC. They showed that the method yielded good results from test data. Simulations conducted by Evanno *et al.* (2005) showed that in most cases this *ad hoc* procedure failed to provide a correct estimate of K . However, using another *ad hoc* statistic ΔK based on the rate of change in $P(\mathbf{X}|K)$ between successive K values, the authors found that STRUCTURE accurately detects the uppermost hierarchical level of structure for the scenarios simulated in their study. For example, when a population is subdivided into K archipelagos, and a number of islands within each archipelago, the method of Evanno *et al.* is more accurate than the method of Pritchard *et al.* in retrieving the K value. Ever since then, Evanno *et al.*'s method has been widely used in estimating K from STRUCTURE analysis. However, more recent studies (e.g. Wang, 2017) have pointed out that Pritchard *et al.*'s original method is more accurate than Evanno *et al.*'s method in many simulated scenarios. Furthermore, the statistic ΔK cannot be calculated when $K=1$ (i.e. panmixia), so Evanno *et al.*'s method is guaranteed to fail when the sampled individuals come from a single random mating population.

Other issues come from the assumptions of STRUCTURE. When these assumptions are met, usually the method works very well. Otherwise, it could give misleading results. One of the assumptions is Hardy–Weinberg equilibrium within populations, and thus the absence of inbreeding. Gao *et al.* (2007) extended the STRUCTURE method by eliminating the assumption of Hardy–Weinberg equilibrium within clusters so that they could infer both inbreeding (or selfing) and population structure simultaneously. They also showed that, in the presence of inbreeding or selfing, STRUCTURE could give grossly incorrect inferences, while their extended method, implemented in a computer program InStruct, gave accurate inference of both population structure and inbreeding. The Hardy–Weinberg equilibrium assumption also implies that there exists no genetic structure among the individuals sampled from a population; they are unrelated. In practice, however, some individuals sampled from a population could be close relatives. This is especially likely for highly fecund species (e.g. fishes) when early life-stage individuals (e.g. tadpoles, juveniles) are sampled locally. Simulations and analysis of empirical data (Anderson and Dunham, 2008; Rodríguez-Ramilo and Wang, 2012) showed that close relatives such as siblings included in a sample could cause STRUCTURE to overestimate K , mistaking a family as a population. Rodríguez-Ramilo and Wang (2012) showed that the multilocus genotype data could first be analysed to identify close relatives, and then cleaned by removing identified relatives before conducting a STRUCTURE analysis. Close relatives could also be identified by standard relatedness inference methods. However, these methods require ample genetic marker data (e.g. many SNPs) and the inference accuracy could be compromised by the presence of samples from multiple populations (Moltke and Albrechtsen, 2014).

As a Bayesian method, STRUCTURE uses quite a few priors together with genotype data to make inferences. For example, it requires a prior for the ancestry of an individual. The default used by STRUCTURE assumes a uniform prior that a sampled individual comes from each of the assumed K source populations with a probability of $1/K$. Wang (2017) showed that, with unbalanced sampling in which many individuals are sampled from one population while few individuals are sampled from another population, STRUCTURE gives poor results (see also Kalinowski, 2011; Neophytou, 2014; Puechmaille, 2016) because its default ancestry prior is severely violated. Using the alternative prior which allows for unequal prior ancestries from

different populations largely solved the problem, giving accurate estimates of both K and population assignments. This demonstrates that one needs to be careful in choosing the right priors and models offered by STRUCTURE; blindly using it with the default settings could yield sub-optimal or incorrect results.

It is worth mentioning that many other model-based (e.g. Dawson and Belkhir, 2001; Corander *et al.*, 2003; Alexander *et al.*, 2009) or non-model-based (e.g. Jombart, 2010; Rodríguez-Ramilo *et al.*, 2009) methods have been proposed to infer hidden population structure from genetic marker data. Some methods run much faster than STRUCTURE, such as the maximum likelihood method ADMIXTURE (Alexander *et al.*, 2009), fastSTRUCTURE (Raj *et al.* 2014) and discriminant analysis of principal components (DAPC) (Jombart, 2010), making it possible to analyse large data sets involving many individuals and many loci. Further developments also make it possible to use low-depth next-generation sequence data (using genotype likelihood to account for the uncertainty of genotypes due to the limited amount of data available) in inferring population structure (Skotte *et al.*, 2013). The non-model-based methods have an additional advantage that they are not based on a population genetics model and thus make fewer assumptions. For example, genotyping errors, inbreeding, and relatedness would have much less impact on these methods than on model-based methods. However, overall, model-based methods are preferred because they are more accurate and produce inferences that are biologically meaningful if the model assumptions (e.g. no inbreeding and no close relatives in the samples) are not much violated.

16.4.4 Inferring Levels of Recent Gene Flow

In a conservation setting there is usually interest in identifying whether populations of interest are currently exchanging migrants. A particular problem is to identify whether there is an ongoing problem of admixture, for example between domesticated and wild populations of a particular species. Thus, there is a need to identify the time-scale on which these processes are happening, and to distinguish between ancient and recent processes. As described in **Chapter 8**, there are many methods that have been developed for assessing gene flow on a number of different time-scales. This section will concentrate on methods that are more focused on recent time-scales, and have a conservation focus.

A progenitor of many of these methods is the study by Pritchard *et al.* (2000), whose admixture model – as previously explained – allows the inference of the proportions of an individual's genome (or ancestry) coming from each of an assumed K populations. Therefore, conceptually an individual is from population X if its estimated ancestral coefficient is 1 for X and is 0 for each of the remaining $K - 1$ populations. Otherwise, the individual is a putative hybrid. However, in practice it is difficult to determine with confidence whether an individual is a hybrid or not, and which hybrid class (say, F1, F2, B1, B2, ...) it belongs to if it is a hybrid. The problem is due to sampling errors of markers as well as inference uncertainties.

Taking this approach further, it is possible to use multilocus genotypes to infer parameters in a demographic model, such as the current levels of gene flow between populations. The model of Wilson and Rannala (2003) assumes that all populations in the system have been sampled, and that the individuals each has at most one immigrant ancestor. The assumption is realistic and works well when hybridization is rare; otherwise the method becomes powerless anyway because populations are little differentiated with high migration rates (see below). The data consist of the genotypes (X) and sampling locations of individuals (S), and the parameters are the population source of the immigrant ancestor for each individual (M), the number of generations back in the past that the immigrant ancestor occurred (t), the probability that two alleles in an individual are IBD from a recent ancestor (F), leading to a departure from Hardy–Weinberg

equilibrium, and the population allele frequencies (\mathbf{p}). The likelihood of the parameters ($\mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}$) is the probability of the data (\mathbf{X}, \mathbf{S}) given the parameters,

$$P(\mathbf{X}, \mathbf{S} | \mathbf{M}, \mathbf{t}, \mathbf{F}, \mathbf{p}) = \prod_{i=1}^n \prod_{j=1}^J P(X_{ij} | S_i, M_i, t_i),$$

where subscripts i and j index individuals ($1, \dots, n$) and loci ($1, \dots, J$) for a sample size of n individuals and J loci.

The prior for \mathbf{M} and \mathbf{t} for each individual can then be written as a function of the immigration rate, m_{lq} , the proportion of individuals in population q that are immigrant from population l . It is this hierarchical parameter that is of most interest. Given this prior and priors for other parameters in the model, MCMC is then used to obtain the posterior distribution of the parameters. Thus, the method can identify the immigrant ancestry of particular individuals, but it also, for example, constructs a matrix with point estimates of current migration rates between populations. The method was implemented in a computer program BayesAss (Wilson and Rannala 2003).

An example where this approach has been used is in the study of movement between orang-utan populations in Borneo (Goossens *et al.*, 2006). In this case the model of Wilson and Rannala (2003) was used to demonstrate that there was very little current gene flow between populations on two sides of a river. Goossens *et al.* note that in more general comparisons the method appeared to converge well when the migration rate is low, but convergence was more problematic on data sets with higher migration rates. Wang (2014) compared the method of Wilson and Rannala (2003) with a method based on parentage assignment, MigEst, in terms of accuracy in estimating migration rate m from simulated multilocus genotype data. He concluded that the population assignment method (Wilson and Rannala, 2003) is accurate and powerful only when populations are highly differentiated, which occurs in the scenario of strong drift (i.e. N_e small) and weak migration (i.e. m small). Otherwise, parentage assignments yield better estimates of m .

Another detailed evaluation of BayesAss has been carried out by Faubet *et al.* (2007), who simulated populations under a variety of demographic assumptions, with a finite allele model of genetic variation. They concluded that the method could moderately reliably estimate migration rates for populations whose levels of differentiation $F_{ST} > 0.05$, provided the assumptions of the model were met. In particular, scenarios in which the population gene frequencies varied over time due to drift and migration in Wright–Fisher simulations led to inaccuracies, as did scenarios where the migration rate was high, but it was still often possible to obtain good estimates of migration rates provided $F_{ST} > 0.1$ and migration rate $m < 0.01$. They noted that a particular problem with the method was lack of convergence in the MCMC simulation, assessed by examining the posterior distributions of several runs with different starting conditions. Even very long runs still led to apparent convergence issues. These points are also highlighted in Meirmans (2014); they carried out a survey of 100 studies that had used BayesAss, and showed clearly that there were strong peaks in the distribution of estimates of the proportion of non-migrants at the prior boundaries of 2/3 and 1, consistent with the behaviour seen in non-converged runs. The main recommendations of Faubet *et al.* (2007) and Meirmans (2014) include that few populations and many individuals should be sampled, scenarios with low levels of differentiation should be avoided, and many independent runs should be made.

One approach to avoiding the assumption that gene frequencies remain unchanged over the last few generations is suggested by Broquet *et al.* (2009), following Vitalis (2002), who observed that for many organisms it is possible to take two samples within one generation – the first taken when individuals are in their natal population before dispersal, followed by a post-dispersal sample. In contrast to the method of Vitalis (2002), who developed a method

based on within-locus F -statistics, Broquet *et al.* (2009) used the likelihood-based modelling formalism of STRUCTURE and BayesAss, solved by MCMC. They implemented their model, essentially similar to BayesAss, within the BUGS (Bayesian inference Using Gibbs Sampling; Lunn *et al.*, 2009) software, via an R interface, illustrating that STRUCTURE-like genotypic models are generally straightforward to implement in high-level MCMC modelling languages. The essential difference from BayesAss is that there are two sets of genotype frequencies, pre and post dispersal, g and G , and two corresponding vectors with location labels, \mathbf{z} and \mathbf{Z} , and within the model there is an additional latent vector $\tilde{\mathbf{z}}$ for the pre-dispersal locations of individuals labelled in \mathbf{Z} . The method, implemented in the software IMIG, aims to infer $\tilde{\mathbf{z}}$ and the associated migration matrix. By using both pre- and post-dispersal samples, the method is expected to extract more information about migration from the data and thus to yield better migration rate estimates. An example study for which this method is applied, and compared with BayesAss and STRUCTURE, is that of Dussex *et al.* (2016), applied to populations of the greater white-toothed shrew (*Crocidura russula*). Broadly, the results from simulation experiments in Broquet *et al.* (2009) and the empirical study in Dussex *et al.* (2016) suggest similar performance of all three methods – although the IMIG software avoids sensitivity to assumptions about dispersal rates and recent drift, a drawback is the sensitivity to the assumption that \mathbf{z} and \mathbf{Z} are observed correctly. A further extension of the two-sample approach is to take samples over multiple generations, and explicitly take into account the effects of drift. Wang and Whitlock (2003), as described in more detail in Section 16.2.4, showed that it is potentially feasible to obtain an estimate of current migration rates from continually monitored populations by examining the fluctuation in gene frequencies under drift in populations that exchange migrants.

A STRUCTURE-like program that has been widely used in conservation and management is NEWHYBRIDS (Anderson and Thompson, 2002), which can be used to identify hybrid individuals in a certain hybrid class (e.g. F1, B1,...) and to infer the proportion of hybrids in a sample. In their method, they consider hybridization over n generations. It is then possible to partition genotype frequencies into classes corresponding to the pedigree of the individuals – for example, when $n = 2$, the typical usage, whether they are purely of one species or the other, or F1 hybrids, or F2 hybrids, or backcrosses. For each individual, the posterior probability of belonging to one of these genotype frequency classes can be computed using MCMC, and hence non-admixed individuals can be detected, given admixture for n generations. A good example application is in the identification of hybridization levels in Swiss wildcat (*Felis silvestris silvestris*) populations using a panel of SNP markers (Nussberger *et al.*, 2013). An empirical Bayes method was also developed to partition sampled individuals into different hybrid classes, allowing for locus-specific dropout rates and other genotyping error rates (Mondol *et al.*, 2015). It relies on the reference allele frequencies estimated from the reference (purebred) individuals identified by STRUCTURE.

As a final note on migration rate estimation, it is worth recalling that there is a connection between migrations rates estimated by some of the methods described in this section and other parameters discussed earlier in this chapter, namely effective population size N_e and F_{ST} . The migration rate estimates from the method of Wilson and Rannala (2003) need not be closely correlated with, for example, pairwise F_{ST} , both because the latter takes some time to equilibrate and also because in a migration model F_{ST} is a function of the product of the effective population size, N_e , and migration rate m . However, it should be possible to obtain estimates of N_e and m given information on the genetic divergence between populations. An early study in this regard was by Vitalis and Couvet (2001) who developed a method-of-moments estimator to infer effective population size and migration rate into a single focal population using single-locus and two-locus functions of identity by state, assuming an infinite island model and an accurate estimate of the gene frequency in the migrant population. Their single-locus statistic is similar to Weir and Cockerham's F_{ST} , and has the same expectation: $F_{ST} = 1/(1 + 4N_e m)$. Their

two-locus statistic is an estimator of the covariance between pairs of loci in the (single-locus) probability of non-identity. They show that there is a unique mapping between the expectations of their two estimators estimates and N_e and m can be obtained by solving a pair of simultaneous equations for the one- and two-locus statistics.

16.4.5 Landscape Genetics

The above-described clustering methods (e.g. Pritchard *et al.* 2000) assign individuals to populations and estimate the number of populations from genetic marker data. In doing so, they do not use spatial data, such as geographical locations of the sampled individuals or the landscape of the habitat of the sampled individuals. They also do not attempt to link the clustering results with the landscape, such as a river or a mountain ridge. However, in molecular ecology and conservation biology, we want to know not only the population structure, but also the causes of the structure. Frequently, population genetic structure and the landscape of the habitat of the population are closely linked. These considerations have motivated the development of methods that identify the number of groups through model selection and also incorporate explicit spatial information into the models. These spatial models tend to borrow heavily from the techniques of geostatistics.

One example of this is the study by Guillot *et al.* (2005), who proposed a model similar to that of Pritchard *et al.* (2000), with K populations, and with the aim of inferring K . However, the prior for K is structured around a Voronoi tessellation. Here, a series of points, the number, n , of which is drawn from a Poisson distribution with parameter λ , are located uniformly at random over a rectangle, covering the geographic area of interest. Around each point it is possible to draw a convex polygon that contains the region of the rectangle that is closer to that point than to any other point. The rectangle can be divided into m such non-overlapping regions. This is a Voronoi tessellation, and, given K populations, each tile of the tessellation is assumed to be assigned uniformly at random to one of the K populations. An example illustrating the model is shown in Figure 16.2, taken from the original work of Guillot *et al.* (2005). One implication of this model is that, although the tessellation provides a spatial element, it only specifies loosely,

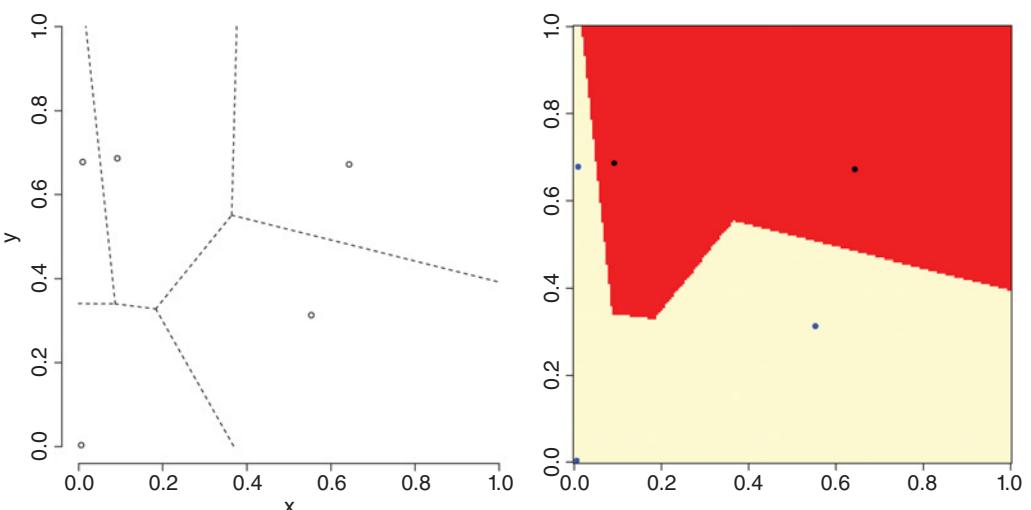


Figure 16.2 Random tessellation of a unit square into two spatial domains through a coloured Voronoi tiling. Left: realization of a Poisson point process with Voronoi tessellation induced. Right: partition obtained after union of tiles belonging to the same population (coded as two colours). From Guillot *et al.* (2005).

via the interaction of K and λ , how the space is subdivided. Thus, for example, if K is 2 and λ is high the two populations are distributed as a mosaic through the region, whereas if λ is low the region is likely to be dominated by two main blocks. In contrast to the method of Pritchard *et al.* (2000), reversible jump MCMC (Green, 1995) is used to infer the posterior distribution of K . The method appears able to recover accurately spatial genetic discontinuities in simulated data sets. In an application of this method (implemented in the program Geneland) to roe deer populations in France (Coulon *et al.*, 2006), the method found weak evidence of spatial discontinuity into two populations, north and south of a region including a highway, canal, and river running close together. In contrast, analysis with STRUCTURE suggested that there was only one population.

An alternative spatial model, specifically for assigning individuals to their geographic origin (also possible with Geneland), is presented in Wasser *et al.* (2004), which includes an application of their method to the problem of locating the geographic origin in Africa of elephants from their DNA samples. Theirs is primarily an assignment method, similar to that of Rannala and Mountain (1997). However, they incorporate a spatial model for the population allele frequencies. Importantly, they also explicitly allow for genotyping errors in the microsatellite data that they analyse. The frequencies f_{jlk} for allele j at locus l in population k are modelled logically $f_{jlk}(\theta) = \exp(\theta_{jlk}) / (\sum_{j'} \exp(\theta_{j'lk}))$, and the θ is modelled by a Gaussian process. A Gaussian process is one that generates random variables, any sample n of which are drawn from an n -dimensional multivariate Gaussian distribution (and hence any linear combination of these random variables is also Gaussian). The mean vector and covariance matrix of this distribution are determined by the problem under consideration. For spatial problems, as here, this is more generally called a Gaussian random field. In the case of the spatial model here, for the same allele at the same locus in two different locations k and k' the covariance depends on d the distance between the two locations, and they use the function given by $(1/\alpha_0) \exp[-(\alpha_1 d)^{\alpha_2}]$. Wasser *et al.* (2004) further extend the model by allowing a location W to be specified so that genetic samples can be assigned to locations that are not specified in advance. This is implemented using MCMC, allowing the posterior distribution of W to be computed. With this method Wasser *et al.* analysed samples from 399 elephants, and demonstrated that they were able to make accurate spatial assignments. The implication of this study is that, in future, DNA from ivory samples could be used to determine their provenance, and thereby help control the trade in ivory.

In a landscape genetics context, spatially localized ‘effective migration rates’ can be estimated by the EEMS software (Estimated Effective Migration Surfaces; Petkova *et al.*, 2016). Petkova *et al.* (2016) use the term ‘effective’ migration because their model makes assumptions (such as equilibrium in time) that may preclude interpreting effective migration as representing historical rates of gene flow. In this method, the spatial distribution of individuals is modelled by mapping the samples onto a dense triangular grid covering the area of interest. The method aims to infer the effective migration rate for the edges connecting the nodes in the grid. The method uses theory for the expected coalescence times between pairs of samples in a stepping-stone model, and uses MCMC to estimate the relevant parameters in the model. Applied to populations of elephants from Africa, surveyed at 16 microsatellite loci, the method is able to highlight the barrier to gene flow between the forest elephant (*Loxodonta cyclotis*) and the savannah elephant (*L. africana*), and also elements of the substructure within each species.

16.5 Deintrogression Strategies

As noted in Chapter 8, given a sufficiently dense set of SNP markers, either from an array or whole-genome genotyping, it is possible to identify tracts of genome that correspond

to introgressed segments of chromosome when genetically differentiated populations meet and hybridize. The interested reader is directed to **Chapter 8** for further details on these techniques (see also **Chapter 9** for methods for testing and identifying archaic introgression, such as Neanderthal introgression into modern humans), which have, for example, been used in studies of horses to assess if different horse breeds have admixed (Orlando *et al.* 2013). In conjunction with these developments a small group of studies have examined the potential effectiveness of ‘deintrogression’ strategies for ameliorating the effects of introgressive hybridization in populations of conservation concern, and these are briefly reviewed here. These authors define deintrogression as the method for removing introgressed genes (in a statistical sense, as explained below) by using selective breeding.

The essence of these approaches is to use the predicted level of individual admixture as a phenotype for artificial selection. For example, Amador *et al.* (2013) investigate a simulated scenario in which individuals are sampled from each of two separate and genetically diverged populations, and these are then admixed to varying degrees. They distinguished between focal (‘native’) and exogenous (‘non-native’) individuals. For each focal individual, they estimated the sum over all exogenous individuals of the molecular coancestry coefficients between the focal individual and each exogenous individual. Specifically, for both focal and exogenous samples they computed a standardized genotype for each SNP/individual,

$$x_{ij} = \frac{g_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}},$$

where p_j is the base population frequency of the reference allele, and g_{ij} is the genotype of individual i for SNP j (scored 0,1,2 with respect to number of copies of the alternate allele). These form the elements of matrices \mathbf{X}_F and \mathbf{X}_E for the focal and exogenous individuals. The matrix of molecular co-ancestry coefficients between focal and exogenous individuals, \mathbf{A} , is then

$$\mathbf{A} = \frac{\mathbf{X}_F \mathbf{X}_E^T}{L},$$

for L SNPs. They then compute $a_i = \sum_{k=1}^{N_E} a_{ik}$ to obtain a score of the introgression level for the i th individual. To reduce the effects of inbreeding, they selected the minimum-ranking 10 males and 10 females to breed. This strategy generally works very well for one generation after the admixture event, even when the proportion of admixed individuals is close to 50%. Their experiments involved sample sizes of 100 individuals and 50,000 SNPs. Performance tails off for further generations of admixture, but for three generations of admixture, 10 generations of selection can still recover close to 100% the native genetic material, up to an admixture rate of 40%. For five generations of admixture, 10 generations of selection can recover close to 100% native genome for 20% admixture.

Further work, in Amador *et al.* (2014), investigates the improvement in efficiency when using haplotypic information from genomically contiguous SNPs. The earlier study (Amador *et al.*, 2013) simulated frequencies of SNPs evenly distributed in 20 linkage groups, each of 1 morgan, but did not specifically use haplotype information. Amador *et al.* (2014), rather than simulate baseline frequencies, used 50,000 SNP array data from two sheep breeds (Merino and Poll Dorset), and then admixed them *in silico*, knowing the recombination map. They took two approaches to predict the breed of origin. In the first method, they divided up the markers into segments of 10 consecutive SNPs. For the j th segment they computed a probability b_{Mj} that the segment was of Merino origin,

$$b_{Mj} = \frac{p_{Mj}}{p_{Mj} + p_{Dj}},$$

for haplotype frequencies p_{Mj} and p_{Dj} in the respective breeds. Their classification rule was

$$\begin{aligned} b_{Mj} < 0.4 &\rightarrow \text{Poll Dorset}, \\ 0.4 < b_{Mj} < 0.6 &\rightarrow \text{unclassified}, \\ b_{Mj} > 0.6 &\rightarrow \text{Merino}. \end{aligned}$$

This rule was performed for each SNP, and the individual admixture proportion (proportion Merino) was taken as the proportion of haplotypes that were classified as Merino. Thus, this method requires that the classification rule is trained with the known breed information before being applied in their simulations of introgression. In the second method, they fit the following mixed model (genomic best linear unbiased prediction (GBLUP)) with the known breed information:

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{Zg} + \mathbf{e},$$

where \mathbf{y} is the proportion of Merino, μ is the intercept, \mathbf{Z} is a matrix with coefficients to be fitted, with $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. Given the relationship matrix \mathbf{G} , estimated by the molecular coancestry (computed as for matrix \mathbf{A} in the Amador *et al.*, 2013 study), and the known breed values (scored as 1 for Merino and 0 for Poll Dorset), and assuming heritability $\sigma_g^2/(\sigma_g^2 + \sigma_e^2) = 0.99$, they are able to fit μ and \mathbf{Z} , which can then be used to predict the admixture proportions of individuals in their simulations of introgression. As with their earlier study, they generally found good performance with one generation of admixture, with lower performance with three and five generations of admixture. Both the GBLUP and haplotypic methods performed similarly, and appeared to be an improvement on the method described in the Amador *et al.* (2013) study, although no direct comparison was made on the same data.

A drawback of all the methods described is the need for predefined pure groups to train the methods. However, Amador *et al.* (2014) note that one could train the GBLUP method with any \mathbf{y} , not necessarily pure, as long as the degree of admixture was known. Methods, such as those described in **Chapter 8**, for identifying admixed tracts, could be useful for the design of *ad hoc* breeding programs. However, it is reassuring that, at least for moderate levels of admixture, the general approaches here could potentially be useful in deintroduction strategies.

16.6 Genetic Species Delimitation

There has been an increasing interest in using genetic information to ‘discover’ and delimit new species, particularly with advent of DNA sequencing data (although the use of genetic methods to distinguish species goes back to the era of allozymes). These methods are used to discover new species by some form of clustering and classification algorithm. The term ‘delimitation’ refers to the rules that are used to break a group into its component species (cryptic species). Mitochondrial sequence DNA has been used to form ‘barcodes’ with applications in both species identification (Hebert *et al.*, 2003), and species discovery (Hickerson *et al.*, 2006). The difference between identification and discovery is that in the case of the former the existence of the species is known *a priori* and barcoding is used to label individuals according to their species. In the latter case barcoding is used to split what was previously considered to be a single species, into component cryptic species. The use of barcoding for species identification is still widely used as a shorthand route for field and forensic taxonomy (Mishra *et al.*, 2016), and also underlies metagenomic methods (Andújar *et al.*, 2015). The use of barcoding for species discovery has increasingly been questioned (Yang and Rannala, 2017) because there has been

a subsequent appreciation of the causes of discrepancy between gene trees and species trees, and among the gene trees at different parts of the genome (Rannala and Yang, 2003; Degnan and Rosenberg, 2009).

Although barcoding for species identification has itself led to interesting statistical developments (Matz and Nielsen, 2005), this section concentrates on model-based methods for species delimitation using multilocus genomic data. For single loci, a method based on constructing gene trees under a mixture of coalescent and Yule process priors has been studied (Pons *et al.*, 2006). They make the assumption, forwards in time up to time T before the present, that the genealogy branches according to a Yule process (constant rate of branching), and that from T until the present there is coalescent process. The critical time T is estimated by maximum likelihood, and species are then defined as those that derive from a single lineage crossing this critical time.

More specific methods for multiple loci are based on the theoretical distribution of genealogies expected under a demographic history of populations that periodically split apart without gene flow, termed the multispecies coalescent (Degnan and Rosenberg, 2009; for more details about the multispecies coalescent, see Chapter 7, this volume). The basic aim of the methods (Yang and Rannala, 2010, 2014), summarized in Rannala (2015), is to compute the posterior distribution of the delimitation model, M , given genomic data, \mathbf{D} ,

$$f_{\beta}(M|\mathbf{D}) \propto \int_{\Omega} \int_{\mathbf{G}} f(\mathbf{G}|M, \Omega) f(M|\Omega) f(\mathbf{D}|\mathbf{G}, \Omega) f_{\beta}(\Omega),$$

with model/prior parameters Ω , and fixed hyperprior parameters β . The genomic data \mathbf{D} is modelled to be generated from the set of latent genealogies \mathbf{G} : $f(\mathbf{D}|\mathbf{G}, \Omega)$. The genealogies are generated under the multi-species coalescent $f(\mathbf{G}|M, \Omega)$. It is assumed that the genomic data consists of sets of sequences (loci) within which there is no recombination. The unknown genealogies are integrated out using the MCMC techniques developed in Rannala and Yang (2003). The delimitation model, including the species tree, is of varying dimension, depending on which nodes are collapsed. Reversible-jump MCMC is used to traverse the set of topologies of the tree. The intuition is that a move that splits one population (either current or ancestral) into two would be favoured if the resulting independent set of coalescences for a genealogy makes the genomic data more probable. Similarly, a move that joins two population lineages implies that a single-population genealogy is more likely.

Convergence in reversible-jump MCMC is notoriously problematic, and in the first publication in this series Yang and Rannala (2010) constrain the set of possible species tree topologies by specifying a guide tree. For example (Figure 16.3), with three populations a , b , and c , a guide tree might be that in the full delimitation case the tree would be $((a, b), c)$, so the possible delimitations are (abc) [no delimitation], $((ab), c)$, or, the same as the guide tree, $((a, b), c)$. Their priors for these three models would be respectively $1/3$, $1/3$, $1/3$ (with more taxa, these may not be

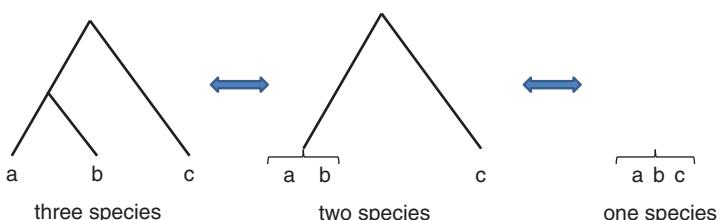


Figure 16.3 Example of species delimitation using a guide tree. The guide tree is shown on the left, fully delimiting the three populations as $((a, b), c)$. Successive grouping of populations gives $((ab), c)$ and (abc) . Reversible-jump MCMC is used to flip between the three possible states, conditional on the guide tree.

uniform because their prior is uniform on the set of labelled histories, which takes into account the temporal order of splitting of different pairs of populations). With this constraint, it is then possible to obtain the posterior probability of all delimitations, given by collapsed nodes. This is implemented in the BPP program. Although the basic principle of the method is straightforward, many technicalities in implementation are needed to ensure good convergence. Rannala and Yang (2013) describe improvements in the prior specification and reversible-jump MCMC algorithm, and Yang and Rannala (2014) introduce further improvements to BPP that eliminate the need for a guide tree, and therefore allow the MCMC algorithm to sample all possible species tree topologies, including collapsed branches, given some genomic data. The main improvement is in the adaptation of a MCMC move, the nearest-neighbour interchange algorithm, so that a change in the species tree includes a modification of the gene trees at multiple loci. The problem being solved here is common generally in MCMC-based treatments of hierarchical Bayesian models, where an update to a parameter needs an update to a hyperparameter to make it more likely to be accepted. To illustrate the method, they give an example from Leaché *et al.* (2009), who studied populations of the coast horned lizard (*Phrynosoma*) in California. The samples were originally split by mitochondrial clades into five groups on a north–south cline (Northern and Southern California; Northern, Central, Southern Baja California: NC, SC, NBC, CBC, SBC), and the question of interest is how many species there are, and what are their historical relationships. On the basis of various criteria Leaché *et al.* (2009) concluded that there are three species ((NC SC NBC), CBC, SBC) – with NC, SC, and NBC grouped into a single species. The data analysed by Rannala and Yang consist of two nuclear loci (of 529 and 1100 bp respectively), genotyped for 130 sequence copies. Rannala and Yang obtained samples from the posterior distribution of phylogenies (including collapsed nodes). The 95% credible set of phylogenies included five-species and four-species phylogenies, with five-species phylogenies having 76% posterior probability. The most probable tree contained the (NC, SC, NBC) clade found in Leaché *et al.*, but all lineages were separated under the model. Furthermore, based on mtDNA, Leaché *et al.* proposed a ((NC, SC), NBC) topology for this group, whereas the most probable grouping in Rannala and Yang's analysis based on nuclear data was (NC, (SC, NBC)). The latter discrepancy is easily accommodated by incomplete lineage sorting; however, the difference between the two studies in the degree of species delimitation (three species versus five) raises the question of how long populations need to be diverged before they are counted as separate species. These populations are of conservation concern, as described in Leaché *et al.*, and this issue lies at the heart of many debates about conservation of species (Sukumaran and Knowles, 2017). Similar questions have been raised about the tendency to subdivide species based on the results of STRUCTURE analyses (Bercovitch *et al.*, 2017).

16.7 Conclusions and Outlook

Estimation of effective population size from genetic marker data has progressed a long way from the initial studies by Krimbas and Tsakas (1971), and it is now possible, particularly with genome-scale data, to obtain accurate estimates of N_e , even from single samples. Over the last 10 years there has also been growing appreciation of the ability to estimate census size based on genotypic information. Similarly, we now have a much improved ability to estimate gene flow and model connectivity of populations. It should be noted that despite these great improvements in estimation, there is still some uncertainty in how these estimated parameters feed into a useful and predictive viability model for populations of conservation concern (Hoffman *et al.*, 2017; Wood *et al.*, 2016; Frankham *et al.*, 2014), and this is an area that needs further development.

With respect to the specific themes of this chapter, areas that could be improved include the joint estimation of demographic and life-history parameters as well as N_e using LD-based methods. Similarly, a temporal method for multiple populations with gene flow is yet to be developed. The application of methods for inferring remote relatives from genomic data is also an area that could be explored in a conservation context for estimating N_e . With the dramatic improvement in genotyping technology we will be able to sharpen the estimates of all relevant parameters through the use of extensive genomic data. Although the primary focus in this chapter has been on likelihood-based statistical methods, moment-based and nonparametric methods may be seen as increasingly attractive in future analyses for computational reasons, as is evident in many studies of human demography (Patterson *et al.*, 2012). However, the strength of Bayesian and likelihood-based methods lies in their flexibility, and their focus on a model-based and falsifiable framework (Gelman and Shalizi, 2013). From the perspective of next-generation sequencing, likelihood-based methods have been shown to be useful for working with low-depth sequencing data (Korneliussen *et al.*, 2014; Skotte *et al.*, 2013), as exemplified in the recent analysis of zebra population structure (Pedersen *et al.*, 2018). An additional consequence of the widespread introduction of next-generation sequencing is that it has been necessary to abandon the assumption that genetic markers can be treated as unlinked and approximately independent, leading to an increase in the application of composite-likelihood methods (Larriba and Fearnhead, 2011).

Acknowledgements

We would like to thank Ida Moltke for her very helpful editing of this chapter and Robin Waples and a reviewer for useful comments on an earlier draft.

References

- Albrechtsen, A., Korneliussen, T.S., Moltke, I., van Overeem Hansen, T., Nielsen, F.C. and Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology* **33**, 266–274.
- Alexander, D.H., Novembre, J. and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664.
- Amador, C., Fernández, J. and Meuwissen, T.H. (2013). Advantages of using molecular coancestry in the removal of introgressed genetic material. *Genetics Selection Evolution* **45**, 13.
- Amador, C., Hayes, B.J. and Daetwyler, H.D. (2014). Genomic selection for recovery of original genetic background from hybrids of endangered and common breeds. *Evolutionary Applications* **7**, 227–237.
- Anderson, E.C. (2005). An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent-based likelihood. *Genetics* **170**, 955–967.
- Anderson, E.C. and Dunham, K.K. (2008). The influence of family groups on inferences made with the program Structure. *Molecular Ecology Resources* **8**, 1219–1229.
- Anderson, E.C. and Thompson, E.A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**, 1217–1229.
- Anderson, E.C., Williamson, E.G. and Thompson, E.A. (2000). Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics* **156**, 2109–2118.

- Andújar, C., Arribas, P., Ruzicka, F., Crampton-Platt, A., Timmermans, M.J. and Vogler, A.P. (2015). Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Molecular Ecology* **24**, 3603–3617.
- Baetscher, D.S., Clemento, A.J., Ng, T.C., Anderson, E.C. and Garza, J.C. (2018). Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Molecular Ecology Resources*, **18**, 296–305.
- Balloux, F., Brunner, H., Lugon-Moulin, N., Haussser, J. and Goudet, J. (2000) Microsatellites can be misleading: An empirical and simulation study. *Evolution*, **54**, 1414–1422.
- Barbato, M., Orozco-terWengel, P., Tapió, M. and Bruford, M.W. (2015). SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in Genetics* **6**, 109.
- Beaumont, M., Barratt, E.M., Gottelli, D., Kitchener, A.C., Daniels, M.J., Pritchard, J.K. and Bruford, M.W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology* **10**, 319–336.
- Beaumont, M.A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.
- Beaumont, M.A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379–406.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.
- Bercovitch, F.B., Berry, P.S.M., Dagg, A., Deacon, F., Doherty, J.B., Lee, D.E., Mineur, F., Muller, Z., Ogden R., Seymour, R. and Shorrocks, B. (2017). How many species of giraffe are there? *Current Biology* **27**, R136–R137.
- Berthier, P., Beaumont, M.A., Cornuet, J.M. and Luikart, G. (2002). Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: A genealogical approach. *Genetics* **160**, 741–751.
- Bollback, J.P., York, T.L. and Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**, 497–502.
- Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. and Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261–3273.
- Broquet, T., Yearsley, J., Hirzel, A.H., Goudet, J. and Perrin, N. (2009). Inferring recent migration rates from individual genotypes. *Molecular Ecology* **18**, 1048–1060.
- Browning, S.R. and Browning, B.L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American Journal of Human Genetics* **97**, 404–418.
- Burren, A., Signer-Hasler, H., Neuditschko, M., Tetens, J., Kijas, J., Drögemüller, C. and Flury, C. (2014). Fine-scale population structure analysis of seven local Swiss sheep breeds using genome-wide SNP data. *Animal Genetic Resources* **55**, 67–76.
- Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity* **73**, 657–679.
- Carreras-Carbonell, J., Macpherson, E. and Pascual, M. (2006). Population structure within and between subspecies of the Mediterranean triplefin fish *Tripterygion delaisi* revealed by highly polymorphic microsatellite loci. *Molecular Ecology* **15**, 3527–3539.
- Cockerham, C.C. (1969). Variance of gene frequencies. *Evolution* **23**, 72–83.
- Cockerham, C.C. (1973). Analysis of gene frequencies. *Genetics* **74**, 679–700.
- Corander, J., Waldmann, P. and Sillanpää, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

- Corbin, L.J., Liu, A.Y.H., Bishop, S.C. and Woolliams, J.A. (2012). Estimation of historical effective population size using linkage disequilibria with marker data. *Journal of Animal Breeding and Genetics* **129**, 257–270.
- Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000.
- Cornuet, J.M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J.M., Balding, D.J., Guillemaud, T. and Estoup, A. (2008). Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**, 2713–2719.
- Coulon, A., Guillot, G., Cosson, J.F., Angibault, J.M.A., Aulagnier, S., Cargnelutti, B., Galan, M. and Hewison, A.J.M. (2006). Genetic structure is influenced by landscape features: Empirical evidence from a roe deer population. *Molecular Ecology* **15**, 1669–1679.
- Creel, S., Spong, G., Sands, J.L., Rotella, J., Zeigle, J., Joe, L., Murphy, K.M. and Smith, D. (2003). Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology* **12**, 2003–2009.
- Crow, J.F. and Denniston, C. (1988). Inbreeding and variance effective population numbers. *Evolution* **42**, 482–495.
- Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- Dawson, K.J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59–77.
- Degnan, J.H. and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24**, 332–340.
- Dussex, N., Broquet, T. and Yearsley, J.M. (2016). Contrasting dispersal inference methods for the greater white-toothed shrew. *Journal of Wildlife Management* **80**, 812–823.
- Eggert, L.S., Eggert, J.A. and Woodruff, D.S. (2003). Estimating population sizes for elusive animals: The forest elephants of Kakum National Park, Ghana. *Molecular Ecology* **12**, 1389–1402.
- Evanno, G., Regnaut, S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* **14**, 2611–2620.
- Ewens, W.J. (2004). *Mathematical Population Genetics: Theoretical Introduction*, vol. 1. Springer, New York.
- Falush, D., Stephens, M. and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587.
- Falush, D., Stephens, M. and Pritchard, J.K. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Resources* **7**, 574–578.
- Faubet, P., Waples, R.S. and Gaggiotti, O.E. (2007). Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular Ecology* **16**, 1149–1166.
- Foll, M., Shim, H. and Jensen, J.D. (2015). WFABC: A Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources* **15**, 87–98.
- Frankham, R., Bradshaw, C.J. and Brook, B.W. (2014). Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biological Conservation*, **170**, 56–63.
- Gao, H., Williamson, S. and Bustamante, C.D. (2007). A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* **176**, 1635–1651.

- Gelman, A. and Shalizi, C.R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, **66**, 8–38.
- Gilbert, K.J. and Whitlock, M.C. (2015). Evaluating methods for estimating local effective population size with and without migration. *Evolution* **69**, 2154–2166.
- Goossens, B., Chikhi, L., Ancrenaz, M., Lackman-Ancrenaz, I., Andau, P. and Bruford, M.W. (2006). Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biology* **4**, 285–291.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo and Bayesian model determination. *Biometrika* **82**, 711–732.
- Guillot, G., Estoup, A., Mortier, F. and Cosson, J.F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.
- Hardy, O.J. and Vekemans, X. (2002). SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Resources* **2**, 618–620.
- Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics* **9**, e1003521.
- Hayes, B.J., Visscher, P.M., McPartlan, H.C. and Goddard, M.E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13**, 635–643.
- Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. (2003). Biological identification through DNA barcodes. *Proceedings of the Royal Society of London B* **270**, 313–321.
- Hedrick, P.W. (2005). A standardized genetic differentiation measure. *Evolution* **59**, 1633–1638.
- Hickerson, M.J., Meyer, C.P. and Moritz, C. (2006). DNA barcoding will often fail to discover new animal species over broad parameter space. *Systematic Biology* **55**, 729–739.
- Hilborn, R. (2002). The dark side of reference points. *Bulletin of Marine Science* **70**, 403–408.
- Hill, W.G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**, 209–216.
- Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.
- Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., Poss, M.L., Reed, L.K., Storfer, A. and Whitlock, M.C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *American Naturalist* **188**, 379–397.
- Hoffmann, A.A., Sgrò, C.M. and Kristensen, T.N. (2017). Revisiting adaptive potential, population size, and conservation. *Trends in Ecology & Evolution*, **32**, 506–517.
- Hollenbeck, C.M., Portnoy, D.S. and Gold, J.R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity* **117**, 207–216.
- Hubisz, M.J., Falush, D., Stephens, M. and Pritchard, J.K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322–1332.
- Jombart, T., Devillard, S. and Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.
- Jones, O.R. and Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* **10**, 551–555.
- Jorde, P.E. and Ryman, N. (1995). Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics* **139**, 1077–1090.
- Jorde, P.E. and Ryman, N. (2007). Unbiased estimator for genetic drift and effective population size. *Genetics* **177**, 927–935.
- Jost, L. (2008). G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**, 4015–4026.

- Kalinowski, S.T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: Simulations and implications for human population structure. *Heredity* **106**, 625–632.
- Kijas, J.W., Lenstra, J.A., Hayes, B.J., Boitard, S., Porto Neto, L.R., San Cristobal, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology* **10**, e1001258.
- Korneliussen, T.S. and Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* **31**, 4009–4011.
- Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356.
- Krimbas, C.B. and Tsakas, S. (1971). The genetics of *Dacus oleae* V. Changes of esterase polymorphism in a natural population following insecticide control: Selection or drift? *Evolution* **25**, 454–460.
- Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica* **21**, 43–69.
- Laval, G., SanCristobal, M. and Chevalet, C. (2003). Maximum-likelihood and Markov chain Monte Carlo approaches to estimate inbreeding and effective size from allele frequency changes. *Genetics* **164**, 1189–1204.
- Leaché, A.D., Koo, M.S., Spencer, C.L., Papenfuss, T.J., Fisher, R.N. and McGuire, J.A. (2009). Quantifying ecological, morphological, and genetic variation to delimit species in the coast horned lizard species complex (*Phrynosoma*). *Proceedings of the National Academy of Sciences* **106**, 12418–12423.
- Lowe, W.H. and Allendorf, F.W. (2010). What can genetics tell us about population connectivity? *Molecular Ecology*, **19**, 3038–3051.
- Luikart, G. and Cornuet, J.M. (1999). Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* **151**, 1211–1216.
- Luikart, G., Ryman, N., Tallmon, D.A., Schwartz, M.K. and Allendorf, F.W. (2010). Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. *Conservation Genetics* **11**, 355–373.
- Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* **28**, 3049–3067.
- Mathieson, I. and McVean, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984.
- Matz, M.V. and Nielsen, R. (2005). A likelihood ratio test for species membership based on DNA sequence data. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1969–1974.
- Meirmans, P.G. (2014). Nonconvergence in Bayesian estimation of migration rates. *Molecular Ecology Resources* **14**, 726–733.
- Meirmans, P.G. and Hedrick, P.W. (2011). Assessing population structure: FST and related measures. *Molecular Ecology Resources* **11**, 5–18.
- Mezzavilla, M. and Ghirotto, S. (2015). Neon: An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. *Journal of Computer Science & Systems Biology* **8**, O37–O44.
- Mishra, P., Kumar, A., Nagireddy, A., Mani, D.N., Shukla, A.K., Tiwari, R. and Sundaresan, V. (2016). DNA barcoding: An efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology Journal* **14**, 8–21.
- Moltke, I. and Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics* **30**, 1027–1028.

- Mondol, S., Moltke, I., Hart, J., Keigwin, M., Brown, L., Stephens, M. and Wasser, S.K. (2015). New evidence for hybrid zones of forest and savanna elephants in Central and West Africa. *Molecular Ecology* **24**, 6134–6147.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321–3323.
- Nei, M. and Chesser, R. (1983). Estimation of fixation indices and gene diversities. *Annals of Human Genetics* **47**, 253–259.
- Nei, M. and Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640.
- Neophytou, C. (2014). Bayesian clustering analyses for genetic assignment and study of hybridization in oaks: Effects of asymmetric phylogenies and asymmetric sampling schemes. *Tree Genetics & Genomes* **10**, 273–285.
- Nussberger, B., Greminger, M.P., Grossen, C., Keller, L.F. and Wandeler, P. (2013). Development of SNP markers identifying European wildcats, domestic cats, and their admixed progeny. *Molecular Ecology Resources* **13**, 447–460.
- Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78.
- Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–354.
- Paetkau, D., Shields, G.F. and Strobeck, C. (1998). Gene flow between insular, coastal and interior populations of brown bears in Alaska. *Molecular Ecology* **7**, 1283–1292.
- Palamara, P.F., Lencz, T., Darvasi, A. and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics* **91**, 809–822.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* **192**, 1065–1093.
- Palstra, F.P. and Ruzzante, D.E. (2008). Genetic estimates of contemporary effective population size: What can they tell us about the importance of genetic stochasticity for wild population persistence? *Molecular Ecology* **17**, 3428–3447.
- Pearse, D.E. (2016). Saving the spandrels? Adaptive genomic variation in conservation and fisheries management. *Journal of Fish Biology* **89**, 2697–2716.
- Pedersen, C.-E.T., Albrechtsen, A., Etter, P.D., Johnson, E.A., Orlando, L., Chikhi, L., Siegismund, H.R. and Heller, R. (2018). A southern African origin and cryptic structure in the highly mobile plains zebra. *Nature Ecology & Evolution*, **2**, 491–498.
- Petkova, D., Novembre, J. and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics* **48**, 94–100.
- Pollak, E. (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**, 531–548.
- Pompanon, F., Bonin, A., Bellemain, E. and Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847–850.
- Pons, J., Barraclough, T.G., Gomez-Zurita, J., Cardoso, A., Duran, D.P., Hazell, S., Kamoun, S., Sumlin, W.D. and Vogler, A.P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* **55**, 595–609.
- Pradel, R. (1996). Utilization of capture-mark-recapture for the study of recruitment and population growth rate. *Biometrics* **52**, 703–709.
- Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

- Pudovkin, A.I., Zaykin, D.V. and Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144**, 383–387.
- Puechmaille, S.J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: Sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources* **16**, 608–627.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.
- Raj, A., Stephens, M. and Pritchard, J.K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589.
- Rannala, B. (2015). The art and science of species delimitation. *Current Zoology* **61**, 846–853.
- Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 9197–9201.
- Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.
- Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for Bayesian species delimitation. *Genetics* **194**, 245–253.
- Ringler, E., Mangione, R. and Ringler, M. (2015). Where have all the tadpoles gone? Individual genetic tracking of amphibian larvae until adulthood. *Molecular Ecology Resources* **15**, 737–746.
- Robertson, A. (1965). The interpretation of genotypic ratios in domestic animal populations. *Animal Production* **7**, 319–324.
- Rodríguez-Ramilo, S.T. and Wang, J. (2012). The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. *Molecular Ecology Resources* **12**, 873–884.
- Rodríguez-Ramilo, S.T., Toro, M.A. and Fernández, J. (2009). Assessing population genetic structure via the maximisation of genetic distance. *Genetics Selection Evolution* **41**, 49.
- Ryman, N. and Leimar, O. (2009). G_{ST} is still a useful measure of genetic differentiation – a comment on Jost's *D*. *Molecular Ecology* **18**, 2084–2087.
- Saura, M., Tenesa, A., Woolliams, J.A., Fernández, A. and Villanueva, B. (2015). Evaluation of the linkage-disequilibrium method for the estimation of effective population size when generations overlap: An empirical case. *BMC Genomics* **16**, 922.
- Schwartz, M.K., Luikart, G. and Waples, R.S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution* **22**, 25–33.
- Skotte, L., Korneliussen, T.S. and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**, 693–702.
- Sukumaran, J. and Knowles, L.L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*, **114**, 1607–1612.
- Sun, M., Jobling, M.A., Taliun, D., Pramstaller, P.P., Egeland, T. and Sheehan, N.A. (2016). On the use of dense SNP marker data for the identification of distant relative pairs. *Theoretical Population Biology* **107**, 14–25.
- Sved, J.A., Cameron, E.C. and Gilchrist, A.S. (2013). Estimating effective population size from linkage disequilibrium between unlinked loci: Theory and application to fruit fly outbreak populations. *PLoS ONE* **8**, e69078.
- Takahata, N. and Nei, M. (1984). F_{ST} and G_{ST} statistics in the finite island model. *Genetics* **107**, 501–504.
- Tallmon, D.A., Koyuk, A., Luikart, G. and Beaumont, M.A. (2008). COMPUTER PROGRAMS: onesamp: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources* **8**, 299–301.

- Traill, L.W., Brook, B.W., Frankham, R.R. and Bradshaw, C.J. (2010). Pragmatic population viability targets in a rapidly changing world. *Biological Conservation* **143**, 28–34.
- Vitalis, R. (2002). Sex-specific genetic differentiation and coalescence times: Estimating sex-biased dispersal rates. *Molecular Ecology* **11**, 125–138.
- Vitalis, R. and Couvet, D. (2001). Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157**, 911–925.
- Waits, L., Luikart, G. and Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: Cautions and guidelines. *Molecular Ecology* **10**, 249–256.
- Wang, J. (1996). Deviation from Hardy–Weinberg proportions in finite populations. *Genetical Research* **68**, 249–257.
- Wang, J. (2001). A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research* **78**, 243–257.
- Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics* **166**, 1963–1979.
- Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1395–1409.
- Wang, J. (2007). Parentage and sibship exclusions: Higher statistical power with more family members. *Heredity* **99**, 205–217.
- Wang, J. (2009). A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Molecular Ecology* **18**, 2148–2164.
- Wang, J. (2011). COANCESTRY: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources* **11**, 141–145.
- Wang, J. (2012). On the measurements of genetic differentiation among populations. *Genetics Research* **94**, 275–289.
- Wang, J. (2014). Estimation of migration rates from marker-based parentage analysis. *Molecular Ecology* **23**, 3191–3213.
- Wang, J. (2015). Does G_{ST} underestimate genetic differentiation from marker data? *Molecular Ecology* **24**, 3546–3558.
- Wang, J. (2016a). A comparison of single-sample estimators of effective population sizes from genetic marker data. *Molecular Ecology* **25**, 4692–4711.
- Wang, J. (2016b). Individual identification from genetic marker data: Developments and accuracy comparisons of methods. *Molecular Ecology Resources* **16**, 163–175.
- Wang, J. (2017). The computer program STRUCTURE for assigning individuals to populations: Easy to use but easier to misuse. *Molecular Ecology Resources* **17**, 981–990.
- Wang, J. and Caballero, A. (1999). Developments in predicting the effective size of subdivided populations. *Heredity* **82**, 212–226.
- Wang, J. and Santure, A.W. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**, 1579–1594.
- Wang, J. and Whitlock, M.C. (2003). Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* **163**, 429–446.
- Wang, J., Brekke, P., Huchard, E., Knapp, L.A. and Cowlishaw, G. (2010). Estimation of parameters of inbreeding and genetic drift in populations with overlapping generations. *Evolution* **64**, 1704–1718.
- Wang, J., Santiago, E. and Caballero, A. (2016). Prediction and estimation of effective population size. *Heredity* **117**, 193–206.
- Waples, R.K., Larson, W.A. and Waples, R.S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity* **117**, 233–240.

- Waples, R.S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379–391.
- Waples, R.S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics* **7**, 167–184.
- Waples, R.S. and Do, C. (2008). LdNe: A program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**, 753–756.
- Waples, R.S. and England, P.R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics* **189**, 633–644.
- Waples, R.S. and Yokota, M. (2007). Temporal estimates of effective population size in species with overlapping generations. *Genetics* **175**, 219–233.
- Waples, R.S., Antao, T. and Luikart, G. (2014). Effects of overlapping generations on linkage disequilibrium estimates of effective population size. *Genetics* **197**, 769–780.
- Wasser, S.K., Shedlock, A.M., Comstock, K., Ostrander, E.A., Mutayoba, B. and Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14847–14852.
- Weir, B.S. and Cockerham, C.C. (1984). Estimating F -statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.
- Weir, B.S. and Hill, W.G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.
- White, G.C. and Burnham, K.P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study* **46**(Supplement), 120–138.
- Whitlock, M.C. (2011). G_{ST} and D do not replace F_{ST} . *Molecular Ecology* **20**, 1083–1091.
- Wilberg, M.J. and Dreher, B.P. (2004). genecap: A program for analysis of multilocus genotype data for non-invasive sampling and capture-recapture population estimation. *Molecular Ecology Notes* **4**, 783–785.
- Williamson, E.G. and Slatkin, M. (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**, 755–761.
- Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**, 1177–1191.
- Wolf, D. E., Takebayashi, N. and Rieseberg, L.H. (2001). Predicting the risk of extinction through hybridization. *Conservation Biology* **15**, 1039–1053.
- Wood, J.L., Yates, M.C. and Fraser, D.J. (2016). Are heritability and selection related to population size in nature? Meta-analysis and conservation implications. *Evolutionary Applications*, **9**, 640–657.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.
- Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.
- Wright, S. (1965). The interpretation of population structure by F -statistics with special regard to systems of mating. *Evolution* **19**, 395–420.
- Wright, S. (1969). *Evolution and the Genetics of Populations, Vol. 2, The Theory of Gene Frequencies*. University of Chicago Press, Chicago.
- Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences* **107**, 9264–9269.
- Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution* **31**, 3125–3135.
- Yang, Z. and Rannala, B. (2017). Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Molecular Ecology* **26**, 3028–3036.

17

Statistical Methods for Plant Breeding

Ian Mackay,¹ Hans-Peter Piepho,² and Antonio Augusto Franco Garcia³

¹ IMplant Consultancy Ltd, Chelmsford, UK

² University of Hohenheim, Stuttgart, Germany

³ University of São Paulo, Piracicaba, Brazil

Abstract

In this chapter we highlight differences in the application of quantitative methods to plant breeding compared to their application in animals and humans. These originate from the very different and diverse mating systems, life histories and genome organisations found in plants compared to animals. In addition, it is common to replicate plant genotypes as clones, inbred lines, or F1 hybrids. This gives the breeder greater control experimentally over the precision with which genotypic values are estimated and reduces gain from incorporating information from relatives, as common in animal breeding. Plants show great plasticity in their response to environmental challenges, and genotype–environment interactions are typically larger than in animals. Consequently, there has been considerable emphasis on improving experimental design for field testing. The implementation of genomic selection and prediction in plants is becoming common, and there are opportunities for its incorporation into plant breeding programmes which differ from those in animals. Ultimately, however, the link between applications of statistical methods to plant and animal breeding remains the breeder's equation.

17.1 Introduction

Much of the current advance in statistical genetics and the resurgence of its application in plant breeding originated with work in animal and human genetics. This is particularly so for genomic prediction and selection (see **Chapter 28**), which is revolutionising the approach to animal breeding, and of genome-wide association mapping, which originated in human genetics (Bodmer, 1986; Risch and Merikangas, 1996) but has become routine for trait mapping in plants too (Huang and Han, 2014; Barabaschi *et al.*, 2016). Any delay in uptake of methods of statistical genetics prevalent elsewhere is not due to ignorance or lack of skills among plant researchers. Rather, plant breeding has a proprietary set of strengths and weaknesses, often not recognised outside the field, which influence the application of statistical methods, and in some cases have resulted in the development of different and novel approaches. These are sometimes specific to individual species: there is much greater variation in genetic systems in domesticated plants than in animals, many more species are domesticated, and research tends to be divided into small communities working on each.

This chapter reviews the application of statistical and quantitative genetics to plant breeding, highlighting differences from applications in animals and humans. We start on common ground by describing the central place of the breeder's equation in plant breeding and its role in assessing and optimising programme design and selection strategy. This is followed by descriptions and consequences of complications in plant breeding arising from, in turn, the diverse breeding systems of plants, the prevalence of polyploidy, and of polymorphic genomic rearrangements. A particular feature of plants compared to most domesticated animals is their plastic response to the environment and the importance of genotype–environment interactions. Methods used by plant researchers to study and model these are presented. Genomic selection will have greater impact on animal and plant breeding than any other area of quantitative genetics in the next decade. This is discussed comprehensively in **Chapter 28**. Here we discuss eight areas in which its application in plant breeding can vary from animal breeding. These are: the use of genomic selection to accommodate genotype–environment interaction; the incorporation of major genes and quantitative trait loci (QTLs) into genomic prediction; the prediction of the merit other than the breeding or trait value of an individual; the use of genomic prediction to avoid cost in phenotyping; mate selection; the development of sequential selection schemes; the prediction of hybrid performance; and heterosis and marker imputation.

We end the chapter with a discussion of experimental design and analysis for the phenotypic assessment of new genotypes and varieties, an area in which plant breeding has a long history of innovation.

17.2 Heritability and the Breeder's Equation in Plant Breeding

For most of the major plant species, the crop in a farmer's field is a single genotype: an inbred line, an F1 hybrid, or a vegetatively propagated clone, though minor crops are often genetically heterogeneous. In contrast, animals raised by farmers are most commonly genetically distinct. This difference has influenced the development and uptake of methods. The focus of plant breeding theory has been on 'transgressive segregation', that is, identifying individuals in crosses with trait values which fall outside the range of the parents and may therefore be developed as improved varieties. Animal breeding considers genetic improvement as a process of increasing the frequency of favourable alleles in a population. The distinction is not perfect; for example, pasture grasses and population varieties of rye are often genetically heterogeneous. Although this distinction characterises a broad difference in philosophy between much of plant and animal breeding, the two approaches are merely different ways of viewing genetic progress, and the inviolable link between all breeding methods remains the 'breeder's equation' (Lush, 1943, Section 2.6),

$$R = \frac{\sigma_g \times i \times r}{L},$$

where R is the change in trait mean per unit of time, σ_g is the amount of genetic variation within the population, i is the selection intensity, r is the accuracy of selection, and L is the interval between successive cycles of selection. The accuracy of selection is the square root of heritability if selection is on an individual's phenotype, or more generally it is the correlation between the mean of the progeny and the criterion on which selection was based, which can include information from other individuals and other traits.

The breeder's equation makes explicit that there are only four ways to improve response to selection: increase accuracy of selection (i.e. increase heritability for direct selection), increase genetic variability, increase intensity of selection, and reduce the time required for a single breeding cycle. Plant breeders have more opportunities to influence this equation than animal breeders. Generation times can be reduced through out-of-season nurseries by moving plant

material between hemispheres, or for example by shuttle breeding, as practised at the International Maize and Wheat Improvement Center (CIMMYT), in which complete wheat and maize breeding programmes are moved between two regions of Mexico to achieve two generations per year (Borlaug, 2007). Controlled environment facilities are used, for example the 'speed breeding' platform (Watson *et al.*, 2018) in Australia in which up to six generations per year are possible for some crops. To date, no equivalent approach is possible routinely in animals. This may be one reason why genomic selection (GS; **Chapter 28**) has been so readily adopted in animal breeding but not in plants: although the decoupling of selection from phenotyping in GS can greatly reduce cycle time, for several crop species a cycle of selection is already fast. There is still a delay, however, often of several years, in which the number of seed or clones of a new variety must be increased before release to growers. This has led to the proposal for a two-stage selection process in which cycles of GS proceed quickly, but product development is slower (Gaynor *et al.*, 2017). This is similar to approaches based solely on phenotypic selection, developed as recurrent selection schemes (Hallauer, 1981).

Plant breeders have greater control over accuracy of selection than animal breeders. Individual genotypes (inbred lines, F1 hybrids, clones) can be replicated at any scale and tested in multiple environments, for multiple traits over several years. As a result, heritability is under the control of the breeder and can range from extremely low values to near 1 if a genotype is raised in a sufficient number of replicate plots over multiple sites and years, with each plot containing several hundred plants. In its simplest form the heritability of a replicated genotype in a single experiment laid out in complete blocks is

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2/n},$$

where σ_g^2 is the genetic variance, σ_e^2 is the residual error variance and n is the number of replicates. The heritability can be interpreted as the proportion of phenotypic variance, $\sigma_p^2 = \sigma_g^2 + \sigma_e^2/n$, that is due to genetic variance. It is also the squared correlation between the observed phenotype (the genotype's mean over replicates, corresponding to the best linear unbiased estimator (BLUE) of the genotypic value) and the unobserved genotypic value. When trials are conducted across multiple environments, the phenotypic variance also comprises variances due to genotype–environment interaction, because the 'phenotype' is a mean over environments. For a trial series replicated over years and locations, the phenotypic variance is

$$\sigma_p^2 = \sigma_g^2 + \frac{\sigma_{gy}^2}{y} + \frac{\sigma_{gl}^2}{l} + \frac{\sigma_{gyl}^2}{yl} + \frac{\sigma_e^2}{yln},$$

where subscripts *gy*, *gl* and *gyl* on the variances indicate genotype–year, genotype–site and genotype–year–site interactions, respectively. The phenotypic variance can be manipulated by the choice of trial design (here we have assumed a randomised complete block design), particularly the number of replicates per trial, n , and the number of years, y , and sites, l (Talbot, 1997).

The most common use of heritability is as a measure of accuracy of selection in the breeder's equation. Heritability is also often used as a measure of the accuracy of a trial or series of trials. It must be realised, however, that both the breeder's equation and the above simple equation for heritability are based on a number of simplifying assumptions, for example, (i) the trial design has equal replication for each genotype and the design is either a completely randomised design or a randomised complete block design; (ii) the genotypic value is identically and independently normally distributed, which precludes usage of pedigree or kinship information; (iii) only a single trait is considered and no information from other traits is involved; and (iv) all trials have the same error variance. Most of the time, at least one of these assumptions is violated and hence the definition of heritability needs to be modified (Oakey *et al.*, 2006; Piepho and Möhring,

2007). Most importantly, in many cases selection will be based on best linear unbiased prediction (BLUP) rather than BLUE of breeding values or genotypic values, and hence definitions of heritability/reliability as used in animal breeding are required (Mrode and Thompson, 2005; Cullis *et al.*, 2006).

An additional means of increasing heritability is through improved trial design and analysis. Historically, research in experimental design was initiated and driven by the requirement to reduce environmental error in large-scale field trials (Gosset, 1931; Fisher, 1935), coupled with the requirement for ease of analysis with little or no electronic assistance. With the restraint of simplicity of analysis now greatly reduced, improvements in the design of field trials continue to be made. This is an area where innovation in plant breeding can be spread elsewhere, for example in the similarity between agricultural field trials and design for microarrays (Kerr and Churchill, 2001).

17.3 The Breeding System of Plants

In contrast to the vast majority of higher animals, which are diploid and dioecious (males and females are separate individuals), flowering plants (angiosperms) have very diverse mating systems. As a consequence of their inability to change location or choose their mates, plants depend on external agents (such as animals, wind and water) to transfer gametes (pollen) between individuals. Most plants have hermaphroditic flowers, which can result in self-pollination. The presence of multiple reproductive structures results in mating systems with considerable complexity. A consequence is a more complicated distribution of gametes within and between plants in populations. Individuals can mate with themselves and with numerous related and unrelated partners (Barrett and Crowson, 2016).

It is useful to distinguish two types of breeding systems: sex systems (hermaphroditic, separate sexes and others), and the mating system of hermaphroditic populations (inbreeding, outcrossing or intermediate) (Charlesworth, 2006). The second classification, together with the possibility of asexual reproduction, is commonly used in plant breeding (Bernardo, 2010). Species that predominantly self from within-flower pollination are called autogamous, and those with predominantly between-flower pollination are described as allogamous or panmictic. Examples of asexually propagated species are sugar cane, potato, and cassava, while maize and rubber tree are allogamous and soybean, wheat and rice are autogamous. Asexually propagated species can also frequently reproduce sexually (e.g. sugar cane, strawberry).

Sexual reproduction generates variability through meiosis and recombination. For asexual reproduction, the whole genome behaves like a single linkage group, and linkage disequilibrium (LD) is total (Richards, 1996). Asexual reproduction involves two main mechanisms: vegetative reproduction through stems, tubers and other structures, and the less common production of seeds without sex, termed agamospermy or apomixy. Species such as maize and squash have populations with unisexual flowers but with male and female flowers present on the same individual (termed monoecy), preventing self-pollination within flowers and favouring outcrossing. Although not common (Renner and Ricklefs, 1995), other species, such as kiwifruit and hops, have unisexual flowers on separated individuals (dioecy). Other combinations of individuals are gynodioecious (females and hermaphrodites) and androdioecious (males and hermaphrodites). To make things even more complicated, these sex phenotypes can sometimes coexist (Barrett and Crowson, 2016).

The mating systems of flowering plants have varied extensively during their evolution, in response to changes in life history, ecology and availability of pollinators. Autogamy has evolved multiple times from outcrossing species. As a result, there is considerable variation in mating systems both within and between species, though with a predominance of hermaphroditic sex expression (cosexuality). There are some common patterns; for example, perennial tree

species with stable communities tend to be predominantly outcrossing, whereas weedy plants in ephemeral colonies tend to be selfing (Barrett and Crowson, 2016). There is an important number of species with mixed mating, a mixture of outcrossing and selfing. Cosexual individuals are not necessarily self-compatible and outcrossing rates are normally inferred using data from molecular markers (Charlesworth, 2006). In natural plant populations, a mixed reproductive strategy has evolved, with habitual selfers occasionally becoming outcrossers, and perennial plant species having some asexual reproduction (Richards, 1996).

The reproductive system influences genetic variability. Selfing leads to reproductive isolation, restricted gene flow, reduced rate of recombination and small effective population size, whereas outcrossing populations maintain high levels of diversity. Outcrossing tends to break up LD. This has positive and negative effects. It can increase the rate at which favourable coadapted alleles at linked loci are selected in coupling, building up adaptive linkage groups and reducing hitchhiking of deleterious alleles, but it also acts to break up existing favourable linkages (Richards, 1996). A population of homozygotes has half of the effective size of an outcrossing population with the same number of individuals (Charlesworth, 2006). Outcrossing species have mechanisms to prevent the occurrence of selfing or crosses between related individuals, most notably homomorphic self-incompatibility systems (Lawrence, 2000; Castric and Vekemans, 2004). When forced to inbreed, these species commonly show strong inbreeding depression. Theoretical panmixia assumes a sexual population of infinite size and random distribution of male and female gametes; this is not achieved in real populations. A common parameter used to study departures from panmixia is F , the fixation index, derived from the observed and expected frequencies of heterozygous individuals for a given locus; t , the outcrossing rate, and its complement s (selfing rate, $t = 1 - s$) are based on comparison of these frequencies.

The mode of reproduction also influences breeding strategy, the types of cultivar and the procedures used in their development. In autogamous species individuals are homozygous, and historically breeding has consisted of making crosses between divergent individuals and then making selections during several generations of selfing until near homozygous cultivars or inbred lines are produced. For allogamous species many loci are heterozygous and breeding can be by recurrent selection in populations, or by obtaining hybrids by crossing inbred lines obtained by selfing or other methods, such as double haploid production. These distinctions are becoming blurred, however; systems to create F1 hybrid cultivars in inbreeding species like barley, wheat and rice are increasingly being used commercially and the breeding processes by which inbred parents of hybrid cultivars are developed are very similar to those used to produce inbred cultivars in other species. Asexual reproduction does not increase genetic variability, but allows the cloning of superior individuals with many heterozygous loci. Schnell (1982) classified plant breeding systems broadly into the four categories described in Table 17.1, depending on the degree of heterozygosity and homogeneity within the resulting varieties, the propagation system for the variety and the importance of non-additive variation in determining yield. The difference in variety type and their heterogeneity and homozygosity are illustrated in Figure 17.1.

17.4 Polyploidy in Plants and Its Genetic Consequences

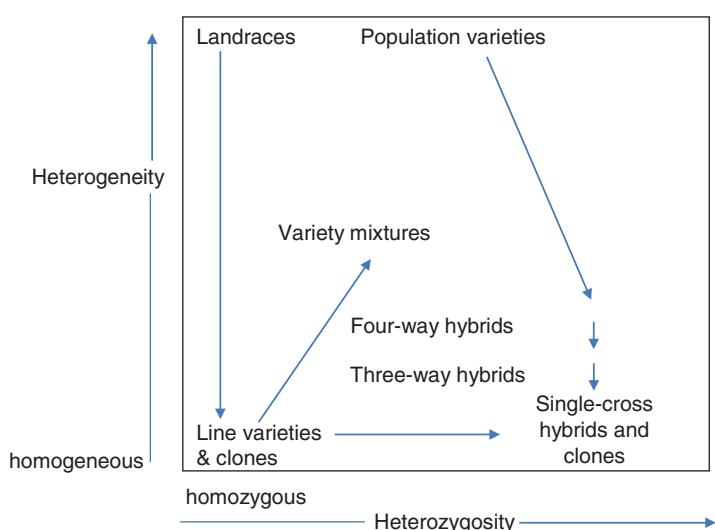
Another feature peculiar to plants is the large number of polyploid species (having more than two sets of chromosomes). Excepting some fish and amphibians, this condition is not common for animals, but it is widespread among plants, including economically important species such as potato, strawberry, wheat and sugar cane. The most common number of sets of chromosomes (the ploidy level) is four (tetraploidy), and the majority of methodological developments are for species in this category. Even numbers of chromosome sets are more

Table 17.1 Four categories of variety (after Schnell, 1982)

	Line breeding	Population breeding	Hybrid breeding	Clonal breeding
Is dominance a major factor determining yield in resulting varieties?	no	yes	yes	yes
Are genetically homogeneous varieties feasible?	yes	no	yes	yes
Can the variety be propagated from itself?	yes	yes	no	yes
Are varieties propagated by seed?	yes	yes	yes	no

common than odd numbers. Polyploids can have some problems during mitosis and meiosis, but polyploidy has been associated with a number of advantages for plants, including heterosis, gene redundancy and a tendency for asexual reproduction with apomixis (Comai, 2005).

Polyploids are classified into two categories: allopolyploids (with sets of chromosomes from different origin, e.g. wheat) and autopolyploids (having sets of chromosomes of same origin and type, e.g. potato). Meiotic pairing is different for these categories. Allopolyploids can exhibit preferential pairing between the chromosomes from the same ancestral species, resulting in what is named disomic inheritance. In practice, this implies that they will segregate as diploids, and so genetic analyses can be based on standard models for diploids. In contrast, autopolyploids typically have multisomic inheritance, where all chromosomes from the same set (homology group) can be associated in pairs in meiosis, forming bivalents. Multivalents (more than two homologous chromosomes pairing) can also be formed, and in this case recombination can take place between the locus and centromere, with sister chromatids migrating to the same pole, causing what is named double reduction, in which a gamete can be formed which contains copies of a single parental allele (Figure 17.2). Double reduction

**Figure 17.1** Variety types classified by their heterogeneity and heterozygosity (after Schnell, 1982).

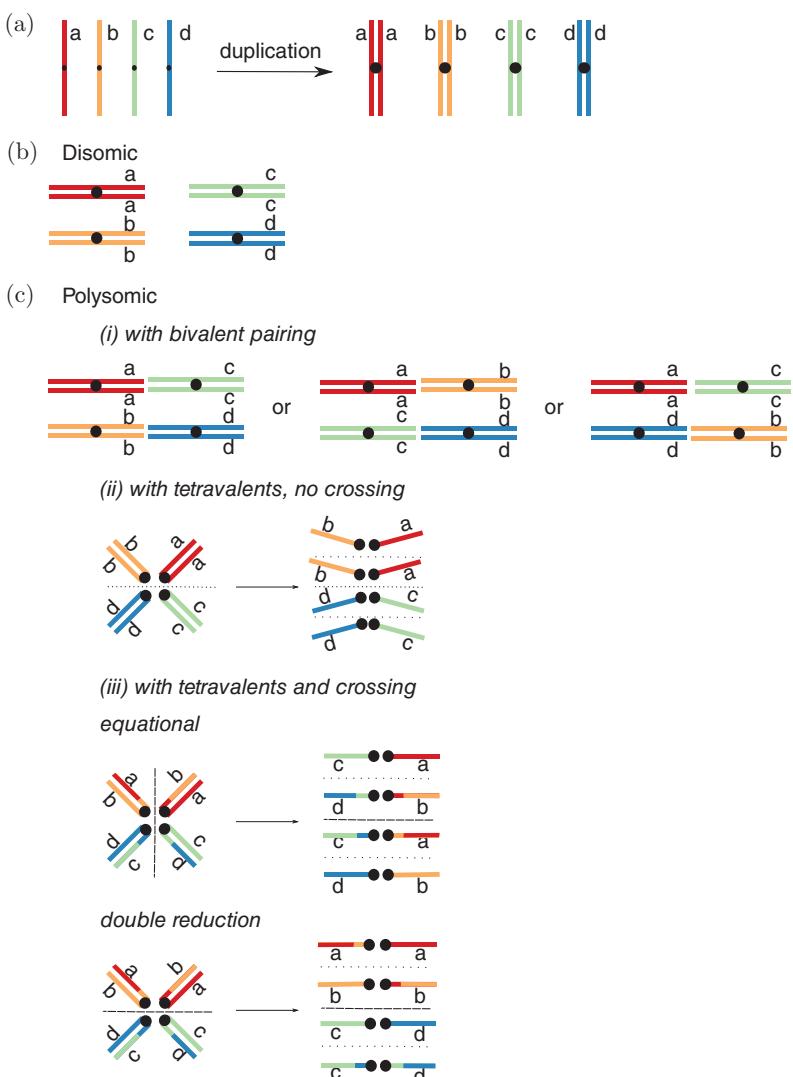


Figure 17.2 A schematic illustration of the meiotic behaviour of polyploids with four sets of chromosomes (tetraploid). (a) Four chromosomes with different alleles, showing DNA duplication for meiosis. (b) If, for example, chromosomes with allele *a* always pair with chromosomes with allele *b*, and the same happens for *c* and *d*, the inheritance is disomic and the species is an allotetraploid, behaving like a diploid. (c) With polysomic inheritance, it is possible to have bivalents and tetravalents. (i) shows all possible bivalents when there is no exclusive pairing. Notice that gametes will have allelic combinations not present in allotetraploids (*ab* and *cd*). (ii) illustrates a possible tetravalent configuration and the gametes; others are possible as well. (iii) shows that the occurrence of crossing-over between the loci and the centromere can result in gametes with two copies of the same allele when the gametes that recombine migrate to the pole of the cell during the division. Therefore, autopolyploids with multivalents can have inbreeding even without mating between relatives.

can be regarded as the result of random chromatid assortment in meiosis rather than random chromosome assortment. For example, in a tetraploid of composition *Aaaa*, random chromosome assortment would produce gametes with frequencies $\frac{1}{6}AA$, $\frac{2}{3}Aa$, $\frac{1}{6}aa$, whereas with random chromatid assortment (double reduction) the frequencies are $\frac{1}{4}AA$, $\frac{1}{2}Aa$, $\frac{1}{4}aa$. Gamete frequencies are different for disomic and multisomic inheritance, although Mendel's

rules still apply and segregation patterns can be predicted. For example, an autohexaploid individual *AAAaaa* carries three *A* alleles. It will produce gametes with frequencies of $\frac{1}{2}$ for each allele; gametes with all three *A* alleles (*AAA*) will then have frequency $\frac{1}{8}$, since they will have independent segregation. An autohexaploid *AAAaaa* with bivalent formation will have gametes *aaa*, *Aaa*, *AAa* and *AAA*, with expected frequencies $\frac{1}{20}$, $\frac{9}{20}$, $\frac{9}{20}$ and $\frac{1}{20}$, respectively. Therefore models developed for diploids cannot be directly used for autopolyploids with polysomic inheritance. Another complication arises from the possibility of populations that can deviate from exclusive disomic or polysomic inheritance.

Even with the enormous contemporary advances in genotyping and sequencing technologies, polyploids are lagging behind diploids in terms of available genomic information, mainly because of the technical and methodological challenges caused by their complexity. One of the most important of these is the difficulty in resolving the allelic dosage of individual loci (Dufresne *et al.*, 2014). For example, it is important to distinguish an *AAAAAaa* individual from *AAAaaa* (for di-allelic loci), or $A_1A_2A_2A_3$ from $A_1A_2A_3A_4$ (for multi-allelic loci). It is difficult to distinguish fully heterozygous individuals (all alleles are different) from partial heterozygotes (at least one allele has two or more copies). Traditional markers such as microsatellites have codominant behaviour only when ploidy level is four or less. SNPs, although abundant in the genome and very promising for studies in polyploids (Garcia *et al.*, 2013), are di-allelic and so not fully informative. There are statistical methods to infer the genotype of polyploids (for example, Voorrips *et al.*, 2011; Serang *et al.*, 2012; Bourke *et al.*, 2018; Gerard *et al.*, 2018), but further advances are needed, because the consequences of assumptions made by these methods demand detailed verification.

A detailed review of the population genetics of polyploids can be found in Dufresne *et al.* (2014). Using allelic and genotypic frequencies, the same principles developed for diploids can be used in principle, but the presence of polysomic inheritance adds complexity. There are more genotypic categories, since the allele dosage varies from zero up to the ploidy level. In the absence of double reduction, expected genotype frequencies under Hardy–Weinberg equilibrium are given by the expansion of $(p_1A_1 + p_2A_2 + p_3A_3 + \dots + p_kA_k)^n$ for a locus with alleles A_i , corresponding frequencies $p_i(i = 1, 2, \dots, k)$ and ploidy level n . On expansion, $A_2A_3^3$ would, for example, represent the tetraploid genotype $A_2A_3A_3A_3$. Equilibrium frequencies are reached slowly in comparison to diploids. Double reduction adds further complexity, and raises the expected frequencies of homozygous genotypes. *F*-statistics used for measuring population differentiation (Weir, 1996) can be adapted to polyploids, but with difficulties caused by uncertainty in estimation of allelic dosage. The same problems affect methods for studying population structure, exacerbated by the tendency of many polyploids to reproduce asexually. Model-free methods (such as principal components analysis) can avoid these problems.

Measuring genetic relatedness for autopolyploids, based on pedigrees or molecular data, demands a number of modifications. In diploids, it is possible to calculate the probabilities that two individuals share 0, 1 or 2 alleles identical by descent (IBD) within a given pedigree, which requires consideration of nine possible IBD configurations among the four alleles. For autotetraploids, autohexaploids and autooctoploids, the corresponding numbers of configurations are 109, 1043 and 8405, respectively (Huang *et al.*, 2015). Since marker systems cannot reveal all possible alleles when the ploidy level is higher than four, measures of relatedness for autopolyploids with high ploidy level suffer from incomplete information, although haplotype prediction might help. Genotype ambiguity, the necessity to separate identity in state from IBD, and the presence of double reduction bring yet another layer of complexity. There are methods and software for estimating relatedness from pedigree and molecular data (Kerr *et al.*, 2012; Huang *et al.*, 2015; Amadeu *et al.*, 2016), but there is room for improvement and this will be the subject of research in upcoming years.

For polyploids, building linkage maps, mapping QTLs in biparental crosses or through association mapping, and performing genomic selection are not as advanced as for diploids, due to the complications mentioned above. For tetraploids, Voorrips *et al.* (2011) developed a method based on fitting mixture models with five components to call genotypes for di-allelic markers (allelic dosages). Schmitz Carley *et al.* (2017) developed a clustering method for the same purpose. A similar method has been developed to call SNPs in crosses of F1 offspring of two autotetraploid parents, prior to creating a linkage map and QTL detection (Hackett *et al.*, 2013). For more complex polyploids, Serang *et al.* (2012) have developed a graphical Bayesian model for genotyping which has been applied successfully to sugarcane (Garcia *et al.*, 2013). The decay of LD is also slower in autoploids, since alleles in different homology groups have the opportunity to recombine only when they pair in meiosis. Asexual reproduction is also common in these species, which can lead to large LD blocks. Estimating LD also depends on inferring the correct genotypes. Mapping approaches for polyploids (Hackett *et al.*, 2017; Grandke *et al.*, 2017; Bourke *et al.*, 2018) assume that the estimated allelic dosages used as input, usually inferred from finite mixture models fitted to quantitative data (Voorrips *et al.*, 2011), are error-free, which they are not. There is room for improvement by developing methods that can take allelic dosage uncertainty into account (Piepho, 2001).

17.5 Genomic Rearrangements in Plants

Many plant species tolerate gross genomic rearrangements. In part, this is due to recent allopolyploid and autoploid ancestry of many species. Large tracts of genome, including whole chromosomes, can often be deleted without noticeable disadvantage. Cultivated sugar cane, for example, is a cross between *Saccharum officinarum* and *S. spontaneum*, and clonal varieties will typically have between 100 and 120 chromosomes. Given the magnitude of variation in plants, the concept of the pangenome, the total set of genes present in a species, is becoming of increasing practical importance (Morgante *et al.*, 2007; Hurgobin and Edwards, 2017). In maize, the reference genome of line B73 was estimated to represent only 70% of the pangenome (Gore *et al.*, 2009).

Historically, this tolerance of chromosome loss and rearrangements has enabled deletion mapping and substitution mapping. For example, in wheat, whole chromosome deletion and substitution stocks were developed through classical cytogenetics and used to locate trait effects to individual chromosomes (Law *et al.*, 1987). More recently, stocks with smaller overlapping chromosome tracts have been used in bin mapping to locate markers and QTLs to shorter deleted tracts of chromosome (Endo and Gill, 1996). At the final extreme, polyploids are more tolerant of mutations, permitting the development of stocks in which an individual plant may carry very many induced mutations. Targeting Induced Local Lesions in Genomes (TILLING: McCallum *et al.*, 2000) is a method that creates mutagenised populations carrying very high numbers of mutations per line; over 5000 per line in hexaploid wheat, and 23–24 missense and truncation alleles per gene across the population (Uauy, 2017). This gives a very high probability of detecting a desired knockout in a specific gene or of identifying a desired mutant phenotype by screening a modest number of lines. In both cases, however, a backcrossing or crossing programme may be required to isolate a mutant or to combine mutations in homoeologous genes. For targeting individual loci, gene editing (Song *et al.*, 2016) may soon supersede tilling, but the very large number of mutations carried by lines in tilling populations will remain an advantage.

More problematically, this great variability in the genome makes the creation of genetic maps and subsequent QTL mapping more complex. In a population derived from a cross between two inbred lines, chromosome rearrangements cause fewer problems: they may result

in segregation distortion or in large non-recombinating regions, but genetic maps can be created without problem. However, other populations may have quite different maps. A practical result is that trait mapping in plants is usually accompanied by the creation of a genetic map *de novo* for each population. In humans, in contrast, a single genetic map has generally been used across multiple studies, for example that based on CEPH families (Dib *et al.* 1996). The more recent uptake of multi-founder populations such as NAM (Yu *et al.*, 2008; McMullen *et al.*, 2009), MAGIC (Cavanagh *et al.*, 2008; Kover *et al.*, 2009) and AMPRIL (Huang *et al.*, 2011) and QTL mapping across multiple populations (Blanc *et al.*, 2006) has increased the problem since the effects of suppression of recombination and viability distortion are not uniform across all recombinant individuals. There remains limited software to create consensus genetic maps in these populations (Stam, 1993; Huang and George, 2011). Shah *et al.* (2014) developed a method that accounts for segregation distortion among tens of thousands of markers in a single linkage group and which can result in a considerable reduction in map length in the presence of a segregating tract of alien chromosome. However, problems remain, as evidenced by the extended map length of most consensus genetic maps compared to that typical of biparental crosses. The presence of high marker density and segregation distortion can be used as a method of detecting new chromosome rearrangements, however, though these should also be confirmed cytogenetically.

17.6 Genetic Architecture of Traits in Plants

The major plant species were domesticated ~10,000 years BP (Doebley *et al.*, 2006), though the process was protracted and may have started much earlier (Allaby *et al.*, 2017). There is a domestication syndrome of genetic changes, which are common across several crops, most notably for those grown for seed. These changes include loss of seed dispersal mechanisms, loss of seed dormancy, increase in seed or fruit size and changes in photoperiod response (Meyer and Purugganan, 2013). Some species have been domesticated much more recently, for example sugar beet within the last 250 years and oil palm which remains very close to the wild form. Other species have come into cultivation through mimicry of previously domesticated species; for example, 'false flax' (*Camelina sativa* subsp. *linicola*), and rye as a mimic of wheat and barley. The population bottleneck of domestication resulted in a major loss of genetic variation in all species, with additional loss occurring with the advent of scientific plant breeding and subsequent very small effective population sizes. Effective population sizes (N_e) among collections of elite cultivars are commonly regarded as low, though we know of only one published, based on 30 microsatellite loci: the wild ancestor of durum wheat has $N_e = 32,500$, falling to 6000 for landraces and 1300 among recent improved varieties (Thuillet *et al.*, 2005).

An interesting case is that of bread wheat. This came into cultivation from crosses between an allotetraploid and a diploid ancestral species to create a new allohexaploid. Few diploid individuals must have been involved in this process as the genetic diversity of the ancestral diploid genome within hexaploid bread wheat is much lower than that of the tetraploid genome (Akhunov *et al.*, 2010). Maize is the most studied crop. It was domesticated from its wild form, teosinte, about 9000 years ago and bears less resemblance to its ancestor than any other edible plant (Kingsbury, 2009). Much of this difference maps to only six major domestication loci (Doebley *et al.*, 1990) though Hufford *et al.* (2012) detected 484 genomic regions with strong evidence of selection during domestication.

Due to small effective population sizes, LD decays at a slower rate in domesticated plants than in most domesticated animals. As a result, genome-wide association study (GWAS) panels can in principle be smaller than in animals and humans yet maintain adequate power, though without the precision possible in humans and animals. However, creating GWAS panels of even

a few hundred elite lines for many crops is difficult: there may not be that number in existence. The use of breeders' early generation lines could augment the number, but in general, to increase size, lines must be included from a greater geographical area or chronological range, though both these processes introduce population structure which acts to reduce power. There is also an additional risk that QTLs are identified that are not relevant to the target environment or for which the favourable allele is already fixed. Consequently, there is a considerable danger that GWASs in crops are too small and are underpowered, leading to many false positives (MacArthur, 2012). Unfortunately, replication studies in plant genetics are rare, though they have been firmly advocated (Ingvarsson and Street, 2011). QTL mapping in biparental populations has already gained a poor reputation (Bernardo, 2008) and there is a risk that GWASs in plants will also be viewed as lacking utility unless the problem of inadequate power and the consequent need for replication studies becomes more widely appreciated.

Much of the genetic improvement seen in plants is attributable to a small number of major genes. Common examples are major genes for dwarf plant habit introduced into cereals during the Green Revolution of the 1960s and 1970s, genes for photoperiod sensitivity in several crops, and genes for disease resistance. The importance of major genes in plant breeding is greater than in animals. In many cases the variant responsible for domestication is known to be present in the ancestral species, but in others the responsible mutations have arisen post-domestication (Meyer and Purugganan, 2013). For highly polygenic traits, such as yield, knowledge of the origins of the variants responsible for the phenomenal improvements that have been made would assist in plant conservation strategies and knowledge of how best to exploit variation stored in genebanks.

17.7 Response to the Environment and Plasticity

Genotype–environment interaction ($G \times E$) in crops is of greater importance than in animals and humans. Plants cannot avoid environmental challenge by moving; they must respond through phenotypic plasticity. While plants are homeostatic to some extent, for example they are able to regulate their gas and water supply, this is not on the same scale as in mammals. Genetic variation in response to environmental challenges, is therefore important. The $G \times E$ variance component is often larger than that for genetic variance for traits like yield (Talbot 1997; Laidig *et al.*, 2008).

Philosophically, $G \times E$ can be treated either as a source of noise which reduces heritability and hence response to selection; or as a manifestation of variation in the adaptation of genotypes to the environment, the drivers of which should be identified and incorporated into selection decisions and variety recommendations. Analytical approaches are available to address each of these objectives. For the first, partitioning trait variance into components for G and $G \times E$, which may be further partitioned into terms for $G \times$ sites, $G \times$ years and $G \times$ sites \times years, allows optimisation of testing programmes to maximise average response to selection across all environments. Of these components, $G \times$ sites is largely reproducible whereas the other two components are not (Annicchiarico, 1997). By treating the same trait scored in different environments as separate, classical multi-trait selection indices can be built to maximise response to selection in any individual environment (Curnow, 1988; Piepho & Möhring, 2005). Using information on all sites, the best estimate of the genetic value g_{ij} of variety i at site j can be expressed as a selection index given by

$$\hat{g}_{ij} = b_1 y_{i1} + b_2 y_{i2} + \cdots + b_m y_{im},$$

where \hat{g} is a weighted mean of the m individual site means (y) and the b are regression coefficients, estimated as

$$\mathbf{b} = \mathbf{aGP}^{-1},$$

where \mathbf{G} and \mathbf{P} are the genetic and phenotypic variance–covariance matrices of variety means across sites. Generally, the covariance terms, but not the variances, of \mathbf{P} are the same as \mathbf{G} (since usually the only cause of correlation between sites is genetic). The vector \mathbf{a} is the ‘economic value’ of each site, in the terminology of selection indices. Here \mathbf{a} would be 1 for the site of interest and 0 for all other sites. However, rather than focusing on individual sites, breeders usually select for a target population of environments, TPE (Comstock & Moll, 1963). If the TPE can be divided into sub-regions, the selection index can be set up for a specific sub-region, considering the phenotypic means in each of the sub-regions as different traits (Atlin *et al.*, 2000; Piepho & Möhring, 2005).

To identify individual drivers of adaptation, traits scored in each environment can be regressed on quantitative and qualitative measures of environment using factorial regression (Malosetti *et al.*, 2013). Environments may be controlled, allowing specific components of the environmental to be varied, or uncontrolled. In the latter case, which typically involves experiments conducted over multiple locations and years, there is a degree of luck as to how large the environmental variance turns out to be and what are the drivers of that variation (e.g. rainfall, temperature). Until recently, data on environmental variables has been hard to collect, but with the advent of cheap and frequent means of recording data such as rainfall, temperature and light intensity, the opportunity to model $G \times E$ as a function of underlying environmental measurements has improved.

Historically, the absence of adequate environmental data and limited understanding of the physiology of adaptation and plasticity in most crops favoured approaches to quantifying $G \times E$ that used trait data alone. The simplest model for a set of varieties grown in multiple environments is

$$y_{ij} = m + g_i + e_j + (ge)_{ij},$$

where y_{ij} is the trait value (henceforth yield) of the i th variety in the j th environment, m is the mean, g_i is the effect of the i th variety, e_j the effect of the j th environment and $(ge)_{ij}$ is the interaction term (here we ignore within-environment error for simplicity).

The $(ge)_{ij}$ terms can be further modelled and interpreted. The simplest approach is to treat $\sum_j (ge)_{ij}^2$ for each variety as a measure of stability. This is the ecovalence (Wricke, 1962). High values indicate varieties which are sensitive (or responsive) to changes in environment. The ecovalence can be partitioned into a component that is linearly dependent on e_j and a remainder as popularised by Finlay and Wilkinson (1963) and first proposed by Yates and Cochran (1938). The model is

$$y_{ij} = m + g_i + b_i e_j + d_{ij}.$$

The linear regression coefficient b_i , with an expected value of 1, is a measure of stability or sensitivity. This model is nonlinear due to the multiplicative term $b_i e_j$ and there are complications in its fitting, but the approximate methods in common use work well (Bulmer, 1980). The interpretation of the regression coefficients is subjective. A coefficient $b_i > 1$ indicates a responsive variety that performs better at high yielding sites or a sensitive variety whose performance falls away at poor sites. Equally, a coefficient $b_i < 1$ could be a stable variety (good) or an unresponsive variety (bad). In practice, varieties must be judged by both g_i and b_i . The remainders, d_{ij} , can be squared and summed over environments for each variety and are also a measure of stability. Early work (Mather and Jinks, 2013) established that g_i and b_i could be selected for independently. More recently, g_i and b_i have been treated as traits and mapped. For example, using the maize NAM population (McMullen *et al.*, 2009), QTLs for these were found to be largely non-overlapping (Kusmec *et al.*, 2017).

Singular value decomposition of the $(ge)_{ij}$ matrix is now more common than Finlay–Wilkinson regression to summarise $G \times E$. Here, the $(ge)_{ij}$ terms are decomposed into one or more (most frequently two) genotype and environmental scores and their respective eigenvalues. The model, generally referred to as AMMI (additive main effects, multiplicative interaction; Gauch, 2013), is

$$y_{ij} = m + g_i + e_j + \sum_{n=1}^N u_{ni} w_n v_{nj} + d_{ij},$$

where n indexes the first N dimensions of the decomposition of $(ge)_{ij}$, $\mathbf{u}_n = (u_{n1}, u_{n2}, \dots)$ is the n th eigenvector of genotype scores, $\mathbf{v}_n = (v_{n1}, v_{n2}, \dots)$ is the n th eigenvector for scores for environments, w_n is the n th singular value, and d_{ij} is the residual $G \times E$ terms not accounted for by inclusion of the first N dimensions.

The first two \mathbf{u} and \mathbf{v} terms are usually depicted in biplots, and clustering within and between genotypes and environments can sometimes be interpreted in terms of underlying genetic relationships and environmental variables (Kempton, 1984). The d_{ij} terms, squared and summed across environments for each variety, can also be considered as measures of stability. The AMMI model has also been advocated for use in animal breeding (Meyer, 2009).

As described earlier, the \mathbf{u} scores can also be used in QTL analyses (Romagosa *et al.*, 1996; Rodrigues *et al.*, 2015).

AMMI has been developed further. Firstly, rather than decomposing the matrix of $(ge)_{ij}$, the matrix of $g_i + (ge)_{ij}$ can be decomposed, which is called a GGE analysis. This may provide a better visual summary and easier interpretation of the performance of each line at each site (Yan, 2014).

Secondly, the AMMI model is suitable for fixed effects only, which is restrictive in many plant breeding contexts and the standard method of fitting also requires a balanced data set (but see Gauch and Zobel, 1990; Paderewski and Rodriguez, 2014). AMMI has been extended in a mixed model framework in which either genotypes or environments are treated as random effects, giving rise to factor-analytic variance–covariance structures for genotype–environment effects (Piepho, 1998; Smith *et al.*, 2001), and QTL and environmental covariates are incorporated (Boer *et al.*, 2007; Crossa, 2013). These models are also directly applicable to animal breeding (Meyer, 2009). Factor-analytic variance–covariance structures can be viewed as low-rank approximations to unstructured variance–covariance matrices. At the same time, these models provide more flexibility than simple random effects analysis-of-variance models. For example, if the effects g_i and $(ge)_{ij}$ are modelled as independent with constant variances σ_g^2 and σ_{ge}^2 , the random vector $y_i = (y_{i1}, y_{i2}, \dots)$, comprising all responses of the i th genotype, has variance–covariance structure with variances $\sigma_g^2 + \sigma_{ge}^2$ on all diagonal positions and covariances σ_g^2 on all off-diagonal positions. This ‘compound symmetry’ structure is not usually a realistic model, however, as genetic variance may depend on the environment, giving rise to heterogeneity of variance, and covariances between environments may depend on the pair of environment being considered. It is therefore natural to assume a completely unstructured model for the variance of y_i , but this comes at the cost of a drastic proliferation of the number of parameters, as each environment has its own variance parameter and each pair of environments has its own covariance parameter. Factor-analytic models cover the range between these two extremes and allow the right balance to be struck between parsimony on the one side and model realism on the other (Gauch, 1992). For example, a factor-analytic model with a single factor (akin to Finlay–Wilkinson regression) assumes a variance of $\lambda_j^2 + \sigma_j^2$ for the j th environment and a covariance of $\lambda_j \lambda_h$ for environments j and h , thus allowing for heterogeneity of both variances and covariances. Adding more factors (λ) further increases flexibility, but this needs to be balanced against

the cost of extra parameters to be estimated. Using model selection criteria such as the Akaike information criterion helps to identify a factor-analytic structure of sufficient complexity.

The availability of more detailed environmental measures, routine genotyping and better statistical techniques is changing the analysis of $G \times E$ from a descriptive to a predictive process, at least for interaction of G with locations rather than across years (Jarquín *et al.*, 2014). Nevertheless, the greatest cause of variability in yields in crops is variation in the environment, and the major cause of this, the weather, cannot be controlled. Improved long-range weather forecasts would help growers avoid some problems through better planning, including selecting varieties to match predicted conditions. Better long-range weather forecasting is important for many reasons and progress is being made (Scaife *et al.*, 2014; Ossó *et al.*, 2017).

In plants, variation between repeating units within a single plant, or between plants of the same genotype, has a long history of being treated as a phenotype. One of the benefits of hybrid cultivars over inbred cultivars has long been recognised as the uniformity of the product – particularly important in vegetable breeding. This can be due to increased repeatability of a trait within individuals, such as fruit size or leaf veining, or reduced phenotypic response to micro-environment variability in the same field or glasshouse. This accords with Lerner's (1954) proposal that increased stability was associated with high heterozygosity. Although experimental evidence generally shows F1s to be less variable than their parents (Oettler *et al.*, 2005), residual heterozygosity among the inbred parents accompanied by dominant gene action of the trait can generate the same effect (Lynch and Walsh, 1997). In principle, developmental stability should be more easily and accurately measured and modelled in plants than in animals, because of the wider availability of genotypes fixed as clones, doubled haploids, fully inbred lines or F1 hybrids which can be replicated to give any desired level of precision. In practice, most work has been on animals, at least partly driven by greater interest in evolutionary aspects of homeostasis and canalisation. There is a broad distinction between response to the environment in plants compared to animals. In animals the essential body plan is laid down during embryogenesis, with scope for some subsequent physiological and behavioural adaptation to the environment. In plants, development is much less fixed at embryogenesis and there is greater opportunity for a plastic response, to environmental challenges and opportunities, during subsequent growth and development (Leyser, 2011).

At their simplest, methods of trait mapping and modelling for plasticity could treat plant to plant variation among replicated genotypes as a trait and map the standard deviation or the coefficient of variation directly. However, since trait means and variances are often correlated, and variances are not themselves normally distributed, this can cause false positives and confusion over interpretation. Ordas *et al.* (2008) mapped QTLs for several morphological traits in a maize biparental mapping population by comparing variances within homozygous marker classes using an F -statistic and assessing significance through a permutation test. Methods for study of environmental variation and for detecting loci affecting phenotypic variability have been reviewed by Hill and Mulder (2010) and Rönnegård and Valdar (2012) who had previously introduced a method (Rönnegård and Valdar, 2011) to detect QTLs that affect trait mean and/or variance simultaneously based on a double generalised linear model.

17.8 Genomic Selection

17.8.1 Genotype–Environment Interaction

Other than for multinational organisations working on the major crop species, it is difficult for breeding programmes to switch wholly to genomic selection (Hickey *et al.*, 2017). However,

the incorporation of genomic prediction into the later stages of cultivar testing and release is more feasible. Multi-environment trial series in most crop breeding programmes are based on a rolling process of between two and five seasons (for annual crops), in which candidate varieties are first tested in a single year at only one or two locations, with those selected for retesting in subsequent years tested at an increasing number of locations. Varieties can be dropped at any stage of testing if, for new varieties, they offer no improvement or, for old varieties, they are outclassed. The optimisation of such sequential testing processes has a long history in plant breeding (Curnow, 1961; Patterson and Silvey, 1980). Candidate varieties entering the trial series are commonly closely related to each other (as full sibs or half sibs) and to the varieties that have been in trial for several years (as parent–offspring and grandparent–offspring). These close relationships enable genomic prediction with relatively small numbers of training population individuals and markers. The prediction of the performance of new candidates from the historical trials data may save a season of testing or be used to increase precision. Since the training population will have been tested over several seasons, predictions are buffered against large variety–season interactions. Year-to-year correlations in phenotypic performance are often low, and selection on predicted performance can be more accurate than selecting on results from a single year of testing. A small number of candidates can therefore be selected first on predicted performance for inclusion in trials at a greater number of locations. Similarly, the number of locations at which a variety needs to be tested can be reduced, with loss of precision compensated for by incorporating information from relatives. The incorporation of genomic prediction into multi-environment trial series is an area of much current research (Malosetti *et al.*, 2016; Oakey *et al.*, 2016; Crossa *et al.*, 2017).

The size of the training set is one of the key determinants of accuracy in genomic prediction. Enhancing the training set size in breeding programmes requires integrating data across multiple years and cycles, and such integration has been demonstrated to improve predictive accuracy (Aunger *et al.*, 2016). For many annual crop species, the same genotype will not be phenotyped in more than one year, as segregation progresses from year to year. Thus, a major challenge is to suitably connect phenotypic data across trials with little or no connectivity and to disentangle genomic estimates of breeding values (GEBVs) from genotype–year interaction effects. Whereas there may be little replication across years for genotypes in a breeding programme, there is always ample replication at the level of alleles, and this can be exploited by modelling both GEBV and genotype–year interaction using marker information, thus permitting these effects to be dissected (Bernal-Vasquez *et al.*, 2017). However, Brandariz and Bernardo (2018) showed empirically in maize that if the training population for the next generation is a selected set of lines from the current cycle of selection, then response to selection and prediction accuracy are considerably reduced compared to an equal sized unselected set of lines. Inclusion of a small number of lines with poor performance in the training population compensated for this loss. In their examples, training population sizes (after selection) ranged from 224 to 2543, with restoration of response coming from the inclusion of the five lines with the poorest response.

Predictive accuracy in genomic selection can be enhanced by making use of biological knowledge as available, for example, in the form of crop growth models (Bustos-Korts *et al.*, 2013). One option is to introduce genetic variation in component traits as modelled by markers into the crop growth model, thus obtaining predictions for the target trait via genomic prediction for the component traits (Cooper *et al.*, 2016; Messina *et al.*, 2018). The approach can be implemented using approximate Bayesian computation or linear mixed models (Bustos-Korts *et al.*, 2013).

17.8.2 Quantitative Trait Loci and Major Genes

For the majority of domesticated plants, biparental mapping populations are easily created and QTLs are routinely detected. Unfortunately, this has not been effectively translated into marker-assisted breeding programmes except for a small number of large and consistent effects (Bernardo, 2008, 2016), often responsible for major gene disease resistance or major changes in plant phenology. However, genomic selection (**Chapter 28**) will become routine in most crops as the appropriate infrastructure is developed and there is better understanding of how best to incorporate it into breeding programmes. Although this will overcome the historical restriction of marker-assisted selection to genes of major effect, the major loci already detected should not be discarded nor simply included as part of a genome-wide marker set. In these circumstances, use of genomic best linear unbiased prediction (GBLUP, in which DNA-based estimates of the relationship among individuals are incorporated into the prediction of trait values), which commonly works well and gives accuracies close to those of more sophisticated methods (Heslot *et al.*, 2012; Ongutu *et al.*, 2012), can perform poorly. For a hypothetical quantitative trait where the majority of the genetic variation is determined by a single major gene, inclusion of increasing numbers of genome-wide markers could reduce rather than increase accuracy as the effect of the major gene is over-penalised. The simplest means of accommodating known QTLs is to include them as fixed effects in the prediction model. This works well provided each individually accounts for less than 10% of the genetic variance (Bernardo, 2014a). Alternatively, known QTLs can be penalised independently of the genome-wide marker set. More sophisticated approaches from the Bayesian alphabet (Juevas *et al.*, 2017) may work better in these circumstances, as will methods that take into account prior knowledge that specific subsets of markers are likely to behave differently; for example, MultiBLUP (Speed and Balding, 2014) in which multiple sets of markers can be penalised independently. Classes of markers could include candidate genes, markers tagging known functional genes or QTLs, and separate classes for each genome in allopolyploid species. With only two classes of predictors, standard software for ridge regression can be used by scaling the variances of markers in the two sets independently, then maximising cross validation accuracy for the variance scaling factor. For example, 10 probable QTLs were penalised independently of a background set of 3046 makers to predict seven traits in 376 wheat varieties with improvements in cross-validation accuracy made using standard ridge regression software (Bentley *et al.*, 2014).

Genomic selection will not replace all forms of marker-assisted selection in plants. There remain traits and objectives where tracking one or a few major genes is more efficient. Successes of marker-assisted backcrossing are described by Hospital (2009) who also reports this to be routine but unpublished in the private sector. There are also many reported successes for marker-assisted gene pyramiding to fix multiple QTLs in one line or variety (Hospital, 2009), most notably of disease resistance genes. The advantage of stacking multiple QTL for disease resistance is that the pathogen must evolve to overcome two or more host resistance loci, which outcome is less likely than overcoming a single resistance locus. Pyramiding of disease resistance cannot easily be accomplished through phenotypic selection alone, since increasing dosages of favourable alleles cannot be distinguished phenotypically: this is duplicate epistasis in which a favourable genotype at one locus masks the effect of favourable genotypes at others. The probability of successful pyramiding depends on recombination fractions between loci, number of generations of crossing, and population sizes in each generation (Servin *et al.*, 2004). The optimum strategy will depend on cost and time. In addition, each QTL must be accurately tagged by a marker or by flanking markers. Identification of multiple QTLs by association mapping in populations of elite lines could provide candidates for immediate gene pyramiding in the same genetic background. This approach might complete with GS for traits such as disease

resistances in many species, in which a large proportion of the variation is determined by a modest number of loci. We are not aware that this has been tested.

Although a consensus is emerging in animal breeding that GBLUP predicts GEBVs with accuracies close to those of more sophisticated methods, and with benefits in ease of implementation and speed of analysis, other methods may be more appropriate for plants. This is particularly so if the training population and the candidates for selection are members of a population with little genetic structure. An extreme example of this would be an F2, or lines derived from an F2, but populations undergoing recurrent selection with little or no immigration are also common. In these cases there is relatively little variation in kinship among individuals, and so GBLUP, which is known to strongly exploit variation in kinship (Habier *et al.*, 2007; Hayes *et al.*, 2009), is less likely to perform well. Methods such as the LASSO and Bayesian LASSO can work better (Heslot *et al.*, 2012; Pasanen *et al.*, 2015). Empirical testing of methods in these circumstances has been limited, partly because of the absence of suitable datasets.

17.8.3 Genomic Selection and Cross Prediction

For plant breeders, rather than directly selecting on GEBV, methods to predict the probability of transgressive segregation in the descendants of an individual would be attractive. These are, in effect, genomically updated versions of older methods to predict probabilities of transgressive segregation from estimates of the mean and variance of crosses (Jinks and Pooni, 1976). To date, this has received little attention in comparison with approaches to predict GEBVs. A simple method is to simulate progeny from a cross and predict the GEBVs of the simulated individuals. The distributions of the simulated crosses are then used as selection criteria (Bernardo, 2014a; Tiede *et al.*, 2015). A similar approach is to select on the optimal haploid value: the predicted performance of the best doubled line which an individual could produce (Daetwyler *et al.*, 2015).

17.8.4 Genomic Selection and Phenotyping Cost

A major use of genomic prediction in plants is to reduce the cost of phenotyping in multi-environment series of trials. These are generally used in the later stages of a breeding programme to identify lines for release to growers, are expensive to run and require investment in seed production (or clonal propagation) to generate enough individuals to test. Substituting GEBVs for direct phenotyping eliminates the requirement for all individuals to be tested at all sites (Heffner *et al.*, 2009), which reduces the scale of trials and also requires less seed. Alternatively, the number of lines tested could be increased for the same cost of phenotyping. Optimal designs for multi-environment trial series are required. This approach is likely to be integrated with the improved modelling of G × E described earlier.

17.8.5 Mate Selection

The intensity of selection in a breeding programme can be restricted by concern over of loss of variation due to small effective population size, or equivalently for outbreeding species by concerns about inbreeding. One way of mitigating the adverse effects of small population size is to move away from simple truncation selection, in which each selected individual has an equal probability of contributing to the next generation (or equal contributions are forced through controlled mating). Schemes can be found which, for the same intensity of selection as truncation selection, have a greater effective population size or (equivalently) rate inbreeding per generation. Most research in this area has focused on animal breeding and is described as

optimal contribution theory (Woolliams *et al.*, 2015). In inbreeding species there has been little interest in these methods, since the equivalent of truncation selection in a closed population is rarely applied, though it has been considered in plants, most notably in tree breeding (Lindgren and Mullin, 1997). Optimal contribution theory has received new impetus from the adoption of genomic selection by animal breeders and accompanying concerns about accelerated rates of inbreeding (Woolliams *et al.*, 2015). As genomic selection is adapted to plant breeding and schemes of rapid population cycling in closed populations are incorporated, including for inbreeding species, application of optimal contribution theory and related methods will become more important (Hickey *et al.*, 2017; Lin *et al.*, 2017).

17.8.6 Sequential Selection

Substitution of GEBVs for direct phenotyping can be done sequentially. First, a small number of individuals from a population are genotyped and phenotyped. Then, a second batch of individuals is genotyped, selection made on GEBVs, and the selections phenotyped. The phenotyped individuals are used to augment the training population and the process is repeated on the next batch. The process stops when a target trait value is reached or a predefined number of individuals tested. This approach is most suitable for traits such as brewing quality in barley, or bread-making quality in wheat, which are expensive, but relatively quick, to measure (assuming seed is available). Tanaka and Iwata (2018) proposed this system as a means of reducing the phenotyping required to screen germplasm collections. They proposed that, rather than selecting on GEBVs alone, selection should be on the probability that an untested individual is better than any with known phenotype. This requires the GEBV of the individual and an estimate of the variance of that GEBV; the probability of improvement can be higher for an individual with higher variance but lower GEBV than another. In practice, this translates into giving additional weight to individuals with lower kinship to others in the current training set.

17.8.7 Genomic Prediction of Hybrid Performance and Heterosis

Hybrid breeding to exploit heterosis is typically based on the development of parental lines in different heterotic pools. A heterotic pool is a set of lines or individuals which tend to show similar levels of heterosis when crossed to members of other pools. The number of possible crosses is usually very large, often too large to permit field-testing of all possible hybrids. Thus, prediction of hybrids based on a subset of tested hybrids is a very promising strategy. This has been pioneered by Bernardo (1992) in maize, first making use of pedigree data for the parental inbred lines and derived matrices of coancestry coefficients, which can be used to model random effects for general combining ability (g.c.a., the average effect of a line in hybrid combination) and specific combining ability (s.c.a., the deviation in effect of a cross from that predicted by the parental g.c.a.). Later, the pedigree-based matrices were replaced by restriction fragment length polymorphism marker-based equivalents (Bernardo, 1994), employing BLUP for prediction. Schrag *et al.* (2010) used amplified fragment length polymorphism and simple sequence repeat markers for predicting hybrid performance in a large unbalanced multi-site, multi-year data set and demonstrated that BLUP outperforms simpler regression-based approaches. Maenhout *et al.* (2010) showed that support vector machine regression is a viable alternative to BLUP for hybrid prediction. Predictive performance can be substantially enhanced by using omics data (Xu *et al.*, 2012; Riedelsheimer *et al.*, 2012), which are closer physiologically to the quantitative traits of interest than genomic markers. There is by now ample evidence from genomic prediction studies that heterosis is governed by additive, (over-)dominance and epistatic effects, but as yet no consensus has emerged as to the relative importance of these types of effect (Li *et al.*, 2008; Larièpe *et al.*, 2012; Zhou *et al.*, 2012; Mäki-Tanila and Hill, 2014; Jiang *et al.*, 2017).

17.8.8 Marker Imputation

Marker imputation is required in plant breeding, as in animal genetics, to reduce the cost of genotyping in genomic selection. Methods developed for imputation of marker data are described in Chapter 3. Highly heterozygous autopolyploids are a problem for imputation methods in plants. In contrast, for inbreeding species, the preponderance of homozygous markers should make imputation easier. Methods are now being developed specifically for plants; Swarts *et al.* (2014) developed a method for populations of inbred lines which exploit the reduced requirement for phasing, and the method of Hickey *et al.* (2015) works on basic plant pedigrees of biparental crosses, selfs, backcrosses and top crosses. However, as yet there are no accurate imputation methods for autopolyploids.

17.9 Experimental Design and Analysis

Agricultural science has always been at the forefront of experimental design. In plant breeding and crop genetics, the scale of variety trials continues to grow, with the testing of more than 1000 lines now common. The historical development of methods has been reviewed, for example, by Edmondson (2005), Smith *et al.* (2005) and Mead *et al.* (2012). The design and analysis of trials were initially restricted by the necessity for simple analysis, often carried out by hand. The classic textbook of Cochran and Cox (1957) describes and catalogues such designs. These placed severe restrictions on the number of genotypes that could be tested, for example to perfect squares. Access to cheap powerful computers and to good algorithms and software means that current designs are much more flexible. Alpha designs, developed in the 1970s (Patterson and Williams, 1976), are in common use. These are circulant partially balanced resolvable incomplete block designs (John and Williams, 1995) with only modest restrictions, primarily that the number of entries per block must not be greater than the square root of the number of varieties, which is never a problem in practice. Resolvable row–column designs (John and Williams, 1995) are used too, and different software packages are available for generating such designs. Unfortunately, many plant breeders and plant researchers continue to use classical designs inappropriately, overcoming limitations, for example, by assigning large numbers of candidate varieties to multiple designs for small numbers of entries, linked by a common set of control varieties. This is demonstrably less efficient than testing all candidates in a single experiment (Piepho *et al.*, 2006). This process of forcing experimental material into inappropriate and inefficient designs has been described as ‘Procrustean design’ by Mead *et al.* (2012) and is no longer necessary.

Typical replicated designs, such as alpha designs, rely on incomplete blocking, in one or two dimensions, with block sizes of the order 10, tempered by knowledge of local field variability. A recommended starting block size is the square root of the number of varieties. The basic model for analysis of a single trial with blocking in two dimensions, in which blocks are arranged physically into complete replicates in the field (the design is said to be resolvable), is

$$y_{ijhk} = \mu + g_i + w_j + r_{jh} + c_{jk} + e_{ijhk},$$

where w_j is the effect of the j th replicate, r_{jh} is the effect of the h th row within the j th replicate, c_{jk} is the effect of the k th column within the j th replicate and e_{ijhk} is a residual error. This model is commonly fitted by restricted maximum likelihood (which was originally published as a method for trials analysis; Patterson and Thompson, 1971). Blocks (rows and columns) are treated as random effects, allowing ‘recovery of interblock information’, giving modest improvements in accuracy of variety differences. If variation among blocks is large, there is little

difference between treating them as fixed or random (Piepho *et al.*, 2013). Varieties are commonly treated as fixed (but see below).

A plot typically contains multiple plants of the same genotype or family. The number of plants per plot varies greatly with species, for example 16 oil palms might occupy about 1000 m² while 500 cereal plants require 12 m². Size of plot is generally determined by dimensions of specialist planting and harvesting equipment, but the requirement to avoid inter-genotype competition effects places a lower limit on plot size while absence of intra-genotype competition places an upper limit on plant spacing. In practice, experimental technique places an implicit balance between bias and precision from these sources. More complex modelling of direct and indirect genetic effects (David *et al.*, 2000, 2001), as used to take account of interactions between animals in herds, or of siblings raised in litters (Bijma, 2014), is known but not generally used; avoidance of problems through adequate plot construction is simpler. Exceptions involve restricted randomisation in some species (e.g. oil-seed rape) in which lines in plots are grouped by height, flowering time or variety type (e.g. hybrid or inbred) in a split-plot type of design with groups as main plots. These factors are known to contribute to inter- and intra-genotype competitive effects. Simple adjustment through regression of plot yields on covariates, such as height of neighbouring plots, has also been used (David *et al.*, 2001), but genetic and environmental competitive effects are confounded, and it is difficult to adjust plot yield to a value expected in the presence of intra-genotype competitive effects but the absence of inter-genotype competition. Designs balanced so that all genotypes are adjacent to all other genotypes an equal number of times have been proposed (Azais *et al.*, 1993), but the number of genotypes that can be included is small.

As breeding programmes and experiments have increased in scale there has been greater interest in designs with variable replication. For traits of high heritability, and where seed or clone numbers are limiting, the greatest response to selection can come from experiments in which each entry occurs only once in a single plot (see below). Precision can be increased if a proportion of the entries are replicated, allowing for adjustment for field fertility effects and also providing an estimate of error. Two common approaches are the augmented design, in which a small number of entries, usually standards or controls, are replicated many times (Federer, 1956, 1961) and partially replicated designs (*p*-rep) designs (Smith *et al.*, 2006; Williams *et al.*, 2014) in which a larger number of experimental entries are present in two replicates. With *p*-rep designs, there is scope for optimisation by making sure that each entry is replicated about the same number of times across locations of a trial series.

Inter-plot competition aside, generally, the closer two plots are, the more closely correlated their performance, since their local environments tend to be more similar. Knowledge of the autocorrelation between adjacent plots has been incorporated into trial design (Cullis *et al.* 2006; Williams and Piepho, 2013). Analysis of trials data to take into account the spatial relationships has a long history, dating back to Papadakis (1937), in which regression of a plot yield on that of neighbouring plots is used in an analysis of covariance to improve precision. More recently, autoregressive methods in one and two dimensions have been developed (Gilmour *et al.*, 1997) and adopted, most notably in Australia. Software to fit these is not readily available or free. The most commonly used programs are ASReml, GenStat and SAS. A recent R package allowing analysis of trials and series of trials by mixed model procedures is sommer (Covarrubias-Pazaran, 2016). Methods of modelling and fitting fertility trends continue to be developed, for example by fitting a two-dimensional spline in the R package SpATS (Rodriguez-Alvarez *et al.*, 2018).

When integrating data across trials (years, sites), there are two basic options. Either the data are analysed in a single stage using a model for the plot data across trials, or analysis is performed in several stages, starting with the analysis of individual trials in the first stage to obtain

genotype means. These means are then analysed across trials using information in the precision of means from the first stage as weights (Smith *et al.*, 2005; Damesa *et al.*, 2017). Analysis can also proceed in more than two stages, for example, with GBLUP or GWAS performed in a third stage after integrating the phenotypic data across trials in the second stage. In stage-wise analysis, it is crucial to model genotypes by fixed effects throughout all stages except the last where genotypes are fixed or random depending on the objective of the analysis (e.g., random for GBLUP, fixed for obtaining variety means, fixed and random for GWAS) (Piepho *et al.*, 2012).

Research on trial design is now taking into account the availability of information from pedigree or realised relationships among entries. For very high (single-plot) heritabilities, trial design is irrelevant, since any layout will give equivalent high response to selection. However, at lower heritabilities, the importance of information from related lines increases and designs that do not take into account these relationships will not be optimal (Bueno Filho and Gilmour, 2003). More recent work (Butler *et al.*, 2014; Feoktistov *et al.*, 2017) describes methods for the optimal spatial arrangement of related varieties in trials, though as far as we are aware software to implement these approaches is not available.

17.10 Conclusions

While heeding the warning of Bernardo (2016) to beware of bandwagons, it is probable that most innovation in plant breeding in the next ten years will be driven by adoption of genomic selection. It now seems tautological that if a trait is heritable, then it can be predicted from a genome-wide set of markers used, most simply, to estimate genetic relationships. Research in crops is increasingly focused on to how best to implement genomic selection in breeding programmes rather than to make improvements in prediction accuracy. This can range from the complete redesign of a breeding programme (Gaynor *et al.*, 2017) to a simple substitution of trait prediction for direct phenotyping of traits which are expensive and time-consuming to measure. The requirement for statistical genetics input in plant breeding will become more important, and this will align plant and animal breeding more closely (Hickey *et al.*, 2017). There remain areas in which approaches differ, however. Plant breeders focus more on major genes, particularly for phenology and disease resistance, have to work with greater genotype–environment interactions, and regard success as identifying transgressive segregants rather than increasing frequencies of favourable alleles.

There is a strong need for better training of plant breeders in basic statistics and quantitative genetics to exploit ever cheaper marker and sequencing platforms (van Eeuwijk *et al.*, 2016). Underpowered studies remain common and carry a risk that researchers can apply sophisticated analyses without sufficient understanding, leading to false positives (MacArthur, 2012). A hypothetical example, closely modelled on published studies, is a ‘genome-wide’ study which used 48 markers on 33 accessions tested for 15 morphological and agronomic traits. This reported 59 significant marker–trait associations involving 30 markers. This improbably high success rate from a small study exemplifies a widespread lack of understanding of statistical power in experimental studies (Kanehman, 2011).

References

- Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., Crossman, C.C., Deal, K.R. and Dubcovsky, J. (2010). Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**, 702.

- Allaby, R.G., Stevens, C., Lucas, L., Maeda, O. and Fuller, D.Q. (2017). Geographic mosaics and changing rates of cereal domestication. *Philosophical Transactions of the Royal Society B* **372**, 20160429.
- Amadeu, R.R., Cellon, C., Olmstead, J.W., Garcia, A.A.F., Resende, M.F.R. and Muñoz, P.R. (2016). AGHmatrix: R package to construct relationship matrices for autotetraploid and diploid species: A blueberry example. *Plant Genome* **9**(3).
- Annicchiarico, P. (1997). Additive main effects and multiplicative interaction (AMMI) analysis of genotype-location interaction in variety trials repeated over years. *Theoretical and Applied Genetics* **94**, 1072–1077.
- Atlin, G.N., Baker R.J., McRae K.B., and Lu X. (2000). Selection response in subdivided target regions. *Crop Science* **40**, 7–13.
- Auinger, H.J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H.H., Piepho, H.-P., Gordillo, A., Wilde P., Bauer, E. and Schön, C.C. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (*Secale cereale* L.). *Theoretical and Applied Genetics* **129**, 2043–2053.
- Azais, J.M., Bailey, R.A. and Monod, H. (1993). A catalogue of efficient neighbour-designs with border plots. *Biometrics* **49**, 1252–1261.
- Barabaschi, D., Tondelli, A., Desiderio, F., Volante, A., Vaccino, P., Valè, G. and Cattivelli, L. (2016). Next generation breeding. *Plant Science* **242**, 3–13.
- Barrett, S.C.H. and Crowson, D. (2016). Mating systems in flowering plants. In D. Wake (ed.), *Encyclopedia of Evolutionary Biology*, Vol. 2. Elsevier, Oxford, pp. 473–479.
- Bentley, A.R., Scutari, M., Gosman, N., Faure, S., Bedford, F., Howell, P., Cockram, J., Rose, G.A., Barber, T., Irigoyen, J. and Horsnell, R. (2014). Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theoretical and Applied Genetics* **127**, 2619–2633.
- Bernal-Vasquez, A.M., Gordillo, A., Schmidt, M. and Piepho, H.-P. (2017). Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC Genetics* **18**, 51.
- Bernardo, R. (1992). Relationship between single-cross performance and molecular marker heterozygosity. *Theoretical and Applied Genetics* **83**, 628–634.
- Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science* **34**, 20–25.
- Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science* **48**, 1649–1664.
- Bernardo, R. (2010). *Breeding for Quantitative Traits in Plants*, 2nd edition. Stemma Press, Woodbury, MN.
- Bernardo, R. (2014a). Genomewide selection when major genes are known. *Crop Science* **54**, 68–75.
- Bernardo, R. (2014b). Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. *Crop Science* **54**, 2586–2595.
- Bernardo, R. (2016). Bandwagons I, too, have known. *Theoretical and Applied Genetics* **129**, 2323–2332.
- Bijma, P. (2014). The quantitative genetics of indirect genetic effects: A selective review of modelling issues. *Heredity* **112**, 61–69.
- Blanc, G., Charcosset, A., Mangin, B., Gallais, A. and Moreau, L. (2006). Connected populations for detecting quantitative trait loci and testing for epistasis: An application in maize. *Theoretical and Applied Genetics* **113**, 206.
- Bodmer, W.F. (1986). Human genetics: The molecular challenge. *Cold Spring Harbor Symposia on Quantitative Biology* **51**(1), 1–13.

- Boer, M.P., Wright, D., Feng, L., Podlich, D.W., Luo, L., Cooper, M. and van Eeuwijk, F.A. (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariates for QTL-by-environment interactions, with an example in maize. *Genetics* **177**, 1801–1813.
- Borlaug, N.E. (2007). Sixty-two years of fighting hunger: Personal recollections. *Euphytica* **157**, 287–297.
- Bourke, P.M., van Geest, G., Voorrips, R.E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R.G.F., Arens, P., Smulders, M.J.M. and Maliepaard, C. (2018). polymapR – linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics* **34**, 3496–3502.
- Brandariz, S.P. and Bernardo, R. (2018). Maintaining the accuracy of genomewide predictions when selection has occurred in the training population. *Crop Science* **58**, 1226–1231.
- Bueno Filho, J.S. de S. and Gilmour, S.G. (2003). Planning incomplete block experiments when treatments are genetically related. *Biometrics* **59**, 375–381.
- Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, Oxford.
- Bustos-Korts, D., Malosetti, M., Chapman, S. and van Eeuwijk, F.A. (2013). Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics. In X. Yin and P. C. Struik (eds.), *Crop Systems Biology: Narrowing the Gaps Between Crop Modelling and Genetics*. Springer, New York, pp. 55–82.
- Butler, D.G., Smith, A.B. and Cullis, B.R. (2014). On the design of field experiments with correlated treatment effects. *Journal of Agricultural, Biological, and Environmental Statistics* **19**, 539–555.
- Castric, V. and Vekemans, X. (2004). Plant self-incompatibility in natural populations: A critical assessment of recent theoretical and empirical advances. *Molecular Ecology* **13**, 2873–2889.
- Cavanagh, C., Morell, M., Mackay, I. and Powell, W. (2008). From mutations to MAGIC: Resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology* **11**, 215–221.
- Charlesworth, D. (2006). Evolution of plant breeding systems. *Current Biology* **16**, 726–735.
- Cochran, W.G. and Cox, G.M. (1957). *Experimental Designs*, 2nd edition. John Wiley & Sons, New York.
- Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**, 836–846.
- Comstock, R.E. and Moll, R.H. (1963). Genotype-environment interaction. In W.D. Hanson and H.F. Robinson (eds.), *Statistical Genetics and Plant Breeding*. National Academy of Sciences-National Research Council, Washington, DC, pp. 164–198.
- Cooper, M., Technow, F., Messina, C., Gho, C. and Totir, R. (2016). Use of crop growth models with whole-genome prediction: Application to a maize multi-environment trial. *Crop Science* **56**, 2141–2156.
- Covarrubias-Pazaran, G. (2016). Genome assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* **11**(6), e0156744.
- Crossa, J. (2013). From genotype × environment interaction to gene × environment interaction. *Current Genomics* **13**, 225–244.
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J.M., Pérez-Elizalde, S., Beyene, Y. and Dreisigacker, S. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science* **22**, 961–975.
- Cullis, B.R., Smith, A. and Coombes, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 381–393.
- Curnow, R.N. (1961). Optimal programmes for varietal selection. *Journal of the Royal Statistical Society, Series B* **23**, 282–318.

- Curnow, R.N. (1988). The use of correlated information on treatment effects when selecting the best treatment. *Biometrika* **75**, 287–293.
- Daetwyler, H.D., Hayden, M.J., Spangenberg, G.C. and Hayes, B.J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* **200**, 1341–1348.
- Damesa, T., Worku, M., Möhring, J., Piepho, H.-P. (2017). One step at a time: Stage-wise analysis of a series of experiments. *Agronomy Journal* **109**, 845–857.
- David, O., Monod, H. and Amoussou, J. (2000). Optimal complete block designs to adjust for interplot competition with a covariance analysis. *Biometrics* **56**, 270–274.
- David, O., Monod, H., Lorgeou, J., Philippeau, G. (2001). Control of interplot interference in grain maize: A multi-site comparison. *Crop Science* **41**, 406–414.
- Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Kazan, J., Seboun, E. and Lathrop, M. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154.
- Doebley, J., Stec, A., Wendel, J. and Edwards, M. (1990). Genetic and morphological analysis of a maize-teosinte F2 population: Implications for the origin of maize. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 9888–9892.
- Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006). The molecular genetics of crop domestication. *Cell* **127**, 1309–1321.
- Dufresne, F., Stift, M., Vergilino, R., Mable, B.K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* **23**, 40–69.
- Edmondson, R. (2005). Past developments and future opportunities in the design and analysis of crop experiments. *Journal of Agricultural Science* **143**, 27–33.
- Endo, T.R. and Gill, B.S. (1996). The deletion stocks of common wheat. *Journal of Heredity* **87**, 295–307.
- Federer, W.T. (1956). Augmented (or hoonuiaku) designs. *Hawaiian Sugar Plantation Records* **55**, 191–208.
- Federer, W.T. (1961). Augmented designs with one-way elimination of heterogeneity. *Biometrics* **17**, 447–473.
- Feoktistov, V., Pietravalle, S. and Heslot, N. (2017). Optimal experimental design of field trials using differential evolution. Preprint, arXiv:1702.00815.
- Finlay, K.W. and Wilkinson, G.N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research* **14**, 742–754.
- Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Garcia, A.A.F., Mollinari, M., Marconi, T.G., Serang, O.R., Silva, R.R., Vieira, M.L., Vicentini, R., Costa, E.A., Mancini, M.C., Garcia, M.O., Pastina, M.M., Gazaffi, R., Martins, E.R.F., Dahmer, N., Sforça, D.A., Silva, C.B.C., Bundock, P., Henry, R.J., Souza, G.M., van Sluys, M-A., Landell, M.G.A., Carneiro, M.S., Vincentz, M.A.G., Pinto, L.R., Vencovsky, R., and Souza, A.P. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific Reports* **3**, 1–10.
- Gauch, H.G. Jr. (1992). *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam.
- Gauch, H.G. Jr. (2013). A simple protocol for AMMI analysis of yield trials. *Crop Science* **53**, 1860–1869.
- Gauch, H.G. Jr. and Zobel, R.W. (1990). Imputing missing yield trial data. *Theoretical and Applied Genetics* **79**, 753–761.
- Gaynor, C.R., Gorjanc, G., Bentley, A.R., Ober, E.S., Howell, P., Jackson, R., Mackay, I.J. and Hickey, J.M. (2017). A two-part strategy for using genomic selection to develop inbred lines. *Crop Science* **57**, 2372–2386.

- Gerard, D., Ferrão, L.F.V., Garcia, A.A.F. and Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics* **210**, 789–807.
- Gilmour, A. R., Cullis, B. R. and Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 269–293.
- Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J. and Ware, D.H. (2009). A first-generation haplotype map of maize. *Science* **326**, 1115–1117.
- Gosset, W.S. ('Student') (1931). Yield trials. In *Bailliere's Encyclopedia of Scientific Agriculture*, London. Reprinted in Pearson, E.S. and Wishart, J. (1942). 'Student's' Collected Papers. Biometrika Office, University College, London.
- Grandke, F., Ranganathan, S., van Bers, N., de Haan, J.R., and Metzler, D. (2017). PERGOLA: Fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics* **18**, 12.
- Habier, D., Fernando, R.L. and Dekkers, J.C.M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.
- Hackett, C.A., McLean, K. and Bryan, G.J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PloS ONE* **8**, e63939.
- Hackett, C.A., Boskamp, B., Vogogias, A., Preedy, K.F., and Milne, I. (2017). TetraploidSNPMap: Software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *Journal of Heredity* **108**, 438–442.
- Hallauer, A.R. (1981). Selection and breeding methods. In K.J. Frey (ed.), *Plant Breeding*. Iowa State University Press, Ames.
- Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. (2009). Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433–443.
- Heffner, E.L., Sorrells, M.R. and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.
- Heslot, N., Yang, H.P., Sorrells, M.E. and Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Science* **52**, 146–160.
- Hickey, J.M., Gorjanc, G., Varshney, R.K. and Nettelblad, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden Markov model. *Crop Science* **55**(5), 1934–1946.
- Hickey, J.M., Chiurugwi, T., Mackay, I. and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics* **49**, 1297.
- Hill, W.G. and Mulder, H.A. (2010). Genetic analysis of environmental variation. *Genetics Research* **92**, 381–395.
- Hospital, F. (2009). Challenges for effective marker-assisted selection in plants. *Genetica* **36**, 303–310.
- Huang, B.E. and George, A.W. (2011). R/mpMap: A computational platform for the genetic analysis of multiparent recombinant inbred lines. *Bioinformatics* **27**, 727–729.
- Huang, K., Guo, S.T., Shattuck, M.R., Chen, S.T., Qi, X.G., Zhang, P. and Li, B.G. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity* **114**, 133–142.
- Huang, X. and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology* **65**, 531–551.
- Huang, X., Paulo, M.J., Boer, M., Effgen, S., Keizer, P., Koornneef, M. and van Eeuwijk, F.A. (2011). Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4488–4493.

- Hufford, M.B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaepller, S.M. and Lai, J. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**, 808–811.
- Hurgobin, B. and Edwards, D. (2017). SNP discovery using a pangenome: Has the single reference approach become obsolete? *Biology* **6**, 21.
- Ingvarsson, P.K. and Street, N.R. (2011). Association genetics of complex traits in plants. *New Phytologist* **189**, 909–922.
- Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* **127**, 595–607.
- Jiang, Y., Schmidt, R.H., Zhao, Y., Reif, J.C. (2017). Quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nature Genetics* **49**, 1741–1746.
- Jinks, J.L. and Pooni, H.S. (1976). Predicting the properties of recombinant inbred lines derived by single seed descent. *Heredity* **36**, 253–266.
- John, J.A. and Williams, E.R. (1995). *Cyclic and Computer Generated Designs*. Chapman and Hall, London.
- Juevas, J., Crossa, J., Montesinos-López, O.A., Burgueño, J., Pérez-Rodríguez, P. and de los Campos, G. (2017). Bayesian genomic prediction with genotype \times environment interaction kernel models. *G3: Genes, Genomes, Genetics* **7**, 41–53.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan, New York.
- Kempton, R. (1984). The use of biplots in interpreting variety by environment interactions. *Journal of Agricultural Science* **103**, 123–135.
- Kerr, M.K. and Churchill, G.A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.
- Kerr, R.J., Li, L., Tier, B., Dutkowski, G.W. and McRae, T.A. (2012). Use of the numerator relationship matrix in genetic analysis of autopolyploid species. *Theoretical and Applied Genetics* **124**, 1271–1282.
- Kingsbury, N. (2009). *Hybrid: The History and Science of Plant Breeding*. University of Chicago Press, Chicago.
- Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C. and Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genetics* **5**, e1000551.
- Kusmec, A., Srinivasan, S., Nettleton, D. and Schnable, P.S. (2017). Distinct genetic architectures for phenotype means and plasticities in *Zea mays*. *Nature Plants* **3**, 715.
- Laidig, F., Drobek, T. and Meyer, U. (2008). Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials. *Plant Breeding* **127**, 541–547.
- Lariépe, A., Mangin, B., Jasson, S., Combes, V., Dumas, F., Jamin, P., Lariagon, C., Jolivot, D., Madur, D., Fiévet, J., Gallais, A., Dubreuil, P., Charcosset, A., Moreau, L. (2012). The genetic basis of heterosis: Multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (*Zea mays* L.). *Genetics* **190**(2), 795–811.
- Law, C.N., Snape, J.W. and Worland, A.J. (1987). Aneuploidy in wheat and its uses in genetic analysis. In F.G.H., Lupton (ed.), *Wheat Breeding*. Chapman & Hall, London, pp. 71–108.
- Lawrence, M.J. (2000). Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. *Annals of Botany* **85**(suppl_1), 221–226.
- Lerner, I.M. (1954). *Genetic Homeostasis*. Oliver and Boyd, London.
- Leyser, O. (2011). Auxin, self-organisation, and the colonial nature of plants. *Current Biology* **21**, R331–R337.

- Li, L., Lu, K., Chen, Z., Mu, T., Hu, Z., Li, X. (2008). Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. *Genetics* **180**, 1725–1742.
- Lin, Z., Shi, F., Hayes, B.J. and Daetwyler, H.D. (2017). Mitigation of inbreeding while preserving genetic gain in genomic breeding programs for outbred plants. *Theoretical and Applied Genetics* **130**, 969–980.
- Lindgren, D. and Mullin, T.J. (1997). Balancing gain and relatedness in selection. *Silvae Genetica* **46**, 124–128.
- Lush, J.L. (1943). *Animal Breeding Plans*, 2nd edition. Iowa State College Press, Ames.
- Lynch, M. and Walsh, B. (1997). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
- MacArthur, D. (2012). Methods: Face up to false positives. *Nature* **487**, 427–428.
- Maenhout, S., De Baets, B., Haesaert, G. (2010). Prediction of maize single-cross hybrid performance: Support vector machine regression versus best linear prediction. *Theoretical and Applied Genetics* **120**, 415–427.
- Mäki-Tanila, A. and Hill, W.G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355–367.
- Malosetti, M., Ribaut, J.M., van Eeuwijk, F.A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology* **4**, 44.
- Malosetti, M., Bustos-Korts, D., Boer, M.P. and van Eeuwijk, F.A. (2016). Predicting responses in multiple environments: Issues in relation to genotype \times environment interactions. *Crop Science* **56**, 2210–2222.
- Mather, K. and Jinks, J.L. (2013). *Biometrical Genetics: The Study of Continuous Variation*. Springer, Berlin.
- McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000). Targeted screening for induced mutations. *Nature Biotechnology* **18**, 455–457.
- McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C. and Brown, P. (2009). Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740.
- Mead, R., Gilmour, S.G., Mead, A. (2012). *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge University Press, Cambridge.
- Messina, C.D., Technow, F., Tang, T., Totir, R., Gho, C. and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: Integrating crop growth models (CGM) with whole genome prediction (WGP). *European Journal of Agronomy* **100**, 151–162.
- Meyer, K. (2009). Factor-analytic models for genotype \times environment type problems and structured covariance matrices. *Genetics Selection Evolution* **41**, 21.
- Meyer, R.S. and Purugganan, M.D. (2013). Evolution of crop species: Genetics of domestication and diversification. *Nature Reviews Genetics* **14**, 840–852.
- Morgante, M., De Paoli, E. and Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* **10**, 149–155.
- Mrode, R.A. and Thompson, R. (2005). *Linear Models for the Prediction of Animal Breeding Values*, 2nd edition. CABI, Wallingford.
- Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. and Kuchel, H. (2006). Joint modelling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics* **113**, 809–819.
- Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C. and Waugh, R. (2016). Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics* **6**, 1313–1326.

- Oettler, G., Tams, S.H., Utz, H.F., Bauer, E. and Melchinger, A.E. (2005). Prospects for hybrid breeding in winter triticale: I. Heterosis and combining ability for agronomic traits in European elite germplasm. *Crop Science* **45**, 1476–1482.
- Ogutu, J.O., Schulz-Streeck, T. and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings* **6**: S10.
- Oradas, B., Malvar, R.A. and Hill, W.G. (2008). Genetic variation and quantitative trait loci associated with developmental stability and the environmental correlation between traits in maize. *Genetics Research* **90**, 385–395.
- Ossó, A., Sutton, R., Shaffrey, L. and Dong B. (2017). Observational evidence of European summer weather patterns predictable from spring. *Proceedings of the National Academy of Sciences of the United States of America* **115**(1), 59–63.
- Paderewski, J. and Rodriguez, P.C. (2014). The usefulness of EM-AMMI to study the influence of missing data pattern and application to polish post-registration winter wheat data. *Australian Journal of Crop Science* **8**, 640–645.
- Papadakis, J. S. (1937). Méthode statistique pour des expériences sur champ. *Bulletin de l'Institut de l'Amélioration des Plantes Thessalonique* **23**.
- Pasanen, L., Holmström, L. and Sillanpää, M.J. (2015). Bayesian LASSO, scale space and decision making in association genetics. *PLoS ONE* **10**, e0120017.
- Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.
- Patterson, H.D. and Williams, E.R. (1976). A new class of resolvable incomplete block designs. *Biometrika* **63**, 83–92.
- Patterson, H.D. and Silvey, V. (1980). Statutory and recommended list trials of crop varieties in the United Kingdom. *Journal of the Royal Statistical Society, Series A* **143**, 219–252.
- Piepho, H.-P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. *Theoretical and Applied Genetics* **97**, 195–201.
- Piepho, H.-P. (2001). Exploiting quantitative information in the analysis of dominant markers. *Theoretical and Applied Genetics* **103**, 462–468.
- Piepho, H.-P. and Möhring, J. (2005). Best linear unbiased prediction for subdivided target regions. *Crop Science* **45**, 1151–1159.
- Piepho, H.-P. and Möhring, J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881–1888.
- Piepho, H.-P., Büchse, A. and Truberg, B. (2006). On the use of multiple lattice designs and α -designs in plant breeding trials. *Plant Breeding* **125**, 523–528.
- Piepho, H.-P., Möhring, J., Schulz-Streeck, T. and Ogutu, J.O. (2012). A stage-wise approach for analysis of multi-environment trials. *Biometrical Journal* **54**, 844–860.
- Piepho, H.-P., Williams, E.R. and Ogutu, J.O. (2013). A two-stage approach to recovery of inter-block information and shrinkage of block effect estimates. *Communications in Biometry and Crop Science* **8**, 10–22.
- Renner, S.S. and Ricklefs, R.E. (1995). Dioecy and its correlates in the flowering plants. *American Journal of Botany* **82**, 596–606.
- Richards, A.J. (1996). Breeding systems in flowering plants and the control of variability. *Folia Geobotanica Phytotaxonomica* **31**, 282–293.
- Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Liseic, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L. and Melchinger, A.E. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* **44**, 217–220.
- Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

- Rodrigues, P., Malosetti, M., Gauch, H.G. and van Eeuwijk, F.A. (2015). A weighted AMMI algorithm to study genotype-by-environment interaction and QTL-by-environment interaction. *Theoretical and Applied Genetics* **54**, 1555–1570.
- Rodriguez-Alvarez, M.X., Boer, M.P., van Eeuwijk, F.A., and Eilers, P.H.C. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics* **23**, 52–71.
- Romagosa, I., Ullrich, S.E., Han, F. and Hayes, P.M. (1996). Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theoretical and Applied Genetics* **93**, 30–37.
- Rönnegård, L. and Valdar, W. (2011). Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics* **188**, 435–447.
- Rönnegård, L. and Valdar, W. (2012). Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics* **13**, 63.
- Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N., Eade, R., Fereday, D., Folland, C.K., Gordon, M. and Hermanson, L. (2014). Skilful long-range prediction of European and North American winters. *Geophysical Research Letters* **41**, 2514–2519.
- Schmitz Carley, C.A., Coombs, J.J., Douches, D.S., Bethke, P.C., Palta, J.P., Novy, R.G. and Endelman, J.B. (2017). Automated tetraploid genotype calling by hierarchical clustering. *Theoretical and Applied Genetics* **30**, 717–726.
- Schnell, F.W. (1982). A synoptic study of the methods and categories of plant breeding. *Plant Breeding* **89**, 1–18.
- Schrag, T.A., Möhring, J., Kusterer, B., Dhillon, B.S., Melchinger, A.E., Piepho, H.P. and Frisch, M. (2010). Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theoretical and Applied Genetics* **120**, 451–473.
- Serang, O., Mollinari, M. and Garcia, A.A.F. (2012). Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE* **7**, e30906.
- Servin, B., Martin, O.C. and Mézard, M. (2004). Toward a theory of marker-assisted gene pyramiding. *Genetics* **168**, 513–523.
- Shah, R., Cavanagh, C.R. and Huang, B.E. (2014). Computationally efficient map construction in the presence of segregation distortion. *Theoretical and Applied Genetics* **127**, 2585–2597.
- Smith, A.B., Cullis, B.R. and Thompson, R. (2001). Analysing variety by environment data using multiplicative mixed models. *Biometrics* **57**, 1138–1147.
- Smith, A.B., Cullis, B.R. and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *Journal of Agricultural Science* **143**, 449–462.
- Smith, A.B., Lim, P. and Cullis, B.R. (2006). The design and analysis of multi-phase plant breeding experiments. *Journal of Agricultural Science* **144**, 393–409.
- Song, G., Jia, M., Chen, K., Kong, X., Khattak, B., Xie, C., Li, A. and Mao, L. (2016). CRISPR/Cas9: A powerful tool for crop genome editing. *Crop Journal* **4**, 75–82.
- Speed, D. and Balding, D.J. (2014). MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Research* **24**, 1550–1557.
- Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant Journal* **3**, 739–744.
- Swarts, K., Li, H., Romero Navarro, J.A., An, D., Romay, M.C., Hearne, S., Acharya, C., Glaubitz, J.C., Mitchell, S., Elshire, R.J. and Buckler, E.S. (2014). Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. *Plant Genome* **7**(3).
- Talbot, M. (1997). Resource allocation for selection systems. In: R.A., Kempton and P.N., Fox (eds.), *Statistical Methods for Plant Variety Evaluation*. Chapman & Hall, London, pp. 162–174.

- Tanaka, R. and Iwata, H. (2018). Bayesian optimization for genomic selection: A method for discovering the best genotype among a large number of candidates. *Theoretical and Applied Genetics* **131**, 93–105.
- Thuillet, A.C., Bataillon, T., Poirier, S., Santoni, S. and David, J.L. (2005). Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics* **169**, 1589–1599.
- Tiede, T., Kumar, L., Mohammadi, M. and Smith, K.P. (2015). Predicting genetic variance in bi-parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. *Molecular Breeding* **35**, 199.
- Uauy, C. (2017). Wheat genomics comes of age. *Current Opinion in Plant Biology* **36**, 142–148.
- van Eeuwijk, F.A., Bustos-Korts, D.V. and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science* **56**, 2119–2140.
- Voorrips, R.E., Gort, G., Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* **12**, 172.
- Watson, A., Ghosh, S., Williams, M., Cuddy, W.S., Simmonds, J., Rey, M.D., Hatta, M.A.M., Hinchliffe, A., Steed, A., Reynolds, D., Adamski, N., Breakspear, A., Korolev, A., Rayner, T., Dixon, L.E., Riaz, A., Martin, W., Ryan, M., Edwards, D., Batley, J., Raman, H., Rogers, C., Domoney, C., Moore, G., Harwood, W., Nicholson, P., Dieters, M.J., DeLacy, I.H., Zhou, J., Uauy, C., Boden, S.A., Park, R.F., Wulff, B.B.H. and Hickey, L.T. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nature Plants* **4**, 23–28.
- Weir, B.S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, MA.
- Williams, E.R., John, J.A. and Whitaker, D. (2014). Construction of more flexible and efficient p-rep designs. *Australian and New Zealand Journal of Statistics* **56**, 89–96.
- Williams, E.R. and Piepho, H.P. (2013). A comparison of spatial designs for field variety trials. *Australian and New Zealand Journal of Statistics* **55**, 253–258.
- Woolliams, J.A., Berg, P., Dagnachew, B.S. and Meuwissen, T.H.E. (2015). Genetic contributions and their optimization. *Journal of Animal Breeding and Genetics* **132**, 89–99.
- Wricke, G. (1962). On a method of understanding the biological diversity in field research. *Zeitschrift für Pflanzenzüchtung* **47**, 92–146.
- Xu, S., Xu, Y., Gong, L., Zhang, X. (2012). Metabolic prediction of yield in hybrid rice. *Plant Journal* **88**, 219–227.
- Yan, W. (2014). *Crop Variety Trials: Data Management and Analysis*. John Wiley & Sons, Hoboken, NJ.
- Yates, F., Cochran, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science* **28**, 556–580
- Yu, J., Holland, J.B., McMullen, M.D. and Buckler, E.S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539–551.
- Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J., Zhang, Q. (2012). Genetic composition of yield heterosis in an elite rice hybrid. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 15847–15852.

18

Forensic Genetics

B.S. Weir

Department of Biostatistics, University of Washington, Seattle, WA, USA

Abstract

The use of DNA profiles for human identification often requires statistical genetic calculations. The probabilities for a crime scene DNA profile can be evaluated under alternative hypotheses about the contributor(s) to a profile, and presented as likelihood ratios. It is conditional probabilities that are needed: the probabilities of unknown people having specific profiles, given the profiles seen in known people. These probabilities depend on the relationships between known and unknown people and the populations to which they belong, and the algebraic treatment is greatly simplified when it can be assumed that allelic frequencies have Dirichlet distributions over populations. The growing sensitivity of genotyping for forensic samples has meant that many profiles represent contributions from more than one contributor. This, in turn, has required new attention to artifacts, such as allele drop-out, drop-in and stutter, in the profiles generated by capillary-gel electrophoresis and a move towards the use of statistical algorithms and software to evaluate the profiles. There is also a growing attention to using DNA sequencing to generate forensic profiles, a methodology with its own statistical challenges. Once appropriate match statistics have been calculated, there is need for care in presenting the evidence of a match between the profiles of a person of interest and an evidentiary item, with distinctions between commission of a crime, the activity leading to deposition of DNA, and identity of the source of the evidentiary item.

18.1 Introduction

Human individualization based on the genome exploits the fact that every pair of people is genetically distinguishable. Moreover, human genetic material is found in every nucleated cell in the body and can be recovered from samples as diverse as bone, blood stains, saliva residues, nasal secretions, and even fingerprints. DNA may be recovered from very old samples that have been well preserved, and some DNA signatures may be preserved over successive generations.

Genetic markers have been used for human individualization since the discovery of blood groups by Landsteiner in 1900 (see Tan and Graham, 2013), and statistical genetic arguments have long played a large part in parentage dispute cases. The role of statistical genetics in forensic science increased sharply in the late 1980s when DNA markers began to be used and emphasis shifted from excluding specific people as being the sources of evidentiary profiles to

making probability statements about genetic profiles if these people were not excluded as the sources. As more markers have been developed, it has become less likely that two people would share the same DNA profile and therefore more likely that people will be convicted on the basis of DNA evidence. Given the serious nature of the charges of many crimes where genetic markers are used, and the serious consequences of conviction for these crimes, there has been considerable scrutiny paid to the statistical genetic arguments upon which forensic probabilities are based. These arguments are reviewed in this chapter.

A simple situation is where DNA is recovered from a biological sample left at the scene of a crime, and there is reason to believe the sample is from the perpetrator of the crime. DNA is also extracted, most likely from a buccal swab, from a suspect in the crime and is found to have the same profile as that of the crime sample. An immediate question is how much evidence against the suspect is provided by this matching, and a naive answer might be that the probability of the crime sample profile, if not from the suspect, is the population proportion of that profile. High values for this proportion would favor the suspect and low values would not. How is the proportion to be estimated?

For genetic profiles based on single genetic loci, it is quite feasible to estimate the population proportion of any genotype on the basis of a moderate-sized sample of profiles from that population. The size of the sample would need to be greater for highly variable loci where there are many different genotypes and there are issues of how the population is to be defined and sampled. As the number of loci used for identification increases, the numbers of genotypes increases substantially and no sample can hope to capture all genotypes. Current practice in the United States is to use a set of 20 short tandem repeat (STR) loci, each with at least 10 alleles and 55 genotypes. The possible number of 20-locus profiles is therefore more than 10^{34} , so that less than one profile in 10^{24} of all possible profiles exists anywhere in the world. Although there may not, therefore, be much value in declaring that a particular profile is rare, a first attempt to attach a probability to a 20-locus profile might be to multiply together the probabilities of the 40 constituent alleles, along with a factor of 2 for every heterozygote. The implied assumption of allelic independence, within and between loci, is a daunting statistical genetic issue to justify. Moreover, it is difficult to describe just how unlikely is an event with a probability of the order of 10^{-30} : even the 'birthday problem' approach of asking the chance that at least two people in a world of 10^{10} people would have the same unspecified profile (as opposed to the same particular profile) produces very small numbers.

A more satisfactory approach has been to ask the question: What is the probability that an unknown person has a particular genetic profile, given that the profile has been seen already for a person of interest (POI) in this crime? Calculating this conditional probability needs to take into account the relationship between the known POI and the unknown person. This relationship may be due to close family membership or to shared evolutionary history. Once again, these are statistical genetic issues. Taking these issues into account can have a significant impact on likelihood ratios for forensic profiles from less than 20 autosomal loci or for mixtures when typing artifacts may be confounded with genetic signals.

18.2 Principles of Interpretation

Evett and Weir (1998) suggested that genetic evidence be interpreted according to three principles:

- *First Principle.* To evaluate the uncertainty of any given proposition it is necessary to consider at least one alternative proposition.

- *Second Principle.* Interpretation is based on questions of the kind ‘What is the probability of the evidence given the proposition?’
- *Third Principle.* Interpretation is conditioned not only by the competing propositions, but also by the framework of circumstances within which they are to be evaluated.

The first two of these principles lead to the use of likelihood ratios, as will soon be shown. The question in the second principle is in contrast to the very common ‘prosecutor’s fallacy’ (Thompson and Schumann, 1987) that answers the transposed question ‘What is the probability of the proposition given the evidence?’ The third principle recognizes the difference, for example, in the strength of evidence of a blood stain at the scene of a crime having a profile matching that of a POI, and the evidence of bloodstain in the clothing of a POI away from the crime scene with a profile matching that of a victim.

The propositions mentioned in the principles refer, in this context, to the source of the genetic profile in the evidentiary stain. For the immediate discussion these will be taken to be those reflecting the views of the prosecution (H_p) and defense (H_d):

H_p : The profile is from the person of interest.

H_d : The profile is from some other person.

Suppose G_S and G_C are the profile types from the POI and the evidentiary stain, and they are found to match. Then the evidence E is these two profiles: $E = (G_S, G_C)$. Consideration of alternative propositions is carried out by comparing the probabilities of E under these propositions by means of the likelihood ratio (LR),

$$\begin{aligned} \text{LR} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_C|G_S, H_p)}{\Pr(G_C|G_S, H_d)} \times \frac{\Pr(G_S|H_p)}{\Pr(G_S|H_d)} \\ &= \frac{1}{\Pr(G_C|G_S, H_d)}, \end{aligned} \quad (18.1)$$

with the last step depending on the assumption that the profiles must be found to match when they have a common source (i.e. ignoring potential DNA profiling errors), and on recognition that G_S does not depend on H_p or H_d .

The forensic question of attaching weight to matching genetic profiles has therefore reduced to the statistical genetic question of determining the probability of a profile given that (the same) profile has been seen already. This conditional probability will be referred to as the match probability. It would be a substantial simplification to assume the two profiles were independent and work only with the probability of the crime stain profile,

$$\text{LR} = \frac{1}{\Pr(G_C|H_d)} = \frac{1}{\Pr(G_C)},$$

but that assumption is not correct for real populations and it can be avoided, as shown below with the ‘ θ correction’.

Bayes’ theorem can be expressed algebraically in odds form as

$$\frac{\Pr(H_p|E)}{\Pr(H_d|E)} = \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \times \frac{\Pr(H_p)}{\Pr(H_d)},$$

and verbally as

Posterior odds on H_p = LR × Prior odds on H_p .

Table 18.1 Hierarchy of propositions for a burglary case

Level	Type	Hypotheses
III	Offense	$H_p(III)$: The defendant committed the burglary. $H_d(III)$: Some other person committed the burglary.
II	Activity	$H_p(II)$: The defendant broke the window. $H_d(II)$: Some other person broke the window.
I	Source	$H_p(I)$: The blood on the broken window is from the defendant. $H_d(I)$: The blood on the broken window is from some other person.

This makes clear that the LR captures the strength of the DNA evidence. The odds expressing belief in the prosecution hypothesis before the evidence is presented are multiplied by the LR to give odds expressing belief after presentation. Although this formalizes the interpretation of evidence, it is unlikely to be articulated in court.

This discussion has been about the source of DNA in the evidentiary profile. At trial, the issue is about who committed the crime for which the DNA profile is relevant. A useful hierarchy of propositions was described by Cook *et al.* (1998) and illustrated in Table 18.1 for a burglary case where the profile was developed from a blood stain on a broken window at the burgled premises. The trier of fact is interested in the level III propositions and will reach a decision after hearing about the evidence of blood on the broken window having the same profile as the defendant. The forensic scientist, and this chapter, can address only the level I propositions.

18.3 Profile Probabilities

Although it is match (conditional profile) probabilities that are needed, we begin by considering (unconditional) profile probabilities. At a single locus **A** with alleles A_u having probabilities p_u , the probabilities $\Pr(A_uA_v)$ for genotypes A_uA_v within a single population may be parameterized as

$$\Pr(A_uA_v) = \begin{cases} p_u^2 + fp_u(1 - p_u), & u = v, \\ 2p_u p_v - 2fp_u p_v, & u \neq v. \end{cases} \quad (18.2)$$

As the loci used for forensic typing are unlikely to be under the influence of selection, the use of a single inbreeding coefficient f for all genotypes may be thought reasonable, although genotype-specific coefficients should strictly be used for loci with allele-specific mutation processes (Graham *et al.*, 2000). Expressing genotype probabilities as functions of allele probabilities is necessary for highly variable loci, when many genotypes are not seen in a sample and their population probabilities are difficult to estimate by direct observation. As an aside, the FBI databases used for calculations in the USA have sample sizes around 200 (Moretti *et al.*, 2016), so that many of the 55 or more possible genotypes at a locus have observed counts of zero.

Equation (18.2) allows for departures from Hardy–Weinberg equilibrium (HWE) within a single population. In practice HWE is assumed when population genotype probabilities are estimated from database allele proportions. Following the early claim by Lander (1989; refuted by Devlin *et al.*, 1990) of HWE departures in a Lifecodes database, thousands of HWE tests in forensic databases have been conducted and generally non-significant results published.

Although the need for the HWE assumption has largely been circumvented by the Balding and Nichols (1994) approach described below, HWE testing still has a role as a quality control measure (Balding and Steele, 2015) but is not as important as was once thought.

HWE refers to independence of the pairs of alleles within a locus. Linkage disequilibrium (LD) refers to dependence of pairs of alleles, one from each of two loci, but the term is more loosely applied to genotypes across many loci. Any statistical issues in testing for HWE would be magnified if it were desired to test for dependencies among all 40 alleles in 20-locus profiles for loci with many possible allelic states. There have been several locus-pair investigations undertaken with the expectation that an absence of detectable associations among these sets of four alleles would support an assumption of little association among alleles at larger numbers of loci (e.g. Moretti *et al.*, 2016). It has long been known (e.g. Cockerham and Weir, 1973) that non-zero identity disequilibrium, reflecting the difference of two-locus genotype probabilities from products of one-locus probabilities, even for unlinked loci in linkage equilibrium, can exist in random mating populations. As with HWE testing, however, there has been a movement to a consideration of match probabilities and an adoption of genetic models, as discussed next.

18.3.1 Genetic Models for Allele Frequencies

The issues of allelic dependence, within and among loci, have been addressed in practice by work of Balding and Nichols (1994). Their work brings an evolutionary perspective to the consideration of allele and genotype probabilities. Equation (18.2) relates genotype to allele probabilities in a single extant population and, although f is referred to as the ‘within-population inbreeding coefficient’ (or F_{IS}), no particular evolutionary model is implied by that parameterization. Relatedness between individuals or similarities among populations, however, do require a genetic model reflecting common ancestral origins of alleles between individuals or populations. If these ancestral alleles were in recent generations we say the individuals carrying the current alleles are related. More distant ancestral alleles could be considered as leading to ‘evolutionary relatedness’. The key point is the explicit recognition that present genotype proportions are shaped by events that have taken place in previous generations.

18.3.1.1 Population Structure

From an evolutionary perspective, an extant population can be regarded as just one realization of a process over time involving the stochastic processes of genetic drift, mutation, migration and so on. Predictions of quantities such as match probabilities are expectations over replicates of the population history. The allele proportions p_u in a specific population have expected values over these replicates of π_u , and equation (18.2) has expectation

$$\mathcal{E}(P_{uv}) = \begin{cases} \pi_u^2 + F\pi_u(1 - \pi_u), & u = v, \\ 2\pi_u\pi_v - 2F\pi_u\pi_v, & u \neq v, \end{cases} \quad (18.3)$$

where F is the ‘total inbreeding coefficient’ (or F_{IT}). The Balding–Nichols formulation can be approached by assuming the population-specific proportions p_u have beta distributions with parameters $(1 - \theta)\pi_u/\theta$ and $(1 - \theta)(1 - \pi_u)/\theta$. The ‘population-structure parameter’ θ (or F_{ST}) can be regarded as a measure of relationship for pairs of alleles within a population where relationship can be a correlation of allelic-state indicators or a probability of allelic identity by descent. The trio F_{IS}, F_{IT}, F_{ST} are referred to as ‘Wright’s F -statistics’ (e.g. Wright, 1951), and are related by $F_{IS} = (F_{IT} - F_{ST})/(1 - F_{ST})$. A random-mating population is expected to be in HWE ($F_{IS} = 0$), but this does not imply allelic independence ($F_{IT} = 0$) in an evolutionary sense. HWE does imply $F_{IT} = F_{ST}$.

The beta distribution leads to probabilities of sets of alleles, nicely summarized by what Balding (e.g. Balding and Steele, 2015) calls a sampling formula for alleles. Suppose n alleles have been sampled from a population and n_u are of type A_u . The probability that the next allele sampled is also of type A_u is

$$\Pr(A_u | n_u \text{ of type } A_u \text{ in } n) = \frac{n_u \theta + (1 - \theta) \pi_u}{1 + (n - 1)\theta}. \quad (18.4)$$

The probability of two A_u alleles given two A_u alleles, or the match probability for $A_u A_u$ homozygotes is, therefore,

$$\Pr(A_u A_u | A_u A_u) = \frac{[3\theta + (1 - \theta)\pi_u][2\theta + (1 - \theta)\pi_u]}{(1 + \theta)(1 + 2\theta)}, \quad (18.5)$$

and the probability of an A_u and an A_v , in either order, given an A_u and an A_v , with $u \neq v$ is the match probability for $A_u A_v$ heterozygotes,

$$\Pr(A_u A_v | A_u A_v) = \frac{2[\theta + (1 - \theta)\pi_u][\theta + (1 - \theta)\pi_v]}{(1 + \theta)(1 + 2\theta)}. \quad (18.6)$$

These results were endorsed by the US National Research Council (1996) and are now used widely. They clarify that the match probabilities $\Pr(A_u A_v | A_u A_v)$ for all u, v are greater than the profile probabilities $\Pr(A_u A_v)$ when $\theta > 0$. Even when the profile probability is low, the chance of finding it again in a population increases after it has been found once. They also clarify that HWE is not assumed for expected genotype proportions.

Equations (18.5) and (18.6) require knowledge of the parameters π_u and θ . Allele probabilities π_u are replaced by sample proportions \tilde{p}_u in databases such as those reported by Moretti *et al.* (2016) and categorized by ancestral origin: ‘Caucasian’, ‘African-American’ and so on. It is recognized that the relevant population for a particular crime, even if the origins of that population are known, is unlikely to be the same as the population from which the database was constructed. Variation among populations within the broad group represented by the database is considered equivalent to variation among evolutionary replicates of a single population and θ represents variation among (sub)populations within the database population. For this reason, θ can be referred to as a population-structure parameter. The National Research Council (1996) suggested values such as $\theta = 0.01$ for major ethnic groups, and higher values, such as $\theta = 0.03$, for smaller groups like Native Americans. Steele *et al.* (2014), on the basis of estimating θ in worldwide national populations relative to continental-scale databases, recommended $\theta = 0.03$ as being almost always conservative, even if the source of the DNA is from a different continent than the suspected source.

Results consistent with those of Steele *et al.* (2014) were found in another comprehensive survey of forensic STR marker allelic data by Buckleton *et al.* (2016). These authors extracted allelic sample proportions from 250 publications and estimated θ values within and between continental-ancestry groups of populations. This survey included Native American populations, and for those populations the relevant θ functions can be as high as 0.10 if world-wide allelic proportions are used. They invoked an estimation procedure based on allelic matching and described in more detail by Weir and Goudet (2017). Those authors wrote the sample proportion of matching pairs of alleles sampled from population i , without regard to which individuals carry those alleles, as \tilde{M}_i , and the sample proportion of matching pairs of alleles, one taken from each of two populations, and averaged over pairs of populations, as \tilde{M}_B . They estimated θ_i , the relationship for pairs of alleles within population i , relative to the relationship of pairs of alleles from pairs of populations, as $\hat{\beta}_i = (\tilde{M}_i - \tilde{M}_B)/(1 - \tilde{M}_B)$. When database sample proportions \tilde{p}_u are used in place of population proportions π_u in equations (18.5) and (18.6), it is appropriate to use the average $\hat{\beta}_W$ of the population-specific values $\hat{\beta}_i$ and the match

Table 18.2 Relationship probabilities for common relatives

Relationship	k_2	k_1	k_0
Identical twins	1	0	0
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Parent–child	0	1	0
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Half sibs*	0	$\frac{1}{2}$	$\frac{1}{2}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$
Unrelated	0	0	1

*Also grandparent–grandchild and avuncular (e.g. uncle–niece).

probability estimates will apply to any population within the group of populations represented by the database. The work of Weir and Goudet (2017) applies to haploid (e.g. Y-haplotypes) and diploid data.

18.3.1.2 Relatedness

Equations (18.5) and (18.6) rest on the joint genotypic probabilities of sets of alleles with a shared evolutionary history. The expressions have an implicit HWE assumption within a single replicate population ($F_{IS} = 0$) and the equivalence of pairs of alleles within individuals with those between individuals within the same population ($F_{IT} = F_{ST}$). Related individuals, however, may share alleles from recent common ancestors: if neither is inbred the probabilities that they share 0, 1 or 2 pairs of alleles identical by descent are written as k_0, k_1, k_2 , respectively. Values for these probabilities for common pairs of relatives are shown in Table 18.2, and expressions for joint genotypic probabilities are shown in Table 18.3. From these tables, the likelihood ratio for the propositions

- H_p : The profile is from the POI,
 H_d : The profile is from a brother of the POI

when the POI and evidence profiles are homozygous A_uA_u at a single locus, ignoring population structure, is

$$\text{LR} = \frac{4}{(1 + \pi_u)^2},$$

and this can be substantially smaller than the $1/\pi_u^2$ for unrelated alternative sources.

Table 18.3 Joint genotypic probabilities for pairs of relatives

Genotypes*	Probability
uu, uu	$k_0\pi_u^4 + k_1\pi_u^3 + k_2\pi_u^2$
uu, vv	$k_0\pi_u^2\pi_v^2$
uu, uv	$2k_0\pi_u^3\pi_v + k_1\pi_u^3\pi_v$
uu, vw	$2k_0\pi_u^2\pi_v\pi_w$
uv, uv	$4k_0\pi_u^2\pi_v^2 + k_1\pi_u\pi_v(\pi_u + \pi_v) + 2k_2\pi_u\pi_v$
uv, uw	$4k_0\pi_u^2\pi_v\pi_k + k_1\pi_u\pi_v\pi_k$
uv, wz	$4k_0\pi_u\pi_v\pi_w\pi_z$

* $u \neq v \neq w \neq z$.

To account for both population structure and relatedness, the four alleles carried by two individuals are reduced to sets of two, three or four alleles identical by descent from family relatedness and then Balding's sampling formula used for evolutionary relatedness. For two related homozygotes or heterozygotes,

$$\Pr(A_u A_u, A_u A_u) = k_0 \Pr(A_u A_u A_u A_u) + k_1 \Pr(A_u A_u A_u) + k_2 \Pr(A_u A_u),$$

$$\Pr(A_u A_v, A_u A_v) = 4k_0 \Pr(A_u A_u A_v A_v) + k_1 [\Pr(A_u A_u A_v) + \Pr(A_u A_v A_v)] + 2k_2 \Pr(A_u A_v),$$

so the match probabilities are

$$\Pr(A_u A_v | A_u A_v) = \begin{cases} k_0 \frac{[2\theta + (1-\theta)\pi_u][3\theta + (1-\theta)\pi_u]}{(1+\theta)(1+2\theta)} + k_1 \frac{2\theta + (1-\theta)\pi_u}{1+\theta} + k_2, & u = v, \\ k_0 \frac{2[\theta + (1-\theta)\pi_u][\theta + 1 - \theta]\pi_v}{(1+\theta)(1+2\theta)} + k_1 \frac{2\theta + (1-\theta)(\pi_u + \pi_v)}{2(1+\theta)} + k_2, & u \neq v. \end{cases}$$

Parameters π_u and θ are assumed to have the same value in successive generations, so that the same level of approximation holds as in equations (18.5) and (18.6).

18.3.1.3 Multi-locus dependencies

The hope has been that using the single-locus 'theta-corrections' at each locus, equations (18.5) and (18.6), would compensate for between-locus dependencies, as discussed by Balding and Steele (2015) and supported empirically by Weir (2004). Dependencies among single-locus match probabilities arising from relatedness were discussed by Donnelly (1995):

after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption.

A theoretical prediction of dependencies in match probabilities, focusing on the joint effects of mutation and genetic drift, was given by Laurie and Weir (2003). Notwithstanding these theoretical predictions, an empirical study of matching at up to six loci was given by Weir (2004): among 15,000 forensic STR profiles the ratios of multi-locus match proportions to products of single-locus proportions were 1.000, 1.000, 1.008, 1.034 and 1.041 for two, three, four, five and six loci. This observation suggests that attempting to test for multi-locus LD is not necessary as dependencies are expected, and it also suggests the need to question the extent to which increasing the number of loci will decrease the probability of matching between pairs of profiles. Even more, it raises the issue of increasing inconsistency between actual match probabilities and values predicted by taking products over loci. Empirical work does support the use of sufficiently large single-locus θ s to address the issue. The upper panels in Figure 18.1 show an increasing excess of multi-locus match proportions over products of one-locus proportions as the number of loci increases from two to four in a set of 2849 STR profiles (Weir and Zhao, in preparation). The lower panels show a decrease in the excess as θ increases from 0 to 0.01 for five-locus matches. It would be prejudicial to present the product of single-locus match proportions when the multi-locus proportions were greater. It is not yet known how large θ should be to prevent prejudice for all 20-locus profiles – the extreme rarity of matches at large numbers of loci rules out empirical studies on real data.

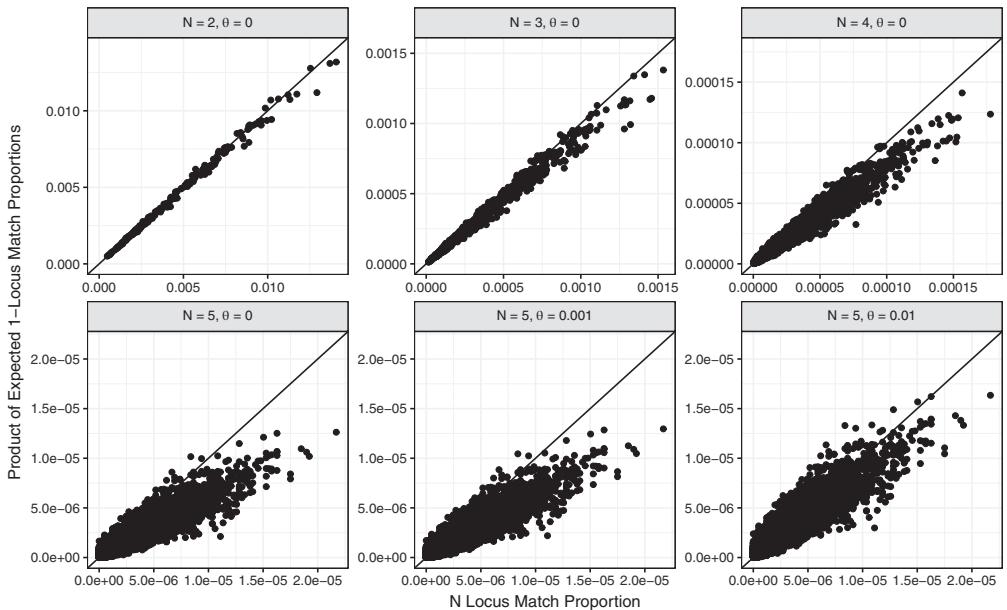


Figure 18.1 Multi-locus match proportions for 2849 STR profiles. The top row of plots compare observed multi-locus match proportions to products of single-locus proportions for two, three and four loci. The bottom row of plots compare five-locus observed match proportions with products of one-locus match probabilities calculated with $\theta = 0, 0.001$ and 0.01 .

18.3.2 Y-STR Profiles

The exclusive paternal transmission of Y chromosomes from fathers to sons makes Y-STR haplotypes of use in associating males to evidentiary profiles, especially in sexual assault cases where most of the DNA in a profile is from the female victim. Y-typing detects only the male chromosomal component of a profile. The absence of recombination in the non-recombining portion of the Y chromosome argues against independence of single-locus Y-STR match proportions, although the independence of mutational events at different loci results in low levels of LD (Hall, 2016).

The simplest alternative to multiplying match proportions over loci is to determine haplotype proportions in a database. The current US Y-STR database (<https://www.usystrdatabase.org/>) has a total of 35,000 profiles from five ancestral groups, scored with varying numbers of 29 Y-STR loci. The YHRD database (www.yhrd.org) has over 200,000 profiles from over 1100 populations (over 30 ancestral groups), typed at varying numbers of 27 Y-STR loci. The limitation of these databases for giving profile proportions is that a profile not observed at some number of loci will also not be observed at an increased number of loci, even though the strength of the evidence would seem to increase with the number of loci.

Just as with autosomal profiles, however, there are diminishing returns from adding loci past some optimal number of loci. A match at 10 or more loci suggests two profiles from the same male lineage and further matching is expected at additional loci apart from some mutations.

The relationship between multi-locus profile probabilities and the product of single-locus probabilities can be addressed with the concept of entropy (Caliebe *et al.*, 2015; Siegert *et al.*, 2015). For a locus with sample frequencies \tilde{p}_u for alleles A_u the entropy is

$$H_A = - \sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

For independent loci, entropies are additive: if haplotypes $A_u B_v$ have sample frequencies \tilde{P}_{uv} the two-locus entropy is

$$H_{AB} = - \sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv}) = - \sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] = H_A + H_B,$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence in the sense $\tilde{P}_{uv} \neq \tilde{p}_u \tilde{p}_v$.

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B , given A , is $H_{B|A} = H_{AB} - H_A$. In constructing Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy or highest discriminatory power. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype $L_1 L_2$, and so on. Some results for the YHRD database obtained by Hall (2016) are shown in Table 18.4. This table shows that the most-discriminating loci, when considered separately, may not contribute appreciably to already discriminating haplotypes. Furthermore, there is little additional discriminating power from Y-STR haplotypes beyond 10 loci.

The match probability formulation of Balding and Nichols (1994) is also relevant for Y-haplotypes. The probability for haplotype A , given that it has already been seen, is

$$\Pr(A|A) = \theta + (1 - \theta)\pi_A, \quad (18.7)$$

where θ is the probability of identity by descent for two haplotypes drawn randomly from a population and π_A is the probability a single haplotype is of type A . The value of θ decreases

Table 18.4 Entropy measures for Y-STR markers

Marker	Entropy		
	Single	Combined	Conditional*
1. DYS385ab	4.750	4.750	4.750
2. DYS570	2.554	8.447	1.474
3. DYS458	2.220	9.741	0.423
4. DYS549	1.719	9.999	0.093
5. DYS19	2.112	10.08	0.028
6. DYS533	1.433	10.11	0.010
7. GATAH4	1.512	10.12	0.005
8. DYS448	1.858	10.13	0.002
9. DYS390	1.844	10.13	0.002
10. DYS481	2.962	6.972	2.222
11. DYS576	2.493	9.318	0.871
12. DYS389II	2.329	9.906	0.165
13. DYS635	2.136	10.05	0.053
14. DYS439	1.637	10.10	0.024
15. DYS456	1.691	10.12	0.006
16. DYS393	1.654	10.13	0.003
17. DYS643	2.456	10.13	0.002
18. DYS391	1.058	10.13	0.002

*Conditional on previously added markers.

Table 18.5 Predicted autosomal and Y-STR θ values

N	μ	θ_Y	$\theta_{Y A}$	θ_A	$\theta_{A Y}$
10^4	10^{-3}	0.00244	0.00370	0.01233	0.01868
	10^{-4}	0.02434	0.02447	0.11110	0.11168
10^5	10^{-3}	0.00024	0.00151	0.00125	0.00768
	10^{-4}	0.00249	0.00262	0.01234	0.01300
10^6	10^{-3}	0.00002	0.00129	0.00012	0.00656
	10^{-4}	0.00025	0.00038	0.00125	0.00191

with the number of loci, and for 20 loci Hall (2016) estimated values of the order of 10^{-5} . For very rare haplotypes, the match probability reduces to θ for the relevant population.

Walsh *et al.* (2008) discussed the interpretation of matching autosomal profiles from men with matching Y-chromosome profiles. Even though the autosomes and the Y chromosome are distinct entities, their profile probabilities are not independent. A high degree of Y-matching suggests the men are related and so are more likely to show autosomal matching, and vice versa. By iterating transition equations for finite random-mating populations (Weir, in preparation) it was possible to generate joint probabilities of identity by descent, shown in Table 18.5. The values are for a single autosomal locus and a 20-locus Y haplotype. The mutation rate was the same at all loci. The joint measure θ_{AY} is the probability that a pair of autosomal alleles, one from each of two men, is identical by descent at the same time as their Y profiles are identical by descent. There are two conditional coancestries: autosomal given Y, $\theta_{A|Y} = \theta_{AY}/\theta_Y$, and Y given autosomal, $\theta_{Y|A} = \theta_{AY}/\theta_A$. The effect of matching in one system on matching in the other increases with mutation rate μ and decreases with population size N .

18.3.2.1 Other Approaches

The difficulty in providing numerical strength for matching Y-STR profiles based on many loci has been addressed in other ways. Two of these alternatives focus on the lack of recombination leading to haplotype lineages with mutation causing changes within lineages. Brenner (2010) used the fraction κ of a database of size n , augmented by the evidence sample, that consists of singletons (once-observed types). He showed that the probability for a random innocent suspect to match a previously unobserved crime scene type is $(1 - \kappa)/n$. This can be substantially less than the database proportion of $1/n$. Brenner (2013) also emphasized the distinction between profile probabilities and match probabilities, and argued against the use of θ for Y-haplotypes.

Andersen and Balding (2017) used population genetic theory and simulations to generate the distribution of the number of males with a matching Y profile. They showed that this distribution is robust to different values for the variance in reproductive success and the population growth rate. They found that conditioning on the number of copies of the profile in a database had only a modest impact. They suggest presenting evidence in the form different from likelihood ratios:

A Y-chromosome profile was recovered from the crime scene. Mr Q has a matching Y profile and so is not excluded as a contributor of DNA. Using population genetics theory and data, we conclude that the number of males in the population with a matching

Y profile is probably less than 20, and is very unlikely (probability < 5%) to exceed 40. These men or boys span a wide range of ages and we don't know where they live. They are all paternal line relatives of Q, but the relationship may extend over many father-son steps, well beyond the known relatives of Q. Since these individuals share paternal-line ancestry with Q, some of them could be similar to Q in ethnic identity, language, religion, physical appearance and place of residence.

The numbers 20 and 40 in this quotation are typical for those they found.

18.4 Mixtures

Evidentiary samples may contain material from more than one contributor. A common situation is for evidence collected in rape cases where material from the victim, possible consensual partners, and the perpetrator(s) may all be present. Even if some of these people contributed only a small proportion of the DNA in the sample, improved technology has made it easier to detect their alleles in the mixed profile. Probabilities for the set of alleles constituting the genetic evidence E are statements about alleles carried by typed people known to have contributed as well as by unknown people. Consideration of the alleles from people who may have been typed even though they are excluded and/or are hypothesized not to have contributed to the sample can be useful in allowing for the effects of population structure.

18.4.1 Combined Probabilities of Inclusion and Exclusion

There have been efforts to simplify the interpretation of mixtures by considering only the set of genotypes included in the evidence profile. The combined probability of inclusion (CPI) refers to a person whose genotypes at all typed loci are included in the evidence profile, whereas the combined probability of exclusion (CPE) is the probability that a person does not have their genotypes included in at least one locus of the evidence profile. If the evidence profile has alleles a_1, a_2, a_3, a_4 at one locus and alleles b_1, b_2, b_3, b_4 at another locus, the CPI is $PE_a = (p_{a1} + p_{a2} + p_{a3} + p_{a4})^2$ times $PE_b = (p_{b1} + p_{b2} + p_{b3} + p_{b4})^2$ and the CPE is $1 - (1 - PE_a)(1 - PE_b)$.

The use of CPI/CPE has been considered simple in that it does not involve the number of contributors to the evidence profile. The claim that it is easy to explain in court is not persuasive because the simplicity masks inefficiencies (Gill *et al.*, 2006). In the extreme situation of the only two alleles at a locus, a, b , being present in a mixture, then the CPI is $(p_a + p_b)^2 = 1$. If the circumstances of a crime suggest there were two contributors then these two people cannot both be homozygous for the same allele and the probability of the evidence is less than 1.

A less extreme situation is when an evidence profile for a crime known to represent the contributors from two perpetrators has alleles a, b, c, d at a locus. The CPI would be $(p_a + p_b + p_c + p_d)^2$. The likelihood ratio for two known people, with genotypes ab and cd , being the source of the evidence versus two unknown people being the source is $1/(24p_a p_b p_c p_d)$, which is larger than $1/(p_a + p_b + p_c + p_d)^2$. The CPI understates the strength of the evidence and it is difficult to justify any method of interpretation that does not follow the principles of interpretation listed earlier. Calculation of likelihood ratios seems appropriate,

18.4.2 Likelihood Ratios

In line with the principles of evidence interpretation, there need to be alternative propositions concerning the contributors to the evidentiary sample. Some of these contributors will be

known and genotyped people, and some will be unknown people. Those contributors, together with any typed people who are known (under the proposition) not to be contributors, contain among them a set of alleles whose probability depends on the separate allele probabilities and the population structure parameter θ . There is also a factor of 2 for each known heterozygote, and a term for the number of ways of arranging all $2x$ alleles from x unknown people into pairs. There may be different sets of alleles from unknown people under some propositions, and the probabilities for these sets must be added together. The likelihood ratio is the ratio of probabilities under alternative propositions.

With increasing sensitivity of forensic DNA typing, many evidence profiles are found to have multiple contributors, and some of these provide only a small fraction of the detected DNA. This has led to deep examination of the capillary electrophoresis technology in current use for STR typing. Three analysis models have emerged: binary, semi-continuous and continuous.

18.4.2.1 Binary Model

Under the binary model, an allele at an STR locus is declared to be either present or absent in an evidentiary profile. This means that the peak in the electropherogram produced by the typing equipment falls within the range determined by the forensic laboratory to be a valid indicator for the presence of that allele. All the analyses presented so far in this chapter have been under the binary model.

Likelihood ratios were introduced in equation (18.1) in terms of the genotypes for a POI and for some other person. For mixtures, the LRs refer to sets of alleles and these may be from known contributors, known non-contributors and unknown people where 'known' depends on the proposition.

Much of the complexity in dealing with mixtures can be removed by a mnemonic notation, as laid out in Table 18.6 (Curran *et al.*, 1999). There are sets of alleles (not necessarily distinct) that occur in the evidentiary sample (C). For a particular proposition there may be alleles (T) carried by typed people declared to be contributors, alleles (W) carried by unknown contributors to the sample, as well as alleles (V) carried by people declared not to have contributed to the sample. There are corresponding sets of distinct alleles, and these sets are indicated by a g subscript. Note that the same person may be declared to be a contributor to the sample under one proposition, and declared not to be a contributor under another proposition. Note also that the word 'known' in Table 18.6 refers to a value specified by the proposition under consideration.

The alleles in the evidence profile are carried either by typed people declared to be contributors or by unknown people, so that C is the combination (union) of sets T and W . For a given proposition, the probability of the evidence profile depends also on the alleles carried by people who have been typed but are declared by that proposition not to have contributed to the profile. For a proposition in which there are x unknown contributors, the probability is $P_x(T, W, V|C_g)$. This probability is for all $2n_C + 2n_V = 2n_T + 2n_W + 2n_V$ alleles in the sets T, W, V , among which allele A_u occurs $c_u + v_u = t_u + w_u + v_u$ times. The probabilities are added over all possible $n_x = (c + r - 1)! / [(c - 1)!r!]$ distinct sets of w_u . As listed in Table 18.6, c is the number of distinct alleles in C_g and r is the number of alleles carried by unknown people that can be any one of these c alleles.

Generating the n_x sets W is a two-stage process. Some of the alleles in each set must be present: these are the alleles in the set C_g that are not in set T_g . Other alleles are not under this constraint because they already occur in T_g , and there are r_u copies of A_u alleles in this unconstrained set. It is a straightforward computing task to let r_1 range over the integers $0, 1, \dots, r$, then let r_2 range over the integers $0, 1, \dots, r - r_1$, then let r_3 range over the integers $0, 1, \dots, r - r_1 - r_2$, and so on. The final count r_c is obtained by subtracting the sum of

Table 18.6 Notation for mixture calculations

Alleles in the profile of the evidence sample	
C	The set of alleles in the evidence profile
C_g	The set of distinct alleles in the evidence profile.
n_C	The known number of contributors to C
h_C	The unknown number of heterozygous contributors
c	The known number of distinct alleles in C_g
c_u	The unknown number of copies of allele A_u in C
	$1 \leq c_u \leq 2n_C, \sum_{u=1}^c c_u = 2n_C$
Alleles from typed people that H declares to be contributors	
\mathcal{T}	The set of alleles carried by the declared contributors to C
\mathcal{T}_g	The set of distinct alleles carried by the declared contributors
n_T	The known number of declared contributors to C
h_T	The known number of heterozygous declared contributors
t	The known number of distinct alleles in \mathcal{T}_g carried by n_T declared contributors
t_u	The known number of copies of allele A_u in \mathcal{T}
	$0 \leq t_u \leq 2n_T, \sum_{u=1}^t t_u = 2n_T$
Alleles from unknown people that H declares to be contributors	
\mathcal{W}	The sets of alleles carried by the unknown contributors to C
x	The specified number of unknown contributors to C : $n_C = n_T + x$
$c - t$	The known number of alleles that are required to be in \mathcal{W}
r	The known number of alleles in \mathcal{W} that can be any allele in C_g , $r = 2x - (c - t)$
n_x	The number of different sets of alleles \mathcal{W} , $n_x = (c + r - 1)! / [(c - 1)!r!]$
r_u	The unknown number of copies of A_u among the r unconstrained alleles in \mathcal{W}
	$0 \leq r_u \leq r, \sum_{u=1}^r r_u = r$
w_u	The unknown number of copies of A_u in \mathcal{W} : $c_u = t_u + u_u, \sum_{u=1}^c u_u = 2x$
	If A_u is in C_g but not in \mathcal{T}_g : $u_u = r_u + 1$. If A_u is in C_g and also in \mathcal{T}_g : $w_u = r_u$
Alleles from typed people that H declares to be non-contributors	
\mathcal{V}	The set of alleles carried by typed people declared not to be contributors to C
n_V	The known number of people declared not to be contributors to C
h_V	The known number of heterozygous declared non-contributors
v_i	The known number of copies of A_i in \mathcal{V} : $\sum_i v_i = 2n_V$

r_1, r_2, \dots, r_{c-1} from r . The total number of A_u alleles in set \mathcal{W} is $\sum_{u=1}^c w_u = 2x$ where $w_u = r_u$ for those alleles in both C_g and \mathcal{T}_g , and $w_u = r_u + 1$ for alleles in C_g but not in \mathcal{T}_g .

For any ordering of the $2x = \sum_u w_u$ alleles in \mathcal{W} , successive pairs of alleles can be taken to represent genotypes and there are $(2x)! / (\prod_{u=1}^c w_u!)$ possible orderings. This is the number of possible sets of unknown genotypes that have each allelic set \mathcal{W} . Although it is the genotypes that correspond to the x unknown people, it is the set of $2x$ alleles that determine the probability, in combination with the $2n_T + 2n_V$ alleles among the known people. Because the n_T typed people all have specified genotypes, there is just a factor of 2 for each heterozygote, and there is a factor of 2 for each heterozygote among the set of n_V non-contributors.

Using Balding's sampling formula, the probability of the set of alleles in the evidence is

$$P_x(\mathcal{T}, \mathcal{W}, \mathcal{V} | C_g) = \sum_{r_1=0}^r \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-\dots-r_{c-2}} \frac{(2x)! 2^{H_T+H_V}}{\prod_{u=1}^c w_u!} \frac{\prod_{u=1}^c \prod_{j=0}^{t_u+w_u+v_u-1} [(1-\theta)p_u + j\theta]}{\prod_{j=0}^{2x+2n_T+2n_V-1} [(1-\theta) + j\theta]}. \quad (18.8)$$

Likelihood ratios are formed as the ratios of two such probabilities.

Every person typed is declared to be either a contributor or a non-contributor. The number of people typed, and the alleles they carry among them, are the same for every proposition. For this reason, $n_T + n_V$, $H_T + H_V$ and $w_u + v_u$ will be the same in the probabilities for each proposition.

If population structure is ignored, and θ is set to zero, equation (18.8) reduces to

$$P_x(\mathcal{T}, \mathcal{W}, \mathcal{V} | C_g) = \sum_{r_1=0}^r \sum_{r_2=0}^{r-r_1} \cdots \sum_{r_{c-1}=0}^{r-r_1-\dots-r_{c-2}} \frac{(2x)! 2^{H_T+H_V}}{\prod_{u=1}^c w_u!} \prod_{u=1}^c p_u^{t_u+w_u+v_u}.$$

The likelihood ratio now depends only on the numbers and probabilities of the alleles carried by unknown contributors. In this situation the genotypes of typed people provide no information about those of untyped people.

With population structure, or relatedness, however, there is information in the genotypes of all people typed during the course of an investigation. Even if they are excluded from being contributors, they provide information for the probability calculations when they can be considered to belong to the same subpopulation as (some of) people not excluded. They make their contribution to the calculation via allelic set \mathcal{V} .

18.4.2.2 Semi-continuous Model

Gill *et al.* (2006) discussed the complications for interpreting mixtures when some of the alleles in the evidence profile may be masked by typing artifacts such as stutter or may have dropped out completely and are not detected. There may also be the possibility of sporadic alleles from fragmented genomes dropping into the evidentiary profile, a different situation from whole genomes contaminating the evidence profile. A complete analysis needs to take into account the relative amounts of DNA inferred to be present at each of the alleles observed to be in the mixture. Having to allow for unseen alleles reduces the possibility of being able to exclude a potential contributor to the mixture simply because that person's alleles are not detected. Great care needs to be taken to avoid prejudicial conclusions if it is decided to ignore those loci in a profile for which interpretation is difficult or alleles are suspected of not being detected.

The essence of both the semi-continuous and continuous models is to condition the sets of alleles for which probabilities are calculated on sets of genotypes that have various probabilities of needing to be considered. Following S. Gittelson (personal communication), the likelihood ratio compares the probabilities of the set G_C of alleles in the evidence profile conditioned on alleles G_K from known contributors under alternative hypotheses by introducing sets S of genotypes for all the contributors to the evidence. Sums are taken over all possible sets of genotypes S consistent with alleles G_C :

$$LR = \frac{\sum_{j=1}^M \Pr(G_C | S_j) \Pr(S_j | G_K^p, H_p)}{\sum_{i=1}^N \Pr(G_C | S_i) \Pr(S_i | G_K^d, H_d)}, \quad (18.9)$$

where H_p, H_d may indicate prosecution and defense propositions. Note that the known alleles and the possible contributor genotypes are different under the two propositions.

For the binary model, all $\Pr(G_C|S_i)$ and $\Pr(G_C|S_j)$ values are set equal to 0 or 1. For the semi-continuous model, they are assigned values depending on the probabilities of allelic drop-out and drop-in. The binary case can allow the use of different electropherogram peak heights, indicating different amounts of DNA in the evidence profile. For an evidence profile with alleles a, b, c, d where the peaks for a, b were judged to be similar to each other but greater than the similar peaks for c, d then only the genotype combination ab (major contributor) and cd (minor contributor) would allow a non-zero probability $\Pr(G_C|S)$. This approach can be problematic as there may be no clear procedures for distinguishing one pair of peak heights from another. Other constraints on probabilities under the binary model reflect heterozygote balance and mixture proportions (Gill *et al.*, 2006). The ratio of peak heights for the two alleles in a single heterozygote should fall within a laboratory-determined range, and the contributions of various contributors to an evidence profile should be similar across loci.

As an example, suppose the evidence profile has peaks corresponding to alleles a, b, c , the victim has alleles a, b and a suspect has alleles a, c . The two propositions may be that the contributors to the evidence were H_p : the victim and suspect, and H_d : the victim and an unknown person. Without considering allelic drop-out or drop-in, the only set of genotypes under H_p is $\{(ab, ac)\}$ whereas under H_d the set is $\{(ab, ac), (ab, bc), (ab, cc)\}$. If allelic drop-in is considered possible, then for H_p any of the victim and suspect alleles may have dropped out from the evidence and the same alleles dropped in, and under H_d any of the victim and unknown person alleles may have dropped out and alleles necessary to complete the evidence profile dropped in. Probabilities for allelic drop-out and drop-in are assigned by the laboratory, and drop-in must be accompanied by the probability of the particular dropped-in allele.

18.4.2.3 Continuous Model

In the continuous model peak heights are treated as continuous variables. The probabilities $\Pr(G_C|S)$ are replaced by probability densities assigned based on models that are fitted using training data from mixtures with contributors of known genotype in known proportions (Perlin *et al.*, 2011; Taylor *et al.*, 2013). The following overview of continuous-model software is from Moretti *et al.* (2017):

The software weighs potential genotypic solutions for a mixture by utilizing more DNA typing information (e.g., peak height, allelic designation and molecular weight) and accounting for uncertainty in random variables within the model, such as peak heights (e.g., via peak height variance parameters and probabilities of allelic dropout and drop-in, rather than a stochastic or dropout threshold). Likelihood ratios ... are generated to express the weight of the DNA evidence given two user-defined propositions. Probabilistic genotyping software has been demonstrated to reduce subjectivity in the interpretation of DNA typing results and, compared to binary interpretation methods, is a more powerful tool supporting the inclusion of contributors to a DNA sample and the exclusion of non-contributors.

There are now commercial and open-source software packages available, and guidelines published for validating them (Scientific Working Group on DNA Analysis Methods, 2015; Coble *et al.*, 2016).

18.5 Behavior of Likelihood Ratio

Although the likelihood ratio is an appropriate quantity to express the probative value of matching DNA profiles, it does not have the intuitive appeal of probability statements about the source

of a crime scene profile. Such statements require prior probabilities for alternative sources, but there has been work to attach some sense to the magnitude of LRs. Should an LR of a billion convince triers of fact? What about an LR of a thousand? Or ten? A narrow answer was provided by Beecham and Weir (2011): they addressed the uncertainty in LRs introduced by the use of sample allele proportions for allele probabilities. A similar treatment by Gittelson *et al.* (2017) accounted for the choice of database on calculated LR values.

Another approach was reviewed by Swaminathan *et al.* (2016). They used the distribution of LRs to assess robustness: for hypotheses that the POI is or is not a contributor to a (mixture) profile, how often will the LR be greater than 1, and so support the POI being a contributor, when in fact he is not? Gill and Hanned (2013) extended this to introduce a '*p*-value' or the probability that a randomly chosen individual results in an LR greater than the LR obtained for a POI. This may be interpreted as the false positive rate resulting from a binary hypothesis test between the prosecution and defense hypotheses. Swaminathan *et al.* (2016) used a continuous model to find results such as *p*-values less than 10^{-9} being found when the LR is greater than 10^8 . This concept of a *p*-value has been criticized by Kruijver *et al.* (2015) who pointed out that the *p*-value, defined as the probability of an LR at least as large as the LR for the POI when the POI is not a contributor, is bounded above by the reciprocal of the LR. Moreover, the *p*-value does not address the alternative hypotheses concerning the POI for which the LR was calculated.

18.6 Single Nucleotide Polymorphism, Sequence and Omic Data

The emergence of sophisticated statistical models and software packages for capillary electrophoretic detection of STR variants represents considerable effort by statisticians, statistical geneticists and forensic scientists. This continues a tradition of attention to the error structure in forensic genetic data. For example, at the time when DNA fragment lengths were being estimated from gel electrophoresis data, a pair of papers using Bayesian techniques were published by Berry *et al.* (1992) and Devlin *et al.* (1992) that modeled the correlation structure of observations on the variants in a profile. A quarter-century later, forensic scientists are gaining access to other new genomic data: single nucleotide polymorphisms (SNPs) at a targeted number of positions, or at all the positions in a targeted region revealed by DNA sequencing. The latter has been called next-generation sequencing or massively parallel sequencing by forensic scientists (e.g. Borsting and Morling, 2015). These sequence-based profiles are not only highly discriminatory, but also may contain information about physical appearance, ethnicity and even health status (e.g. Kidd *et al.*, 2015).

For the forensic science community to embrace SNP typing there would need to be compatibility with large STR-profile databases such as the national database, currently containing 15 million profiles, maintained by the FBI (e.g. CODIS: <https://www.fbi.gov/services/laboratory/biometric-analysis/codis/>). The most complete genomic data come from DNA sequencing, and whole-genome sequencing is becoming common for human genetic studies (e.g. TOPMed: <http://nhlbiwgss.org>). The forensic applications of DNA sequencing currently rest on the sequencing of relatively short regions of DNA including the current set of STR markers. There are software packages to convert these sequence data to STR profiles (e.g. Woerner *et al.*, 2017) and also to reveal single nucleotide variation in the sequenced region to augment the length variation among STR variants.

In addition to genomic data, other omic data are being introduced to forensic science. Parker *et al.* (2016) used amino acid variation in protein sequences to infer SNP profiles as a way to avoid problems with degradation of DNA, and to be able to generate profiles from hair shafts. There is considerable activity (e.g. Chong *et al.*, 2015) in using RNA assays to identify the types

of the source organs for biological tissue or body fluids. Interpreting these new forensic profiles introduces new challenges for statistical genetics.

References

- Andersen, M.M. and Balding, D.J. (2017). How convincing is a matching Y-chromosome profile? *PLoS Genetics* **13**(11), e1007028.
- Balding, D.J. and Nichols, R.A. (1994). DNA match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.
- Balding, D.J. and Steele, C.D. (2015). *Weight-of-Evidence for Forensic DNA Profiles*. Wiley, Chichester.
- Beecham, G.W. and Weir, B.S. (2011). Confidence intervals for DNA evidence likelihood ratios. *Journal of Forensic Sciences Supplement Series* **1**, S166–S171.
- Berry, D.A., Evett, I.W. and Pinchin, R. (1992). Statistical inference in crime investigation using deoxyribonucleotide-acid profiling. *Applied Statistics* **41**, 499–531.
- Borsting, C. and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics* **18**, 78–89.
- Brenner, C.H. (2010). Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics* **4**, 281–291.
- Brenner, C.H. (2013). Understanding Y haplotype matching probability. *Forensic Science International: Genetics* **8**, 233–243.
- Buckleton, J.S., Curran, J.M., Goudet, J., Taylor, D., Thiery, A. and Weir, B.S. (2016). Population-specific F_{ST} values: A worldwide survey. *Forensic Science International: Genetics* **23**, 91–100.
- Caliebe, A., Jochens, A., Willuweit, S., Roewer, L., Krawczak, M. (2015). No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Science International: Genetics* **15**, 69–75.
- Chong, K.W.Y., Wong, Y.X., Ng, B.K., Thong, Z.H. and Syn, C.K.C. (2015). Development of a RNA profiling assay for biological tissue and body fluid identification. *Forensic Science International: Genetics Supplement Series* **5**, E196–E198.
- Coble, M.D., Buckleton, J., Butler, J.M., Egeland, T., Fimmers, R., Gill, P., Gusmao, L., Guttman, B., Krawczak, M., Morling, N., Parson, W., Pinto, N., Schneider, P.M., Sherry, S.T., Willuweit, S., and Prinz, M. (2016). DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications. *Forensic Science International: Genetics* **25**, 191–197.
- Cockerham, C.C. and Weir, B.S. (1973). Descent measures for two loci with some applications. *Theoretical Population Biology* **4**, 300–330.
- Cook, R., Evett, I.W., Jackson, G., Jones, P.J. and Lambert, J.A. (1998). A hierarchy of propositions: Deciding which level to address in casework. *Science and Justice* **38**, 231–239.
- Curran, J., Triggs, C.M., Buckleton, J. and Weir, B.S. (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Sciences* **44**, 987–995.
- Devlin, B., Risch, N. and Roeder, K. (1990). No excess of homozygosity at loci used for DNA fingerprinting. *Science* **249**, 1416–1420.
- Devlin, B., Risch, N. and Roeder, K. (1992). Forensic inference from DNA fingerprints. *Journal of the American Statistical Association* **87**, 337–350.
- Donnelly, P. 1995. Nonindependence of matches at different loci in DNA profiles – Quantifying the effect of close relatives on the match probability. *Heredity* **75**, 26–34.

- Evett, I.W. and Weir, B.S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Science*. Sinauer Associates, Sunderland, MA.
- Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.R., Morling, N., Prinz, M., Schneider, P.M. and Weir, B.S. (2006). DNA Commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of mixtures. *International Journal of Legal Medicine* **160**, 90–101.
- Gill, P. and Haned, H. (2013). A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Science International: Genetics* **7**, 251–263.
- Gittelson, S., Moretti, T.R., Onorato, A.J., Budowle, B., Weir, B.S. and Buckleton, J. (2017). The factor of 10 in forensic DNA match probabilities. *Forensic Science International: Genetics* **28**, 178–187.
- Graham, J., Curran, J. and Weir, B.S. (2000). Conditional genotypic probabilities for microsatellite loci. *Genetics* **155**, 1973–1980.
- Hall, T.O. (2016). The Y-chromosome in forensic and public health genetics. PhD dissertation, University of Washington, Seattle.
- Kidd, K.K., Speed, W.C., Wootton, S., Lagace, R., Langit, R., Haigh, E., Chang, J. and Pakstis, A.J. (2015). Genetic markers for massively parallel sequencing in forensics. *Forensic Science International: Genetics Supplement Series* **5**, E677–E679.
- Kruijver, M., Meester, R. and Slooten, K. (2015). p-values should not be used for evaluating the strength of DNA evidence. *Forensic Science International: Genetics* **16**, 226–231.
- Lander, E.S. (1989). DNA fingerprinting on trial. *Nature* **330**, 501–505.
- Laurie, C. and Weir, B.S. (2003). Dependency effects in multi-locus match probabilities. *Theoretical Population Biology* **63**, 207–219.
- Moretti, T.R., Moreno, L.I., Smerick, J.B., Pignone, M.L., Hizon, R., Buckleton, J.S., Bright, J.A. and Onorato, A.J. (2016). Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Science International: Genetics* **25**, 175–181.
- Moretti, T.R., Just, R.S., Kehl, S.C., Willis, L.E., Buckleton, J.S., Bright, J.A., Taylor, D.A. and Onorato, A.J. (2017). Internal validation of STRmixTM for the interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics* **29**, 126–144.
- National Research Council (1996). *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC.
- Parker, G.J., Leppert, T., Anex, D.S., Hilmer, J.K., Matsunami, N., Baird, L., Stevens, J., Parsawar, K., Durbin-Johnson, B.P., Johnson, B.P., Rocke, D.M. et al. (2016). Demonstration of protein-based human identification using the hair shaft proteome. *PLoS ONE* **11**(9), e0160653.
- Perlin, M.W., Legler, M.M., Spencer, C.E., Smith, J.L., Allan, W.P., Belrose, J.L. and Duceman, B.W. (2011). Validating TrueAllele[®] DNA mixture interpretation. *Journal of Forensic Sciences* **56**, 1430–1447.
- Scientific Working Group on DNA Analysis Methods (2015). Guidelines for the validation of probabilistic genotyping systems. <https://bit.ly/2CiZ83W>.
- Siegert, S., Roewer, L. and Nothnagel, M. (2015). Shannon's equivocation for forensic Y-STR marker selection. *Forensic Science International: Genetics* **16**, 216–225.
- Steele, C., Syndercombe Court, D. and Balding, D.J. (2014). Worldwide F_{ST} estimates relative to five continental-scale populations. *Annals of Human Genetics* **78**, 468–477.
- Swaminathan, H., Garg, A., Grgicak, C.M., Medard, M. and Lun, D.S. (2016). CEESIt: A computational tool for the interpretation of STR mixtures. *Forensic Science International: Genetics* **22**, 149–160.
- Tan, S.Y. and Graham, C. (2013). Karl Landsteiner (1868–1943): Originator of ABO blood classification. *Singapore Medical Journal* **54**, 243–244.

- Taylor, D., Bright, J.A. and Buckleton, J. (2013). The interpretation of single source and mixed DNA profiles. *Forensic Science International: Genetics* **7**, 516–528.
- Thompson, W.C. and Schumann, E.L. (1987). Interpretation of statistical evidence in criminal trials – The prosecutors fallacy and the defense attorneys fallacy. *Law and Human Behavior* **11**, 167–187.
- Walsh, B., Redd, A.J. and Hammer, M.F. (2008). Joint match probabilities for Y chromosomal and autosomal markers. *Forensic Science International* **174**, 234–238.
- Weir, B.S. (2004). Matching and partially-matching DNA profiles. *Journal of Forensic Sciences* **49**, 1009–1014.
- Weir, B.S. and Goudet, J. (2017). A unified characterization for population structure and relatedness. *Genetics* **206**, 2085–2103.
- Woerner, A.E., King, J.L. and Budowle, B. (2017). Fast STR allele identification with STRait Razor 3.0. *Forensic Science International: Genetics* **30**, 18–23.
- Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

19

Ethical Issues in Statistical Genetics

Susan E. Wallace¹ and Richard Ashcroft²

¹ Department of Health Sciences, University of Leicester, Leicester, UK

² School of Law, Queen Mary University of London, London, UK

Abstract

The ethical, legal and social issues (genethics) concerning genetic research are extensive and complex. This chapter reviews some of the central issues in current debates. The first part of the chapter considers the scope of genethics and the relationship between ethics, morality, professional conduct and genetics research. It then considers the relationship between risk-control and benefit-maximising models of governance in genetics research. The second part of the chapter, using case studies, looks at the main issues in governance of genetic databases, including the scientific value of the research, recruitment, consent and mental capacity, voluntariness and incentives to participate, feedback of research results, confidentiality and security. The third part of the chapter looks at issues in the conduct of research, concentrating on stewardship and wider social issues.

19.1 Introduction

The ethical, social and legal issues arising in genetic research and its applications are so extensive that they encompass their own field of research and scholarship, often referred to as ELSI (ethical, legal and social implications of genetics) or ‘genethics’ (Clarke and Ticehurst, 2006; Sherlock and Morrey, 2002). Among the topics discussed in this literature are the moral limits on modification of the human genome in light of new technologies such as CRISPR-Cas9; whether prenatal and preimplantation genetic testing are ways to improve quality of life or new variants of eugenics; consent, privacy and confidentiality of genetic data; ethical debates about biosafety of genetically modified crops; the morality of patenting genetically engineered living creatures; the changing nature of the obligations clinicians and researchers have regarding sharing genetic information with patients and participants; recontacting patients who have had tests whose interpretation has subsequently changed in light of new knowledge; and the obligations researchers in human biodiversity may have to share the benefits of any commercialisable discoveries with the donors of samples and the communities of which they are members.

These issues are so diverse and complex that to cover them all adequately would require a book-length treatment. Therefore, this is a necessarily selective survey. In the present chapter we concentrate on issues of primary concern to statistical geneticists. ‘Big data’ approaches are being used to gather huge data sets of medical, administrative, and social media information through national and international collaboration. These data are being used in ‘personalised’

or 'targeted' approaches in medical treatment as well as in public health strategies. Individuals are finding out more about their own genomic make-up and using it, with clinicians, to make health-care decisions and lifestyle choices. Yet this preponderance of data raises many questions such as governance and consent (how are these data being collected and used?), privacy and confidentiality (how are these data being protected?) and public interest (how do we balance the wishes of individuals regarding their own genetic information with the potential benefits of using that information for public – and potentially private – good?) We will explore some of these questions by focusing on two large-scale research projects, UK Biobank and the 100,000 Genomes Project, to look more closely at how we judge between acceptable and unacceptable conduct in genetic research.

Before we begin an examination of these issues, it will be useful to review, in this first section, what the main sources of ethical argument are in the literature.

19.1.1 What Is Ethics?

Ethics can be defined as philosophical inquiry into the values, rules of conduct and character traits which are involved in right action, doing good, and living well. It is often contrasted with morality, which is the commonly shared set of rules and principles shared within a community and taken for granted in assessing one's own behaviour and that of others. As defined, ethics can be thought of as systematic inquiry into the foundations of morality and – where necessary – correction of the principles of morality in the light of reason and evidence (Benn, 1998).

Ethics in this philosophical sense should not be confused with 'professional ethics'. Some professionals may refer to conduct as 'unethical', by which they mean that it violates the formal or informal norms of expected behaviour by members of that profession, as laid out in codes of conduct or as inculcated through professional training. This usage of the term 'ethics' is analogous to the term 'morality' as defined above. Professional ethics (or, as we would prefer to say, professional morality) is referred to as such to distinguish it from common morality, which is the morality shared by most members of a community whether or not they are members of a profession. Philosophical ethics may address questions of professional ethics, but professional ethics need not be philosophical. Professional ethics is developed in close liaison with the legal and regulatory requirements on professionals and is a part of training in many fields – for example, medical ethics, which is taught formally in all medical schools in the United Kingdom.

One important part of philosophical ethics is bioethics. Bioethics can be defined as the application of principles and methods of analysis of philosophical ethics to the analysis of moral and social problems arising in the life sciences and medicine. Genethics is therefore part of bioethics. The methods of bioethics are generally analytic and philosophical. Nevertheless, in recent years considerable attention has been paid to the need for ethical arguments to make use of the best-quality social scientific evidence: good ethics needs good facts. This is particularly important in population and public health genetics, since studies involve large numbers of participants, and are occasions of potential controversy. Empirical research can both clarify what issues are at stake in the participant community and also help to build trust and confidence in the aims and processes of research. In addition to analytical and empirical methods, research in genethics necessarily overlaps with research in law and public policy. Research in the fields of medical and biotechnology law has increased over the past thirty years, using both standard case and statute law resources and, increasingly, the jurisprudence of human rights. Significant international human rights declarations concerning genetics are directly relevant to population genetic research.

19.1.2 Models for Analysing the Ethics of Population Genetic Research

There are a number of different ways of conceiving what ethically is at stake in population genetics research, each of which goes some way towards shaping the regulatory and ethical framework in which contemporary genetic research is done. We can start by categorising these into two broad groups: benefit maximisation models and risk control models. Benefit maximisation models of the ethics of population genetic research focus on the benefits of such research, and seek ways to maximise these. Risk control models identify specific risks which may be involved in genetic research and seek to control these. While we might see these as complementary in that benefit maximisation models recognise risk-based constraints, and risk control models try to avoid strong risk aversion which might dilute any benefits accruing to such research to the point of futility, nevertheless they differ in emphasis and orientation.

19.1.2.1 Risk Control Models

For historical reasons, it is arguable that risk control models of genetics have dominated discussion. We can identify a number of factors explaining this. The first is the history of eugenics, especially in the way eugenic ideas were used to justify forced sterilisations, barriers to ‘mixed’ or ‘inappropriate’ marriages, racial discrimination, abuse of the mentally ill and learning disabled, up to mass killings of the genetically ‘unfit’ under the German Third Reich (Paul, 2002). Any genetic research in humans since the Second World War has needed to establish the distance between its aims and those of eugenics. Genetic research and genetic medicine are widely seen as an area in which serious risks of personal harm and social injustice need to be forestalled or overcome. A second factor is the history of the ethics of research on human subjects. Again because of the Nazi experience (and parallel episodes in the Japanese empire, and subsequent morally problematic experiments carried out during the Cold War by NATO and Warsaw Pact countries and elsewhere), research on human beings has been considered intrinsically risky (Rhodes, 2005). To the extent that population genetics for medical or non-medical purposes involves engaging with human subjects, it is seen as a form of research on human subjects and is governed by the research ethics risk control paradigm.

Because of the scale of modern population genetic research two further types of risk have come to the fore. The first is the role of the state in funding, facilitating and regulating genetics research. Concerns of modern citizens about state interference in their lives are read across into the field of genetics research. The second is the role of the commercial private sector in developing technologies using genetic research and its applications. Concerns about the self-interested behaviour of corporate actors are similarly read across into the field of genetics research. These types of risk are exemplified by concerns about data protection, confidentiality and privacy, intellectual and real property rights in samples, data and discoveries, benefit sharing with donors of samples or data, exploitation of poor individuals or groups (especially transnationally), state coercion and surveillance, sharing of information with agencies or individuals for non-medical purposes, and so on (Bauer and Gaskell, 2002). Finally, there is a general concern with any kind of research into the biological characteristics of human beings, such as the creation of chimeras and gene editing, that it undermines human identity or human dignity. This kind of risk is often invoked in religious contexts, but it has also been influential in the framing of certain kinds of regulation, most notably international patent law, and law and regulations prohibiting germline gene therapy and reproductive cloning (Beyleveld and Brownsword, 2001). Unlike the first four kinds of risk, this is not clearly linked to any specific historical or political experience, but may be linked to a more general historical epoch often referred to as ‘modernity’, which many intellectual historians associate with secularisation or with the ‘scientific revolution’ of the sixteenth century. Further, this sort of risk is generally

considered qualitatively rather than quantitatively. Notionally it cannot be traded off against other benefits or risks, because undermining human dignity (for instance, by undermining the grounds for considering human beings fundamentally equal and members of the same human family) would be absolutely wrong. This sort of risk is held in mind especially in international human rights declarations. Concentrating on these kinds of risk makes sense to many analysts of the ethics of genetics for historical and political reasons. Because they are seen to be salient, they frame and shape much of the discussion of the ethics of genetic research, and much of the regulatory framework for genetic research is fixed by the concern for these risks and solutions developed in other contexts for managing them.

As well as identifying these five kinds of risk by the context in which they arise and the sorts of harm or moral danger involved, we can also categorise them by the level at which they operate: individual, family, social group (such as ethnic group or gender), or as a society as a whole. There is a pronounced tendency in the literature to concentrate on risks of harm to identifiable individuals, and to regulate that risk of harm through a dependency on the informed consent mechanism. As we shall see, it is harder to regulate risks that operate at group level, and the troubles attending the extension of informed consent models to group protection are well known and difficult.

19.1.2.2 Benefit Maximisation Models

Perhaps surprisingly, benefit maximisation models are rarely articulated formally, but are instead the 'common sense of science'. Benefit maximisation arguments can appear in various forms, from appeals to the future health benefits that will accrue if particular research lines are (successfully) pursued, to appeals to a right to freedom of scientific inquiry, to more bluntly economic arguments about the inefficiency of research ethics regulations. In the current state of ethical debate it is arguable that benefit maximisation models function in two ways: firstly, as a corrective to (excessive) caution in risk control models; and secondly, as substantive arguments about the best way to get maximum value out of particular research resources. An example of the former type of argument is Steven Pinker's (2015) claim that the most ethically useful thing bioethicists can do in relation to genome editing is to 'get out of the way' of impeding its advances, which he claims will be overwhelmingly beneficial for human health. An example of the latter type of argument is whether open access publishing for journal articles or the more traditional 'pay per view' style of access better balances the need to support a process that is necessary for the dissemination of the results of research (Parker, 2013).

19.2 Ethics and Governance in Population Genetics Research: Two Case Studies

In order to understand the ways in which the different kinds of research risk frame the governance of population genetic research, while retaining an intention to do research that is maximally beneficial, it is helpful to consider case studies. We will first present the backgrounds of two research projects, UK Biobank and the 100,000 Genomes Project. UK Biobank is an example of a longitudinal cohort study recruiting healthy volunteers with the understanding that they will not benefit individually from research using their samples and data. The 100,000 Genomes Project, on the other hand, recruits patients – those with cancer and their relatives, and those with rare diseases. Many of these patients have already benefited from participating through information on potential treatment options for their condition found through genomic sequencing and associated health data. We will use these two different genomic research

studies to compare and contrast different approaches to ethics and governance and the complexities surrounding them.

19.2.1 'Healthy Volunteer' Longitudinal Cohort Studies: UK Biobank

UK Biobank is a major initiative in studying the interaction between genes and environment in a health context, using a large cohort study drawn from the UK population.

UK Biobank aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia. It is following the health and well-being of 500,000 volunteer participants and provides health information, which does not identify them, to approved researchers in the UK and overseas, from academia and industry. (UK Biobank, 2018d)

UK Biobank is currently funded until 2022. It was established by the Wellcome Trust, the Medical Research Council, the Department of Health, the Scottish Government and the Northwest Regional Development Agency, and is supported by the UK National Health Service (NHS). Additional funding comes from the Welsh Government, British Heart Foundation, Cancer Research UK and Diabetes UK.

People recruited to UK Biobank were identified from central registries. Names, addresses, sex, date of birth, and name of their general practice were processed centrally (in accordance with the UK Data Protection Act 1998) and UK Biobank sent an invitation letter to those in England, Scotland and Wales potentially eligible to participate who lived within a reasonable travelling distance of an assessment centre. General practitioners (GPs) were advised that people registered with their practices were invited to take part; participants could notify their GP of their acceptance. Those excluded from the study are those unable to give informed consent (for example, because of diminished mental capacity), those too ill to take part in data collection, and those whom study recruiters deemed uncomfortable with any of the conditions of participation. The latter exclusion was intended to exclude people who it is felt would prefer not to consent, or seemed not to understand what is required of them, but who nonetheless seemed for one reason or another to be giving their consent. Recruitment was completed in 2010 and it has become a significant, valuable, and much accessed resource.

At the initial recruitment participants:

- attended a local study assessment centre for about 90 minutes to answer some simple questions, to have some standard measurements, and to give small samples of blood (about 2 tablespoons) and urine;
- agreed to allow their health to be followed for many years by UK Biobank directly through routine medical and other records;
- agreed to be recontacted to answer additional questions, attend a repeat assessment visit or to join new approved studies.

Since the creation of the initial baseline dataset, a considerable amount of new data has been added through ongoing recontact of the participants. Additional measurements have covered aspects of the eye and included saliva samples. Activity monitors have provided lifestyle data on 100,000 participants. Online questionnaires on diet, cognitive function, work history and digestive health have been completed or are ongoing. After a successful pilot study, an imaging survey will scan the brain, heart, abdomen, bones and carotid artery of each of 100,000 participants. Researchers from anywhere in the world, working in academia or private industry,

can apply to UK Biobank to access these data for approved research purposes (UK Biobank, 2018a).

UK Biobank has a detailed Ethics and Governance Framework and an Ethics Advisory Committee to consider ethical issues. UK Biobank provides updates on its activities through its website, emails, newsletters, an annual general meeting for the public and at participant meetings held periodically in different cities across the country (UK Biobank, 2018b). It has a policy of no return of individual results from research using samples and data.

19.2.2 Precision Medicine Approaches: 100,000 Genomes Project

In contrast to UK Biobank, the 100,000 Genomes Project has the goal of using genomic sequencing together with NHS record data to better diagnose and treat patients in the future. Treatment options are explored and, using results from the sequencing of their DNA, patients may benefit from new or improved treatment options. The overall goal of the project is to better integrate genomic research and data into health care in the UK. The project is being run under the auspices of Genomics England, a company wholly owned by the Department of Health (Genomics England, 2018c).

Recruitment was completed in December 2018 with 85,000 participants, of whom approximately 40,000 are patients with cancer and rare diseases. Each of 25,000 cancer patients will have had two genomes sequenced – one of their cancer tumour and one of healthy tissue as a comparison. Approximately 17,000 patients with a rare disease will have been sequenced, together with two of their blood relatives who will act as comparators. Further, ‘the Scottish Genomes Partnership, in collaboration with Genomics England, will analyse the entire genetic make-up of 330 people with rare diseases and members of their family’ (Genomics England, 2017).

This is an example of a ‘precision medicine’ (alternatively targeted, strategic, or personalised medicine) approach where the ideal is to individualise the treatment to the patient based on their specific genomic and other health information. While ‘precision medicine’ has been a popular buzz phrase for some years, examples of efficacy are now appearing, especially in cancer. By examining the genome sequence data, mutations may be found that could indicate the cause of the condition and suggest treatment options specifically targeted at that change in the DNA. In addition, lessons learned from following the patient through treatment and outcomes could inform the treatment of other sufferers of that condition. There also can be a psychological benefit for those afflicted and their families from simply having a diagnosis, even if no effective treatment is currently available. Access to the data is provided for approved research, and many groups focusing on various cancers have already begun work using the data (Genomics England, 2018a).

19.2.3 The Scientific and Clinical Value of the Research

Both case study projects emphasise the *expected value* of the research. Participants are invited to take part, and the public (and scientific community) are invited to support the project, on the basis that the project will produce important new knowledge that will make a significant contribution to the understanding, prevention, diagnosis and treatment of major diseases. To succeed scientifically they must recruit a large number of people, and secure consent to a variety of investigations, some of which may be painful or inconvenient, and some of which may require long-term contact. Large numbers are needed to achieve statistical power as ‘[i]nadequate statistical power increases not only the probability of missing genuine associations but also the

probability that significant associations represent false-positive findings' (Sham and Purcell, 2014).

Healthy volunteer studies, where one participates on the basis of the potential benefits for others and not oneself, without consideration of the relationship one may have to those beneficiaries, can be described as 'altruistic participation'. Studies like UK Biobank are presented as a moral enterprise. While it will take many years for Biobank research to be translated into clinical practice, it is potentially useful and helpful in advancing a vital interest we all share, our interest in being healthy and in receiving good medical care. A potential participant might ask why they should volunteer. One answer might be: to help scientists do something they think is scientifically interesting. For some participants, this would be a sufficient reason to take part. Many people do think that science is intrinsically valuable, and that it is exciting and honourable to play a part in its advance. Yet in fact this appeal to shared scientific curiosity is relatively rare in modern biomedical research.

A risk of selling science to the public on the grounds of utility is that scientific projects may fail, or may fail to produce the expected or hoped-for results (Fortun, 2008). Participants in research like Biobank may perceive this utility in one of three ways. Firstly, they may perceive participation as useful to them personally, here and now. This may be akin to 'therapeutic misconception', believing that, even when told to the contrary, they will experience some benefit from participating. Secondly, they may perceive participation as generating benefits that will be useful to people like them (possibly including themselves or their relatives and descendants personally) in future. Thirdly, they may perceive participation as generating benefits that may benefit people in future, without consideration of their own personal interests.

Participation on the basis of benefits to people like oneself (possibly including oneself) in future could be called 'solidary participation': through this 'sense of community' we take part out of solidarity with people we recognise as our peers and whose suffering we want to ease, on the basis that they would do the same for us, or that we hope that they would. Participation on the basis of solidarity is a theme that has long been emphasised in epidemiology, at least in the postwar period in which epidemiology in the UK and other welfare states was linked institutionally to social medicine, to socialised health care, and to social movements such as the trades union movement (Ashcroft *et al.*, 2000; Prainsack and Buyx, 2017). Solidarity may be presented in a more or less moralised form. On the one hand, many commentators would see solidarity as the basis of any moral relationship with others, since what motivates us to help others is an understanding of the plight of others based on empathy. On the other hand, solidarity could be a much more pragmatic motive, where we reason on a quasi-contractual basis that we are obliged to put into a social relationship more or less what we get out of it. Some argue that we have a duty to participate in research, because we have benefited to date from research in which others have participated (Harris, 2005; Stjernschantz Forsberg *et al.*, 2014).

It is important also to emphasise the other side to the argument, that even though an individual may not benefit, they should still support research because these projects, in their goal to improve health and society, represent good value for money and a worthwhile investment of public funding. In general terms, it is accepted that scientific projects must be well founded in terms of posing a well-defined question, to which the answer is not already known, and which, if answered, would generate new knowledge of scientific importance. The twin mechanisms of systematic review (to establish what is already known, and to what degree of methodological confidence) and peer review (to establish the credentials of the research team, the likely scientific value of the research, and the value for money the project offers) are intended to ensure that scientific projects put to the public to invite participation are well founded scientifically, and

thus that the claims made about the utility of the research are as well supported as is possible. In addition, greater participant engagement, from providing updates on the research to using participant panels to provide feedback, is another way in which the utility is judged (McCarty *et al.*, 2011).

We have seen how research projects start their approach to ethical governance and recruitment by emphasising the scientific and public value of their projects. The next topic we address is the methodology of recruitment.

19.2.4 Recruitment of Participants

As discussed above, recruitment will vary based on differing inclusion and exclusion criteria. For UK Biobank they are healthy individuals (that is, they were not recruited through clinics in which they were current patients) within a specified age range who live in the specific locations chosen for recruitment. For the 100,000 Genomes Project, they are patients with certain rare diseases, their close relatives, and patients with cancer who are registered with certain GPs. Other studies might have differing inclusion criteria. Case–control studies, for instance, where ‘cases’ must meet defined clinical or other criteria for inclusion and where ‘controls’ must be comparable to selected cases, will have much more specific inclusion and exclusion criteria. In the context of controlled clinical trials there are important issues concerning the fairness of inclusion and exclusion criteria in terms of the distribution of the risks and benefits of research to (non-)participants. This is less of a significant issue in epidemiological research, but does arise in the context of the interpretation of findings, which we will return to below.

A more important question concerns how individuals eligible for participation are identified as suitable for recruitment, and how they are then approached. Some studies simply ask people to volunteer, through letters or placing advertisements. It is up to each individual to decide whether they are eligible and interested, and then to contact researchers. For most purposes this is not adequate in terms of constructing an unbiased sample for research, so the practical issue is how to identify potential research participants, and then to contact them, without breaching norms of confidentiality and law relating to data protection. If potential participants are identified through public information (such as electoral rolls or telephone directories), this is rarely a significant issue. However, where participants may be identified through information not in the public domain, such as personal medical records, the situation may be complicated. For a researcher who would not have access to this information routinely (for instance, in the case of medical records, because he or she does not have clinical care of these patients), individual consent might be required for researcher access to these records; but the researcher can only know which records (or which patients) he or she would like to see, and thus whose consent is needed, when he or she has had the chance to look at the records. In practice, pragmatic solutions to this catch-22 need to be found: either someone who has right of access to the set of records preselects the individuals for the researcher and makes the first approach to them to obtain consent to the researcher approaching them, or actually to enrol them into the study, or the researcher is given a contract with the record-holding organisation which allows them to see records and places them under a contractual duty not to misuse the records outside the terms of the contract. Under some circumstances, it might be that a research project is of such public importance that enrolment of participants may take place without their consent (if this involves data extraction from records only). Many countries have regulations in place to permit this. However, in the case of genetics research, while this method might be used to allow data extraction to identify individuals who are potential research participants, actual recruitment into the research would almost always involve a direct approach to the individuals and the seeking of their consent.

The principal exception to these rules about recruitment of participants is where the samples to be used are de-identified, in such a way that re-identification of individuals is impossible. Normally this would occur when a sample bank is used, rather than samples being collected for a specific research project.

19.2.5 Consent

The central element to the ethics of genetic sample based research is – inevitably – consent. Consent in epidemiological and genetic research has been much debated in recent years (O'Neill 2002; McHale, 2004; Gibbons *et al.*, 2005). This is in part because genetic research poses new issues, and in part because of the social and economic trends providing the context against which genetic research now takes place. The standard requirements for a valid informed consent concern the mental capacity of the individual to consent to the research participation; the voluntariness of the decision; and the information necessary to make a decision (Jackson, 2006).

Mental capacity is a complex topic, and could be treated at length, but for present purposes a few simple points are sufficient. Mental capacity is the ability to understand and retain the information one is given to make a decision, to believe it, and to weigh it up and make a choice. A consequence of this definition is that capacity is relative to the nature of the decision being made: it is much easier to show that someone has the capacity to buy a newspaper than it is to show that they have the capacity to consent to a heart transplant. So, one does not simply have ‘mental capacity’, but rather, under a particular set of circumstances, one can be said to have the capacity to make this sort of decision. In a medical context, adults are presumed to have mental capacity to make any decision that they may be asked to make, and considerable evidence and inquiry may be needed to show that they lack mental capacity to make that decision. This is true even of people with a psychiatric disorder. In the case of children, for the most part, the opposite is true: a child under 18 (or, for many purposes, under 16) has to be shown to have the ability to make a decision, and is presumed to lack that capacity. For medical research purposes, these assumptions about adults, children and mental capacity are controversial. Should we, for instance, assume that research is necessarily harder to understand than ordinary clinical treatment? Surely not. But some research is highly complex. Moreover, what may be difficult to understand is that enrolment in research is voluntary, that according to all research guidance it is clear that patients are not to be compelled to take part in research, or threatened with poor treatment if they refuse, or forbidden to leave research once it is started. Patients (in particular) may believe that participation is either obligatory, or necessarily in their best interests, or that they ‘owe’ their doctor something, when none of these need be true. (This, again, is the ‘therapeutic misconception’ mentioned above.) So, what may be the stumbling block in determining capacity is not the complexity of the technical information, but the changed relationship between researcher and participant (where it was one of doctor and patient, it is now a rather different relationship between researcher and participant).

We can debate whether this is an issue of capacity, or rather one of voluntariness. In practice, since most epidemiological research is non-therapeutic, people who lack capacity are simply excluded from prospective research (unless the research links specifically to the reasons for their incapacity – as in some psychiatric genetic research), since participation in research can rarely be shown to be in the individual’s best interest. This would be the test of whether someone can be enrolled in research when they lack mental capacity. A finer-grained interpretation of this condition can be developed, as in guidelines from the UK Medical Research Council (2007). On this interpretation, a person lacking capacity to consent to a research project can be

included in the research if it is in their best interests, or if the risks are minimal and this research cannot be conducted in people with capacity and this research would benefit others in future with a similar condition and participation is not *against* the individual's interests. Under some circumstances research which is not of minimal risk can be permitted, but a compelling case would have to be made about why this research was essential for the welfare of people suffering from the condition under investigation and there was no other way to carry it out or to resolve the problem under study.

A related difficult issue concerns the achievement at a later date of capacity by someone enrolled in a study while lacking capacity. For example, a child might be enrolled in a study at birth. When the child reaches majority, it is generally agreed that the now adult individual should have the power to withdraw the consent given on their behalf, or to ratify it. This raises complex issues concerning what it actually means to leave a study in this context, as we will discuss in a moment, but also the wider question of how far 'proxy consent' can really be considered valid, and how far parents (or carers) can choose authoritatively for their children (or incapacitated relative) (Ross, 1998; Archard, 2004). There is also a potential impact on a longitudinal study if participants in, for example, a birth cohort decide to withdraw (Wallace *et al.*, 2016). Those leading such projects seek to engage participants as part of a study 'family' so as to keep them informed of the progress and results of the study. This enables a gradual understanding of the project and encourages continued support.

This leads us to voluntariness. In practice, people may be happy to participate, while some may feel pressured. Requests to participate would not normally be seen as coercive. However, since the Nuremberg Code of 1947, all research ethics guidelines have insisted that people should be able to choose freely whether or not to participate, and those who take part in research should be free to leave it at any time. This allows people to change their minds or go back on decisions they might regret or – in the worst case – were suborned into making (Marks, 2006). This poses a difficult challenge in epidemiological research for two reasons. Since such research depends on the collection of large data sets and the analysis of such sets as collections of data, often longitudinally, it is not always clear what joining or leaving a research study really means, for example, when secondary research, unknown to the participant and possibly years later, is conducted. UK Biobank has a solution to this which we discuss below. The data subject may not know that their data is being collected and used, if an opt-out consent is used; and it is not clear what 'withdrawing consent' means when data are incorporated in an aggregated data set. Participation hinges on the nature of the information participants should receive in order that they can give a valid consent and whether it is 'sufficient'. The importance of information is twofold. Firstly, the quality of the information and style in which it is presented has a major influence on the potential participant's ability to understand what is being proposed, and consequently on the likelihood that they agree to take part. Secondly, the information given defines what it is the participant is consenting to, and hence what that consent authorises the researcher to do in terms of investigations, and how data and samples may be used in research. Recently, in response to changing attitudes towards what constitutes valid consent, options have become more granular, giving more specific options to participants; this issue is discussed in more detail later in this section.

One important feature of much epidemiological research is that data and samples hold much of their value because they can be reused and reanalysed, either directly or in combination with data collected for other purposes. Thus, the tissue samples collected, and the data derived from them, for a study on the consumption of salt and cardiovascular disease could be useful to a researcher interested in the vascular aspects of Alzheimer's disease. Yet the consent taken for building the collection for the first purpose would not necessarily authorise the use of the samples and data for the second purpose. Genetic sample collections are now built up not in the

context of a specific research project, with tightly designed objectives and research questions, but as resources for use in future unspecified research (Gibbons *et al.*, 2007). Instead of being able to give consent to a specific project, participants are asked to give it for future unknown research, governed by appropriate ethical and regulatory mechanisms to protect the interests of the participant. Researchers (and indeed ethics committees and research funders) welcome this approach as a useful way to allow future flexibility in what research can be done with samples and data, and to get the best value out of the funding. Others question whether this is a valid consent as it is difficult (some argue impossible) to clearly understand to what one is consenting (Greely, 2007).

The problem here is that consent, to be seen as valid when considering if someone has been potentially negligent, needs to be quite specific, whereas what researchers typically need, because the science moves so swiftly, is a consent which is quite broad and durable in terms of what it authorises in the short and long term. Both UK Biobank and the 100,000 Genomes Project are designed to be resources which will grant access for approved research purposes and use a 'broad' consent which allows participants to consent to a broad purpose for use of their data and samples, rather than to only a narrow and specific use within a tightly defined protocol. In these cases, the questions of utility of the future research may be more diffuse, and oversight, if practised at all, would focus more on ensuring that proper governance structures are in place, that data access procedures are robust and transparent, and that there is continuing communication with participants.

Another approach is to require a new consent each time a new use of the samples is proposed. In practice this would normally require each individual to be recontacted by the research team. This is advantageous in terms of ensuring that each new investigation is formally authorised by each participant, and it provides a mechanism for the participant to leave, by reminding them that their continued participation is optional. On the other hand, it is cumbersome and expensive, may cause attrition in the study population, and may even be burdensome on participants who are willing to continue as long as they are not bothered too often. For a longitudinal study such as UK Biobank, recontact would be agreed in the original consent, for new questionnaires or for participation in new sub-studies. But in much epidemiological research, where direct contact with participants is unusual after initial recruitment, this approach has been unpopular for pragmatic reasons, such as cost. The advent of electronic methods may change this in the future. Dynamic consent models create a way to engage participants by using personalised interfaces on mobile devices or computers that will allow them 'to alter their consent choices in real time' (Kaye *et al.*, 2015).

While broad consent may be useful for researchers, it may not address the issues of most concern to research participants. Consent to participate in research, being developed on the basis of the risk control model appropriate to clinical trials, tends to focus on the risks and benefits of research participation, the nature of the investigations a participant will undergo, and other issues concerned with the personal safety and integrity of the individual. However, for many research participants these are not the only important issues. At least some research participants are interested in issues such as the possible commercialisation of research findings or the possible uses of their data or samples in research of which they may not approve (see Section 19.3). For instance, while a participant might be entirely happy for their samples to be used in genetic research in cardiovascular disease, they might disapprove of research into the genetic basis of intelligence, and thus the reuse of their samples donated for the former by researchers working on the latter. Now, it is reasonable to say that although participants are the donors of the samples, their donation of the samples involves ceding control of those samples (ownership, if you like, although notions of property in biological samples are controversial too), and beyond certain limits they have no further say in what is done with those samples. Nevertheless,

protecting and promoting trust in the research enterprise and in specific researchers may involve giving assurances to donors about the kinds of use which are foreseen in the long term, beyond the terms of narrow and specific consent in the short term, and what kinds of use are excluded. The broad consent is supported by independent oversight by a research ethics committee and a governance process that protects the interests of sample donors or data subjects in place of the protection which a series of narrow and specific consents would give. This process includes very clear rules on data access. Data access committees review applications to see if the requested use of the data is in keeping with ethos of the consent agreed with participants; requests for samples are judged on whether they are a legitimate use of a limited resource. Researchers and their institutions sign a contract agreeing a set of provisions including that the data will only be used for the research proposed and that the results will be made publicly available. This form of consent relies on a robust governance system and the trust of the participants (Dixon-Woods and Tarrant, 2009).

Another challenge is judging what leaving a study actually means and whether one is happy to join on that basis. UK Biobank has three withdrawal options (UK Biobank, 2018c). A person can choose to leave the study and no new data on them can be collected, but existing data can continue to be used. Or, more permissively, they can choose that once they leave, existing data can be used and further data on them may be collected from routine data, but UK Biobank would no longer contact them. Or, more restrictively, if they choose 'no further use' any information and samples collected previously would no longer be available to researchers. Samples will be destroyed but the data cannot be removed from research that has already taken place. The data will have been linked to and incorporated into other data sets, making it impossible to retrieve. In more principled terms, it is not clear that the person has a right to request that their data should be removed: firstly, that they agreed to that data being used and cannot retrospectively revoke that agreement; and secondly, that although the data may be derived from them, they are not owned by them, but by the researcher in whom intellectual property in the data resides. What is being negotiated here is the meaning of 'leaving the study'. The normal approach here is to be reasonably explicit about what agreeing to take part in the study amounts to, and about what leaving the study amounts to. Nonetheless, it is possible to imagine circumstances under which someone loses trust in the researcher or the research project to the extent that they feel that they have been misled. They might then reasonably say that the data were collected under false pretences and should be deleted. This might arise in a study of race/ethnic differences, for instance, where a person feels that the study undermined the dignity or reputation of their race/ethnic group in a way they could not foresee and were not warned about. Similarly, it might arise where a child was recruited into a study by his or her parents in early childhood and at adulthood may wish to remove his or her data from the study, as it was collected without his or her consent.

Another issue related to voluntariness concerns incentives to participate. These may be both formal and informal. Formal incentives, in the form of payments or offers of services in exchange for agreement to participate, are controversial for two main reasons. Firstly, many epidemiologists would feel that people should participate in research for solidary or altruistic reasons, and offering payments, gifts or in-kind exchanges devalues such reasons for participation. They would argue that this discourages people from participating in research unless there is something in it for them. Secondly, ethicists tend to worry that people who take part when there is an incentive scheme are doing so in order to get the incentive, rather than with full understanding of the risks and benefits of participation. The greater the incentive, the greater the risk that this may occur, up to the point where people put themselves knowingly in danger merely because they want or need the incentive. Of course, this is the situation with much phase I drug research and many kinds of employment. But this is felt to be

regrettable, if necessary, and not something to be expanded. One hard issue here is that it is almost, if not actually, impossible to draw a line separating reasonable from coercive levels of inducement.

Informal incentives to participate comprise reasons to participate in the research which are formally part of the research protocol itself, but which are attractive to participants for reasons unconnected with the research. For example, in many epidemiological studies, such as UK Biobank, participation in the research involves the collection of vital statistics and medical examinations for the purposes of collecting baseline and study time-point data, but which are also (potentially) useful health screening data for the research participants. Many studies are attractive to participants because they believe they are getting free, or more than usually convenient, 'health checks', although studies mostly will explicitly state in consent materials that this is not the case. Sometimes, participants may also believe that they are getting access to tests or checks on their health that are not normally available. For some studies, when initial quality control analyses are being carried out on samples, if something is found that suggests that the participant would be wise to see a doctor, this is passed on (Knoppers *et al.*, 2013).

19.2.6 Returning Individual Genetic Research Results

As more has become known about our genomic data, we now know it to be predictive of our potential for health, well-being, and disease. Because of the ability to see our health 'future', there is now a stronger call to report findings arising from genomic research to participants (Wolf *et al.*, 2012), but this has been controversial. In a doctor–patient relationship, there is an expectation that such data will be passed on for their clinical care. In most epidemiological research, however, the researcher is not seeing the participant as a patient, and the researcher has no specific legal duty of care to the participant. In the past, should it appear, through a research study, that a person had a particular mutation that might give them a higher than normal risk of suffering from a condition, there was not thought to be any obligation to tell those individual research participants with this genotype that they need to see their doctor for further counselling and possible investigations. In part, this is because the quality of such genetic testing in research is not at clinical grade. As well, because most conditions are caused by the interaction of many different mutations, the finding does not conclusively indicate that the individual will suffer from the condition. Another key consideration is whether the genetic mutation found is 'actionable' – is there is an intervention that will improve the participant's health and/or well-being? In studies where data are shared with outside researchers, especially internationally, it will be difficult, if not impossible, to return such information as the researcher will not have access to the identity of the donor as the data will have been de-identified (Wallace, 2011).

Funding bodies now require researchers to be explicit about what information will be given back and what will not, and what participants should do if they have further (medical) questions (see, for example, Wellcome Trust, 2017). However, this does not entirely address the point of principle: do individuals have a right to data concerning their health and genetic constitution? Do the researchers have a duty to give them such data? Within a medical relationship, doctors are entitled to withhold information from patients if it may be misinterpreted or if it would be psychologically damaging to the patient to receive a piece of information which they are ill-prepared to cope with. This entitlement is somewhat controversial even within medicine (it is the so-called 'therapeutic privilege'). Where the researcher has no medical care relationship with the participant, what responsibilities do they owe and what should participants expect? What does the 'social licence to operate' for medical researchers require here? (Dixon-Woods and Ashcroft, 2008; Carter *et al.*, 2015).

The practice of not feeding back individual data directly to the patient or participant is now changing for many studies. There is a greater blurring of the clinical and research worlds where clinicians are more involved in the research process, which further muddies the question of where there is a duty of care. However, in some areas such as cancer studies, genomic research has progressed to a point where precision medicine approaches ('the right drug for the right patient at the right time') are more feasible. As discussed, the 100,000 Genomes Project is a case in point. Patients agreeing to participate are given a clear policy regarding the return of findings related to the reason for participation (to diagnose more specifically the genomic make-up of their cancer or to diagnose the rare disease causing their condition). These findings can lead to an intervention or a change in existing treatment. But participants also can be given 'additional findings' that arise. These are an agreed list of actionable changes in their genome that might lead to increased risk of suffering a disease or condition. This list is discussed at the time of consent and the participant needs to agree to receive additional findings or opt out if they wish. This of course does not solve the problem of when an 'incidental' finding is made which was not on this list, and not consented to in advance because novel or unexpected. And this further raises the ethical question of what happens if something very serious is found and the individual has chosen not to be contacted, a dilemma that has not been solved.

This 'right not to know' balances a trend that when it comes to genomic knowledge, individuals have a 'right to know'. Some argue that this reflects a move over the last 50 years towards individuals taking greater control of aspects of their lives, such as their health. People no longer need to wait to participate in a research project to have the possibility of knowing more about their genomic make-up; by providing a sample of their saliva, they can now have their complete or partial genome sequenced and the sequence data and analysis sent to them. The difficulty is in knowing what the analysis means, and therefore many are happy for an expert clinician or researcher to provide that to them and then only if there is a successful intervention available. In the present context, participants need to be clearly told what they can and cannot expect to receive and why.

19.2.7 Confidentiality and Security

One of the more practically important issues of concern to participants, which can have major consequences for the governance of research sample and data collections, is confidentiality (Laurie, 2002). A general principle of the management of confidential information, such as medical records and genetic information, is that it should only be accessible on a need-to-know basis. This protects the subject of the information from disclosure of information to parties who should not have access to particular items of information. In the context of research, this means that personal information collected or extracted for research purposes should be recorded in a way that minimises the extent to which the information allows identification of the individual data subjects. There is a tension in epidemiological research between protecting individual privacy and confidentiality by de-identification such as full anonymisation, and retaining sufficient information to allow informative analysis of data sets, linkage of different items of information, and (re)contact of individuals (for research purposes, or to disclose clinically relevant information to the individual as discussed).

Two approaches are popular. One is to protect privacy by removing identifying elements from data sets (such as name, address, and postal code) and only keeping the kind of information required for the research project (or sample or data bank). This approach appears to build in privacy protection, but it is relatively inflexible to changes of protocol or for reuse or reanalysis. The other approach, often combined to some extent with the first, is to protect privacy through access controls and through coding so that linkage between records can be managed on a

limited basis for approved purposes, but with researchers agreeing to not attempt to re-identify individuals. The key that allows re-linkage will be secured in a separate file or database from the de-identified data set, with only the anonymised data being shared. If necessary, any linking back to the participant is done only in accordance with a defined protocol, such as returning results as discussed. One concern with this approach is that technology-based solutions may be subject to attack, and that in some ways they provide false reassurance about the security of a system that is subject to human intervention or human error.¹ It has been shown that by combining a de-identified data set with data publicly available, such as through social media, an individual can be identified. Researcher Latanya Sweeney was able to identify specific data about the then Governor of Massachusetts by cross-referencing publicly available aggregate census data with pseudonymised health records (Ohm, 2010).

In response, researchers have been seeking to create protected environments where 'data can be analysed – either locally, or remotely via secure privacy protecting mechanisms, but cannot physically be removed from that setting' (Burton *et al.*, 2015) This is the approach Genomics England has chosen (Genomics England, 2018b). While access is granted only for approved research following agreed data access and data sharing policies, as with UK Biobank, instead of receiving the data and analysing it at their home institution, Genomics England data do not leave the Genomics England data centre: they are analysed through access to the data centre. In this way institutions can protect their data under their own country's data protection regulations, while permitting access to the global research community. While there will always be those who attempt to circumvent security features, it is hoped that this new approach of sending researchers to the data rather than data to the researchers, will provide greater privacy protections while still enabling access to these important data. However, some researchers may find this approach to accessing data insufficiently flexible for their needs.

19.3 Stewardship and Wider Social Issues

Moving away from research participant related ethical issues, we turn to the value of a resource created in research, which could be a data set, a sample collection, or indeed a research protocol which partners can sign up to in whole or in part. Much debate has centred on the question of whether and how intellectual property rights should be vested in such resources, and how the value of these resources can be maintained in the long term. There is no straightforward answer to these questions, but there is consensus on one point, which is that resources created with public or charitable money should normally be regarded as open access resources, and that fees (if any) should only be levied to cover the costs of processing the requests and upkeep of the research resource, on a non-profit basis. Resources such as these typically also stipulate that exclusive access licences will not be granted, and often some requirement is made that research findings and sometimes any additional data sets or remaining samples should be returned to the resource for future users to access; this is the case for UK Biobank (see UK Biobank, 2017). Some of the challenges here lie in finding ways to guarantee long-term viability of such resources. Samples need to be stored, which costs money; and access systems also need maintenance and oversight. At the end of a project's lifetime, or when key personnel leave or retire, sample or data collections can fall into disuse or disrepair. Most public research funders (such as the UK Medical Research Council) now require detailed plans to be made for the stewardship of research resources beyond the lifetime of their established funding or identified management.

1 For a treatment of this issue in a crime novel, see Indriðason (2004).

The cost of genomic sequencing has decreased rapidly in recent years as a result of improved technologies, which has allowed more of this kind of work to be done by publicly funded projects. However, large-scale research endeavours are increasingly likely to seek industry partnerships for some aspects of the work. For example, UK Biobank partnered with companies to perform sequencing of the exomes of all its 500,000 participants (Philippidis, 2018). Likewise, Genomics England, which oversees the 100,000 Genomes Project, is a limited liability company and partners with companies that specialise in large-scale data sequencing and analysis (Genomics England, 2018c). In both cases, the reasoning is that industrial partners have the finances to conduct work not possible for publicly funded projects, and without them information leading to new medicines or treatments would not available as quickly. In addition, in a competitive situation, lower prices for services can be negotiated in exchange for exclusivity. While there is a recognition that industry usually conducts the development of pharmaceuticals, there is still some confusion among the public as to what commercial access to research data actually means. A study carried out by Ipsos MORI (2016) for the Wellcome Trust found that a slight majority were happy for commercial organisations to use their data for research and there was significant approval of industry working in partnership with academia and charities. However, a significant minority of respondents (17%) did not want any commercial entity having access to their health data. Only if it were shown that there would be strong public benefit would the respondents see industry involvement as acceptable. The use of their data by marketing and insurance companies was especially discomforting, and safeguards and regulations were necessary for acceptance.

This issue of the stewardship of a collection relates to wider issues of scientific research integrity: scientists are generally expected to share their data and teach their techniques. This is for three reasons. Firstly, this provides external researchers with a way of checking that research results are valid and non-fraudulent. Second, it allows for the sharing of best practice, so as best to allow the rapid development of science. Third, it encourages the sense of there being a 'community of science' engaged in a common endeavour to advance humanity. Many critics of the commercialisation of science, especially the use of restrictive contracts and intellectual property rights, are generally concerned that this process undermines the three objectives stated above. On the other hand, commercialisation provides its own incentives for entrepreneurship and ingenuity, and can provide wider social benefits in terms of stimulating the technological application of scientific knowledge.

19.3.1 Benefit Sharing

As well as the value of the research resource to researchers, and in many cases to commercial companies, a database or sample collection represents an investment of time and effort on the part of the participants. In many settings, particularly in the developing world, there is a strong sense that researchers owe a duty of reciprocity to their participants to share any financial or clinical benefits of research with their host communities. This may be negotiated as part of 'community consent'. In developed world settings, researchers normally argue that they are dealing with individuals, that the contribution of any given individual on an identifiable basis to a particular project is minimal, and that it is the collected research resource that has value as a composite. In addition, the research may have been hosted using public infrastructure, and any commercial benefits deriving from the infrastructure are taxed in such a way as to ensure the reinvestment of part of the proceeds in the state. However, these arguments carry less weight where researchers are doing research in a resource-poor setting in another country. In such situations it is arguable that the ratio between benefit accruing to the researcher and sponsor and that accruing to research participants is too high, and to expect participants to

donate samples altruistically when this may be the only exploitable resource they have is unreasonable and unfair. Benefit sharing agreements may be quite complex and difficult to enforce, especially when they are made between groups and researchers rather than between individuals and researchers. Sometimes benefit sharing may involve money payments, but more often involve benefits in kind, such as the provision of hospital facilities to a community, or other medical services (Parry, 2004).

19.3.2 Community Involvement and Public Engagement

Community involvement has been advocated increasingly in recent years (Hansson, 2005; Haimes and Whong-Barr, 2004). In some developing world settings, community consultation (sometimes involving 'community consent') has been seen as essential to the 'licence to do research' in a setting, partly in view of the history of colonial exploitation of poor or vulnerable communities. In the developed world, community consultation is a useful method for building support for a project, encouraging recruitment, and allowing general feedback to participants. Some funders see involving the 'public' as so important that it is required that researchers requesting grant money provide a plan for public engagement or must justify why it is not needed or possible for their project. On occasion, researchers allow community consultation to play a part in the governance of a project, although a more common approach is to have a participants panel such as in the 100,000 Genomes Project or one or two community members as members of the project steering committee or ethics governance board. Careful consideration of the best way to engage the public can be crucial for the success of some projects, but it does raise the question of who is the 'public' in any situation.

19.3.3 Race, Ethnicity and Genetics

Aside from the impact of the research on individual participants or on host communities, there are wider social issues raised by genetics research. One example is scientific research on racial or ethnic differences. It is inherently controversial, but genetic research, because of the history of eugenics and biological racism, is especially so (Macbeth and Shetty, 2001; Ellison and Goodman, 2006). Central elements of the controversy include the following. Firstly, there is controversy over whether any biological sense can be attributed to socially prevalent conceptions of race, and hence whether scientific inquiry into purported biological concepts of race can have any rational justification (Marks, 1995; *Nature Genetics*, 2004). Secondly, even allowing that some biological concept of human variation along race-like lines can be defended, there is the vexed question whether the biological concepts map onto the concepts used in ordinary social life (Smart *et al.*, 2006; Royal, 2006). If they do, can science be seen as supporting social attitudes to racial difference which may be morally and politically problematic, and if they do not, does using a language so open to misinterpretation not confuse issues in a dangerous way (Ashcroft, 2006)? Thirdly, assuming some rigorous and value-neutral concept of racial or ethnic difference can be established, which is biologically useful, do the ways in which these findings are then applied in medicine and applied science make sense?

Some might argue that these issues should not be studied at all, but there are also issues relating to social justice and personal choice. It has been shown that certain drugs act better (and worse) in certain ethnic populations. Large-scale data sets on which decisions are being made regarding susceptibility to conditions or success of drugs are largely made up of Caucasian participants. More research groups around the world are now creating new data sets that focus on their ethnic populations. Likewise, more people are interested in their ethnicity and are having genetic tests to 'find their roots'. Personal data are shared online through social media and

genealogy websites, which contrasts with the efforts researchers make to protect participants' confidentiality. One of the lessons of this complex debate is that genetic research into the biological bases of traits that are complex in their structure and in their meanings in society is very difficult to carry out with 'clean hands'. This raises large issues about whether certain questions should not be investigated at all, or only with great care, and what 'approaching an issue with great care' means in terms of the social responsibilities of scientists. Moreover, ethnic labels are arguably poorly defined, and may not be of much practical use to geneticists, even where they may sometimes be helpful in clinical service planning and health policy.

19.4 Conclusion

This chapter has necessarily been highly selective and rather discursive. It is intended to give a summary overview of some of the more important practical and social issues in the conduct of statistical genetic and genetic sample-based research. Although some of the issues discussed are complex and confusing, three things remain clear. Firstly, public trust in medical and biological research remains relatively high, in part because of the care taken to engage with the public about science and to maintain high ethical standards. Although many of the issues in this chapter may seem frustratingly unresolved at the level of theory, in practice there is considerable consensus about many of them. For example, the UK Biobank Ethics and Governance Framework, written following extensive consultation with the academic community, commissioned surveys and focus groups, and direct public consultation (see UK Biobank Ethics and Governance Council, 2007), supports and guides the work of UK Biobank. Secondly, there is now an extensive and growing literature of philosophical, ethical, legal, and empirical research that can help frame and illuminate the issues and help policy-makers, scientists and the public resolve them. Third, that it remains crucial that scientists remain engaged with these debates, both in informing them and in steering them in directions which will both control risks to participants and society, and maximise the benefits that genetic research will generate.

Acknowledgements

The authors would like to thank David Balding, Adrienne Hunt and Michael Parker for comments on a draft of this chapter.

References

- Archard, D. (2004). *Children: Rights and Childhood*. Routledge, London.
- Ashcroft, R.E. (2006). Race in medicine: From probability to categorical practice. In G.T.H. Ellison and A.H. Goodman (eds.), *The Nature of Difference: Science, Society and Human Biology*. CRC Press/Taylor & Francis, Boca Raton, FL, pp. 135–153.
- Ashcroft, R.E., Jones, S. and Campbell, A.V. (2000). Solidarity in the UK welfare state reforms. *Health Care Analysis* 8, 377–394.
- Bauer, M.W. and Gaskell, G. (eds.) (2002). *Biotechnology: The Making of a Global Controversy*. Cambridge University Press, Cambridge.
- Benn, P. (1998). *Ethics*. UCL Press, London.
- Beyleveld, D. and Brownsword, R. (2001). *Human Dignity in Bioethics and Biolaw*. Oxford University Press, Oxford.
- Burton, P.R., Murtagh, M.J., Boyd, A., Williams, J.B., Dove, E.S., Wallace, S.E., et al. (2015). Data Safe Havens in health research and healthcare. *Bioinformatics* 31(20), 3241–3248.

- Carter, P., Laurie, G. and Dixon-Woods, M. (2015). The social license for research: Why *care.data* ran into trouble. *Journal of Medical Ethics* **41**, 404–409.
- Clarke, A. and Ticehurst, F. (eds.) (2006). *Living with the Genome: Ethical and Social Aspects of Human Genetics*. Palgrave Macmillan, Basingstoke.
- Dixon-Woods, M. and Ashcroft, R.E. (2008). Regulation and the social license for medical research. *Medicine, Health Care and Philosophy* **11**, 381–391.
- Dixon-Woods, M. and Tarrant, C. (2009). Why do people cooperate with medical research? Findings from three studies. *Social Science & Medicine* **68**(12), 2215–2222.
- Ellison, G.T.H. and Goodman, A.H. (eds.) (2006). *The Nature of Difference: Science, Society and Human Biology*. CRC Press/Taylor & Francis, Boca Raton, FL.
- Fortun, M. (2008). *Promising Genomics: Iceland and deCODE Genetics in a World of Speculation*. University of California Press, Berkeley.
- Genomics England (2017). Scotland study to probe causes of rare diseases. <https://www.genomicsengland.co.uk/scotland-study-to-probe-causes-of-rare-diseases/> (accessed 12 January 2018).
- Genomics England (2018a). Current projects approved for access to the 100,000 Genomes Project data set. <https://www.genomicsengland.co.uk/the-100000-genomes-project/data/research/> (accessed 12 January 2018).
- Genomics England (2018b). Frequently asked questions. <https://www.genomicsengland.co.uk/faqs-about-gecip/> (accessed 24 February 2018).
- Genomics England (2018c). The 100,000 Genomes Project. <https://www.genomicsengland.co.uk/the-100000-genomes-project/> (accessed 12 January 2018).
- Gibbons, S.M.C., Helgason, H.H., Kaye, J., Nõmper, A. and Wendel, L. (2005). Lessons from European population genetic databases: Comparing the law in Estonia, Iceland, Sweden and the United Kingdom. *European Journal of Health Law* **12**, 103–133.
- Gibbons, S.M.C., Kaye, J., Smart, A., Heeney, C. and Parker, M. (2007). Governing genetic databases: Challenges facing research regulation and practice. *Journal of Law and Society* **34**(2), 163–189.
- Greely, H.T. (2007). The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annual Review of Genomics and Human Genetics* **8**, 343–364.
- Haines, E. and Whong-Barr, M.T. (2004). Key issues in genetic epidemiology: Lessons from a UK based empirical study. *TRAMES* **8**, 150–163.
- Hansson, M.G. (2005). Building on relationships of trust in biobank research. *Journal of Medical Ethics* **31**, 415–418.
- Harris, J. (2005). Scientific research is a moral duty. *Journal of Medical Ethics* **31**, 242–248.
- Indriðason, A. (2004). *Tainted Blood*. Harvill, London.
- Ipsos MORI (2016). Public attitudes to commercial access to health data. <https://wellcome.ac.uk/sites/default/files/public-attitudes-to-commercial-access-to-health-data-summary-wellcome-mar16.pdf> (accessed 12 January 2018).
- Jackson, E. (2006). *Medical Law: Text, Cases and Materials*. Oxford University Press, Oxford.
- Kaye, J., Whitley, E.A., Lund, D., Morrison, M., Teare, H. and Melham, K. (2015). Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics* **23**(2), 141–146.
- Knoppers, B.M., Deschenes, M., Zawati, M.H., Tasse, A.M. (2013). Population studies: Return of research results and incidental findings Policy Statement. *European Journal of Human Genetics* **21**, 245–247.
- Laurie, G. (2002). *Genetic Privacy: A Challenge to Medico-legal Norms*. Cambridge University Press, Cambridge.
- Macbeth, H. and Shetty, P. (eds.) (2001). *Health and Ethnicity*. Routledge, London.

- Marks, J. (1995). *Human Biodiversity: Genes, Race, and History*. Aldine de Gruyter, New York.
- Marks, S.P. (ed.) (2006). *Health and Human Rights: Basic International Documents*. Harvard School of Public Health and Harvard University Press, Cambridge, MA.
- McCarty, C.A., Garber, A., Reeser, J.C. and Fost, N.C. (2011). Study newsletters, community and ethics advisory boards, and focus group discussions provide ongoing feedback for a large biobank. *American Journal of Medical Genetics Part A* **155**(4), 737–741.
- McHale, J.V. (2004). Regulating genetic databases: Some legal and ethical issues. *Medical Law Review* **12**, 70–96.
- Medical Research Council (2007). Medical research involving adults who cannot consent. <https://www.mrc.ac.uk/documents/pdf/medical-research-involving-adults-who-cannot-consent/> (accessed 12 January 2018).
- Nature Genetics (2004). Genetics for the human race. *Nature Genetics* **36**, S1–S60.
- Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* **57**, 1701–1719.
- O'Neill, O. (2002). *Autonomy and Trust in Bioethics*. Cambridge University Press, Cambridge.
- Parker, M. (2013). The ethics of open access publishing. *BMC Medical Ethics* **14**(1), 16.
- Parry, B. (2004). *Trading the Genome: Investigating the Commodification of Bioinformation*. Columbia University Press, New York.
- Paul, D.B. (2002). Is human genetics disguised eugenics? In D.L. Hull and M. Ruse (eds.), *The Philosophy of Biology*. Oxford University Press, Oxford, pp. 536–551.
- Philippidis, A. (2018). Regeneron to lead \$50 M Exome Sequencing Consortium with UK Biobank. <https://www.genengnews.com/topics/omics/regeneron-to-lead-50m-exome-sequencing-consortium-with-uk-biobank/> (accessed 13 January 2018).
- Pinker, S. (2015). The moral imperative for human genetics. *Boston Globe*, 1 August. <https://www.bostonglobe.com/opinion/2015/07/31/the-moral-imperative-for-bioethics/JmEkoyzITAu9oQV76JrK9N/story.html> (accessed 26 November 2017).
- Prainsack, B. and Buyx, A. (2017). *Solidarity in Biomedicine and Beyond*. Cambridge University Press, Cambridge.
- Rhodes, R. (2005). Rethinking research ethics. *American Journal of Bioethics* **5**(1), 7–28.
- Ross, L.F. (1998). *Children, Families and Health Care Decision-Making*. Oxford University Press, Oxford.
- Royal, C.D.M. (2006). 'Race' and ethnicity in science, measurement and society. *Biosocieties* **1**, 325–328.
- Sham, P.C. and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15**(5), 335–346.
- Sherlock, R. and Morrey, J.D. (eds.) (2002). *Ethical Issues in Biotechnology*. Rowman and Littlefield, Lanham, MD.
- Smart, A., Tutton, R., Ashcroft, R.E., Martin, P.A. and Ellison, G.T.H. (2006). Can science alone improve the measurement and communication of race and ethnicity in genetic research? Exploring the strategies proposed by *Nature Genetics*. *Biosocieties* **1**, 313–324.
- Stjernschantz Forsberg J., Hansson M.G. and Eriksson S. (2014). Why participating in (certain) scientific research is a moral duty. *Journal of Medical Ethics* **40**, 325–328.
- UK Biobank (2017). Return of results data: Guidance note for approved projects. http://www.ukbiobank.ac.uk/wp-content/uploads/2017/08/Return-of-Results_Guidance-Note_aug17.pdf (accessed 13 January 2018).
- UK Biobank (2018a). About UK Biobank. <https://www.ukbiobank.ac.uk/about-biobank-uk>. (accessed 12 January 2018).
- UK Biobank (2018b). Ethics. <https://www.ukbiobank.ac.uk/ethics/> (accessed 12 January 2018).

- UK Biobank (2018c). Frequently asked questions. <https://www.ukbiobank.ac.uk/all-faqs/> (accessed 12 January 2018).
- UK Biobank (2018d). UK Biobank. <https://www.ukbiobank.ac.uk/> (accessed 12 January 2018).
- UK Biobank Ethics and Governance Council (2007). *UK Biobank Ethics and Governance Framework, version 3.0*. <https://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf> (accessed 12 March 2019).
- Wallace, S.E. (2011). The needle in the haystack: International consortia and the return of individual research results. *Journal of Law, Medicine & Ethics* **39**(4), 631–639.
- Wallace, S.E., Gourna, E.G., Laurie, G., Shoush, O. and Wright, J. (2016). Respecting autonomy over time: Policy and empirical evidence on re-consent in longitudinal biomedical research. *Bioethics* **30**(3), 210–217.
- Wellcome Trust (2017). Policy on data, software and materials management and sharing. <https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing> (accessed 13 January 2018).
- Wolf, S.M., Crock, B.N., Van Ness, B., Lawrenz, F., Kahn, J.P., Beskow, L.M., et al. (2012). Managing incidental findings and research results in genomic research involving biobanks and archived data sets. *Genetics in Medicine* **14**(4), 361–384.

20

Descent-Based Gene Mapping in Pedigrees and Populations

E.A. Thompson

Department of Statistics, University of Washington, Seattle, WA, USA

Abstract

Linkage analysis is the analysis of the dependence in inheritance of genes at different genetic loci, on the basis of observations on individuals. Typically these observations will be of single nucleotide polymorphism (SNP) marker genotypes and a trait of interest, and the goal is to determine the locations of DNA underlying the trait relative to the genetic marker map. This chapter considers genetic linkage mapping, taking a descent-based approach applicable to analyses of data on individuals whose pedigree structure is unknown, as well as to data on individuals in a known pedigree structure. The first sections consider the processes of meiosis and the descent of DNA, and consequent probabilities of identity by descent (IBD) among individuals and across genomes. An important result is that segments of IBD in remotely related individuals are rare but not short. While data at each SNP marker are quite uninformative as to shared descent, segments of IBD will contain many SNP markers. This leads to computationally practical and effective methods for inferring IBD among the four genomes of two diploid individuals, whether or not their pedigree relationship is known. However, inference of IBD among multiple genomes remains a challenge. Once IBD at locations across the genome is inferred, the patterns of trait similarity among individuals may be related to this location-specific IBD. Model-based probabilities of trait data lead to a likelihood-based approach, while patterns of IBD among affected individuals lead to alternate test statistics.

20.1 Introduction to Genetic Mapping and Genome Descent

20.1.1 Genetic Mapping: The Goal and the Data

The last century has seen enormous change in methods of inference from genetic data, from the rediscovery of Mendel's laws in 1900, to near completion of Phase I of the Human Genome Project 100 years later, and since 2000 the pace of genomic technologies has increased again. However, the scientific questions remain surprisingly constant: Where are the genes? What do they do? Genetics is the science of heritable variation, meiosis is the biological process whereby genetic information is transmitted from parent to offspring, and genetic analysis involves inferences concerning the outcomes of meioses from data on the genetic characteristics of individuals.

Human individuals are diploid. We each have two copies of the human genome of approximately 3×10^9 base pairs (bp) of DNA. Leaving aside the sex chromosomes, this DNA comes

packaged into 22 pairs of autosomal chromosomes (autosomes) which range in size from 51 to 245 million base pairs (Mbp). The goal of genetic mapping is to determine the locations in the genome (the *locus* or *loci*) of causal DNA that underlies traits of interest. For the purposes of this chapter we will generically denote the values of these qualitative or quantitative observable characteristics of individuals by \mathbf{Y} .

Locations are defined with reference to a genetic marker map. In comparing the complete DNA sequence of any two human genomes, the nucleotides of two genomes will differ at approximately every 1000 bp. Of course, in considering large numbers of genomes from worldwide populations, there will be many more such locations with variable nucleotides. These variable sites are known as *single nucleotide polymorphisms* (SNPs). These SNPs are at known locations in the genome, and provide the framework against which the locations of causal DNA are to be mapped.

For genetic mapping, SNP marker data on individuals are required. These data typically come in the form of SNP genotypes: a specification of the unordered pair of alleles (nucleotides) at each marker (SNP) in each observed individual. Although data may come in the form of phased haplotypes (see **Chapter 3**) or local DNA sequences, for the purposes of this chapter we will assume marker data are these genotypes. Generically we will denote these genetic marker data by \mathbf{X} . We denote the possible alleles at a genetic marker locus by a_1, \dots, a_v . For SNP markers typically $v = 2$, and the alleles may be denoted by a_1 and a_2 , or simply by '1' and '2'. The genotypes are then '1 1', '1 2', and '2 2', where '1' normally denotes the reference allele of the Human Genome Project, and '2' the alternate allele. Not all SNPs may be typed in all observed individuals: we denote missing marker genotypes by '0 0'.

20.1.2 The Process of Meiosis and the Descent of DNA

The genetic material that underlies inherited characteristics consists of the chromosomes: linear structures of DNA in the form of a double helix contained within each cell nucleus. In a diploid organism, chromosomes come in pairs, one deriving from the genetic material of the mother and the other from the father. DNA is copied from parents to offspring over generations, through the process of *meiosis*, the process of formation of a gamete (sperm or egg cell). Each parent provides one chromosome from each chromosome pair. In the formation of this chromosome, several *crossover* events may occur between the two parental chromosomes, such that the transmitted chromosome consists of alternating segments of these two chromosomes. A functional gene consists of a much smaller segment of DNA at a specific location (*locus*) on a pair of chromosomes. Simple genetic characteristics are determined by the DNA types (*alleles*) in these small gene segments.

At every location in the genome, according to Mendel's first law, which dates back to 1866 (see Mendel, 1866), when a (diploid) individual has an offspring, a copy of a randomly chosen one of his two DNA alleles is transmitted (or *segregates*) to the offspring, independently of the other parent and independently for each child. Thus, enshrined in this first law, probability models are fundamental to genetic analyses. In fact, of the biological sciences, genetics is the one with the most clearly defined probability models, and hence the one in which classical parametric statistical inference has had the biggest role.

The 50-50 probabilities of Mendelian segregation lead to rapid decay in the probability of transmission of a particular DNA allele over multiple generations. The probability of transmission through m meioses is $(1/2)^m$, and the probability that two descendants separated by m meioses share DNA by descent from a single diploid ancestor is $(1/2)^{m-1}$. Such DNA, shared by current individuals through successive meioses from a single DNA copy in their common ancestor, is said to be *identical by descent* (IBD). The pointwise probability of IBD between

two current genomes is also the expected proportion of the two genomes that is IBD. Thus, while first- and second-degree relatives may share a high proportion of their genomes IBD, this proportion decays exponentially with the degree of the relationship.

However, the process of meiosis also results in chromosomes being inherited in large chunks, of the order of 100 Mbp in length. With very high probability, DNA at nearby locations is from the same parental chromosome. Indeed, for the smaller chromosomes, there is probability close to 50% that one of the two parental chromosomes will be inherited intact, and close to the remaining 50% probability that there will be just one *crossover* between the two parental chromosomes in formation of the chromosome that is inherited. With small probability, or with larger probability on the larger chromosomes, there may be more such crossovers.

Haldane (1919) defined genetic map distance between two loci as the expected number of crossover events between them in an offspring gamete. This distance measure is additive regardless of dependencies of crossovers in gamete formation. One morgan is the length of chromosome in which one crossover event between the two parental chromosomes is expected. Map distances are normally given in centimorgans ($1\text{ cM} = 0.01\text{ morgan (M)}$). In terms of physical distance (bp), many factors influence crossover rates. A major one is the sex of the parent; in humans, the total female map length of the 22 autosomal chromosome pairs is about 39 M, or 1.5 times the male map length (26.5 M). However, this ratio is not constant over the genome. Increasingly, as the results of meioses over multiple generations through ancestors of unknown gender are analyzed, the differences between male and female genetic maps are ignored.

A simple meiosis model assumes absence of *genetic interference*. Under this model, crossovers occur randomly and independently: that is, as a Poisson process at rate 1 per morgan. There is said to be *recombination* between two loci if the DNA in the offspring gamete at those two locations derives from different parental chromosomes: that is, from different grandparents. For small recombination frequencies, there is little difference between recombination frequency r and map distance d . At larger distances, the relationship depends on the pattern of interference: the extent to which one crossover event inhibits or enhances the occurrence of others nearby. In the absence of interference, the relationship is $r = \frac{1}{2}(1 - e^{-2d})$, the probability of an odd number of events in distance d morgans when events occur as a Poisson process. This relationship is known as the *Haldane genetic map* (Haldane, 1919). Genetic interference does exist, and other map functions accommodate it. For example, the map function due to Kosambi (1944) was often used classically. However, for all natural map functions, $r \approx d$ when both are small. Thus, with modern dense genetic marker maps, interference is generally ignored.

The inheritance of DNA from parents to offspring is in segments of average length of the order of 100 cM. As a result, segments of IBD between the genomes of remotely related individuals are rare but not short (Donnelly, 1983). Over a single chain of m meioses, the probability of IBD is $(1/2)^m$, but, conditional on being IBD, that IBD segment will end when one of the m aligned meioses undergoes a crossover event. This happens at rate m per morgan, so the expected length of an IBD segment is proportional to m^{-1} . Table 20.1 shows the probabilities and expected segment lengths for relatives with a single diploid ancestor separating them by

Table 20.1 Probabilities of IBD and expected segment lengths (for details, see text)

# meioses	$m = 12$	$m = 20$
Probability of IBD at point	0.0005	2×10^{-6}
Probability of any autosomal IBD ($L = 30$)	0.148	0.001
Expected length of IBD segment	8.5 cM	5 cM

$m = 12$ and $m = 20$ meioses. Also shown is the probability of any segment of IBD in an autosomal genome of length L morgans (Donnelly, 1983). At $m = 12$, the pointwise probability of IBD is small, but there is almost 15% chance of some segment of autosomal IBD. At $m = 20$ even the probability of any segment of IBD in the autosomal genome is small, but if there is such a segment it is of the same order of magnitude as for $m = 12$, namely several million base pairs. Such a segment will encompass many SNP markers. It is this that provides the power to detect IBD from genetic marker data, and hence use remote relatives in genetic mapping analyses.

20.1.3 Genetic Linkage Mapping: Association or Descent?

Because DNA descends in blocks, genetic variants that are close together on a chromosome tend to be inherited together. The locations of these variants are said to be (*genetically*) *linked*. This phenomenon of *genetic linkage* enables the locations of causal DNA underlying a trait Y to be mapped relative to the known locations of genetic markers. The procedures are those of *genetic linkage mapping*.

One widely used strategy is association mapping, where a subset of the SNPs in the genome are genotyped, providing data X , and each of these is tested for a statistical association with the trait of interest Y (for details, see **Chapter 21**). An SNP for which the allelic variation is causal will obviously show such an association, but so also will other SNPs nearby in the genome. Any new causal variant arises by mutation on a particular chromosome carrying a particular set of marker alleles at nearby loci. If the causal variant becomes established in the population it will remain associated with this local background over many generations because of the segmental descent of the chromosome. Any SNPs on this background will also be associated with the trait because it is associated to the causal SNP. Then, if one or more of these SNPs are genotyped and tested, they can reveal the locus of the causal variant. The presence of genetic linkage means that association mapping can work as long as one or more SNPs that are close to the causal SNP are genotyped and tested. These associations between allelic types and genotypes at different genetic loci are also known as *linkage disequilibrium* (LD) (see **Chapter 2**).

Association mapping indirectly exploits that DNA is inherited in blocks, but it does not take full advantage of it via direct modeling of this process. It also ignores the fact that functional genes are blocks of DNA, and disregards the fact that there may be many different mutations affecting gene function, arising at different sites within a gene. This variety of causal mutations or *allelic heterogeneity* has been one of the major issues for association mapping studies. Where there are multiple causal rare variants within a functional gene, arising on different genetic backgrounds, associations will not be apparent. Another major issue has been the burden of multiple testing, when very large numbers of SNPs across the genome are tested. Also there are issues of population history and structure. Although genetic linkage maintains associations or LD over generations, it does not cause them. Associations even at unlinked loci may be observed if the typed individuals derive from a mix of populations in which variants differ in frequency (Pritchard and Rosenberg, 1999).

Mapping based on identity by descent (IBD), which is the focus of this chapter, also aims to consider associations between marker data X and trait data Y . However, it aims to do so through the lens of descent Z . Similarity of phenotype Y increases the probability of shared descent Z in causal regions, relative to that expected

- given pedigree relationships, if a pedigree is known;
- in similarly related control individuals, if such can be appropriately defined;
- among the same individuals in non-causal regions of the genome.

The idea of IBD-based mapping is to associate location-specific shared descent of genome inferred from marker data \mathbf{X} with similarity in trait values \mathbf{Y} . The excess shared descent inferred from common SNP variation \mathbf{X} provides the signal. Causal variants need not be pre-identified, hypothesized, or even typed. Since segments of IBD contain many SNPs, the burden of multiple testing is much reduced (Browning and Thompson, 2012). While allelic heterogeneity remains an issue, excess IBD can be detected provided there are groups of individuals sharing some of the different causal variants. In this sense, IBD-based genetic mapping test integrates across (rare) variants, addressing allelic heterogeneity.

To summarize the overall goal, in order to be able to perform descent-based mapping one must be able to (i) estimate IBD sharing along the genome and (ii) detect local associations between marker genotypes and trait phenotypes via the estimated IBD sharing patterns. These two items are the topics of Sections 20.2 and 20.3, respectively.

20.2 Inference of Local IBD Sharing from Genetic Marker Data

IBD is inferred from regions of genotypic similarity. This can either be via *ad hoc* rules for detecting stretches of genetic identity as in Gusev *et al.* (2009), or by using a probability model as in Leutenegger *et al.* (2003). Here we focus on the probabilistic approach for which several components are needed. Marker data \mathbf{X} and IBD \mathbf{Z} are related by the equation

$$\Pr(\mathbf{Z}|\mathbf{X}) = \Pr(\mathbf{X}|\mathbf{Z})\Pr(\mathbf{Z})/\Pr(\mathbf{X}).$$

If it is assumed that the probability of marker data X_j at locus j ($j = 1, \dots, \ell$) depends only on the latent IBD state Z_j at that marker location, and on parameters, such as the allele frequencies, specific to that marker then

$$\Pr(\mathbf{Z}|\mathbf{X}) = \left(\prod_{j=1}^{\ell} \Pr(X_j|Z_j) \right) \Pr(\mathbf{Z})/\Pr(\mathbf{X}). \quad (20.1)$$

Note that this assumption implies absence of LD. In Section 20.2.4 we note the impact of this assumption on inference of IBD. The work of Browning (see Browning, 2006, and many subsequent papers) does model LD, taking it into account in the inference of IBD (Browning and Browning, 2010), but those models are beyond the scope of this chapter.

Thus to implement inference of IBD from genetic marker data, we will need (i) a working definition of IBD and a useful way to represent IBD sharing patterns Z_j at each locus j ; (ii) a way to calculate the probability $\Pr(X_j|Z_j)$ of the marker genotypes X_j given the underlying IBD pattern; (iii) a model that provides the probability $\Pr(\mathbf{Z})$ of a given IBD sharing pattern \mathbf{Z} along the genome; and (iv) a way to combine all these components using equation (20.1) to perform the inference. In the rest of this section each of these items will be described in turn.

20.2.1 Identity by Descent at a Locus

Even at a single locus, the definition of *identity by descent* at the population level is unclear. In an assumed pedigree structure it may be defined relative to the genomes of the founders of the pedigree. Segments of DNA in current individuals are then IBD if they descend from a single copy of that DNA in a common ancestor who is a founder of the pedigree. This founder is not necessarily the *most recent common ancestor* (MRCA) of those current DNA segments; the individuals may share this DNA by descent from a more recent non-founder. Given a pedigree structure, probabilities of IBD at any genome location can be computed, or in more

complex cases realized by Monte Carlo simulation. There is, however, high variance in the meiosis process, and, due to the descent of genome in large segments from generation to generation, the human genome is relatively short (Thompson, 2013). While the pointwise probability provides the expected proportion of the genome shared by current individuals, the actual realized IBD varies (Hill and Weir, 2011). Genome-wide, exact probability computations are complex (Donnelly, 1983; Stefanov, 2002). However, even genome-wide, simulation of descent in a defined pedigree is straightforward, and accurate estimates of probabilities of numbers and lengths of segments of IBD can be obtained from Monte Carlo realizations.

Even if a pedigree is not assumed, of course there is a pedigree of any population of extant individuals. At any point in the genome, DNA may be (conceptually) traced back from a collection of current haploid genomes to the MRCA of the collection. The segment around that point in the genome of the MRCA that has descended to the current individuals without recombination is well defined (Palamara *et al.*, 2012). However, this definition becomes complicated when considering many haploid genomes jointly, and does not extend well to considering IBD across the genome, rather than around a focal point.

Alternatively, in a population without an assumed pedigree, IBD can be defined relative to a given time-depth or founder population. This definition is also problematic, since in any analysis of data we have only information on lengths of shared haplotypes, and the high variance in the meiosis process makes the time-origin of such sharing uncertain (Thompson, 2013). Moreover, allelic association across loci (LD) also reflects the demographic history and structure of a population. There is no clear dividing line between haplotypic similarities among genomes that should be ascribed to the founding population LD and haplotypic similarities that result from IBD from relatively recent common ancestors. This is largely an issue of the goal of an analysis, and the time-depth of interest in the study, as well as of the genetic information available which places a threshold on the lengths of detectable IBD segments (Browning and Browning, 2012). Note that, since lengths of segments of IBD decay linearly in the number of meioses, even segments separated by 100 meioses (50 generations) are expected to be 1 cM long. However, segments separated by 150 meioses (75 generations) have a probability of 22% of being at least this long and those at only 50 meioses separation have 40% probability of being shorter than 1 cM. Given this high variance, too great an emphasis on the relationship between lengths of segment and time-depth of coancestry is not useful. It suffices to recognize that, for a human population, at a 1 cM length, we are seeing the outcome of over 1000 years of history.

Once IBD is defined, the next issue is how it should be represented. At any point in the genome, the IBD state among haploid genomes is a partition: those genomes in the same subset of the partition are IBD. For a pair of individuals, there are four genomes, and thus 15 partitions (Figure 20.1). Note that if we do not distinguish the maternal and paternal genomes of individuals, states 3 and 4 are equivalent, as are 6 and 7, 9 and 10, and the four states 11, 12, 13, and 14. This reduces the 15 states to the nine states classes classically considered in analyses of genotypic data (Thompson, 1974). If, further, we ignore the possibility that the two genomes within *A* or *B* are IBD, we have only the seven states in the right-hand part of Figure 20.1. If we make both reductions we have only three state classes: states 9 and 10, where *A* and *B* share two genomes IBD; states 11–14 where they share one; and state 15 where there is no IBD. While this case of two individuals, with 3, 9, or 15 states, has been very widely considered in the literature, a major challenge is the rapid increase in the number of partitions with an increasing number of haploid individuals. For 3 diploid individuals (6 genomes) there are 203 states. For 6 individuals (12 genomes) there are more than 4 million states.

An additional issue is that representation as a partition does not fit well with analyses of diploid genotypic data, where the pair of genomes of each individual must be considered

1		$\{A_p, A_m, B_p, B_m\}$	1
2		$\{A_p, A_m\}, \{B_p, B_m\}$	0
3		$\{A_p, A_m, B_p\}, \{B_m\}$	1/2
4		$\{A_p, A_m, B_m\}, \{B_p\}$	1/2
5		$\{A_p, A_m\}, \{B_p\}, \{B_m\}$	0
6		$\{A_p, B_p, B_m\}, \{A_m\}$	1/2
7		$\{A_p\}, \{A_m, B_p, B_m\}$	1/2
8		$\{A_p\}, \{A_m\}, \{B_p, B_m\}$	0
9		$\{A_p, B_p\}, \{A_m, B_m\}$	1/2
10		$\{A_p, B_m\}, \{A_m, B_p\}$	1/2
11		$\{A_p, B_p\}, \{A_m\}, \{B_m\}$	1/4
12		$\{A_p, B_m\}, \{A_m\}, \{B_p\}$	1/4
13		$\{A_p\}, \{A_m, B_p\}, \{B_m\}$	1/4
14		$\{A_p\}, \{A_m, B_m\}, \{B_p\}$	1/4
15		$\{A_p\}, \{A_m\}, \{B_p\}, \{B_m\}$	0

Figure 20.1 The 15 states of IBD among the four genomes of two diploid individuals, A and B : paternal genomes have subscripts p , and maternal m . In each icon, A is on the left and B on the right; paternal genomes are above and maternal below. The bullets represent the four genomes of two individuals, and the shading shows the ones that are IBD. The central column shows the state, specified as a partition, and the final column gives the state-dependent kinship, the probability that DNA segregating from A and from B is IBD.

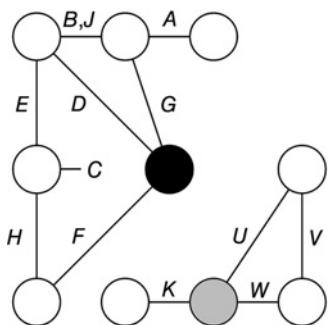


Figure 20.2 An example IBD graph on individuals A, B, \dots . The edges represent the individuals joining the two nodes that represent the genomes the individual carries at this locus. The black and gray nodes are marked only to identify them. The black node indicates that individuals D, G , and F share one of their genomes IBD. The gray node indicates the same for K, U and W . Individuals B and J share both their two genomes IBD at this locus, while the two genomes of C are IBD and IBD also with one of the two genomes of each of E and H .

jointly. An alternative representation of the IBD at a locus among a set of diploid individuals is the descent graph (Sobel and Lange, 1996) or IBD graph (Thompson, 2011). Here the edges represent observed individuals, and the nodes represent the latent DNA that individuals may share. Figure 20.2 shows an example.

20.2.2 Probabilities of Marker Data Given IBD Pattern

We now turn to the second component required by equation (20.1): the computation of probabilities of marker genotypes X_j given the IBD state, Z_j , at a locus j . The IBD graph representation above provides a framework for this computation. In this section, we assume marker genotypes are observed without error. The marker alleles at the locus are denoted by a_1, \dots, a_v , with population frequencies q_1, \dots, q_v , $\sum_{i=1}^v q_i = 1$. Each node k of the IBD graph represents, by definition, an independent genome. Denoting the allelic type of this DNA node at this locus by $a(k)$,

$\Pr(a(k) = i) = q_i$ for $i = 1, \dots, v$. Thus the joint probability of the genotypes of the individuals in any connected component of the IBD graph is (Thompson, 1974)

$$P(X_j|Z_j) = \sum_{\mathcal{A}} \left(\prod_k q_{a(k)} \right) = \sum_{\mathcal{A}} \left(\prod_{i=1}^v q_i^{n_i(\mathcal{A})} \right), \quad (20.2)$$

where \mathcal{A} denotes an allocation of allelic types to the nodes k of the IBD graph which is consistent with the observed genotypes of the individuals, and $n_i(\mathcal{A})$ is the number of nodes in allocation \mathcal{A} that are assigned allelic type i .

Computation requires an efficient algorithm for the summation over all feasible allocations \mathcal{A} , but for genotypes observed without error this is straightforward (Sobel and Lange, 1996). If any individual is homozygous, $a_i a_i$, then his two nodes are of type a_i . Proceeding through individuals in the connected component of the IBD graph, this immediately determines the single (if any) feasible allocation \mathcal{A} that is consistent with the observed marker genotypes. If all individuals are heterozygous at this locus, there may be two feasible allocations, depending on which allelic type is assigned to which node of the initiating individual. There can never be more than two feasible allocations. For example, in Figure 20.2, C must be homozygous, determining the type of his single node, and proceeding sequentially via H to F , and from E to B and A , all the node types are determined. On the smaller component, if U , W , and K are all $a_1 a_2$, either the gray node is a_1 and the other three nodes are a_2 , or the gray node is a_2 and the others are a_1 . However, at most one of these allocations will be consistent with the genotype of V .

20.2.3 Modeling the Probabilities of Patterns of IBD

We now consider the next component required in equation (20.1), the modeling that will provide probabilities $\Pr(\mathbf{Z})$ of IBD states across a chromosome. Because of the huge number of possible trajectories (Z_1, \dots, Z_ℓ) over multiple markers, and the need to compute the normalizing constant

$$\Pr(\mathbf{X}) = \sum_{\mathbf{Z}} \Pr(\mathbf{X}|\mathbf{Z}) \Pr(\mathbf{Z})$$

in equation (20.1), simplifying assumptions must be made about $\Pr(\mathbf{Z})$.

In the case of an assumed pedigree structure, Z_j at any locus j is determined by the *inheritance vector* or vector of meiosis indicators

$$S_j = (S_{i,j}, i = 1, \dots, m), \quad (20.3)$$

where i indexes the m meioses of the pedigree structure and $S_{i,j} = 0$ or 1 according as in meiosis i at locus j the maternal or paternal gene (respectively) of the parent is transmitted to the offspring. The IBD pattern Z_j among pedigree members is determined by S_j but, in general, Z_j does not have a Markov conditional independence structure across loci j . However, in the absence of genetic interference, S_j does have this Markov structure. That is, given all the preceding S_1, \dots, S_{j-1} , S_j depends only on the immediately preceding S_{j-1} :

$$\Pr(S_j|S_1, \dots, S_{j-1}) = \Pr(S_j|S_{j-1}).$$

Then

$$\Pr(\mathbf{X}) = \sum_{\mathbf{S}} \Pr(\mathbf{X}|\mathbf{S}) \Pr(\mathbf{S}) = \sum_{\mathbf{S}} \left(\Pr(S_1) \prod_{j=2}^l \Pr(S_j|S_{j-1}) \prod_{j=1}^l \Pr(X_j|S_j) \right). \quad (20.4)$$

In the absence of a defined pedigree, we need an alternate probability model $\Pr(\mathbf{Z})$ for \mathbf{Z} in equation (20.1). At a point in the genome, the coalescent ancestry of a sample of gametes (Kingman, 1982) defines the partition of genomes into the subsets that are IBD. The ancestral recombination graph (ARG) (Hudson, 1991; Griffiths and Marjoram, 1996) provides the most complete description of the ancestry of the DNA across the genome, defining the IBD partitions relative to any past time point. However, even the sequential Markov coalescent approximation to the ARG (McVean and Cardin, 2005) is too complex a model to use as a model for $\Pr(\mathbf{X})$ in IBD analyses of multiple genotypes or haplotypes across large segments of genome (Zheng *et al.*, 2014a).

Instead, therefore, we use much simpler Markov models, and specify states Z_j with marginal probabilities $\Pr(Z_j)$ and transition probabilities $\Pr(Z_j|Z_{j-1})$. The earliest approaches (Leutenegger *et al.*, 2003; Browning, 2008) considered pairs of haploid genomes, and so only two latent states at a locus: IBD and not IBD. The Markov model for \mathbf{Z} then required two parameters: a pointwise probability of IBD, β , and a change-rate parameter, α . Purcell *et al.* (2007) considered inference of IBD from pairs of diploid genotypes, modeling them as two independent pairs of haplotypes. The IBD states at each locus are summarized as 0, 1, or 2 shared IBD between the two individuals. However, the inbreeding coefficient of offspring is the kinship coefficient of parents, and in most populations IBD within individuals is at least as great as IBD between. Han and Abney (2011) addressed this, providing an estimate of the probability of each of the nine genotypically distinguishable states (see the text describing this grouping of the 15 states following Figure 20.1). However, their model for state transitions is not based on any model of changes in IBD due to ancestral recombination events.

In general, for multiple genomes jointly, we require a model for the IBD partition at a single point in the genome (Section 20.2.1). A useful model for a random sample of genomes from a population is the single-parameter model provided by Ewens (1972). In terms of the pointwise pairwise IBD probability β , this may be written as

$$\Pr(\mathbf{Z}) = (1 - \beta)^{k-1} \beta^{n-k} \left(\prod_{j=1}^{n-1} (1 + (j-1)\beta) \right)^{-1} \prod_{w \in \mathbf{Z}} (|w| - 1)! , \quad (20.5)$$

where n is the total number of haploid genomes, k is the number of IBD subsets in \mathbf{Z} , w is one such IBD subset, and $|w|$ is the number of genomes in the subset w .

A Markov model for transitions among IBD states that has equation (20.5) as its equilibrium distribution is then required. One such was provided by C. Zheng in Brown *et al.* (2012). This model has the single scale parameter α of Leutenegger *et al.* (2003) that scales the rate of potential changes in IBD state along the chromosome. Although this model has clear differences from the pointwise distributions and sequential chromosomal values of IBD states that arise from the ARG, the models are close enough that the model provides a flexible and useful prior for \mathbf{Z} (Zheng *et al.*, 2014b).

Then, analogously to equation (20.4),

$$\Pr(\mathbf{X}) = \sum_{\mathbf{Z}} \Pr(\mathbf{X}|\mathbf{Z}) \Pr(\mathbf{Z}) = \sum_{\mathbf{Z}} \left(\Pr(Z_1) \prod_{j=2}^l \Pr(Z_j|Z_{j-1}) \prod_{j=1}^l \Pr(X_j|Z_j) \right) . \quad (20.6)$$

20.2.4 Inferring Local IBD from Marker Data

Now that we are able to compute all the components of equation (20.6), or of equation (20.4) in the case where the pedigree is known, we can implement a hidden Markov model (HMM) to provide probabilities of IBD, Z_j (equation (20.1)). (For a general discussion of HMMs, see

Chapter 1.) In the current HMM, $\Pr(X_j|Z_j)$ are the emission probabilities, and $\Pr(Z_j|Z_{j-1})$ (or $\Pr(S_j|S_{j-1})$) are the transition probabilities. Standard HMM methods (Rabiner, 1989) allow the computation of $\Pr(Z_j|\mathbf{X})$, the probabilities of IBD states at each locus j given marker data \mathbf{X} jointly at all loci, at least if the number of individuals, and hence IBD states at a locus, is small. When this is not the case, the standard HMM methods are not computationally feasible. However, the HMM still allows for sampling realizations of \mathbf{Z} across all loci from the full conditional distribution $\Pr(\mathbf{Z}|\mathbf{X})$ (equation (20.1)). The exact details of these inference methods depend on whether or not the pedigree is known.

If the pedigree is known, standard HMM inference methods can be developed using the Markov structure of inheritance vectors S_j (equation (20.3)) embodied in equation (20.4) (Kruglyak *et al.*, 1996; Gudbjartsson *et al.*, 2000; Abecasis *et al.*, 2002; Fishelson and Geiger, 2004). The IBD at locus j is a function of S_j . The usefulness of these methods is limited to relatively small pedigrees, since there are 2^m possible values of S_j (equation (20.3)). For larger pedigrees, several different Markov chain Monte Carlo (MCMC) methods have been proposed based on this same HMM dependence structure. Rather than computing probabilities of IBD, these methods sample realizations of descent and IBD, given the genetic marker data \mathbf{X} . The most efficient of these use a combination of block-Gibbs samplers, sampling over loci (Heath, 1997) and over blocks of meioses (Tong and Thompson, 2008). Details of these methods for sampling descent and IBD on pedigrees conditional on marker data can be found in the papers cited, and in the previous edition of this handbook (Thompson, 2007).

If the pedigree is unknown, we instead use a population model for IBD (equation (20.5)) and the HMM structure of equation (20.6). As noted following equation (20.1), this structure assumes absence of LD, and LD can cause problems in IBD inference (Albrechtsen *et al.*, 2009). Thus, assuming no LD in this HMM approach limits the density of markers than can be successfully used in inference. However, IBD segments are generally large relative to the extent of population-level LD, and there is a limit to additional information gained by increasing SNP density. Brown *et al.* (2012) used the HMM model and studied the effects of LD on inference, comparing results with the approach of Browning and Browning (2010) that accommodates LD in the model.

For the 15 states among the four genomes of two diploid individuals the HMM is readily implemented, and Brown *et al.* (2012) carried out an extensive study of IBD inference from both phased and unphased genotypes on pairs of individuals. In addition to considering the impact of LD, they also included an error model for marker data, and other generalizations. For pairs of genomes and/or pairs of individuals, more details of these models and methods are provided in Thompson (2013) and Browning and Browning (2012). Note also that once we have estimates of local pairwise IBD, these can be averaged across the genome to obtain estimates of realized proportion of genome shared IBD. Of course, the HMM is more computationally intensive than standard methods of genome-wide relatedness estimation (see, for example, Visscher *et al.*, 2008). However, Wang *et al.* (2017) showed that the HMM approach provides more accurate estimates.

A major difficulty in inferring IBD states among multiple gametes is simply the large number of possible IBD states (partitions of n haploid genomes (equation (20.5))). Computations using the HMM are feasible, although relatively intensive, for the six genomes of three diploid individuals (203 IBD partitions at a locus), but beyond that HMM computation is not practical. In the absence of a pedigree structure, inference of IBD among multiple genomes remains an intractable problem, although methods have been developed that are useful for up to about 40 genomes (Moltke *et al.*, 2011; Zheng *et al.*, 2014*b*; Glazner and Thompson, 2015).

There is a fundamental difference between pedigree-based probabilities of descent that provide \mathbf{Z} through the processes of recombination and meiosis, and the population-based priors

such as equation (20.5). Given sufficient marker data \mathbf{X} , a flexible approximate population-based prior can work well, allowing the marker data to inform the IBD inference. Indeed, with a high density of markers, the pedigree-based probabilities are too constraining, if many pedigree members are unobserved. A population-based analysis of IBD in observed individuals may then be preferred (Glazner and Thompson, 2015). On the other hand, only with a pedigree-based approach can inferences be made about the genomes of individuals who are not observed. More importantly, a pedigree-based probability $\Pr(\mathbf{Z})$ not only is a prior that can be used to infer IBD from marker data \mathbf{X} (equation (20.1)), but also provides a valid null model for \mathbf{Z} . A population-based prior can be used to infer IBD, but does not provide a realistic null model of probabilities of IBD states. This becomes an issue in the assessment of population-based linkage likelihoods (see Section 20.3.3).

20.3 IBD-Based Detection of Associations between Markers and Traits

Recall that our overall goal is to view associations between marker data \mathbf{X} and trait data \mathbf{Y} through the lens of shared descent of genome \mathbf{Z} . Associations between \mathbf{X} and \mathbf{Y} can be viewed either through $\Pr(\mathbf{X}|\mathbf{Y})$ or $\Pr(\mathbf{Y}|\mathbf{X})$. For the quantitative traits considered in this section, it is more natural to consider $\Pr(\mathbf{Y}|\mathbf{X})$; for the alternate view, see Section 20.4.1. For the purposes of this chapter we assume that \mathbf{X} and \mathbf{Y} are conditionally independent given \mathbf{Z} . Then

$$\Pr(\mathbf{Y}|\mathbf{X}) = \sum_{\mathbf{Z}} \Pr(\mathbf{Y}|\mathbf{Z})\Pr(\mathbf{Z}|\mathbf{X}). \quad (20.7)$$

We have seen in Section 20.2.4, that, given genotypes \mathbf{X} at marker loci across a chromosome, marginal probabilities $\Pr(Z_j|\mathbf{X})$ can be computed, or that realizations of \mathbf{Z} jointly across loci can be obtained from $\Pr(\mathbf{Z}|\mathbf{X})$. Here, we turn to the other factor in equation (20.7), the probabilities $\Pr(\mathbf{Y}|\mathbf{Z})$. We consider two classes of simple models in this section, considering first the probabilities $\Pr(\mathbf{Y}|\mathbf{Z})$ (Sections 20.3.1 and 20.3.2) and then the use of these probabilities in genetic mapping (Sections 20.3.3 and 20.3.4).

20.3.1 Trait Data Probabilities for Major Gene Models

We consider here models where the probability of the observed phenotype of each individual depends only on that individual's (unknown) genotype at a single trait locus. That is, even for individuals whose genomes are IBD at the locus, phenotypes are conditionally independent given the allelic types of the shared IBD node (Figure 20.2). This class of models includes the classical Mendelian models, where each potential genotype provides *penetrance probabilities* (Elston and Stewart, 1971) for a qualitative (usually binary) or quantitative trait. These penetrance probabilities may also depend on individual covariates: for example, age and sex. Also included are models for marker genotypes that include an error model. The 'phenotype' is then the observed marker genotype of the individual, and the 'genotype' is the latent true marker genotype.

This general class of models was termed 'OCIGGOP' by Cannings *et al.* (1978): offspring conditionally independent given genotypes of parents. From Haldane and Smith (1947) onwards, this conditional independence was used to develop computationally feasible ways of computing probabilities of observed data \mathbf{Y} on known pedigree structures. These methods are described in the literature, and summarized in Thompson (2007). They are all basically forms of the now well-known algorithms for computing probabilities in graphical models (Lauritzen, 1992). Here we use the same approach to compute $\Pr(\mathbf{Y}|\mathbf{Z})$ using the IBD graph (Figure 20.2). Generally, the

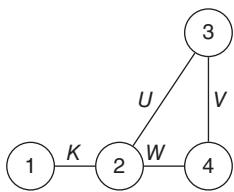


Figure 20.3 Component of IBD graph on individuals K, U, V, W , with four DNA nodes labeled $1, \dots, 4$. The nodes represent the genomes the individuals carry at this locus.

components of an IBD graph are much smaller than a pedigree. The IBD graph contains only observed individuals, whereas a pedigree-based computation often requires summation over the genotypes of many unobserved connecting ancestors of the observed individuals. In almost all cases, computation on the IBD graph is far more efficient, and in fact, graphs are sufficiently small that traits determined by two loci can also be feasibly modeled (Su and Thompson, 2012).

We illustrate the general principles of the computation by the example of the smaller component of the IBD graph of Figure 20.2, shown again in Figure 20.3. The joint probability of phenotypes Y_K, Y_U, Y_V and Y_W given the IBD graph (i.e. given \mathbf{Z}) may be written as

$$\begin{aligned} \Pr(Y_K, Y_U, Y_V, Y_W | \mathbf{Z}) &= \sum_{a(1)} \left(q(a(1)) \sum_{a(2)} P(Y_K | a(1), a(2)) \Pr(Y_U, Y_V, Y_W | a(2)) q(a(2)) \right) \\ &= \sum_{a(1)} \left(q(a(1)) \sum_{a(2)} \left(P(Y_K | a(1), a(2)) q(a(2)) \right. \right. \\ &\quad \times \sum_{a(4)} \left(\Pr(Y_W | a(2), a(4)) q(a(4)) \right. \\ &\quad \times \left. \left. \sum_{a(3)} \Pr(Y_U | a(2), a(3)) \Pr(Y_V | a(4), a(3)) q(a(3)) \right) \right) \right). \end{aligned} \quad (20.8)$$

We see in equation (20.8) that, in this example, no more than two DNA nodes need to be considered jointly at any stage in the summation. Proceeding from the right-hand end of the equation, first data on U and V are included, and the first summation is over the allelic types $a(3)$ for each type of $(a(2), a(4))$. Then the data on W are included, allowing summation over the allelic types of $a(4)$. Next the data on K are included, allowing summation over $a(2)$ for each type $a(1)$. Finally, the summation is completed by summation over $a(1)$. The population allele frequencies $q(a)$ are included for each DNA node, before summation over the allelic types for that node. These are the same computational methods that apply to any graphical structure with implied conditional independencies (Lauritzen, 1992), and many algorithms for efficient computation have been implemented.

20.3.2 Quantitative Trait Data Probabilities under Random Effects Models

Association-based models for the genetic analysis of quantitative trait data \mathbf{Y} focus on the effects attributable to SNP genotypes \mathbf{X}_j at marker j , where each component single-locus genotype is scored as 0, 1, or 2 copies of the reference allele (Wang *et al.*, 1998). The model used, for example, in studies of heritability, is

$$\mathbf{Y} = \mu \mathbf{1} + \sum_j \gamma_j \mathbf{X}_j + \sigma_e \mathbf{e}. \quad (20.9)$$

In this equation, μ is here is mean trait value, but more generally other covariate fixed effects could be included. The last term \mathbf{e} is the vector of individual residuals: $\text{var}(\mathbf{e}) = \mathbf{I}$. The key term in equation (20.9) is the term in \mathbf{X}_j , and in determining the SNPs that show non-zero effects ($\gamma_j \neq 0$). Of course, not all SNPs across the genome can be analyzed jointly. In some cases, methods are developed to ensure sparsity (see, for example, Lango Allen *et al.*, 2012). Others take a Bayesian approach, with a prior on SNP effects, turning these fixed SNP effects into random effects (Yang *et al.*, 2010), allowing many more SNPs across the genome to contribute to covariances among trait values.

In the mapping context, it is usual to consider each SNP j separately, combining the others into a single random effect that summarizes the covariances due to other genetic effects or population structure (Kang *et al.*, 2010):

$$\mathbf{Y} = \mu \mathbf{1} + \gamma_j \mathbf{X}_j + \sigma_a \mathbf{g} + \sigma_e \mathbf{e}. \quad (20.10)$$

The term \mathbf{g} is a random effect modeling genome-wide additive effects. One choice for $\text{var}(\mathbf{g})$ is the realized genomic relatedness matrix Visscher *et al.* (2008). Equation (20.10) is a basic form of the classic mixed model for a quantitative trait.

In the descent-based linkage studies on known pedigrees, the trait-value dependencies among individuals are modeled by IBD, \mathbf{Z} , rather than marker genotypes, \mathbf{X} . The vector of fixed effects $\gamma_j \mathbf{X}_j$ is replaced by a random effect:

$$\mathbf{Y} = \mu \mathbf{1} + \tau_j \mathbf{w}_j + \sigma_a \mathbf{g} + \sigma_e \mathbf{e}, \quad (20.11)$$

where $\text{var}(\mathbf{w}_j) = 2\Phi_j$, the matrix of realized kinships at location j , which must be estimated from the genetic marker data \mathbf{X} (Almasy and Blangero, 1998). In the case of a known pedigree, it is usually assumed $\text{var}(\mathbf{g}) = 2\Psi$, where Ψ is the pedigree-based kinship matrix (Henderson, 1976). Thus \mathbf{g} is the classic additive *polygenic effect* which, as in the association studies case, is assumed to subsume any genetic effects on the trait other than at genome location j . Note that with regard to genetic effects, equation (20.11) is a random-effects model, although there may still be fixed effects due to modeled covariates.

Classically, the genome-wide kinship matrix Ψ could only be based on an assumed pedigree, and marker data were sparse and only highly polymorphic markers gave useful information about Φ_j (see, for example, Haseman and Elston, 1972). With the advent of better markers, pedigree-based HMM methods could give much more accurate estimates of Φ_j using jointly the data across multiple markers (Abecasis *et al.*, 2002), but a pedigree-based computation of Ψ was still required. Now, with modern genetic data, both local and genome-wide estimates of pairwise kinships many be obtained from population samples (Section 20.2.4). Model (20.11) can be used in mapping causal DNA underlying quantitative traits observed in individuals whose pedigree relationships are unknown (Day-Williams *et al.*, 2011).

Equation (20.11) has advantages relative to the models of Section 20.3.1. There are no penetrance parameters and allele frequencies to be specified: obtaining likelihood-based estimates of variance parameters ($\tau_j, \sigma_a, \sigma_e$) is relatively straightforward. This model also has advantages for IBD-based mapping (see Section 20.3.4).

20.3.3 IBD-Based Linkage Likelihoods for Major Gene Models

We consider first the classical linkage likelihood used in mapping genes for single-locus traits on a known pedigree. This likelihood is the probability of the data \mathbf{X} and \mathbf{Y} . The model for trait data \mathbf{Y} specifies the penetrance probabilities and allele frequencies of equation (20.8) and will be denoted by Γ_Y . The model for the SNP marker data \mathbf{X} includes the genetic marker map and will be denoted by Γ_X . The marker model is assumed known. The models for \mathbf{X} and \mathbf{Y} are

connected through the hypothesized location of the causal DNA for the trait, against the fixed known map of the genetic markers. This location is the parameter λ_j , and for each hypothesized value the linkage likelihood is then $\Pr(\mathbf{Y}, \mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j)$.

The *lod score* is the classical tool used to map the genes affecting a trait against a known genetic marker map. These lod scores are computed assuming known pedigree relationships among individuals, and use descent in the pedigree to define the IBD. The lod score is defined as

$$\text{lod}(\lambda_j) = \log_{10} \frac{\Pr(\mathbf{Y}, \mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j)}{\Pr(\mathbf{Y}, \mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j = \infty)} = \log_{10} \frac{\Pr(\mathbf{Y}|\mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j)}{\Pr(\mathbf{Y}; \Gamma_Y)}. \quad (20.12)$$

Here $\lambda_j = \infty$ denotes that the causal DNA is unlinked to the chromosome of the markers. That is, the lod score compares the likelihood that the causal DNA is at a specific location λ_j to that under the same trait and marker models but with the trait-related DNA unlinked to the markers. The normalizing term in the denominator is the marginal probability of trait data \mathbf{Y} . This term requires not only a trait model Γ_Y but also a defined pedigree. Given a pedigree, there are many classic methods developed for computation (Thompson, 2007), but without a pedigree or other relatedness structure it is not defined. We return to this issue below, but first consider the key mapping term: the numerator of the final expression of equation (20.12).

Recall that our goal is to study the relationship between \mathbf{X} and \mathbf{Y} through the medium of IBD \mathbf{Z} (Section 20.1.3), and that, at least in principle, \mathbf{Z} can be inferred from genetic marker \mathbf{X} (equation (20.1)). We also retain our basic assumption that \mathbf{Y} and \mathbf{X} are conditionally independent given \mathbf{Z} . We therefore write the numerator likelihood of equation (20.12) as

$$\Pr(\mathbf{Y}|\mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j) = \sum_{\mathbf{Z}} \Pr(\mathbf{Y}|\mathbf{Z}; \Gamma_Y, \lambda_j) \Pr(\mathbf{Z}|\mathbf{X}; \Gamma_X). \quad (20.13)$$

Even for a few observed individuals and at a single locus, the number of IBD states is huge. Thus equation (20.13) is not practical. However, methods to obtain realizations \mathbf{Z} conditionally on marker data \mathbf{X} have been developed both on pedigrees (Tong and Thompson, 2008) and in populations (Glazner and Thompson, 2015). Following an approach originally due to Lange and Sobel (1991), we may write

$$\hat{\Pr}(\mathbf{Y}|\mathbf{X}; \Gamma_Y, \Gamma_X, \lambda_j) = N^{-1} \sum_{k=1}^N \Pr(\mathbf{Y}|\mathbf{Z}^{(k)}; \Gamma_Y, \lambda_j), \quad (20.14)$$

where $\mathbf{Z}^{(k)}$, $k = 1, \dots, N$ are realizations from $\Pr(\mathbf{Z}|\mathbf{X}; \Gamma_X)$.

Glazner and Thompson (2015) considered an example of data simulated on a large pedigree for a simple single-locus quantitative trait, with data assumed on 31 individuals in three clusters at the bottom of the pedigree. At the trait locus, the causal allele descended to members of each of the three clusters. Conditional on the descent at the trait locus, descent at marker loci across the chromosome was simulated. Conditional on this descent, marker data at 10,188 SNP markers were generated for the 31 individuals, using real human haplotypes. IBD was then inferred, using population models and not assuming any pedigree knowledge, and the base-10 log-likelihood (20.14) computed at 334 points across the 200 cM chromosome. The result is shown in Figure 20.4. Since the data are simulated, we know the log-likelihood curve that would be obtained if the true IBD were known without error (the gray line). Using phased data (dashed line), we come close to this, but unphased genotypes provide less information (dotted line). Even so, in this very clean example, good results are obtained.

However, there are three issues. First, there is no absolute baseline value, since the pedigree is unknown. In fact, in this simulated example the pedigree is known providing a baseline value of -33.02, but in real examples this is not available. Second is the issue of whether any constant baseline is appropriate. Where there is more IBD, likelihoods such as equation (20.8) tend to

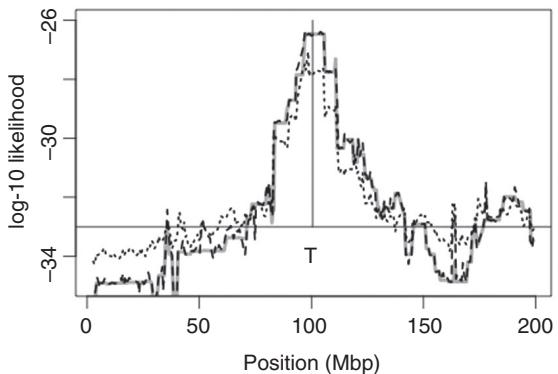


Figure 20.4 Unnormalized linkage likelihoods computed for the example of data on 31 related individuals of Glazner and Thompson (2015). The gray line shows the value that would be obtained if the true realized pairwise IBD were known at each genome location. The dashed and dotted lines show (respectively) the results of based on IBD estimated from the phased haplotypes and unphased genotypes of the 31 individuals. No pedigree information is assumed in this estimation.

be higher, because there are fewer allele frequency terms $q(a)$ to be incorporated. Without the constraints of a pedigree the level of inferred IBD can vary widely across a chromosome, and likelihood may reflect this IBD level rather than trait-related DNA (Glazner, personal communication). Finally, although good results were obtained here, estimation of IBD among multiple individuals in the absence of pedigree information remains a very challenging problem. The methods of Glazner and Thompson (2015) seem to work well for groups of up to eight diploid individuals. However, where the linkage signal is less clear than in this example, 31 diploid individuals is beyond the limit for useful results.

Additional issues arise in assessing the significance of a linkage peak. A classical tool in pedigree-based linkage analysis is the *elod* or expected lod score (Thompson *et al.*, 1978; Ott, 1999), which measures the information to be expected from data on a set of pedigrees, given a trait model. A more relevant measure of expected information for linkage is the elod conditional on specific observed trait data. Ploughman and Boehnke (1989) provided a Monte Carlo method to obtain these conditional expected lod scores. Implemented in their program SIMLINK, this has become a mainstay for practical pedigree-based linkage studies. Given a pedigree structure, simulation provides an elod either for a pedigree design or conditionally on trait data. However, without a pedigree there is no relevant simulation, and hence no elod measure for Mendelian models.

20.3.4 IBD-Based Linkage Likelihoods for Random-Effects Models

In this section we consider mapping causal DNA underlying a quantitative trait using the random-effects model of equation (20.11). For this model we use a slightly different form of likelihood ratio. Rather than comparing to a model with the same trait locus specification, but at a locus unlinked to the currently considered markers, we compare the general model of equation (20.11) to the model in which there is no effect at location j ($\tau_j^2 = 0$),

$$\ell_j = \log \left(\frac{\max_{\mu, \sigma_a^2, \tau_j^2, \sigma_e^2} \Pr(\mathbf{y}; \mu, \sigma_a^2, \tau_j^2, \sigma_e^2; \Phi_j, \Psi)}{\max_{\mu, \sigma_a^2, \sigma_e^2} \Pr(\mathbf{y} | \mu, \sigma_a^2, \tau_j^2 = 0, \sigma_e^2; \Psi)} \right), \quad (20.15)$$

where \mathbf{y} is the observed value of \mathbf{Y} , the vector of trait values. The likelihood ratio is a function of the local and genome-wide kinship matrices, so there is no explicit pedigree involved. Maximization is straightforward, and not computationally intensive unless very large numbers of interrelated individuals are observed.

A major statistical issue is how Φ_j is to be estimated from genetic marker data \mathbf{X} . Day-Williams *et al.* (2011) used a method-of-moments-based estimator based on pairwise genotypic

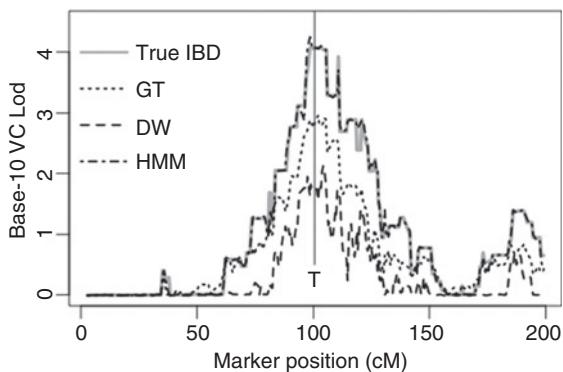


Figure 20.5 Comparison of estimators of the log-likelihood ratio (20.15) in the example of Glazner and Thompson (2015). The gray line shows the value that would be obtained if the true realized pairwise IBD were known as each genome location. The three dashed lines show the results from IBD inferred using the Day-Williams (DW), Glazner–Thompson (GT), and pairwise HMM methods.

similarities. The estimates are smoothed across the chromosome, and at any point the local pairwise value is constrained to be 0, 1/4, 1/2, or 1 (Figure 20.1). Glazner and Thompson (2015) applied their methods for realization of multi-genome IBD states using unphased genotypic data. Each multi-genome state determines the full set of pairwise states (Figure 20.1) for the set of individuals, and hence a collection of matrices Φ_j over all j in which the individual terms are automatically 0, 1/4, 1/2, or 1. However, a major advantage of model (20.11) is that only pairwise estimates of IBD are needed. The HMM methods of Brown *et al.* (2012) (see Section 20.2.4) provide accurate probability estimates or realizations of the 15 pairwise states for all pairs of individuals, and hence also estimates of Φ_j across locations j . Note that the two latter approaches do not require smoothing of estimates of Φ_j across the chromosome. The HMM approach automatically integrates across markers, producing a smooth curve. Test positions j for computation of ℓ_j (equation (20.15)) can be spaced at no more than 2 per centimorgan without loss of information, thus reducing the burden of multiple testing relative to an association test.

The same simulated example data set as in Section 20.3.3 is used to illustrate these three alternatives. As in the example of Day-Williams *et al.* (2011) there was no evidence for a genome-wide polygenic effect, so for simplicity it was assumed $\sigma_a^2 = 0$. The log-likelihood curves are shown in Figure 20.5. The gray line again shows the curve that would be obtained if the IBD were known. It is seen that the model-based IBD realization procedures of Glazner and Thompson (2015) (denoted GT) outperform the method-of-moments estimators of Day-Williams *et al.* (2011) (denoted DW). However, for this simple example, the pairwise method of Brown *et al.* (2012) outperforms both, providing almost perfect estimates of the pairwise IBD states, and hence a log-likelihood curve essentially identical to that which would be obtained if IBD were known without error.

It is an open question whether, in some cases, estimation of IBD jointly among multiple individuals may improve pairwise estimates. In the example data set of this section there are 31 observed individuals and no pedigree structure is assumed in the analysis. This provides an extreme challenge: the number of IBD states among 31 individuals (62 genomes) is gigantic. Even though many of these will be excluded by the marker data and never realized, it is perhaps only surprising that the joint method of Glazner and Thompson (2015) performed as well as it did. However, even for smaller groups of individuals, it seems joint analysis does not improve pairwise estimation. For sets of three individuals, the HMM method of Section 20.2.4 allows computation of the exact posterior probability distribution over the 203 IBD states at each locus given the marker data X on the trio at markers across the chromosome. It is found that, unless additional relationship structure is assumed, the pairwise estimates obtained from the trio analysis are usually almost identical, and in general no more accurate, than the results of running the HMM method on the three pairs separately (results not shown).

In addition to having a more tractable likelihood function, with few parameters, and to avoiding problems of inferring joint IBD among multiple individuals, there is another major advantage to the variance component approach. This is that an expected lod score (or elod) is available, even in the absence of pedigree information. As seen previously, in the absence of pedigree constraints, levels of IBD inferred from marker data can vary widely across a chromosome, and likelihoods may depend more on the level of IBD at a location than on whether that IBD is associated with trait similarity. It is therefore of interest to consider the expected value of the lod score (20.15) conditional on the IBD matrix Φ_j at each location j .

Consider two alternative variance matrices for \mathbf{Y} in equation (20.10), V_1 and V_0 . Then the Kullback–Leibler information (Kullback and Leibler, 1951) or *elod* is

$$0 \leq E_{V_1} (\log \Pr(Y; V_1) - \log \Pr(Y; V_0)) = \frac{1}{2} (\log(|V_0|/|V_1|) + \text{tr}(V_1 V_0^{-1}) - n), \quad (20.16)$$

where $|V|$ denotes the determinant of V , $\text{tr}(\cdot)$ denotes the trace, and n is the dimension of \mathbf{Y} (the number of observed individuals). Given values of trait heritability $h^2 = (\tau^2 + \sigma_a^2)/(\tau^2 + \sigma_a^2 + \sigma_e^2)$ and the relative contributions of major gene and polygenic background τ^2/σ_a^2 , the value of equation (20.16) may be computed using the kinship matrices Φ_j and Ψ . At each location j , it provides the expected lod score, under that assumed model, given the IBD at that location. It shows to what extent an apparent linkage signal (high value of the lod score (20.15)) is due only to the IBD inferred from marker data, and to what extent it is due to the actual observed data \mathbf{y} . Values of equation (20.15) are always non-negative, but only values above those of equation (20.16) are of possible interest in mapping causal DNA.

20.4 Other Forms of IBD-Based Genetic Mapping

20.4.1 IBD-Based Case–Control Studies

In any model-based analysis of the relationship between variables \mathbf{Y} and \mathbf{X} , one may consider $\Pr(\mathbf{Y}|\mathbf{X})$ or $\Pr(\mathbf{X}|\mathbf{Y})$. The models of Sections 20.3.1 and 20.3.2 model trait data \mathbf{Y} conditionally on marker genotypes \mathbf{X} , or on IBD \mathbf{Z} inferred from \mathbf{X} . Case–control studies take the reverse approach. Individuals are sampled based on their status as a case or as a control, and the frequencies of marker alleles, marker genotypes, or local marker haplotypes, compared between the two groups. In the basic form, the frequency of a SNP allele in N_1 cases is compared with that in N_2 controls. The test statistic is

$$T_j = \frac{1}{2N_1} \sum_{\text{cases}} X_{ij} - \frac{1}{2N_2} \sum_{\text{controls}} X_{ij}, \quad (20.17)$$

where $X_{ij} = 0, 1, 2$ is the genotype of individual i at marker j , that is, the number of locus- j alleles of specified type in i . In any population, there may be differences in allele frequencies between cases and controls for reasons that are unrelated to the trait. Generalizations of the basic association test that allow for population structure and heterogeneity have been developed (see Chapter 21).

A major theme of this chapter is that any test of the relationship between \mathbf{Y} and \mathbf{X} that aims to map causal DNA should rather be based on the relationship between \mathbf{Y} and the descent of that DNA \mathbf{Z} . By analogy with equation (20.17), in an IBD-based test, we compare the frequency of IBD between M_1 case–case pairs and M_2 other pairs (case–non-case or non-case–non-case):

$$T_j^* = \frac{1}{M_1} \sum_{\text{case-case}} Z_{pj} - \frac{1}{M_2} \sum_{\text{other}} Z_{pj}, \quad (20.18)$$

where $Z_{pj} = 0, 1$, or 2 as the pair p shares $0, 1$, or 2 , DNA copies IBD at a test location, j . That is, the excess shared descent among cases at causal locations provides the linkage signal (see Section 20.1.3). Just as in an association test, we must allow for population heterogeneity or structure. In an IBD-based test, there may be different degrees of relatedness among cases from among controls, due to the methods of sampling or ascertainment. The average IBD scores within each group in equation (20.18) may be adjusted for the genome-wide average within in each group.

To assess significance a null distribution is required. Whereas, in a known pedigree, Mendelian segregation provides an appropriate null distribution, in a population there is no such framework. However, just as in an association test, permutation of case/control labels provides a null distribution of the test statistic (20.18) under which there is no association between IBD at the test location and the case/control status of individuals. Since IBD is on a scale of millions of base pairs (Table 20.1), at most 3000 tests can cover the genome. This results in a multiple testing burden that is several orders of magnitude less than that for SNP-based genome-wide association studies.

Browning and Thompson (2012) studied under what situations an IBD-based case–control study using equation (20.18) will have greater power than an the allele-based test (20.17). They used population simulations including mutation and selection (Hernandez, 2008) to generate ‘genes’ of length 200 kb, with causal variants in the central 10 kb. In each realization, the most informative SNP in each of 100 blocks distributed across the gene was chosen, forming a set of 100 marker SNPs. When selection against causal variants is weak, the total frequency of causal haplotypes is higher. The association test (20.17) then performed well. In most cases, there was a high association between at least one of the causal variants and one of the marker SNPs.

However, when selection is stronger, the frequencies of causal variants are much lower, and the total frequency of haplotypes carrying causal alleles is of the order of 1%. In this case there is rarely a detectable association between any causal variant and any of the 100 common marker SNPs. In this case, an IBD-based test performs better than an association test. Allelic heterogeneity is a major problem for association testing, unless there is at least one variant with sufficiently high frequency to show association. On the other hand, an IBD-based test is not directly affected by allelic heterogeneity, since each case–case pair has a higher chance of carrying the same causal allele, even though this allele may differ among pairs.

The study of Browning and Thompson (2012) made a number of simplifying assumptions. For an IBD-based test to work, there must be sufficient IBD in the sample of cases, and this IBD must be able to be accurately inferred from genetic marker data. Although the future of this approach remains uncertain, it does seem that the IBD approach can integrate over allelic heterogeneity, and detect distinct groups of chromosomes in affected individuals that each share descent from an ancestral haplotype carrying a distinct causal mutation. An association test may not detect this signal. Albrechtsen *et al.* (2009) give a real data example in which an IBD-based mapping approach provides a large increase in power compared to standard association mapping methods.

20.4.2 Patterns of IBD in Affected Relatives

Analysis of IBD in affected relatives has a long history both in understanding the genetic basis of a trait, and in mapping causal DNA. These methods rely on the fact that, regardless of the trait model, provided it has some genetic component, related affected individuals or related individuals exhibiting extreme trait values will share genes IBD at trait loci with some increased probability. Hence also they will share genes IBD with increased probability at marker loci linked to those trait loci (Section 20.1.3). While genome-sharing methods on individuals with known

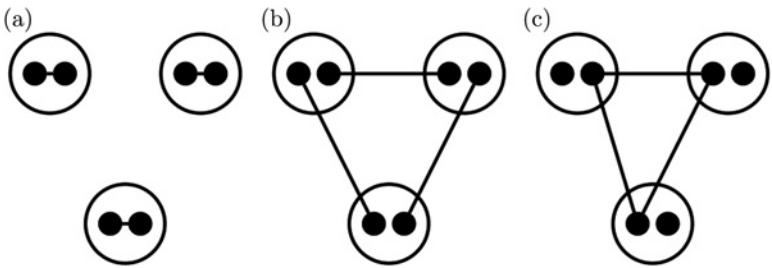


Figure 20.6 Possible patterns of IBD in sets of three affected individuals: The small circles denote the two genes of each individual, and the connecting lines indicate IBD. (a) A classic recessive. (b) Recessive at the gene level: compound heterozygotes. (c) A dominant causal allele.

relationship derive first from Penrose (1935), and are best known in the form of sib pair analyses (Suarez *et al.*, 1978) and affected-relative methods (Weeks and Lange, 1988), they can also involve probability computations on a pedigree structure. For example, the method of *homozygosity mapping* for rare recessive traits developed by Lander and Botstein (1987) can be viewed either as computation of a linkage likelihood (Smith, 1953) or as an inference based on the genome sharing between the two haplotypes of inbred affected individuals.

In the population context, we again have the issue of no clear null value when assessing the extent of IBD among affected individuals. However, as noted in Section 20.1.3, genomic control can always be used. That is, where there is no pedigree to provide a null distribution, and where there is no way to obtain control individuals who can provide a valid basis for levels of IBD, one may still compare levels of IBD across the genomes of affected individuals, and find those regions showing significantly greater levels of IBD.

The pattern of IBD among affected individuals can also provide information about the genetic basis of a trait. In Figure 20.6(a), we see the classic pattern of a simple recessive trait. For homozygosity mapping (Lander and Botstein, 1987), affected individuals whose parents are known to be closely related are genotyped. With very high probability, at the locus being considered, each affected individual is homozygous for an allele received IBD from a recent common ancestor. That is, the two homologues of each affected individual are IBD at this locus. Of course, there may be occasional recombinants between the marker loci and the causal gene, which will result in not every individual being homozygous at even a closely linked marker. Additionally, there may be allelic heterogeneity among cases, but this does not affect the clear signal of IBD between the two homologues of the majority of affected individuals. The occasional impact of deletions resulting in hemizygosity also does not impact this signal of a causal locus. While homozygosity mapping both in pedigrees and in populations has been important in the past, novel such findings are unlikely.

Many Mendelian disorders caused by variants in one or both homologues of a single gene have eluded pedigree-based linkage mapping due to their rarity, and IBD-based population approaches are resolving some of these. Where individuals with rare traits are ascertained from a population, for traits known or thought to be recessive, exclusion mapping (Edwards, 1987) may eliminate the entire genome. This may be due to locus heterogeneity, but more often it may be due to allelic heterogeneity. Rather than having two IBD homologues, affected individuals may be compound heterozygotes, carrying two different but non-normal alleles. Such situations are well known in the case of ‘Mendelian’ traits such as the first to be positionally cloned, the CFTR locus for cystic fibrosis, where now hundreds of alleles are known which have different geographic origins (Estivill *et al.*, 1997), and in their various combinations have varied trait effects (Chillón *et al.*, 1995). Figure 20.6(b) shows an example of IBD that might be seen for

such a trait. More generally, an affected individual who shares one of his allele copies IBD with another affected individual and shares his other allele copy IBD with a third affected individual may provide a mapping signal if the test statistics exploit this expected IBD pattern.

By contrast, Figure 20.6(c) shows the pattern of IBD that might result among affected individuals at a causal locus if a deleterious mutation in a single homologue is sufficient to cause the trait, providing a ‘dominant’ rather than ‘recessive’ effect. *Dominant* is here used in a general sense: not all individuals carrying these variant alleles may be affected, and there may be age-of-onset effects or other covariates that affect the trait status of individuals. Note that Figures 20.6(b) and 20.6(c) exhibit exactly the same pattern of pairwise IBD. Each pair of individuals in the trio shares one homologue IBD at the locus. For additive models pairwise IBD is sufficient, and joint IBD inference does not improve pairwise IBD estimates (Section 20.3.4). In contrast, for non-additive traits even for three individuals, the joint pattern of IBD can convey information about the genetic parameters of trait determination, or, given the trait mode, about the location of causal DNA.

20.5 Summary

This chapter has presented ideas and methods for the mapping of DNA underlying traits of interest observed in individuals. The focus is on the descent of DNA from common ancestors to currently observed individuals. It is this IBD among individuals that gives rise to phenotypic similarities and genotypic correlations, and to the associations in descent between marker genotypes and trait phenotypes that are the basis of genetic mapping. Individuals who share genome IBD are, of course, related. However, it makes no intrinsic difference whether the relationship is assumed known and used in the analysis of descent, or instead the individuals are modeled as sampled from a population. In both cases, patterns of descent and IBD are inferred from genetic marker data, and this inferred descent is then used in analysis of the trait data to infer the genomic locations of causal DNA.

In the case of a known pedigree, that pedigree structure provides both a prior model for inference of IBD from genetic marker data, and also a null model for probabilities of trait and marker data in the absence of genetic linkage. However, the prior may be over-constraining, and MCMC-based methods cannot be used effectively for dense markers or if there are generations of ancestral individuals without observed data. In the case of population data, a flexible prior distribution for IBD may be assumed, and dense data provide clear detection of IBD segments among small numbers of genomes. However, the prior does not provide a null model, and inference of IBD among multiple genomes across a chromosome remains a significant challenge. Where model likelihoods or test statistics depend only on pairwise relatedness (the four genomes of two diploid individuals) the population-based priors provide computationally practical methods and accurate IBD inferences. However, there are situations in which the joint pattern of IBD among the genomes of more than two individuals provides important additional information.

Acknowledgements

This research was supported in part by NIH grant R37 GM-46255. I am grateful to former and current students Chris Glazner and Bowen Wang, and colleague Sharon Browning, for many discussions on IBD inference and QTL mapping, and to Dr. Ida Moltke for detailed helpful comments on an earlier draft. Figures 20.1, 20.4 and 20.5 are based on work with Chris Glazner.

References

- Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002). Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.
- Albrechtse, A., Korneliussen, T.S., Moltke, I., van Overeem Hansen, T., Nielsen, F.C. and Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology* **33**, 266–274.
- Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198–1211.
- Brown, M.D., Glazner, C.G., Zheng, C. and Thompson, E.A. (2012). Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics* **190**, 1447–1460.
- Browning, S.R. (2006). Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics* **78**, 903–913.
- Browning, S.R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* **178**, 2123–2132.
- Browning, S.R. and Browning, B.L. (2010). High-resolution detection of identity by descent in unrelated individuals. *American Journal of Human Genetics* **86**, 526–539.
- Browning, S.R. and Browning, B.L. (2012). Identity by descent between distant relatives: Detection and applications. *Annual Review of Genetics* **46**, 617–633.
- Browning, S.R. and Thompson, E.A. (2012). Detecting rare variant associations by identity by descent mapping in case-control studies. *Genetics* **190**, 1521–1531.
- Cannings, C., Thompson, E.A. and Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10**, 26–61.
- Chillón, M., Casals, T., Mercier, B., Bassas, L., Lissens, W., Silber, S., Romey, M., Ruiz-Romero, J., Verlingue, C., Claustrès, M., Nunes, V., Férec, C. and Estivill, X. (1995). Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *New England Journal of Medicine* **332**, 1475–1480.
- Day-Williams, A.G., Blangero, J., Dyer, T.D., Lange, K. and Sobel, E.M. (2011). Linkage analysis without defined pedigrees. *Genetic Epidemiology* **35**, 360–370.
- Donnelly, K.P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology* **23**, 34–63.
- Edwards, J.H. (1987). Exclusion mapping. *Journal of Medical Genetics* **24**, 539–543.
- Elston, R.C. and Stewart, J. (1971). A general model for the analysis of pedigree data. *Human Heredity* **21**, 523–542.
- Estivill, X., Bancelis, C. and Ramos, C. (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Human Mutation* **10**, 135–154.
- Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.
- Fishelson, M. and Geiger, D. (2004). Optimizing exact linkage computations. *Journal of Computational Biology* **11**, 263–275.
- Glazner, C.G. and Thompson, E.A. (2015). Pedigree-free descent-based gene mapping from population samples. *Human Heredity* **80**, 21–35.
- Griffiths, R.C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology* **3**, 479–502.
- Gudbjartsson, D., Jonasson, K., Frigge, M. and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* **25**, 12–13.
- Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M. and Pe'er, I. (2009). Whole population genome-wide mapping of hidden relatedness. *Genome Research* **19**, 318–326.

- Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 229–309.
- Haldane, J.B.S. and Smith, C.A.B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics* **14**, 10–31.
- Han, L. and Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype. *Genetic Epidemiology* **35**, 557–567.
- Haseman, J.K. and Elston, R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19.
- Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61**(3), 748–760.
- Henderson, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**, 69–83.
- Hernandez, R.D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**, 2786–2787.
- Hill, W.G. and Weir, B.S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetical Research Cambridge* **93**, 47–64.
- Hudson, R. (1991). Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics (eds.), *Oxford Surveys in Evolutionary Biology*, Vol. 7. Oxford University Press, Oxford, pp. 1–44.
- Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354.
- Kingman, J.F.C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27–43.
- Kosambi, D.D. (1944). The estimation of map distances from recombination values. *Annals of Eugenics* **12**, 172–175.
- Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics* **58**(6), 1347–1363.
- Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Statistics* **22**, 79–86.
- Lander, E.S. and Botstein, D. (1987). Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570.
- Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49**, 1320–1334.
- Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., et al. (2012). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838.
- Lauritzen, S.J. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87**, 1098–1108.
- Leutenegger, A., Prum, B., Genin, E., Verny, C., Clerget-Darpoux, F. and Thompson, E.A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics* **73**, 516–523.
- McVean, G. and Cardin, N. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1387–1393.
- Mendel G. (1866). Experiments in Plant Hybridisation, in J. H. Bennett, ed., ‘English translation and commentary by R. A. Fisher’, Oliver and Boyd, Edinburgh, 1965.
- Moltke, I., Albrechtsen, A., Hansen, T., Nielsen, F.C. and Nielsen, R. (2011). A method for detecting IBD regions simultaneously in multiple individuals – with applications to disease genetics. *Genome Research* **21**, 1168–1180.

- Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edn. Johns Hopkins University Press, Baltimore, MD.
- Palamara, P.F., Lencz, T., Darvasi, A. and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics* **91**, 809–822.
- Penrose, L.S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics* **6**, 133–138.
- Ploughman, L.M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44**, 543–551.
- Pritchard, J.K. and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**, 220–228.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. and Sham, P.C. (2007). PLINK: A tool-set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575.
- Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE* **77**, 257–286.
- Smith, C.A.B. (1953). Detection of linkage in human genetics. *Journal of the Royal Statistical Society, Series B* **15**, 153–192.
- Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.
- Stefanov, V.T. (2002). Statistics on continuous IBD data: Exact distribution evaluation for a pair of full (half)-sibs and a pair of a (great-) grandchild with a (great-) grandparent. *BMC Genetics* **3**, 7.
- Su, M. and Thompson, E.A. (2012). Computationally efficient multipoint linkage analysis on extended pedigrees for trait models with two contributing major loci. *Genetic Epidemiology* **38**, 602–611.
- Suarez, B.K., Rice, J. and Reich, T. (1978). The generalized sib pair IBD distribution: Its use in the detection of linkage. *Annals of Human Genetics* **42**, 87–94.
- Thompson, E.A. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.
- Thompson, E.A. (2007). Linkage analysis. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edn. Wiley, Chichester, pp. 1141–1167.
- Thompson, E.A. (2011). The structure of genetic linkage data: From LIPED to 1M SNPs. *Human Heredity* **71**, 86–96.
- Thompson, E.A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics* **194**, 301–326.
- Thompson, E.A., Kravitz, K., Hill, J. and Skolnick, M.H. (1978). Linkage and the power of a pedigree structure. In N.E. Morton (ed.), *Genetic Epidemiology*. Academic Press, New York, pp. 247–253.
- Tong, L. and Thompson, E.A. (2008). Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations, **65**, 142–153.
- Visscher, P.M., Andrew, T. and Nyholt, D.R. (2008). Genome-wide association studies of quantitative traits with related individuals: Little (power) lost but much to be gained. *European Journal of Human Genetics* **16**, 387–390.
- Wang, B., Sverdlov, S. and Thompson, E.A. (2017). Efficient estimation of realized kinship. *Genetics* **205**, 1063–1078.
- Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lander, E.S., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

- Weeks, D.E. and Lange, K. (1988). The affected pedigree member method of linkage analysis. *American Journal of Human Genetics* **42**, 315–326.
- Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E. and Visscher, P.M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.
- Zheng, C., Kuhner, M.K. and Thompson, E.A. (2014a). Bayesian inference of local trees along chromosomes by the sequential Markov coalescent. *Journal of Molecular Evolution* **78**, 279–292.
- Zheng, C., Kuhner, M.K. and Thompson, E.A. (2014b). Joint inference of identity by descent along multiple chromosomes from population samples. *Journal of Computational Biology* **21**, 185–200.

21

Genome-Wide Association Studies

Andrew P. Morris¹ and Lon R. Cardon²

¹ Department of Biostatistics, University of Liverpool, Liverpool, UK

² BioMarin Pharmaceutical, Novato, CA, USA

Abstract

Population-based genome-wide association studies (GWASs) have been extremely successful in identifying loci that contribute to complex human traits and common diseases. In this chapter, we discuss GWAS design issues and describe protocols for the assessment of genotype quality from GWAS arrays. We introduce the statistical models utilised in detecting association between single genetic variants and quantitative traits or disease phenotypes, and describe how these methods can be extended to account for population structure and relatedness between individuals. We describe approaches that evaluate association with a trait using genotype data from multiple variants simultaneously. We also summarise available software for each stage of GWAS analysis. Finally, we discuss the continued prospects of GWASs for detecting regions of the genome associated with complex traits and their utility in predicting disease risk and developing targeted treatment interventions.

21.1 Introduction

In the last century, the traditional approach to mapping regions of the genome contributing to human traits and diseases was through linkage analysis in pedigrees or smaller family units. These studies trace the co-segregation of genetic markers (such as microsatellites) with the trait under investigation through families (discussed in detail in **Chapter 20**), and proved to be extremely effective in localising the genes that are causal for Mendelian disorders, such as cystic fibrosis (Kerem *et al.*, 1989) and Huntington's disease (MacDonald *et al.*, 1992). However, linkage studies proved less successful for complex traits and diseases, such as body mass index and type 2 diabetes, where large numbers of variants, genome-wide, contribute to the phenotype, which may be modified by non-genetic risk factors, such as diet and smoking (Altmuller *et al.*, 2001). Consequently, individuals affected by a complex disease, for example, are less concentrated within families, and are less likely to share the same causal alleles than for Mendelian disorders.

Conversely, population-based association studies focus on identifying genetic markers, usually single nucleotide polymorphisms (SNPs), for which genotypes are associated with the trait under investigation, that is, have different frequencies in individuals affected and unaffected by disease, or have different mean quantitative measures. Association studies are typically undertaken in samples of unrelated individuals or small family units, as opposed to large pedigrees.

In this way, association studies exploit the fact that it is easier to ascertain large samples of individuals sharing a causal genetic polymorphism across a population than within families, which would be required for linkage analysis. Theoretical arguments have been proposed to demonstrate that population-based association studies are more powerful than linkage studies for identifying genetic polymorphisms contributing to complex traits, providing the underlying causal variants are not rare (Risch and Merikangas, 1996). The cornerstone of association studies is thus the 'common disease common variant' (CDCV) hypothesis, which states that complex traits are determined by genetic polymorphisms of modest effect that occur with more than 1% minor allele frequency (MAF) in the population (Reich and Lander, 2001).

Despite the potential power advantage of population-based association studies for identifying genetic polymorphisms contributing to complex traits, the success of initial screens, which focused on SNPs in candidate genes or larger candidate regions, was limited to a handful of major gene effects, including *APOE* for Alzheimer's disease (Corder *et al.*, 1993) and *NOD2* for Crohn's disease (Hugot *et al.*, 2001). However, many other reported causal candidate genes proved difficult to replicate across multiple studies, highlighting the importance of study design to detect and validate more modest effects. Furthermore, a 'hypothesis-free' approach that interrogates hundreds of thousands of SNPs across the whole genome, rather than focusing on suspected candidate genes, might provide additional insight into underlying biology of complex human traits (Tabor *et al.*, 2002).

One of the first success stories from population-based genome-wide association studies (GWASs) was the identification of a causal variant for age-related macular degeneration (AMD) among Europeans in the complement factor H (*CFH*) gene (Klein *et al.*, 2005). An initial scan of more than 100,000 SNPs, genome-wide, in just 96 cases and 50 controls revealed a strong association of a common intronic variant in the gene. Resequencing of the gene region identified a strongly associated coding variant that represents a tyrosine–histidine change at amino acid 402. The coding variant had a larger effect on AMD than we would expect for most complex traits, with a relative risk of 7.4 for individuals homozygous for the causal allele compared to those homozygous for the wild-type allele, but nevertheless highlighted the potential of GWASs for understanding the genetic contribution to disease.

The Wellcome Trust Case Control Consortium (WTCCC) subsequently undertook a landmark GWAS of seven complex diseases (type 1 diabetes, type 2 diabetes, coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis and inflammatory bowel disease) in the UK population (Wellcome Trust Case Control Consortium, 2007). A total of 2000 cases of each disease and 3000 'shared' controls (from the 1958 British Birth Cohort and the National Blood Service) were genotyped for more than 500,000 SNPs genome-wide. The study identified 24 association signals (at $p < 5 \times 10^{-7}$) across the seven diseases, most represented by common SNPs with modest effect sizes, highlighting the need for large sample sizes. The WTCCC also established protocols for all aspects of GWAS design and analysis, including quality control, association testing and accounting for population structure, which are still employed by the genetics research community.

In this chapter we begin by reviewing the concepts underlying GWAS design, including phenotype definition, genotyping technologies, sample size calculations, significance thresholds and replication. We then review protocols for genotype calling and GWAS quality control, including procedures for excluding low-quality SNPs and samples from downstream investigations. Next, we describe GWAS methodology for assessing the evidence of association of SNPs with complex traits, focusing on binary and quantitative outcomes, describing alternative genetic models, and approaches to account for interactions with non-genetic risk factors. We then consider approaches to detect and account for population structure in GWASs, including the identification of related individuals, multivariate approaches to detect ethnic outliers

and adjust for stratification, and the use of mixed models to allow for kinship. Next, we review GWAS methods for detecting complex trait associations with multiple SNPs through haplotype approaches, pairwise interaction testing and gene-based analyses. Although the focus of the chapter is the analysis of GWASs, the methods described for association testing are also applicable to whole-genome sequencing. Throughout the chapter, we highlight the latest methodological innovations and software implementations. Finally, we consider the prospects for GWASs to deliver on their promise of personalised medicine and development of novel therapeutic interventions for common disease.

21.2 GWAS Design Concepts

The initial focus of a GWAS should be on the study design, determining the specific trait for investigation, the ascertainment of samples, and the choice of SNPs to be interrogated for association. Without thoughtful consideration of these design issues, which we review here, the results of any downstream association analyses will be meaningless (Bush and Moore, 2012).

21.2.1 Phenotype Definition

The two most widely considered categories of traits in GWASs are: binary disease phenotypes, such as coronary heart disease, where individuals are classified as affected cases or unaffected controls; and quantitative (continuous) measurements, such as lipid profiles. Other categories of traits include categorical phenotypes, such as severity of disease, or time-to-event outcomes, which are most common in pharmacogenetic GWASs. The category of trait has important implications for downstream quality control, association analysis and interpretation of findings.

For disease outcomes, careful consideration of phenotype definition is essential because non-specific case-control definitions can increase heterogeneity in the underlying causal genetic polymorphisms (and non-genetic risk factors), leading to decreased power for detection (Zondervan *et al.*, 2007). Phenotype definition might reflect a trade-off of outcomes that are clinically and biologically relevant, and will often change over time. Increases in power to detect association can be achieved through selection of cases with a greater expected genetic load, for example those with an early age of onset or family history of disease, referred to as enrichment sampling (Antoniou and Easton, 2003). It is also essential to make an appropriate selection of controls, so that they are ascertained from the same population as cases, and are representative of the population who could have become cases according to the phenotype definition (Rothman and Greenland, 1998). Inappropriate selection of controls can lead to unmeasured confounding, which can increase false positive error rates and reduce power. Some disease outcomes can be defined on the basis of 'intermediate' quantitative traits, which are often preferred if they can be easily, accurately and cost-effectively measured with minimal error in large samples (Bush and Moore, 2012). Importantly, we might expect larger effects of causal variants on these intermediate traits than on the downstream manifestation of the disease, thereby increasing power.

21.2.2 Structure of Common Genetic Variation and Design of GWAS Genotyping Technology

Under the CDCV model, the success of GWASs depends, in part, on efficient study design that is informed by understanding the structure of common genetic variation throughout the genome in different populations. High-density genotype data from the International HapMap

Project (International HapMap Consortium, 2005, 2007) demonstrated that common SNPs are arranged in blocks of strong linkage disequilibrium (LD) within populations, maintained by low levels of recombination, but separated by hotspots of meiotic crossover activity. Within LD blocks, genotypes at pairs of SNPs are strongly correlated with each other, and genetic variation can be arranged on relatively small numbers of common haplotypes. Here, we provide relevant background to LD in the context of GWASs: for a more detailed description, see **Chapter 2**.

There are numerous measures of LD between a pair of SNPs (Devlin and Risch, 1995). We denote alleles at the first SNP by M_1 and m_1 and at the second SNP by M_2 and m_2 . Most measures of LD are based around the statistic $D = q_{12} - q_1 q_2$, where q_1 and q_2 denote the population frequencies of alleles m_1 and m_2 , respectively, and q_{12} denotes the population frequency of the haplotype $m_1 m_2$. Under linkage (or gametic phase) equilibrium, alleles are randomly assigned to haplotypes, so that $q_{12} = q_1 q_2$ and $D = 0$. More generally, D takes values in the range $[-1, 1]$, but is highly dependent on population allele frequencies. To reduce this dependence, two commonly used measures of LD have been proposed, given by

$$r^2 = \frac{D^2}{q_1(1-q_1)q_2(1-q_2)}$$

and

$$D' = \begin{cases} \frac{D}{\min[q_1 q_2, (1-q_1)(1-q_2)]} & \text{if } D < 0, \\ \frac{D}{\min[q_1(1-q_2), (1-q_1)q_2]} & \text{if } D \geq 0. \end{cases}$$

The statistic D' can take values in the range $[0, 1]$, where $D' = 1$ is consistent with no recombination between the pair of SNPs since the mutations that generated the polymorphisms. This is referred to as complete LD, and implies that at least one of the four possible haplotypes of alleles at the SNPs has zero frequency. The squared correlation coefficient, r^2 , between a pair of SNPs can also take values in the range $[0, 1]$. Perfect LD corresponds to $r^2 = 1$, for which genotypes at one SNP can serve as perfect proxies for genotypes at the second SNP. For $r^2 < 1$, genotypes at the two SNPs are imperfect proxies: the stronger the LD, the better the proxy.

Knowledge of the patterns of LD throughout the genome has helped in the design of efficient GWAS genotyping products, since SNPs can be selected that guarantee coverage of all common polymorphisms in the population at some predetermined threshold of r^2 . The advantage of this approach is that we need not genotype all common polymorphisms genome-wide. Instead, GWAS genotyping products focus on a smaller number of so-called ‘tag SNPs’, from which we can recover much of the information about common genetic variation across the genome (Carlson *et al.*, 2004). GWASs thus rely on identifying tag SNPs which have ‘indirect’ association with the trait under investigation, which might not themselves be causal, but which have genotypes that are proxies for those at the causal polymorphism, and are located within the same block of LD (Hirschhorn and Daly, 2005).

GWASs have been enabled by the availability of chip-based microarray technology for assaying hundreds of thousands or millions of SNPs genome-wide. The most widely used platforms include products from Illumina and Affymetrix. The most recent generation of products have been designed to provide comprehensive coverage of common variation across populations, by taking advantage of high-density genotype data from the International HapMap Project (International HapMap Consortium, 2005, 2007, 2010) and the 1000 Genomes Project

(1000 Genomes Project Consortium, 2012, 2015). Specialised arrays have been designed for GWASs in African populations (where the extent of LD is much less than in other ethnicities), and to provide improved coverage of lower frequency variation (particularly within exons). Whole-genome sequencing provides complete coverage of variation in a sample of individuals, irrespective of allele frequency, and is becoming increasingly financially feasible in large-scale population-based association studies. While such studies are not the focus of this chapter, the key design issues, quality control protocols and analytical methods are equally relevant to traditional GWAS genotyping array and whole-genome sequence data.

21.2.3 Sample Size Considerations

An essential component of GWAS design is the consideration of the sample size, which is a key determinant of the statistical power to detect association of the trait of interest with a causal variant. In addition to sample size, power depends on several factors, including: the level of significance, or false positive error rate; the effect size of the causal polymorphism; the frequency of the causal allele; and the LD between the causal polymorphism and a tested tag SNP (as measured by r^2), as determined by coverage of the GWAS array (Klein, 2007). In deriving the sample size for a well-powered study, it is important to remember that the effect size on a complex trait for any single causal polymorphism will be small, otherwise it would explain a large proportion of the genetic contribution and would resemble a Mendelian phenotype (Zondervan and Cardon, 2007).

21.2.4 Genome-Wide Significance and Correction for Multiple Testing

The primary GWAS analysis typically focuses on testing association of the trait under investigation with each SNP in turn. Statistical theory states that, for tests at a significance level α , we would expect $100\alpha\%$ of SNPs to yield a significant association by chance. It is crucial, therefore, to correct for multiple testing to maintain false positive error rates for the experiment overall. The simplest method to allow for multiple testing is to make use of a Bonferroni correction. Under this approach, each test is treated as independent, and the SNP-wise significance level is adjusted to achieve an overall experimentwise false positive error rate of $100\alpha\%$. Specifically, when testing N SNPs, we use a significance level of α/N at each SNP. The disadvantage of this approach is that each test is assumed to be independent, whereas for GWASs we expect that SNPs will be correlated owing to LD across the genome, and the Bonferroni correction will be conservative. Enhancements to this correction attempt to estimate the effective number of independent SNPs tested (Dudbridge and Gusnanto, 2008). However, a widely accepted genome-wide significance threshold of $p < 5 \times 10^{-8}$ has now been established, which corrects for 1 million blocks of LD across the genome, within which common SNPs are assumed to be strongly correlated (Pe'er *et al.*, 2008).

Alternative corrections for multiple testing have also been previously considered for GWASs that take account of the correlation of association between SNPs due to LD. The false discovery rate (FDR), for example, fixes the expected number of false positives among significant associations (Benjamini and Hochberg, 1995). Specifically, if we select an uncorrected SNP-wise significance level of α , the FDR is given by $N\alpha/k$, where k is the number of SNPs with $p < \alpha$. Using these relationships, we can define the appropriate SNP-wise significance threshold to obtain an overall FDR at an appropriate experimentwise error rate. Permutation procedures can be used to generate an empirical null distribution of association test statistics across the genome, while preserving the LD structure between SNPs. Empirical p -values for association for each SNP, correcting for multiple testing genome-wide, can then be calculated by comparing the

observed test statistic with the distribution of the maximum tests statistic across the genome from each permutation. However, permutation approaches are computationally demanding, and are generally considered impractical for large-scale GWASs.

21.2.5 Replication

Most GWASs have only marginal power to detect association with causal variants because of the small effect sizes for complex traits, stringent significance thresholds to correct for multiple testing, and the limited availability of samples and genotyping resources. As a result, many true positive signals of association may be difficult to distinguish from 'chance' findings in GWASs. For example, one of the first GWASs of Crohn's disease (Duerr *et al.*, 2006) identified several association signals that did not meet genome-wide significance, but which were later demonstrated to be genuinely associated with the disease (Cardon, 2006). These challenges highlight that GWASs are best considered as hypothesis-generating mechanisms, the results of which require validation through replication studies (discussed in detail in **Chapter 22**).

In theory, replication studies follow the same design principles as GWASs, and should match the conditions of the 'discovery' as closely as possible. SNPs identified in the initial GWAS can be genotyped in the replication study, with sample size calculations dependent on the observed effect sizes and allele frequencies, and appropriate correction for multiple testing. However, it is important to note that the selection of SNPs for validation that show the strongest signals of association will be subject to the 'winner's curse' (Lohmueller *et al.*, 2003), whereby their estimated effect sizes from GWASs will be overstated, and resulting in underpowered replication studies. It is important, for true validation of association signals, that the direction of effect on the trait under investigation is the same in both GWAS discovery and replication (discussed in detail in **Chapter 22**).

21.3 GWAS Quality Control

Data filtering to identify genotype errors is a critical aspect of GWAS analyses. No experimental process involving biological material is without inaccuracies, and, with large numbers of both SNPs and samples in a study, even modest error rates can be important, and have the potential to introduce systematic biases, increase false positive error rates and reduce power (Anderson *et al.*, 2010). Quality assurance steps in study design and data generation can help to reduce errors, for example by ensuring that high-quality DNA samples are collected, that reliable protocols for DNA extraction and preparation are used, and that samples are randomly assigned to genotyping plates and batches (Weale, 2010).

Irrespective of the GWAS microarray technology used, genotypes at a SNP are typically determined, or 'called', on the basis of probe intensities for the two possible alleles (Figure 21.1). If the technology works as expected, a scatter plot of intensities should reveal three clusters of samples that represent the three possible genotypes, with the heterozygous cluster positioned between the two homozygous clusters. For small-scale candidate gene studies, with relatively small numbers of SNPs and samples, genotypes could be called via manual inspection. However, with the advent of genome-wide genotyping technologies with hundreds of thousands of SNPs, automated genotype-calling algorithms have been developed to cluster genotypes. These algorithms are often specific to the technology used, but will typically provide some measure of confidence in the genotype call, and an overall assessment of the discriminatory success of the clustering process (often referred to as the 'cluster score'). A threshold value of the confidence

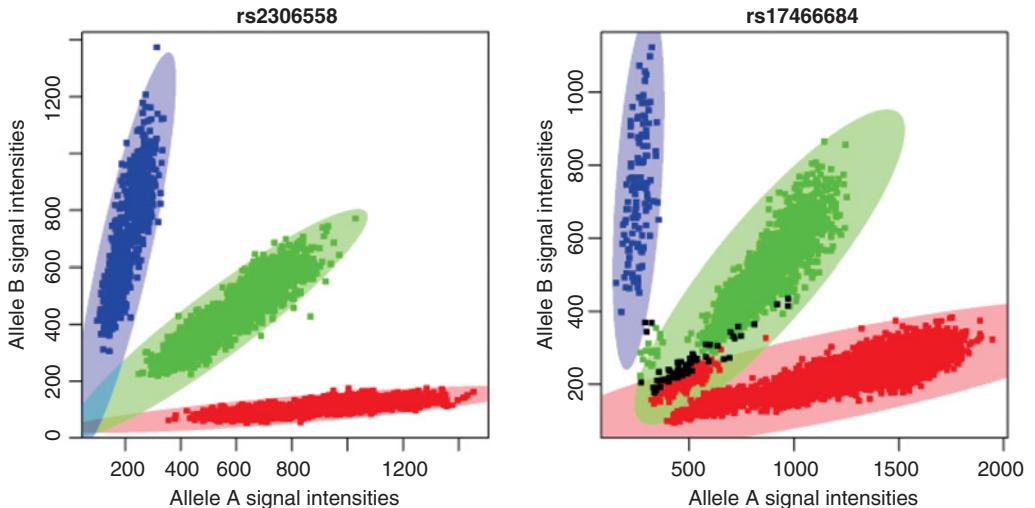


Figure 21.1 Examples of SNP cluster plots. Each point corresponds to the genotype of an individual at the SNPs, plotted according to their intensities of the two possible alleles. Homozygous genotypes, AA and BB , are coloured in red and blue, respectively, and heterozygous genotypes, AB , are coloured in green. The shaded ellipses correspond to genotype clusters as defined by the genotype calling algorithm. Missing genotypes are coloured in black. Reproduced with permission from BioMed Central from Schillert *et al.* (2009).

score can then be used to determine whether a genotype is called within a specific cluster, or is treated as missing (Figure 21.1). The choice of threshold for calling can have major implications for the quality of the genotype data. If the threshold is set too low, genotypes that do not clearly belong to one cluster are more likely to be miscalled. However, if the threshold is set too high, failure to call a genotype may result in ‘informative missingness.’ For example, heterozygotes or rare homozygotes often have lower confidence scores than common homozygotes, introducing bias in the genotyping calling (Hong *et al.*, 2010).

Automated GWAS quality control procedures focus on using called genotypes from the clustering process to identify the poorest-quality samples and SNPs that should be excluded from downstream association analyses, without ‘throwing out the baby with the bathwater’ by being over-stringent in the filtering process. Of course, the ultimate assessment of genotype quality is through manual inspection of signal intensity plots, which is essential for all SNPs identified in downstream association analyses. Quality control is usually performed for samples and SNPs, and the order in which these procedures are applied can have an impact on the genotypes retained for downstream association analyses. There is no ‘best practice’ for the order in which to apply quality control procedures. However, one approach is to first remove the worst-quality SNPs, then assess sample quality for the retained SNPs, and finally employ a more stringent investigation of SNPs in the retained samples.

21.3.1 SNP Quality Control Procedures

The exact criteria used to exclude low-quality SNPs will vary from one GWAS to another, but are typically based on call rate (i.e. the proportion of individuals for which a genotype has been called) and deviation from Hardy–Weinberg equilibrium (HWE) (Weale, 2010; Anderson *et al.*, 2010). The choice of thresholds for these filters is often based on excluding clearly outlying SNPs by plotting the distributions of these statistics genome-wide.

Low SNP call rate is a reflection of poor-quality clustering of genotypes, suggesting that any calls are more likely to be prone to error. Poor-quality clustering can also be reflected in the cluster score, which can also be used to exclude SNPs in addition to call rate. SNPs with call rate below 95% are very often excluded from downstream association analysis. Variable thresholds are often used by MAF, using more stringent thresholds for call rate for rarer SNPs. In GWASs of binary disease outcomes, call rates are also typically compared, for each SNP, between cases and controls (Clayton *et al.*, 2005; Moskvina *et al.*, 2006). This is particularly important if cases and controls have been genotyped in different batches, or from different sources of DNA, that might result in different success of the clustering according to disease status, leading to bias and increased false positive error rates if not excluded from downstream association analyses.

Extreme deviation from HWE is also widely accepted as an indication of poor-quality clustering of genotypes (Wittke-Thompson *et al.*, 2005). Extreme deviation from HWE can occur, for example, if heterozygous or rare homozygous genotypes have been less well called than the common homozygote. The extent of deviation from HWE is generally assessed by means of an exact test (Wigginton *et al.*, 2005). The definition of 'extreme' will depend on the number of SNPs tested, but typical thresholds exclude SNPs with exact *p*-value for deviation from HWE of the order of 10^{-6} . Deviation from HWE for SNPs mapping to the X chromosome should be assessed only in females. In GWASs of binary disease outcomes, deviation from HWE is often evaluated separately in cases and controls. SNPs can be excluded from downstream association analyses if they demonstrate extreme deviation in either group, which will be particularly relevant if cases and controls have been genotyped in different batches. Exclusions are sometimes based only on controls, because under some genetic models (e.g. recessive), deviations from HWE are expected in cases.

MAF (or minor allele count) is also sometimes used as a quality control filter for several reasons. First, SNP quality decreases with MAF because it is more difficult to distinguish the three genotypes. Second, we expect to have little power to detect association with lower-frequency SNPs because effect sizes on complex traits are modest. Third, the asymptotic properties of the statistical tests of SNP association with the trait under investigation are violated when the minor allele count is low, leading to increased false positive error rates. The appropriate threshold for MAF will depend on the sample size, and, for binary disease outcomes, could be applied separately in cases and controls.

21.3.2 Sample Quality Control Procedures

In the same way as for SNPs, the exact criteria used to exclude poor-quality samples will vary from one GWAS to another. However, widely used filters are based on call rate (i.e. the proportion of SNPs for which a genotype has been called), heterozygosity (i.e. the proportion of called autosomal genotypes that are heterozygous) and discordance between reported gender and genetic sex from X chromosome genotypes (Weale, 2010; Anderson *et al.*, 2010). Samples might also be excluded on the basis of relatedness or outlying ancestry, depending on the methods employed for testing association with the trait, as described in detail in Section 21.5. As for SNPs, there are not predetermined thresholds for these quality control measures, but the choice of sample exclusions will be determined by plotting distributions of the relevant statistics genome-wide.

Low sample call rate can be a reflection of poor-quality DNA (such as low concentration), suggesting that any genotypes are more likely to be incorrectly called. Samples with call rate below 95% are very often excluded from downstream association analyses. Excessive autosomal heterozygosity can indicate DNA contamination (i.e. from a mixture of two or more samples that

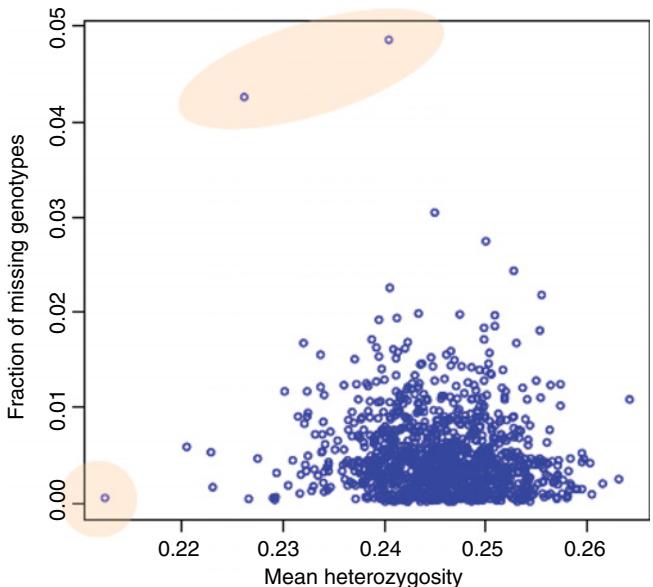


Figure 21.2 Sample quality control: call rate and heterozygosity. Each point corresponds to an individual, plotted according to their mean autosomal heterozygosity on the x-axis and proportion of missing genotype calls on the y-axis. The orange-shaded areas highlight samples with outlying call rate and heterozygosity that are excluded from downstream association analyses.

leads to an excess of heterozygous genotype calls). Wright's (1922) inbreeding coefficient, F , is often used as an alternative to heterozygosity because it is not dependent on allele frequencies: $F > 0$ indicates an excess of homozygous genotypes (low heterozygosity) and $F < 0$ indicates an excess of heterozygous genotypes (high heterozygosity). Plotting call rate against autosomal heterozygosity (or the inbreeding coefficient) is a useful approach to identify poor-quality samples on the basis of the two metrics (Figure 21.2).

Genotype data from the X chromosome are usually used to check for discordance of genetic sex with reported gender, which can be an indication of plating errors and/or sample mix-ups. Males only have one copy of the X chromosome, and thus cannot be heterozygous for SNPs that map outside of the pseudo-autosomal region, (i.e. heterozygosity of 0). Some genotype calling algorithms are informed by reported sex. If a reported male is actually female, any heterozygous genotypes on the X chromosome will be called as missing by these algorithms, resulting in a low sample call rate. Other calling algorithms are agnostic to reported sex, so misreported males will have an excess of heterozygous genotypes, and misreported females will have an excess of homozygous genotypes. Wright's inbreeding coefficient calculated from X chromosome data is also useful for identifying sex discordance: for males, we expect $F = 1$, while for females, we expect $F = 0$. Allowing for low rates of genotyping errors, discordant sex is generally considered as $F < 0.8$ in reported males and $F > 0.2$ in reported females. It could be argued that discordant sex is not important, unless the trait differs between males and females, and gender will be adjusted for (or stratified by) in association analyses. However, of more concern, discordant sex might indicate that the wrong DNA sample and record have been matched, meaning that other trait and covariate data might also be incorrectly assigned, and will impact on association results. Unless the sample can be correctly identified using other additional genotype data, or confirmation of misreported sex can be obtained, individuals with discordant sex should be excluded from any downstream investigations.

21.3.3 Software

PLINK (Purcell *et al.*, 2007) is the most widely used software application for GWAS quality control. PLINK provides all the SNP and sample quality metrics described above, although other software (such as R), is required to produce graphical summaries of these metrics. PLINK can also be used to automatically exclude SNPs on the basis of call rate, extreme deviation from HWE and MAF. Samples can also be automatically excluded on the basis of call rate and sex discordance. Other samples can be removed, for example based on outlying heterozygosity, through the provision of exclusion lists of individual identifiers. The latest version of the software, PLINKv2 (<https://www.cog-genomics.org/plink2>), offers substantial improvements in computational efficiency for large-scale GWASs. Other software applications for GWAS quality control include GenABEL (Karssen *et al.*, 2016) and QCTOOL (<http://www.well.ox.ac.uk/~gav/qctool/>).

21.4 Single SNP Association Analysis

The current standard practice to identify promising association signals for the trait under investigation that warrant further investigation involves testing each individual SNP, in turn, that is typed on the GWAS array and has passed quality control (Clarke *et al.*, 2011). For most traits, including binary disease outcomes and quantitative measures, statistical tests of association are developed within a generalised linear modelling framework.

21.4.1 Generalised Linear Modelling Framework

For the i th individual, we denote their phenotype by y_i , and their genotype at the j th SNP by $\mathbf{G}_{ij} = \{MM, Mm, mm\}$, where M and m denote the major and minor alleles, respectively. We can then represent the relationship between phenotype and genotype such that

$$g(E[y_i]) = \mu + \beta \mathbf{G}_{ij}. \quad (21.1)$$

In this expression, $g(\cdot)$ is the link function, μ is an intercept, and β is a vector of the effects of genotypes at the j th SNP on the phenotype. For GWASs of binary disease outcomes, the logit link function is typically used (resulting in a logistic regression model). For quantitative outcomes, the identity link function is most appropriate (resulting in a linear regression model). The effects of the SNP, β , measure the ‘strength’ of the association, and their interpretation will depend on the trait under investigation and the ‘coding’ of genotypes, which is described in detail below. However, in any case, we test the null hypothesis of no association of the trait of interest with the j th SNP by comparing the deviance of model (21.1) with $\beta = \mathbf{0}$ to that for which β is unconstrained. There are a range of alternative approaches to construct the test statistic, including the likelihood ratio test (or Wald test) and score test, and in the case of binary disease phenotypes, the bias-corrected Firth test. Score tests are computationally less demanding to calculate than others. For common SNPs the results of these tests will be strongly correlated, but for lower-frequency variants they have different advantages, including control of false positive error rates and increased power (Ma *et al.*, 2013).

21.4.2 Accounting for Confounding Factors as Covariates

A key advantage of the generalised linear model is that we can easily account for measured ‘confounding’ factors that might contribute to the association between SNPs and the trait under

investigation. A confounding factor is independently correlated both with SNP genotypes and with the trait under investigation, but is not on the causal path between them. For example, a SNP may demonstrate association with lung cancer. However, cigarette smoking is the most important risk factor for lung cancer, for which there is also a genetic predisposition. If the SNP is also associated with cigarette smoking, it is more likely that the effect of the SNP on lung cancer is not directly causal, from a biological perspective. Instead, the effect of the SNP on lung cancer is mediated through cigarette smoking, which would be considered as a confounder. The generalised linear model (21.1) can be extended by including the confounding factor as an additional covariate. This model assumes that the effect of the SNP is constant across the confounding factor, and estimated genotype effects are adjusted for the confounding factor. Returning to our example, if the association between the SNP and lung cancer is mediated through cigarette smoking, including this potential confounder as a covariate in the generalised linear regression model would eradicate the association signal.

The choice of covariates included in the regression model requires careful consideration. For example, the prevalence of many diseases differs between males and females, and sex is consequently often included as a covariate in GWAS analyses, despite the fact that we expect genetic effects to be homogenous. Intuitively, the inclusion of such ‘non-confounding’ covariates in the regression model will increase power to detect SNP associations by accounting for a proportion of the trait variability. However, while this is true for quantitative traits and common binary phenotypes, adjustment can reduce power to detect association with a rare disease, particularly as the variability explained by the covariate increases (Pirinen *et al.*, 2012).

21.4.3 Coding of SNP Genotypes

The simplest and most commonly applied coding of SNP genotypes assumes an ‘additive’ effect of alleles contributing to a genotype. Under this model, the effect of the heterozygote, Mm , is exactly intermediate between the common and rare homozygotes, MM and mm , respectively, so that genotypes are coded according to the number of minor alleles, denoted G_{ijA} (Table 21.1). The generalised linear model can then be expressed as

$$g(E[y_i]) = \mu + \beta_A G_{ijA}, \quad (21.2)$$

where β_A is the additive effect on the trait of alleles at the SNP. We test the null hypothesis of no association of the trait of interest with the j th SNP assuming additive allelic effects by comparing the deviance of model (21.2) with $\beta_A = 0$ to that for which β_A is unconstrained, with the resulting test statistic having an approximately chi-squared distribution with one degree of freedom. For binary disease outcomes, the score test derived from the logistic regression model is the Cochran–Armitage trend test (Armitage, 1955).

The maximum likelihood estimate, $\hat{\beta}_A$, represents the effect on the trait of the minor allele relative to the major allele at the SNP, adjusted for the effect of any confounding factors included as

Table 21.1 Coding of additive and dominance components of SNP genotypes

Genotype	Additive component G_{ijA}	Dominance component G_{ijD}
MM	0	0
Mm	1	1
mm	2	0

M and m denote the major and minor alleles at the SNP, respectively.

covariates in the regression model. The precise interpretation of the effect estimate will depend on the outcome under investigation and the study design. For quantitative measures, for which association is evaluated via a linear regression model, $\hat{\beta}_A$ can be interpreted as the mean change in trait value for each copy of the minor allele carried by an individual. For binary disease outcomes, for which association is evaluated via a logistic regression model, $\hat{\beta}_A$ can be interpreted as the log odds ratio of the minor allele relative to the major allele. The odds ratio is defined as the odds of the minor allele in cases versus controls relative to the odds of the major allele in cases versus controls. Note that an additive model on the log odds scale is equivalent to a multiplicative model on the odds ratio. For binary disease outcomes, a more natural assessment of the strength of an association is the relative risk, which measures the change in penetrance (i.e. probability that an individual is affected) for each copy of the minor allele, which cannot be obtained directly from case-control studies, for example. However, provided that the disease is not very common (<10% prevalence), the relative risk can be approximated by the odds ratio.

A more general coding of genotypes allows for deviations from the additive effects of alleles at the SNP by including a 'dominance' component, denoted G_{ijD} (Table 21.1). Under this parameterisation of SNP genotypes, often referred to as the 'general genotype' model, the generalised linear model (21.1) can be expressed as

$$g(E[y_i]) = \mu + \beta_A G_{ijA} + \beta_D G_{ijD}. \quad (21.3)$$

In this expression, β_A is the additive effect on the trait of alleles at the SNP, while β_D is the dominance (non-additive) effect. We test the null hypothesis of no association of the trait of interest with the j th SNP by comparing the deviance of model (21.3) with $\beta_A = \beta_D = 0$ to that for which the parameters are unconstrained, with the resulting test statistic having an approximately chi-squared distribution with two degrees of freedom. Furthermore, we can test the null hypothesis of no dominance at the j th SNP by comparing the deviance of model (21.2) with $\beta_D = 0$ to that for which β_D is unconstrained, with the resulting test statistic having an approximately chi-squared distribution with one degree of freedom.

Under this parameterisation of the general genotype model, we obtain maximum likelihood estimates of the additive and dominance effects, denoted by $\hat{\beta}_A$ and $\hat{\beta}_D$, respectively. We can then estimate the effect of the heterozygote, Mm , relative to the common homozygote, MM , by $\hat{\beta}_A + \hat{\beta}_D$, and the effect of the rare homozygote, mm , relative to the common homozygote by $2\hat{\beta}_A$. For quantitative outcomes, these effects are interpreted in terms of differences in mean trait values between genotypes, while for binary disease outcomes, they can be interpreted as log odds ratios.

Note that 'dominance' is a general term to describe a deviation from additivity, and should not be confused with a 'dominant' mode of inheritance, in which the effects on the trait under investigation are the same for the heterozygous and rare homozygous genotypes. If it is of interest to evaluate association with a SNP under a dominant mode of inheritance, we can consider a restricted parameterisation of the generalised linear model (21.3) for which $\beta_A = \beta_D$, with the resulting test statistic having an approximately chi-squared distribution with one degree of freedom. Similarly, for a recessive mode of inheritance, in which the effects on the trait under investigation are the same for the heterozygous and common homozygous genotypes, we can consider a restricted parameterisation of the generalised linear model (21.3) for which $\beta_A = -\beta_D$, with the resulting test statistic having an approximately chi-squared distribution with one degree of freedom.

The additive model has been demonstrated to be more powerful than the general genotype model, provided the effect of the heterozygous genotype is intermediate between that of the two homozygotes. Furthermore, dominance will be weakened at tag SNPs that are not in perfect LD with an untested causal variant, even if the true underlying association model is non-additive

at the causal variant (Spencer *et al.*, 2009). Consequently, a widely used approach is to use an additive model to identify associated SNPs, and then test for evidence of dominance at these selected SNPs. It is not recommended to consider multiple models (e.g. additive, dominant and recessive) genome-wide, since this will incur a penalty for multiple testing that will reduce power to detect association.

21.4.4 Imputed Genotypes

In the context of GWASs, imputation (Marchini and Howie, 2010) describes the process by which genotypes at SNPs not directly typed on a microarray, but present in a high-density reference panel, such as those from the International HapMap Consortium (International HapMap Consortium, 2007, 2010) or 1000 Genomes Project (1000 Genomes Project Consortium, 2012, 2015), are predicted for each individual in the study (see Chapter 3). For each sample at each SNP, imputation provides a probability distribution for possible genotypes, denoted ρ_{ij0} , ρ_{ij1} , and ρ_{ij2} , for the common homozygote, heterozygote and rare homozygote, respectively. The same generalised linear regression framework can be used to test for association of the trait of interest with imputed SNP genotypes. However, when testing for association of the trait of interest with an imputed SNP, it is essential that the uncertainty in the genotype is taken into account in the analysis. This can be achieved in a ‘missing data’ likelihood, averaged over the possible genotypes at the SNP, weighted by their probabilities from imputation. Alternatively, we can consider ‘expected’ genotypes across the imputation probability distribution, the additive and dominance components of which are given by $\hat{G}_{ijA} = \rho_{ij1} + 2\rho_{ij2}$ and $\hat{G}_{ijD} = \rho_{ij1}$. For further details, see Chapter 3.

21.4.5 Visualisation of Results of Single SNP GWAS Analyses

The most common tool for the visualisation of the results of GWAS analyses is the Manhattan plot (Figure 21.3(a)). Each point corresponds to a SNP, plotted according to genomic position on the x -axis (generally with separate chromosomes highlighted) and the evidence in favour of association (typically $-\log_{10} p$ -value) on the y -axis. The ‘skyscrapers’ in the Manhattan skyline represent clusters of SNPs in strong LD with each other that demonstrate strong evidence of association with the trait under investigation. Localised visualisation of the Manhattan plot (Figure 21.3(b)) is useful for investigating patterns of association signals in the context of local genetic variation, and can be generated by LocusZoom (Pruim *et al.*, 2010). Each SNP is coloured according to the magnitude of LD with the lead SNP, usually obtained from reference data such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). SNPs can also be plotted with different symbols to represent categories of annotation. The rate of recombination across the region, estimated from the International HapMap Consortium (2007), is plotted, together with local genes as defined by the UCSC genome browser (Kent *et al.*, 2002).

21.4.6 Interactions with Non-Genetic Risk Factors

For complex human disease, we expect there to be widespread interplay between genetic and non-genetic risk factors, such as environmental exposures. The generalised linear regression model can be easily extended to allow for interaction of SNP genotypes with a non-genetic risk factor, denoted x_i for the i th individual. For example, under the additive coding of SNP genotypes, the generalised linear regression model (21.2) extends to

$$g(E[y_i]) = \mu + \beta_A G_{ijA} + \gamma x_i + \theta_A G_{ijA} x_i. \quad (21.4)$$

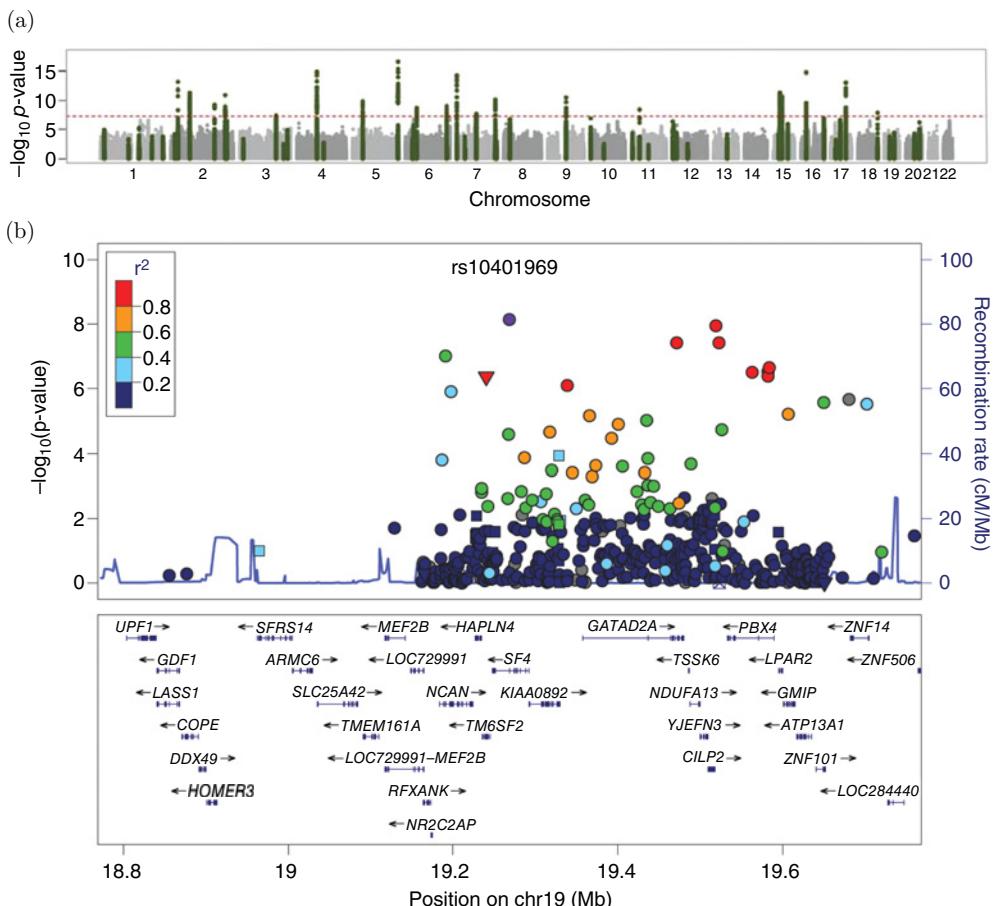


Figure 21.3 Examples of genome-wide and local visualisation of the results of GWAS analyses. (a) Manhattan plot. Each point corresponds to a SNP, plotted according to genomic position on the x-axis and the evidence in favour of association ($-\log_{10} p\text{-value}$) on the y-axis. SNPs highlighted in green map to loci previously reported for the trait. (b) Signal plot. Each point represents a SNP, plotted with its $p\text{-value}$ (on a $-\log_{10}$ scale) as a function of genomic position. The index SNP is represented by the purple symbol. The colour coding of all other variants indicates LD with the index variant in European ancestry haplotypes from the 1000 Genomes Project reference panel: red, $r^2 \geq 0.8$; gold, $0.6 \leq r^2 < 0.8$; green, $0.4 \leq r^2 < 0.6$; cyan, $0.2 \leq r^2 < 0.4$; blue, $r^2 < 0.2$; grey, r^2 unknown. The shape of the symbol corresponds to the annotation of the variant: upward triangle for frameshift, stop or splice; downward triangle for non-synonymous; square for synonymous or untranslated region (UTR); and circle for intronic or non-coding. Recombination rates are estimated from Phase II HapMap and gene annotations are taken from the UCSC genome browser.

In this expression, γ denotes the main effect of the non-genetic risk factor, while θ_A denotes the additive interaction with the SNP. We test the null hypothesis of no association of the trait of interest with the j th SNP, allowing for an interaction with the non-genetic risk factor, by comparing the deviance of model (21.4) with $\beta_A = \theta_A = 0$ to that for which the parameters are unconstrained, with the resulting test statistic having an approximately chi-squared distribution with two degrees of freedom. Alternatively, we can test the null hypothesis that the effect of the j th SNP on the trait of interest is the same across exposures to the non-genetic risk factor (i.e. no interaction), by comparing the deviance of model (21.4) with $\theta_A = 0$ to that for which the parameters are unconstrained, with the resulting test statistic having an approximately chi-squared distribution with one degree of freedom.

21.4.7 Bayesian Methods

The most widely implemented methods for testing association of SNPs with a trait under investigation have generally been derived in a frequentist statistical paradigm, enabling computation of a *p*-value in favour of the null hypothesis. However, a wide range of Bayesian methods have also been proposed which, at the cost of some modelling assumptions, have been suggested to overcome the limitation that a *p*-value alone is insufficient to quantify the confidence we should have that a SNP is truly associated with the trait under investigation (Stephens and Balding, 2009). Bayesian methods require specification of a prior model for allelic or genotypic effect sizes, and a prior probability that each SNP is associated with the trait under investigation, which has the advantage that it can take account of annotation, for example, to upweight signals mapping to functionally important regions of the genome (including exons and promoter/enhancer elements) (Sveinbjornsson *et al.*, 2016). These approaches also have the advantage that they can average over genetic models. Nevertheless, Bayesian methods have often been overlooked because of concerns over computational costs and a lack of guidance on thresholds for declaring association.

21.4.8 Software

There is a wide range of software developed for single-variant association analyses that can deal with the scale and complexity of GWAS data. PLINK (Purcell *et al.*, 2007) (and PLINKv2) offers flexible options for linear and logistic regression modelling of quantitative and binary disease outcomes, considering additive and dominance effects, and allowing for covariate adjustment and interaction with non-genetic risk factors. Other software for single-variant analysis of quantitative and binary disease outcomes includes SNPTEST (Marchini and Howie, 2010) and GenABEL (Karssen *et al.*, 2016). For time to-event outcomes, SurvivalGWAS_SV can be employed to test for SNP association under Cox proportional hazards or Weibull models (Syed *et al.*, 2017). Software implementing Bayesian approaches to detecting SNP associations include SNPTEST (Marchini and Howie, 2010) and BIMBAM (Servin and Stephens, 2007).

21.5 Detecting and Accounting for Genetic Structure in GWASs

Genetic structure in GWASs arises from population stratification and/or relatedness between samples. Such unmeasured confounding, if not accounted for in the analysis, can increase false positive error rates and lead to spurious association signals (Freedman *et al.*, 2004; Marchini *et al.*, 2004). For example, consider a case–control GWAS in which samples are ascertained at random from a population consisting of two underlying strata. If the disease is more prevalent in the first stratum, cases will more often be selected from this stratum than the other. As a result, any SNP that differs in allele/genotype frequency between the strata will appear to be associated with disease, even if there is no association within each stratum. In this simple example, one obvious solution to the problem is to match cases and controls by stratum. However, in most populations, strata are not clearly defined because of increasing levels of migration and admixture, and defining matched samples is difficult for such ‘fine-scale’ genetic structure.

Family-based association studies provide a design-based solution to the problem of structure. The simplest designs ascertain trios that consist of a proband and their two parents, and record the alleles transmitted to the child. The non-transmitted alleles can then be considered as an ‘internal’ control for the proband, and will be perfectly matched in terms of structure because they are derived from the same parents. However, family-based studies are expensive,

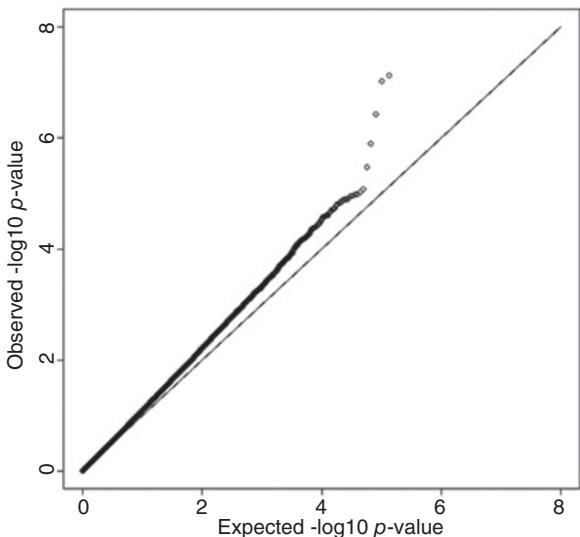


Figure 21.4 Example of GWAS quantile–quantile plot. Each point corresponds to a SNP, plotted according to the ranked $-\log_{10} p$ -value for association on the y-axis against the expected ranked $-\log_{10} p$ -value under the null hypothesis of no association on the x-axis. Inflation of $-\log_{10} p$ -values above the $y = x$ line is indicative of population structure that has not been accounted for in the association analysis.

since two parents are genotyped to obtain the internal control. Furthermore, the design may be impractical, for example for late-age-onset disease, where parental DNA might not be available.

One of the simplest approaches to assess the evidence for false positive associations due to genetic structure in GWASs is by means of genomic control (Devlin and Roeder, 1999). Under the null hypothesis of no association of the trait of interest with SNPs, genome-wide, test statistics under an additive coding of genotypes, for example, have a chi-squared distribution with one degree of freedom. In the presence of genetic structure, not accounted for in the analysis, we would expect many more signals of association, throughout the genome, than we would expect by chance. The distribution of observed test statistics, commonly summarised by means of a quantile–quantile plot (Figure 21.4) would therefore be inflated over that expected under the null hypothesis. The genomic control inflation factor, λ_{GC} , can be estimated by comparing the median of observed test statistics, genome-wide, with that of the null distribution. An inflation factor $\lambda_{GC} > 1$ is indicative of unmeasured confounding due to genetic structure, which will require additional investigation before interpretation of the GWAS findings. A simple correction for structure is to adjust test statistics by dividing by the inflation factor. However, such an approach assumes that all SNPs, genome-wide, are equally confounded by structure, which is unlikely to be true, and thus can result in a loss in power to detect association.

21.5.1 Identification of Related Individuals

One potential source of unmeasured confounding due to genetic structure is through the inclusion of related (or duplicate) samples in the GWAS (Anderson *et al.*, 2010). Relatedness between pairs of samples, typically represented by means of a genetic relationship matrix (GRM), can be assessed by the ‘identity by state’ (IBS) metric, defined as the proportion of the genome at which they share the same alleles. The IBS between a pair of individuals, i and k , over N SNPs is given by

$$d_{ik} = \frac{1}{2N} \sum_{j=1}^N |G_{ijA} - G_{kjA}|.$$

IBS is typically calculated over the autosomes, focusing on common SNPs ($\text{MAF} > 5\%$) that are pruned for LD. IBS has the disadvantage that it is dependent on allele frequency. However, we can also calculate an alternative metric of relatedness between pairs of individuals that is independent of allele frequency, referred to as ‘pi-hat’, and given by $\hat{\pi}_{ik} = 2z_{ik2} + z_{ik1}$, where z_{ikl} denotes the proportion of SNPs, genome-wide, at which they share l alleles ‘identical by descent’ (IBD). IBD represents the proportion of the genome of a pair of individuals that descend from a common ancestor, which can be estimated from IBS and is independent of allele frequency (Purcell *et al.*, 2007). Larger values of pi-hat indicate a greater extent of relatedness between pairs of individuals. Allowing for low rates of genotyping error, $\hat{\pi}_{ik} \approx 1$ for duplicate samples and monozygotic twins; $\hat{\pi}_{ik} \approx 0.5$ for first-degree relatives, including full siblings and parent–child pairs; and $\hat{\pi}_{ik} \approx 0.25$ for second-degree relatives, including half-siblings, grandparent–grandchild pairs and avuncular pairs.

To account for confounding due to relatedness among study samples, it is common to identify pairs (or larger sets) of individuals that exceed a predetermined threshold of pi-hat, typically 0.125 to allow consideration of third-degree relatives. Only one sample from each related pair (or set) of individuals is retained for the downstream association analysis. This sample will usually have the highest call rate. Alternatively, for binary disease outcomes, cases may be preferentially retained if the case–control ratio is low.

21.5.2 Multivariate Approaches to Identify Ethnic Outliers and Account for Population Stratification

A second potential source of unmeasured confounding due to genetic structure is through the inclusion of samples of outlying ethnicity or because of population stratification. Multivariate statistical techniques, such as principal components analysis (PCA) and multidimensional scaling (MDS), have been widely used in population genetics to enable visualisation of genome-wide genotype differences between samples in few dimensions. These techniques can be applied via eigenvalue decomposition of a data correlation or covariance matrix, such as a GRM, to derive principal components onto which samples are projected (Price *et al.*, 2010).

PCA and MDS can be used to identify samples of outlying ethnicity in GWASs by aggregating observed genotype data, genome-wide, with high-density reference genotype data from diverse populations available from the International HapMap Project (International HapMap Consortium, 2005, 2007, 2010) or the 1000 Genomes Project (1000 Genomes Project Consortium, 2012, 2015). The first two principal components from PCA/MDS of these combined data will discriminate samples of European, African and East Asian ancestry (Paschou *et al.*, 2007) (Figure 21.5). GWAS samples would be expected to cluster closely with the reference populations from which they have been ascertained, and can thus be used to highlight ethnic outliers that should be removed from downstream association analyses.

The same techniques can be used to define principal components that describe population stratification after the removal of ethnic outliers (Price *et al.*, 2006) (Figure 21.6). For example, in a landmark investigation of genetic diversity across Europe, PCA of 1387 samples typed at ~200,000 SNPs, genome-wide, generated two principal components that clearly reflected North–South and East–West allele frequency clines (Novembre *et al.*, 2008). Principal components can thus be used to adjust for unmeasured confounding due to population stratification by their inclusion as covariates in the generalised linear regression model (21.1). Including larger numbers of principal components will account for genetic structure due to population stratification at a finer scale (Jallow *et al.*, 2009).

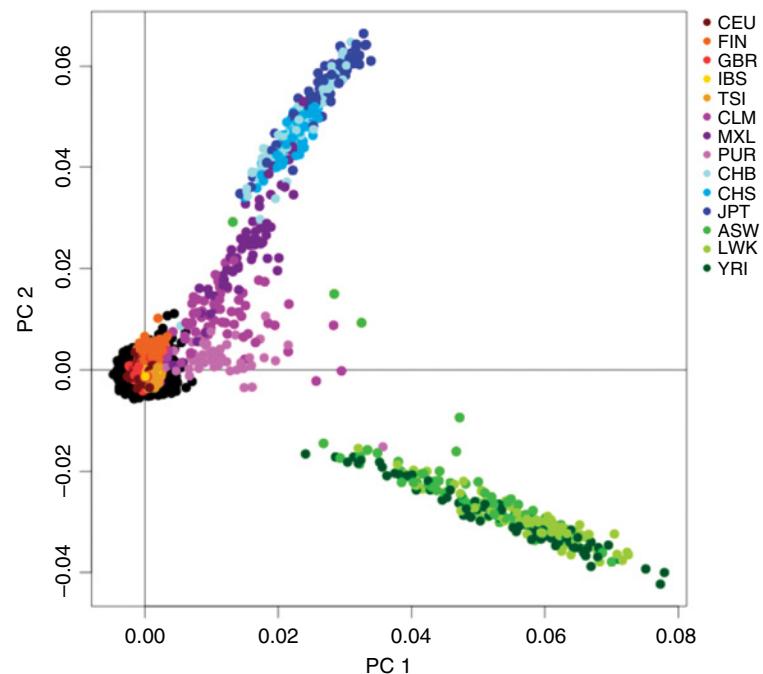


Figure 21.5 Making use of the 1000 Genomes Project reference panel to identify samples of outlying ancestry. Each point corresponds to a sample, plotted for the first two principal components from multidimensional scaling of the genetic relatedness matrix. GWAS samples are coloured in black. Samples from the 1000 Genomes Project are coloured according to the population from which they are ascertained: ASW, African ancestry in Southwest USA; LWK, Luhya in Webuye, Kenya; YRI, Yoruba in Ibadan, Nigeria; CLM, Colombian in Medellín, Colombia; MXL, Mexican ancestry in Los Angeles, California; PUR, Puerto Rican in Puerto Rico; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese in China; JPT, Japanese in Tokyo, Japan; CEU, Northern/Western European ancestry in Utah; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian populations in Spain; TSI, Toscani in Italy. In this example, GWAS samples have been ascertained from a European population, so would be expected to cluster closely with European ancestry 1000 Genomes Project reference populations (CEU, FIN, GBR, IBS and TSI). GWAS samples with outlying ethnicity would be expected to cluster more closely with other ancestry groups represented in the 1000 Genomes Project reference panel, and should be excluded from downstream association analyses.

21.5.3 Mixed Modelling Approaches to Account for Genetic Structure

An alternative approach to accounting for structure is to directly model the genetic correlation between samples, for example as measured by the GRM (Kang *et al.*, 2010; Zhang *et al.*, 2010; Lippert *et al.*, 2011; Listgarten *et al.*, 2012; Zhou and Stephens, 2012). A generalised linear mixed model includes a random effect for relatedness in addition to a fixed effect for the genotype at the j th SNP. The linear component of this model is given by

$$y_i = \mu + \beta \mathbf{G}_{ij} + \mathbf{u}_i + \varepsilon_i \quad (21.5)$$

In this expression, \mathbf{u} is a vector of random effects, defined by $\mathbf{u} \sim \text{MVN}(0, \lambda_U \mathbf{D})$, for the variance component λ_U and GRM \mathbf{D} , and ε is a vector of residual errors, defined by $\varepsilon \sim \text{MVN}(0, \lambda_E \mathbf{I})$. We test the null hypothesis of no association of the trait of interest with the j th SNP by comparing the deviance of model (21.5) with $\beta = \mathbf{0}$ to that for which β is unconstrained.

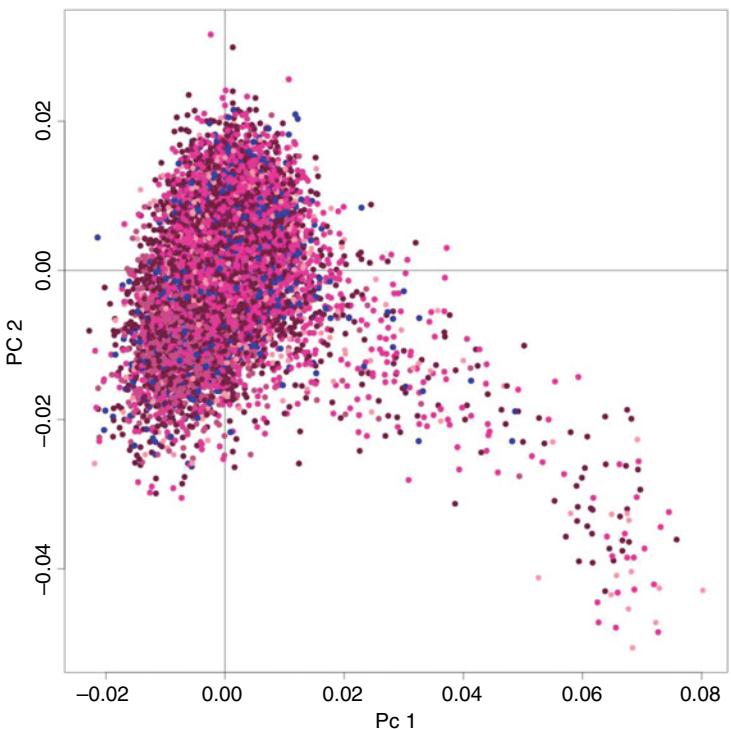


Figure 21.6 Population structure defined by two principal components. Each point corresponds to a GWAS sample, plotted for the first two principal components from multidimensional scaling of the genetic relatedness matrix. Cases are coloured in blue, while controls from four different sources are plotted in shades of pink/purple. Including the two principal components as covariates in the regression analysis can be used to account for the population structure.

For binary disease outcomes, the most appropriate model to assess association with SNPs is via a logistic mixed model (Chen *et al.*, 2016), the fitting of which can be computationally demanding in the context of GWASs. Alternatively, it has been demonstrated that allelic/genotype effects from the less computationally expensive linear mixed model, appropriate for quantitative outcomes, can be transformed onto the log odds scale through a Taylor's expansion (Pirinen *et al.*, 2013; Cook *et al.*, 2017).

21.5.4 Software

The identification of related individuals can be performed using PLINK (Purcell *et al.*, 2007), which constructs a GRM (based on LD-pruned SNPs) and calculates pi-hat between each pair of samples. PRIMUS can also be used to identify related individuals in GWASs, and has the advantage that it can select retained samples using information on call rates and phenotype (Staples *et al.*, 2013). The programs relateAdmix (Moltke and Albrechtse, 2014) and PC-Relate (Conomos *et al.*, 2016) offer powerful implementations to detect relatedness in the presence of population structure and admixture.

PLINK (and PLINKv2) can also be used to perform PCA/MDS applied to the GRM to identify ethnic outliers. PLINK can seamlessly include principal components derived from MDS of the GRM as covariates to account for population structure in linear and logistic regression models to test for association of SNPs with quantitative and binary disease outcomes.

Alternatively, SMARTPCA (Price *et al.*, 2006) has been designed to derive principal components from genome-wide genotype data and has the following advantages: utilisation of a model to account for LD between SNPs; automatic exclusion of samples based on distance from the GWAS cluster along principal components; and implementation of a test of association of the trait under investigation with each principal component to determine those that are potentially confounded.

There are several alternative software tools that implement linear mixed models in the context of GWASs, including EMMAX (Kang *et al.*, 2010), GEMMA (Zhou and Stephens, 2012) and BOLT-LMM (Loh *et al.*, 2015). These tools can be used to derive the GRM and test for association of SNPs, genome-wide, accounting for the genetic correlation between samples. For binary disease outcomes, GMMAT implements a logistic mixed model for GWAS analysis (Chen *et al.*, 2016).

21.6 Multiple SNP Association Analysis

One of the most attractive features of SNPs for identifying genetic variation contributing to complex traits is their abundance throughout the genome. However, each SNP will contribute relatively little information about disease association since effect sizes are expected to be small and might only be in LD with a causal variant. Joint analyses of multiple SNPs simultaneously have thus been proposed as an approach to increase power to detect association with complex human traits by: modelling the LD between variation on a local scale through haplotype-based methods; allowing for epistasis between causal polymorphisms in SNP-SNP interaction models; or aggregating the joint effects of SNPs within the same ‘functional unit’ via gene-based tests.

21.6.1 Haplotype-Based Analyses

Much of common genetic variation can be structured into haplotypes within blocks of strong LD, which are rarely disturbed by recombination, and transmitted intact from one generation to the next. Haplotypes are attractive because the functional properties of a protein are determined by the linear sequence of amino acids, corresponding to DNA variation on a chromosome (Clark, 2004). For example, there is evidence that a combination of causal variants, *in cis*, at the *HPC2-ELAC2* locus increases the risk of prostate cancer (Tavtigian *et al.*, 2001). A lower-frequency causal allele might also reside on a specific haplotype background that might not otherwise be identified through single SNP analysis methodologies. Haplotype-based analyses are appropriate in small genomic regions, such as blocks of LD, that will have been subject to limited ancestral recombination. Haplotype analyses may provide additional information with higher-density genotyping in a follow-up study of associated regions from an initial GWAS.

The generalised linear regression model (21.1) can easily be extended to the haplotype paradigm, such that

$$g(E[y_i]) = \mu + \beta \mathbf{H}_i. \quad (21.6)$$

In this expression, \mathbf{H}_i denotes the pair of haplotypes (i.e. the diplotype) for the i th individual across multiple SNPs in a block of LD, and β is a vector of the effects of haplotypes on the phenotype. It is common to assume an additive model of haplotype effects: each diplotype is then coded according to the number of copies of each possible haplotype carried by an individual. We test the null hypothesis of no association of the trait of interest with haplotypes

in the block by comparing the deviance of model (21.6) with $\beta = \mathbf{0}$ to that for which β is unconstrained. Under an additive model, the resulting test statistic has an approximately chi-squared distribution with $h - 1$ degrees of freedom, where h denotes the number of distinct haplotypes observed in the sample.

One of the obvious challenges of haplotype-based analyses is that we do not observe the diplotype, \mathbf{H}_i , of an individual directly from the unphased SNP genotype data generated by GWAS arrays. The best solution to the problem is to use statistical phasing methods (such as those reviewed in **Chapter 3**) to reconstruct the distribution of possible haplotypes carried by each individual, given their observed unphased genotypes (Stephens *et al.*, 2001; Stephens and Donnelly, 2003; Excoffier and Slatkin, 1995; Delaneau *et al.*, 2012; O'Connell *et al.*, 2014). Uncertainty in phasing can then be taken into account in a 'missing data' likelihood, averaged over the possible haplotype pairs for each individual, weighted by their probabilities from the reconstruction (Schaid *et al.*, 2002).

One potential problem with haplotype-based analyses is a lack of parsimony, since many haplotypes may be consistent with the observed unphased genotype data, particularly as the number of SNPs increases. In the generalised linear regression model, a parameter is required for each haplotype, leading to a test for association with the trait of interest with many degrees of freedom. The effects of rare haplotypes will be difficult to estimate, and there will be a lack of power to detect association if only one or two haplotypes impact on the trait. We could resolve this problem by combining rare haplotypes (say, with frequency less than 1% in the population) into a 'pooled' category, each assigned the same effect in the generalised linear regression model, and therefore requiring only a single degree of freedom. However, a more satisfactory approach to reduce dimensionality is to take advantage of the expectation that similar SNP haplotypes in a small genomic region tend to share recent common ancestry, and hence are more likely to share the same alleles at the underlying causal variant(s). Under this assertion, more similar haplotypes are likely to have more comparable effects on the trait under investigation, and thus could naturally be combined in the analysis. Several statistical approaches have been developed that group haplotypes according to some similarity metric and then ascribe the same effect to all those assigned to the same cluster, thereby reducing the number of required parameters in the generalised logistic regression model (Molitor *et al.*, 2003a,b; Durrant *et al.*, 2004; Morris, 2005, 2006). Haplotype clustering can be thought of as an approximation to the more complex population genetics processes underlying their evolution, providing computationally efficient approaches that can be applied to large numbers of SNPs, while maintaining the most relevant features of the shared ancestry of the chromosomes in a sample of individuals.

21.6.2 SNP–SNP Interaction Analyses

The traditional definition of epistasis is the masking, or modification, of the effects of genotypes at one variant by genotypes at a second. The existence of epistasis underlying association with complex human traits is not surprising, since we expect the underlying biological mechanisms to be extremely intricate, incorporating the joint effects of multiple genetic risk factors. Furthermore, there is increasing evidence from model organisms, including *Drosophila melanogaster* and *Saccharomyces cerevisiae*, that epistasis occurs frequently, involves multiple genetic polymorphisms, and may contribute large effects to the genetic component of phenotypic variation (Mackay, 2001; Brem and Kruglyak, 2005; Brem *et al.*, 2005; Storey *et al.*, 2005).

From a statistical perspective, epistasis can be represented by an interaction between genotypes at two more SNPs. Thus, the generalised linear regression model (21.1) for single SNP association analysis naturally extends to account for interactions (Cordell, 2002). For example,

to model the interaction between two SNPs, allowing for deviations from additivity within and between genotypes, it follows that

$$\begin{aligned} g(E[y_i]) = & \mu + \beta_{1A}G_{i1A} + \beta_{1D}G_{i1D} + \beta_{2A}G_{i2A} + \beta_{2D}G_{i2D} + \beta_{12AA}G_{i1A}G_{i2A} \\ & + \beta_{12AD}G_{i1A}G_{i2D} + \beta_{12DA}G_{i1D}G_{i2A} + \beta_{12DD}G_{i1D}G_{i2D}. \end{aligned} \quad (21.7)$$

In this expression, G_{i1A} and G_{i2A} denote the additive coding of genotypes at the two SNPs, respectively, and G_{i1D} and G_{i2D} denote the dominance coding, respectively. The parameters β_{jA} and β_{jD} correspond to the additive and dominance *main effects*, respectively, of the j th SNP. The four interaction terms, β_{12AA} , β_{12AD} , β_{12DA} and β_{12DD} , correspond to additive–additive, additive–dominance, dominance–additive and dominance–dominance components to epistasis between the two SNPs. We test the null hypothesis of no association of the trait under investigation with the SNPs, allowing for interaction between them, by comparing the deviance of model (21.7) with $\beta = \mathbf{0}$ to that for which β is unconstrained, with the resulting statistic having an approximately chi-squared distribution with eight degrees of freedom. We can also test the null hypothesis of no interaction between genotypes at the two SNPs by comparing the deviance of model (21.7) with $\beta_{12AA} = \beta_{12AD} = \beta_{12DA} = \beta_{12DD} = 0$ to that for which β is unconstrained, with the resulting statistic having an approximately chi-squared distribution with four degrees of freedom. The interaction model can also be simplified to consider only additive main effects and additive–additive epistasis by adding the constraint that $\beta_{1D} = \beta_{2D} = \beta_{12AD} = \beta_{12DA} = \beta_{12DD} = 0$, requiring fewer degrees of freedom in the tests of association and interaction. With more than two SNP, the generalised linear regression model can also be extended to incorporate higher-order interaction terms, although these effects are often difficult to estimate without large samples sizes, and are difficult to interpret.

In the presence of epistasis between SNPs, we would expect modelling interaction effects to lead to increased power over tests that only consider main effects. However, researchers have often been reluctant to consider epistasis in GWASs because: the less parsimonious model will lack power unless interaction effects are large; stringent significance thresholds are required to account for the number of tests in pairwise genome-wide scans; and the complexity of the model and numbers of possible SNP pairs adds to the computational burden of the analysis. However, allowing for additive main effects and additive–additive epistasis for all pairs of SNPs, genome-wide, has been demonstrated to increase power to detect association over single SNP analyses for a range of interaction models, despite the additional burden of multiple testing (Marchini *et al.*, 2005). Two-stage interaction analysis strategies have also been proposed that suffer only a negligible loss in power, while minimising the computational burden (Bhattacharya *et al.*, 2011). Other computational advances make use of data compression and parallelisation techniques (Steffens *et al.*, 2010), graphical processing units (Kam-Thong *et al.*, 2012) or ‘approximately’ exhaustive searches (Prabhu and Pe’er, 2012).

Data mining approaches offer alternative computationally efficient algorithms to search through a space of association models that involve multi-SNP interactions. These approaches typically use cross-validation to avoid over-fitting, and permutation to assess significance. Generalised linear regression will then be used to refit the final model to obtain main effect and interaction estimates. Examples of model search based approaches include multifactor dimensionality reduction (Ritchie *et al.*, 2001) and entropy-based methods (Lee *et al.*, 2016). Bayesian model selection techniques can also be used to investigate the evidence for interaction between genotypes at SNPs in GWASs. The methods typically make use of Markov chain Monte Carlo techniques to search through the space of possible models, given a pre-specified prior model for the number of loci and their effect sizes (Zhang and Liu, 2007; Jiang and Neapolitan, 2015).

21.6.3 Gene-Based Analyses

For lower-frequency and rare variants, typically defined to have MAF less than 5%, single SNP analyses lack power to detect association with the trait under investigation unless effect sizes are large. As a consequence, there has been an exciting period of methodological development focused on the analysis of rare variants within the same ‘functional unit’ (such as an exon, gene, or pathway), increasing power to detect association over single SNP approaches by considering their joint effects in gene-based tests (Moutsianas and Morris, 2014). Most of these methods are developed in a generalised linear regression framework, which enables incorporation of covariates to allow for adjustment for confounders, including non-genetic risk factors and indicators of population structure, and can be extended to a mixed model to include a random effect to account for relatedness.

Most methods for gene-based analyses can be classified as: burden tests, which assume the same direction of effect on the trait of all SNPs in the functional unit (Morris and Zeggini, 2010); dispersion tests, such as the sequential kernel association test (SKAT) (Wu *et al.*, 2011), which allow for deviations from this unidirectional assumption; or a combination of the two approaches, such as SKAT-O (Lee *et al.*, 2012). SNPs can be weighted according to annotation (i.e. how deleterious the mutation is), or according to allele frequency, such as the Madsen–Browning scheme (Madsen and Browning, 2009) that assigns greater weight to SNPs of lower frequency. The relative power of these gene-based methods depends on the alignment of the underlying genetic architecture of the trait under investigation with the modelling assumptions, such as robustness to neutral variants and an assumption of all causal alleles having the same magnitude and/or direction of effect. Empirical investigations and detailed simulation studies have highlighted that there is no uniformly most powerful gene-based test, but that methods that combine burden and dispersion tests, such as SKAT-O, are least sensitive to the underlying genetic architecture (Ladouceur *et al.*, 2012; Moutsianas *et al.*, 2015).

The gold standard technology for accessing low frequency and rare SNPs is through whole-genome (or whole-exome) sequencing, which is still expensive in comparison to GWAS genotyping arrays. However, simulations have demonstrated that imputation of a GWAS scaffold up to a large, population-matched reference panel results in relatively little loss of power compared to sequencing (Mägi *et al.*, 2012). SNPs not present in the reference panel (such as private mutations or very rare variants) are less likely to have a major impact on the trait under investigation, and thus would not be expected to lead to a dramatic reduction in power. Larger reference panels provide more comprehensive coverage of genetic variation in a functional unit, and higher-quality imputation, allowing recovery of genotypes at SNPs with MAF as low as 0.3% (Zheng *et al.*, 2012). However, imputation of variation of lower MAF remains a considerable challenge, and thus could never replace sequencing, although it currently provides a financially feasible, complementary strategy to detecting gene-based association signals with complex traits.

21.6.4 Software

There are a variety of software tools for haplotype reconstruction in GWASs, including PLINK (Purcell *et al.*, 2007) (and PLINKv2), UNPHASED (Dudbridge, 2008), PHASE (Stephens *et al.*, 2001; Stephens and Donnelly, 2003), and SHAPEIT (Delaneau *et al.*, 2012; O’Connell *et al.*, 2014). PLINK can also be used to assess for evidence of association of haplotypes with binary disease outcomes or quantitative traits in a generalised linear regression model. A range of methods for haplotype clustering have been developed, including GENECLUSTER (Su *et al.*, 2009) and GENEBPM (Morris, 2005, 2006). GENECLUSTER assesses the evidence that

haplotype clusters represent multiple underlying causal polymorphisms. GENEBPM clusters haplotypes according to a Bayesian partition model, and has been extended to allow for deviations from additivity. There are a wide range of software tools that enable testing for association allowing for epistasis, including INTERSNP (Herold *et al.*, 2009), SIXPAC (Prabhu and Peer, 2012), MECPM (Miller *et al.*, 2009), BEAM (Zhang and Liu, 2007; Zhang *et al.*, 2011; Zhang, 2012) and LEAP (Jiang and Neapolitan, 2015). The most popular software implementations for gene-based association testing enable analyses with burden and dispersion methods, and include RVTESTS (Zhan *et al.*, 2016), EPACTS (<https://genome.sph.umich.edu/wiki/EPACTS>) and RAREMETAL (Feng *et al.*, 2014).

21.7 Discussion

Over the last ten years, population-based GWASs have become the most popular design to investigate the genetic contribution to complex human traits and diseases. GWASs have successfully identified thousands of regions of the genome that have been robustly associated with a wide range of complex traits, including common diseases such as type 2 diabetes, cardiovascular disease and various forms of cancers (MacArthur *et al.*, 2017).

The success of GWASs can be attributed to several important factors. First, empirical evidence of polymorphisms associated with complex traits suggests that much of the genetic contribution adheres to the CDCV model (Schork *et al.*, 2009). Improved understanding of the structure of genetic variation across the genome in different ethnic groups has enabled the design of highly efficient genotyping arrays that can capture the majority of common polymorphisms in diverse populations. Further refinements in array technologies have dramatically reduced genotyping costs, allowing investigation of the genetic contribution to complex traits in the large sample sizes needed to detect modest allelic effects. Second, increased power to detect association signals, and reduced false positive error rates, have been achieved through appreciation of all aspects of study design, including the need for a stringent threshold for genome-wide significance, and the importance of replication and robust quality assurance and quality control protocols, as described in this chapter. Effective increases in sample size have also been achieved through the aggregation of association summary statistics from multiple GWASs via meta-analysis through international collaborative efforts including GIANT (anthropometric traits) (Locke *et al.*, 2015; Shungin *et al.*, 2015; Wood *et al.*, 2014), GLGC (lipid profiles) (Willer *et al.*, 2013), DIAGRAM (type 2 diabetes) (Scott *et al.*, 2017; Gaulton *et al.*, 2015), ICBP (blood pressure and hypertension) (Ehret *et al.*, 2016) and the Breast and Prostate Cancer Cohort Consortium (Al Olama *et al.*, 2014; Garcia-Closas M, *et al.*, 2013). Third, there has been considerable development of statistical methodologies to maximise the power to detect association with complex traits, including imputation and multi-SNP approaches, and their implementation in user-friendly software that can deal with the scale and complexity of GWAS data.

One of the major promises of GWASs is the delivery of personalised medicine, in which preventative and treatment interventions are designed for individuals based on their genome (Guttmacher and Collins, 2005). An essential requirement for personalised medicine is the development of predictive models of future disease in unaffected individuals, which are of particular importance when interventions are invasive, expensive or have major side effects (Janssens and van Duijn, 2008). However, investigations of the utility of SNPs identified through GWASs for the prediction of common diseases, including type 2 diabetes (Wang *et al.*, 2016) and coronary heart disease (Humphries *et al.*, 2010), have demonstrated limited predictive power in comparison with traditional clinical and lifestyle risk factors. The vast majority of SNPs identified through GWASs have modest effects on complex traits, and thus have minimal

predictive utility in isolation. Even genetic risk scores, constructed by aggregating effects across GWAS SNPs, explain relatively little of the trait variance. There is evidence that much of the 'missing heritability' is due to common variants of increasingly modest effect, which could be detected through GWASs in larger sample sizes (Yang *et al.*, 2010; Lee *et al.*, 2011), and would be expected to increase predictive power. Stratifying individuals by traditional risk factors may also reveal groups for which the genetic contribution to the trait under investigation is greatest, and there will be most predictive power of GWAS SNPs. There has also been some debate over the use of GWAS SNPs identified in one ethnic group for personalised medicine in other populations. Recent evidence suggests that common variant GWAS signals for many complex traits are shared across ethnicities, with minimal evidence for heterogeneity in allelic effects between populations (Li and Keating, 2014). However, the transferability of genetic risk scores built in one ethnic group into other ancestries is not so clear (Martin *et al.*, 2017). Causal variants may differ substantially in allele frequency between populations or may be population-specific, making them less relevant to disease prediction across ancestries, highlighting the importance of undertaking GWASs across diverse ethnicities. For example, a GWAS of type 2 diabetes susceptibility undertaken in the isolated Greenlandic population highlighted a missense variant in *TBC1D4* that confers muscle insulin resistance and explains more than 10% of disease cases, but which is rare or monomorphic in other ancestry groups (Moltke *et al.*, 2014).

Despite the limited predictive power of GWAS SNPs, genetic risk scores do have utility as instruments in Mendelian randomisation (MR) as an approach to examine the causal effect of a modifiable exposure on disease in non-experimental studies (Davey Smith and Hemani, 2014). The underlying principle of MR is that SNPs that are associated with the exposure should be correlated with disease risk to the extent predicted by their impact on the exposure. MR has been used to establish the causal effects on disease of biomarkers and lifestyle factors (Chen *et al.*, 2008; IL6R Genetics Consortium Emerging Risk Factors Collaboration, 2012; Pichler *et al.*, 2013). The ideal MR experiment would utilise a single genetic variant with known mechanistic effect on the trait for which it is an instrument. However, the joint effects of GWAS SNPs provide better predictors of potential causal exposures and a stronger instrument, although they also have greater potential for pleiotropy, which can lead to erroneous risk estimation (Davey Smith and Hemani, 2014). Recent developments include two-sample MR (Pierce and Burgess, 2013), which makes use of association summary statistics from GWAS SNPs for exposure and outcome traits, without the need for phenotype data in the same samples.

GWASs have also promised to revolutionise the development of treatment interventions through the identification of causal genes for complex traits that are drug targets. However, while GWASs have been successful in identifying regions of the genome contributing effects to complex traits, they have not typically pinpointed the causal variants that drive the association signals, precisely because of the strong LD between common variants that has enabled efficient genotyping technology development. Translation is further complicated by the fact that association signals often map to non-coding sequence, making inference of the genes and biological processes through which their effects are mediated even more challenging (McCarthy and Hirschhorn, 2008). Translational progress thus requires fine-mapping of association signals to localise the underlying causal variants, improved annotation to establish the impact of genetic variation on gene function and regulation, and more informative functional studies that will determine mechanism from gene identification.

Fine-mapping of regions of the genome that contribute to complex traits typically considers dissection of the association signal to evaluate the evidence of multiple causal variants, and localisation of the causal variants that drive each distinct association signal. The dissection of signals can be achieved through conditional analyses (Cordell and Clayton, 2002; Spain and Barrett, 2015), for example through backward elimination or stepwise selection, until the

association is fully explained. Each distinct association signal is then obtained by conditioning on all other ‘index variants’ at the locus by including their genotypes as covariates in the generalised linear regression model. Approximate conditional analyses can also be undertaken with association summary statistics and a reference of LD between SNPs that can be used to estimate the expected covariance in allelic effect estimates (Yang *et al.*, 2012). Association summary statistics from (approximate) conditional analyses can then be used to derive 99% (or 95%) ‘credible sets’ of variants that account for 99% (or 95%) of the probability of driving each distinct signal (Maller *et al.*, 2012).

One approach that shows great promise for the localisation of causal variants underlying association signals is the aggregation of GWAS data through trans-ethnic meta-analysis (Rosenberg *et al.*, 2010; Zaitlen *et al.*, 2010; Morris, 2011). Trans-ethnic fine-mapping is enabled by the observation that many GWAS SNPs for complex traits are shared across diverse populations, which is consistent with a model for which the underlying causal variants are the same across ancestry groups and were derived from mutations that occurred before human population migration out of Africa. This approach then takes advantage of the observation that patterns of LD among common variants vary between populations (Wang *et al.*, 2012). As a result, across ethnic groups, we would expect to see strong association signals only at the causal variant and those SNPs in strong LD with the causal variant in all population studied: the greater the diversity of GWASs, the more refined the localisation will be. For example, in a fine-mapping study of type 2 diabetes in the genomic region mapping to *CDKAL1*, the credible set of variants identified in a trans-ethnic meta-analysis was the exact intersection of credible set variants identified in East Asian and Eurasian population groups (Horikoshi *et al.*, 2016).

Integration of genetic fine-mapping data from GWAS with functional and regulatory annotation from large-scale publicly available tissue-specific molecular profiling initiatives, such as GTEx (GTEx Consortium, 2013), Epigenome Roadmap (Bernstein *et al.*, 2010) and ENCODE (ENCODE Project Consortium, 2012), has great potential to provide insight into causal mechanisms underlying complex traits (Pickrell, 2014). Enrichment of associated variants in specific markers of regulatory activity can point to relevant upstream biological mechanisms, while colocalisation with expression quantitative trait loci (eQTL) in relevant tissues can pinpoint causal genes. For example, in a trans-ethnic GWAS meta-analysis of kidney function (Mahajan *et al.*, 2016), credible variants were enriched for DNase I hypersensitivity sites in human kidney cells, and were eQTL for *NFATC1* and *RGS14*. Subsequent functional experimentation highlighted that loss-of-function mutations in ancestral orthologues of both genes in *Drosophila melanogaster* were associated with altered sensitivity to salt stress, and renal mRNA expression of *Nfatc1* and *Rgs14* in a salt-sensitive mouse model was also reduced after exposure to a high-salt diet. Together these data highlighted that salt sensitivity may be an important marker for biological processes that affect kidney function and chronic kidney disease in humans.

Looking forward, GWAS will undoubtedly continue to expand the catalogue of regions of the genome contributing to complex human traits. The availability of deeply phenotyped population biobanks, such as the UK Biobank (Sudlow *et al.*, 2015), with linkage to electronic medical records, offers exciting opportunities to evaluate causal relationships between traits. Discovery will be enhanced by the development of methods that leverage multi-trait data in these large cohorts, increasing power to detect association by modelling the correlation between phenotypes, and offering insight into the shared genetic contribution to human diseases (Yang and Wang, 2012; Mägi *et al.*, 2017). Prospects for fine-mapping will continue to improve with the increasing availability of GWASs in diverse populations, and expanded higher-density reference panels for imputation, such as that from the Haplotype Reference Consortium (McCarthy *et al.*, 2016). Understanding of the biological mechanisms underpinning the effect of GWAS signals on complex traits will continue to be enhanced by improved genomic annotation,

particularly in non-coding regions, and expression data from densely genotyped human samples in diverse tissues, together with methodological development to enable integration of these data resources. Finally, the development of high-throughput and tractable animal models and relevant *in vitro* models will allow the functional impact of potential causal genes and variants to be exhaustively assessed. To fully attain the potential of GWAS in the coming years will require coordinated collaboration between researchers over a wide range of disciplines, including human genetics, functional genomics, computational biology and statistical modelling, and offer exciting and realistic prospects for the prediction and treatment of human disease.

References

- 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.
- 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.
- Al Olama, A.A., Kote-Jarai, Z., Berndt, S.I., et al. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature Genetics* **46**, 1103–1109.
- Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H. and Wjst, M. (2001). Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics* **69**, 936–950.
- Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols* **5**, 1564–1573.
- Antoniou, A.C. and Easton, D.F. (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. *Genetic Epidemiology* **25**, 190–202.
- Armitage, P. (1955). Test for linear trend in proportions and frequencies. *Biometrics* **11**, 375–386.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.
- Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**, 1045–1048.
- Bhattacharya, K., McCarthy, M.I. and Morris, A.P. (2011). Rapid testing of gene-gene interactions in genome-wide association studies of binary and quantitative phenotypes. *Genetic Epidemiology* **35**, 800–808.
- Brem, R.B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572–1577.
- Brem, R.B., Storey, J.D., Whittle, J. and Kruglyak, L. (2005). Genetic interaction between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703.
- Bush, W.S. and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology* **8**, e1002822.
- Cardon, L.R. (2006). Genetics: Delivering new disease genes. *Science* **314**, 1403–1405.
- Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004). Selecting a maximally informative set of single nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**, 106–120.
- Chen, H., Wang, C., Conomos, M.P., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics* **98**, 653–666.

- Chen, L., Davey Smith, G., Harbord, R.M. and Lewis, S.J. (2008). Alcohol intake and blood pressure: A systematic review implementing Mendelian randomisation approach. *PLoS Medicine* **5**, 461.
- Clark, A.G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology* **27**, 321–333.
- Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols* **6**, 121–133.
- Clayton, D.G., Walker, N.M., Smyth, D.J., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**, 1243–1246.
- Conomos, M.P., Reiner, A.P., Weir, B.S. and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. *American Journal of Human Genetics* **98**, 127–148.
- Cook, J.P., Mahajan, A. and Morris, A.P. (2017). Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *European Journal of Human Genetics* **25**, 240–245.
- Cordell, H.J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**, 2463–2468.
- Cordell, H.J. and Clayton, D.G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. *American Journal of Human Genetics* **70**, 124–141.
- Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L. and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923.
- Davey Smith, G. and Hemani, G. (2014). Mendelian randomisation: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**, R89–R98.
- Delaneau, O., Marchini, J. and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181.
- Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.
- Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.
- Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human Heredity* **66**, 87–98.
- Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology* **32**, 227–234.
- Duerr, R.H., Taylor, K.D., Brant, S.R., et al. (2006). A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463.
- Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P. and Morris, A.P. (2004). Linkage disequilibrium mapping via cladistics analysis of SNP haplotypes. *American Journal of Human Genetics* **75**, 35–43.
- Ehret, G.B., Ferreira, T., Chasman, D.I., et al. (2016). The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nature Genetics* **48**, 1171–1184.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.
- Excoffier, L. and Slatkin, M. (1995). Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921–927.
- Feng, S., Liu, D., Zhan, X., Wing, M.K. and Abecasis, G.R. (2014). RAREMETAL: Fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829.
- Freedman, M.L., Reich, D., Penney, K.L., et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics* **36**, 388–393.

- Garcia-Closas, M., Couch, F.J., Lindstrom, S., *et al.* (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nature Genetics* **45**, 392–398.
- Gaulton, K.J., Ferreira, T., Lee, Y., *et al.* (2015). Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics* **47**, 1415–1425.
- GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **29**, 580–585.
- Guttmacher, A.E. and Collins, F.S. (2005). Realising the promise of genomics in biomedical research. *Journal of the American Medical Association* **294**, 1399–1402.
- Herold, C., Steffens, M., Brockschmidt, F.F., Baur, M.P. and Becker, T. (2009). INTERSNP: Genome-wide interaction analysis guided by a priori information. *Bioinformatics* **25**, 3275–3281.
- Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.
- Hong, H., Su, Z., Ge, W., Shi, L., Perkins, R., Fang, H., Mendrick, D. and Tong, W. (2010). Evaluating variations of genotype calling: A potential source of spurious associations in genome-wide association studies. *Journal of Genetics* **89**, 55–64.
- Horikoshi, M., Pasquali, L., Wiltshire, S., *et al.* (2016). Transancestral fine-mapping of four type 2 diabetes susceptibility loci highlights potential causal regulatory mechanisms. *Human Molecular Genetics* **25**, 2070–2081.
- Hugot, J.P., Chamaillard, M., Zouali, H., *et al.* (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603.
- Humphries, S.E., Drenos, F., Ken-Dror, G. and Talmud, P.J. (2010). Coronary heart disease risk prediction in the era of genome-wide association studies: Current status and what the future holds. *Circulation* **121**, 2235–2248.
- IL6R Genetics Consortium Emerging Risk Factors Collaboration (2012). Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *Lancet* **379**, 1205–1213.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.
- International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–61.
- International HapMap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58.
- Jallow, M., Teo, Y.Y., Small, K.S., *et al.* (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature Genetics* **41**, 657–665.
- Janssens, A.C.J.W. and van Duijn, C. (2008). Genome-based predictions of common diseases: Advances and prospects. *Human Molecular Genetics* **17**, R166–R173.
- Jiang, X. and Neapolitan, R.E. (2015). LEAP: Biomarker inference through learning and evaluating association patterns. *Genetic Epidemiology* **39**, 173–184.
- Kam-Thong, T., Azencott, C.A., Cayton, L., Putz, B., Altmann, A., Karbalai, N., Samann, P.G., Scholkopf, B., Muller-Myhsok, B. and Borgwardt, K.M. (2012). GLIDE: GPU-based linear regression for detection of epistasis. *Human Heredity* **73**, 220–236.
- Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354.
- Karssen, L.C., van Duijn, C.M. and Aulchenko, Y.S. (2016). The GenABEL project for statistical genomics. *F1000 Research* **5**, 914.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research* **12**, 996–1006.

- Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. and Tsui, L.-C. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science* **245**, 1073–1080.
- Klein, R.J. (2007). Power analysis for genome-wide association studies. *BMC Genetics* **8**, 58.
- Klein, R.J., Zeiss, C., Chew, E.Y., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.
- Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T. and Richards, J.B. (2012). The empirical power of rare variant association methods: Results from Sanger sequencing in 1,998 individuals. *PLoS Genetics* **8**, e1002496.
- Lee, S., Wu, M.C. and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775.
- Lee, S.H., Wray, N.R., Goddard, M.E. and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**, 294–305.
- Lee, W., Sjolander, A. and Pawitan, Y. (2016). A critical look at entropy-based gene-gene interaction measures. *Genetic Epidemiology* **40**, 416–424.
- Li, Y.R. and Keating, B.J. (2014). Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Medicine* **6**, 91.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835.
- Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E. and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526.
- Locke, A.E., Kahali, B., Berndt, S.I., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206.
- Loh, P.-R., Tucker, G., Bulik-Sullivan, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290.
- Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* **33**, 177–182.
- Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. and GoT2D Investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology* **37**, 539–550.
- MacArthur, J., Bowler, E., Cerezo, M., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research* **45**, D896–D901.
- MacDonald, M.E., Novelletto, A., Lin, C., et al. (1992). The Huntington's disease candidate region exhibits many different haplotypes. *Nature Genetics* **1**, 99–103.
- Mackay, T.F. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**, 303–339.
- Madsen, B.E. and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics* **5**, e1000384.
- Mägi, R., Asimit, J.L., Day-Williams, A.G., Zeggini, E. and Morris, A.P. (2012). Genome-wide association analysis of imputed rare variants: Application to seven common complex diseases. *Genetic Epidemiology* **36**, 785–796.
- Mägi, R., Suleimanov, Y.V., Clarke, G.M., Kaakinen, M., Fischer, K., Prokopenko, I. and Morris, A.P. (2017). SCOPA and META-SCOPA: Software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC Bioinformatics* **18**, 25.
- Mahajan, A., Rodan, A.Y., Le, T.H., et al. (2016). Trans-ethnic fine-mapping highlights kidney function genes linked to salt sensitivity. *American Journal of Human Genetics* **99**, 636–646.

- Maller, J.B., McVean, G., Byrnes, J., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics* **44**, 1294–1301.
- Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511.
- Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512–517.
- Marchini, J., Donnelly, P. and Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413–417.
- Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D. and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics* **100**, 635–649.
- McCarthy, M.I. and Hirschhorn, J.N. (2008). Genome-wide association studies: Past, present and future. *Human Molecular Genetics* **17**, R100–R101.
- McCarthy, S., Das, S., Kretzschmar, W., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283.
- Miller, D.G., Zhang, Y., Yu, G., Liu, Y., Chen, L., Langefeld, C.D., Herrington, D. and Wang, Y. (2009). An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics* **25**, 2478–2485.
- Molitor, J., Marjoram, P. and Thomas, D. (2003a). Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. *Genetic Epidemiology* **29**, 91–107.
- Molitor, J., Marjoram, P. and Thomas, D. (2003b). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *American Journal of Human Genetics* **73**, 1368–1384.
- Moltke, I. and Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics* **30**, 1027–1028.
- Moltke, I., Grarup, N., Jorgensen, M.E., et al. (2014). A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193.
- Morris, A.P. (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes. *Genetic Epidemiology* **29**, 91–107.
- Morris, A.P. (2006). A flexible Bayesian framework for modelling haplotype association with disease allowing for dominance effects of the underlying causative variants. *American Journal of Human Genetics* **79**, 679–694.
- Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology* **35**, 809–822.
- Morris, A.P., Zeggini, E. (2010). An evaluation of statistical approaches to rare variant association analysis in genetic association studies. *Genetic Epidemiology* **34**, 188–195.
- Moskvina, V., Craddock, N., Holmans, P., Owen, M.J. and O'Donovan, M.C. (2006). Effects of differential genotyping error rate on the type I error probability of case-control studies. *Human Heredity* **61**, 55–64.
- Moutsianas, L. and Morris, A.P. (2014). Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits. *Briefings in Functional Genomics* **13**, 362–370.
- Moutsianas, L., Agarwala, V., Fuchsberger, C., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genetics* **11**, e1005165.

- Novembre, J., Johnson, T., Bryc, K., *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**, 98–101.
- O'Connell, J., Gurdasani, D., Delaneau, O., *et al.* (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics* **10**, e1004234.
- Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. and Drineas, P. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics* **3**, e160.
- Pe'er, I., Yelensky, R., Altshuler, D. and Daly, M.J. (2008). Estimation of the multiple testing burden for genome-wide association studies of nearly all common variants. *Genetic Epidemiology* **32**, 381–5.
- Pichler, I., Del Greco, F., Gogele, M., *et al.* (2013). Serum iron levels and the risk of Parkinson disease: A Mendelian randomisation study. *PLoS Medicine* **10**, e1001462.
- Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics* **94**, 559–573.
- Pierce, B.L. and Burgess, S. (2013). Efficient designs for Mendelian randomisation studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology* **178**, 1177–1184.
- Pirinen, M., Donnelly, P. and Spencer, C.C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* **44**, 848–851.
- Pirinen, M., Donnelly, P. and Spencer, C.C.A. (2013). Efficient computation with a linear mixed model on large-scale genetic data sets with applications to genetic studies. *Annals of Applied Statistics* **7**, 369–390.
- Prabhu, S. and Pe'er, I. (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Research* **22**, 2230–2240.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.
- Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463.
- Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010). LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **15**, 2336–2337.
- Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559–575.
- Reich, D.E. and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends in Genetics* **17**, 502–510.
- Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.
- Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* **69**, 138–147.
- Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I. and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics* **11**, 356–366.
- Rothman, K.J. and Greenland, S. (1998). Case-control studies. In K.J. Rothman and S. Greenland (eds.), *Modern Epidemiology*. Lippincott-Raven, Philadelphia, pp. 93–114.
- Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. and Poland, G.A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

- Schillert, A., Schwarz, D.F., Vens, M., Szymczak, S., König, I.R. and Ziegler, A. (2009). ACPA: Automated cluster plot analysis of genotype data. *BMC Proceedings* **3**(Suppl. 7), S58.
- Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. (2009). Common vs rare allele hypotheses for complex diseases. *Current Opinion in Genetics and Development* **19**, 212–219.
- Scott, R.A., Scott, L.J., Mägi, R., et al. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902.
- Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.
- Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196.
- Spain, S.L. and Barrett, J.C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics* **24**, R111–R119.
- Spencer, C.C.A., Su, Z., Donnelly, P. and Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics* **5**, e1000477.
- Staples, J., Nickerson, D.A. and Below, J.E. (2013). Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic Epidemiology* **37**, 136–141.
- Steffens, M., Becker, T., Sander, T., et al. (2010). Feasible and successful: Genome-wide interaction analysis involving all 1.9×10^{11} pair-wise interaction tests. *Human Heredity* **69**, 268–284.
- Stephens, M. and Balding, D.J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* **10**, 681–690.
- Stephens, M. and Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genetic data. *American Journal of Human Genetics* **73**, 1162–1169.
- Stephens, M., Smith, N.J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**, 978–989.
- Storey, J.D., Akey, J.M. and Kruglyak, L. (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology* **3**, e267.
- Su, Z., Cardin, N., Wellcome Trust Case Control Consortium, Donnelly, P. and Marchini, J. (2009). A Bayesian method for detecting and characterising allelic heterogeneity and boosting signals in genome-wide association studies. *Statistical Science* **24**, 430–450.
- Sudlow, C., Gallacher, J., Allen, N., et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12**, e1001779.
- Sveinbjornsson, G., Albrechtsen, A., Zink, F., et al. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics* **48**, 314–317.
- Syed, H., Jorgensen, A.L. and Morris, A.P. (2017). SurvivalGWAS_SV: Software for the analysis of genome-wide association studies of imputed genotypes with ‘time-to-event’ outcomes. *BMC Bioinformatics* **18**, 265.
- Tabor, H.K., Risch, N.J. and Myers, R.M. (2002). Candidate-gene approaches for studying complex genetic traits. *Nature Reviews Genetics* **3**, 391–397.
- Tavtigian, S.V., Simard, J., Teng, D.H., et al. (2001). A candidate prostate cancer susceptibility gene at chromosome 17p. *Nature Genetics* **27**, 172–180.
- Wang, X., Liu, X., Sim, X., et al. (2012). A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations. *European Journal of Human Genetics* **20**, 469–475.
- Wang, X., Strizich, G., Hu, Y., Wang, T., Kaplan, R.C. and Qi, Q. (2016). Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *Journal of Diabetes* **8**, 24–35.

- Weale, M.E. (2010). Quality control for genome-wide association studies. *Methods in Molecular Biology* **628**, 341–372.
- Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
- Wigginton, J.E., Cutler, D.J. and Abecasis, G.R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.
- Willer, C.J., Schmidt, E.M., Sengupta, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1283.
- Wittke-Thompson, J.K., Pluzhnikov, A. and Cox, N.J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 967–986.
- Wood, A.R., Esko, T., Yang, J., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186.
- Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338.
- Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**, 82–93.
- Yang, J., Ferreira, T., Morris, A.P., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369–375.
- Yang, J., Benyamin, B., McEvoy, B.P., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569.
- Yang, Q. and Wang, Y. (2012). Methods for analysing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics* **2012**, 652569.
- Zaitlen, N., Pasaniuc, B., Gur, T., Ziv, E. and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *American Journal of Human Genetics* **86**, 23–33.
- Zhan, X., Hu, Y., Li, B., Abecasis, G.R. and Liu, D.J. (2016). RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426.
- Zhang, Y. (2012). A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genetic Epidemiology* **36**, 36–47.
- Zhang, Y. and Liu, J.S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**, 1167–1173.
- Zhang, Y., Zhang, J. and Liu, J.S. (2011). Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. *Annals of Applied Statistics* **5**, 2052–2077.
- Zhang, Z., Ersoz, E., Lai, C.-Q., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–360.
- Zheng, H.F., Ladouceur, M., Greenwood, C.M.T. and Richards, J.B. (2012). Effect of genome-wide genotyping and reference panels on rare variants imputation. *Journal of Genetics and Genomics* **39**, 545–550.
- Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824.
- Zondervan, K.T. and Cardon, L.R. (2007). Designing candidate gene and genome-wide case-control association studies. *Nature Protocols* **2**, 2492–2501.

Replication and Meta-analysis of Genome-Wide Association Studies

Frank Dudbridge¹ and Paul Newcombe²

¹ Department of Health Sciences, University of Leicester, Leicester, UK

² MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Abstract

Replication and meta-analysis are routinely conducted when following up genome-wide association studies. Replication is a mandatory step in confirming the validity of associations. Technical replication verifies the integrity of the original data, while direct replication reproduces the association as closely as possible in independent data. Indirect replication, while not providing full support for the original association, provides evidence for its generalisability. A statistical effect that bears on the power of replication studies is the *winner's curse*, whereby the effect size seen in a discovery study tends to be more extreme than in truth. Several methods are available to correct for the *winner's curse*, depending on how associations are first identified and whether replication data are already available. Meta-analysis, in which results from several studies are combined, is increasingly important as individual studies have limited power to detect genetic associations. Although standard fixed and random effects models can be used, more powerful methods are available for detecting effects when there is heterogeneity across studies. Heterogeneity may arise from variation in linkage disequilibrium patterns, case ascertainment and interaction effects. Imputation allows the combination of studies that have genotyped different markers. Increasingly, consortia are being formed in advance of initial data analysis, allowing harmonisation of quality control criteria and sharing of individual-level data.

22.1 Introduction

Performing a genome-wide association study (GWAS), while a substantial task in itself, is but the first step towards fully explaining the genetic architecture of a trait and the mechanisms by which genetic variants exert their effects. Several analyses are commonly performed in the follow-up phases of a GWAS. *Replication* is concerned with repeating, in various ways, the results of a GWAS, so as to provide greater confidence that the observed results are genuine. An aspect that bears on the power of replication analyses is the *winner's curse*, a statistical effect whereby the strength of an association tends to be exaggerated in the study that discovered it. *Meta-analysis* is the combination of several GWASs into a single overall result, allowing the discovery of further associations through the increased sample size, and increasing the weight of evidence for known GWAS associations. This chapter reviews these three aspects of

post-GWAS analysis, which each aim to clarify and increase the statistical evidence for GWAS results.

22.2 Replication

22.2.1 Motivation

Before GWASs became feasible, association studies were limited to candidate genes, but those studies were badly affected by poor reproducibility (Hirschhorn *et al.*, 2002). With hindsight this was due to statistical significance levels that were too liberal, as GWAS data later revealed that candidate genes were only slightly more likely to be associated than non-candidate loci (Vimaleswaran *et al.*, 2012). Traditional statistical significance (typically $p < 0.05$) is predicated on a plausible alternative hypothesis, whereas significance for GWASs (typically $p < 5 \times 10^{-8}$) recognises that individual variants are *a priori* unlikely to be associated (Dudbridge and Gusnanto, 2008). Although candidate gene studies were generally held to traditional significance, perhaps after limited adjustment for multiple testing within genes, it is now clear that genome-wide significance would have been more appropriate, and that there was a high false positive rate in candidate gene studies. In some instances, replication studies may have been underpowered, leading to false negative results, but the vast majority of GWAS associations have proved to be outside previously proposed candidate genes.

The poor track record of replication in candidate gene studies led to a consensus that replication should be mandatory before regarding any GWAS association as genuine (NCI-NHGRI Working Group on Replication in Association Studies, 2007). Furthermore, owing to their considerable expense it is important that GWAS associations are robust; ideally therefore, a replication phase should be designed into a GWAS from the outset. Indeed, even if independent replication data are acquired concurrently with GWAS data, it is preferable to analyse the replication data separately from the GWAS, as opposed to combining the data into a larger sample with greater power. This stands in contrast to a statistical dogma that more data is better, but is important because significant results in GWASs may be the results of various biases arising from data handling, population structure, genotyping chemistry and so on. Such biases would persist in combined data sets, whereas associations that are replicated in independent analysis of independent data are unlikely to arise from such biases and are therefore more likely to be genuine.

22.2.2 Different Forms of Replication

Different forms of replication are possible, having different roles in the confirmatory process. Below we will briefly describe the most important forms.

22.2.2.1 Technical Replication

Technical replication involves the reanalysis of all or a subset of genetic variants investigated in the GWAS, and is carried out on the original samples, perhaps using a different genotyping technology. The aim of technical replication is to detect errors in genotyping that could lead to differential genotype calling and false positive associations. When associated markers have been imputed rather than directly genotyped (see Chapter 3), or have been genotyped from low-depth sequencing, technical replication by direct genotyping is important for increasing confidence in the accuracy of the data.

In GWASs, in which large numbers of samples are analysed in a single experiment, technical replication is also used to confirm sample identity through the genotyping process. For example, it is not unknown for DNA samples to be accidentally switched or misplaced when preparing batches for genotyping. Technical replication of a subset of variants, using a different genotyping platform, lends reassurance that the genotyped samples are indeed those that were intended. Note that it is not necessary to replicate the most strongly associated markers. A random subset of markers is sufficient since the aim is simply to confirm the integrity of the original data.

22.2.2.2 Direct and Indirect Replication

A *direct replication* reproduces the original association in every possible respect up to the actual subjects genotyped. The subjects, while independent of the original study, should come from the same source population, and the same genetic variant should be associated – or, if this is not possible, a variant in very strong linkage disequilibrium (LD) with it (*a proxy* with squared correlation $r^2 \sim 1$ with the original variant). Furthermore, the association should be in the same direction as the original study, and consistent with the same genetic model, be it additive, recessive or dominant. The effect size (often an odds ratio) should be comparable in the replication study, although one should allow for the winner's curse effect, to be described below.

An example of direct replication can be found in the GWAS of type 1 diabetes conducted within the Wellcome Trust Case Control Consortium (2007). Among the associated markers was the single nucleotide polymorphism (SNP) *rs17696736* for which the G allele had an odds ratio of 1.37 ($p = 7.27 \times 10^{-14}$) in a UK sample assuming a multiplicative effect. The same allele was then found to be associated in a larger sample of UK subjects, with odds ratio 1.16 ($p = 1.82 \times 10^{-6}$) (Todd *et al.*, 2007). Further studies have also replicated this association, establishing without doubt that the *C12orf30* gene on chromosome 12 is implicated in this condition.

Indirect replication can be claimed when the replication involves a different allele in the same gene or genomic locus, or a phenotype that is correlated with but not identical to that in the original study. Often this is an intermediate quantitative trait, also known as an endophenotype. For example, a SNP associated with heart disease might be replicated by association with cholesterol levels, or a SNP associated with osteoporosis might be replicated by association with bone mineral density. Replication in a different population is also a form of indirect replication, which could be considered as providing stronger evidence for a genuine effect. Replication involving the same allele, but with opposite direction of effect, might also be claimed as an indirect replication, although one should generally be sceptical about such a result.

As an example of indirect replication, a GWAS of breast cancer found association of the G allele of *rs10941679* (odds ratio 1.19, $P = 2.9 \times 10^{-11}$) in samples of European ancestry (Stacey *et al.*, 2007). A subsequent GWAS in UK subjects did not genotype this SNP, but found associations of other SNPs in the same region of chromosome 5p12 (Turnbull *et al.*, 2010). The most significant was *rs7716600* (odds ratio 1.11, $p = 0.0034$), which is in strong but not complete LD with *rs10941679* ($r^2 = 0.75$). The association of other markers in the region, in a population with similar but not identical ancestry, lends support to the presence of a genuine risk variant.

Direct replication confirms an association as originally reported, whereas indirect replication addresses the generalisability of an association. Both are desirable. Direct replication can be hard to achieve because studies differ subtly in aspects such as phenotype definition and LD structure. Indirect association does not confirm a specific association, but provides stronger evidence that a SNP in a genomic region is implicated in some aspect of the trait. Often, indirect replications offer insights into the allelic architecture of a locus of interest, or of shared or unique genetic determinants of correlated traits. For example, variants in the *FTO* locus were first associated with type 2 diabetes and then replicated by association with body mass index

(BMI) in several populations (Loos and Yeo, 2014), indicating that the effect of *FTO* on diabetes risk acts via its effect on obesity.

22.2.2.3 *In silico replication*

With large-scale data collections becoming increasingly available, *in silico replication* is becoming popular, in which the replication data are already available and possibly embedded within a larger data set. For example, Erdmann *et al.* (2009) performed a GWAS for coronary artery disease, and, having identified a significant association on chromosome 3, they were immediately able to reference the same SNP in three other completed GWASs and claim replication at a nominal significance level. *In silico* replication can be a reciprocal arrangement, in which each study acts as replication sample for the discoveries made by the other.

22.2.3 Two-Stage Genome-Wide Association Studies

The two-stage GWAS design is related to but distinct from that of replication studies. In these studies, an initial GWAS is performed on a subset of subjects. A reduced number of the most significant SNPs are then taken forward for genotyping in the remaining subjects, with the number of SNPs ideally large enough to include most of the true positives, but small enough to exclude most of the true negatives. Typically, SNPs are selected at the first stage according to a *p*-value threshold short of genome-wide significance, or by using a false discovery rate criterion, with the balance between the two sample sizes and numbers of genotyped SNPs determined by budgetary constraints. The two-stage approach can be more cost-effective than performing a GWAS on the entire sample (Thomas *et al.*, 2004; Wang *et al.*, 2006; Easton *et al.*, 2007) and was often employed in early GWASs, though nowadays less often as the cost of genotyping has subsequently decreased.

The two-stage approach is designed as an economically efficient implementation of a GWAS, and is only intended to generate hypotheses at its conclusion. The first stage of a two-stage approach does not generate strong hypotheses, so it is inappropriate to analyse the second stage in isolation, and fallacious to claim that the multiple testing is reduced (Wason and Dudbridge, 2012). The use of the term ‘replication’ for the second stage is discouraged. Methods are available to test both stages jointly, and in fact are more powerful than analysis of the second stage alone (Prentice and Qi, 2006; Skol *et al.*, 2006). In contrast, replication studies are intended to confirm the results of an earlier GWAS, and as such should be analysed separately. While stronger significance levels might be obtained by combined analysis of GWAS and replication data, they would not be free from biases in the GWAS data, and true independent replication carries greater weight in establishing the association of a genetic variant.

22.2.4 Significance Thresholds for Replication

The significance threshold for a replication study has not been discussed as thoroughly as for GWASs. It is generally agreed that it need not be as stringent as for GWASs, but that $p < 0.05$ is too lenient. If several SNPs are being replicated simultaneously, then there is a multiple testing problem, suggesting a Bonferroni-type correction based on the number of SNPs replicated. This conservative position would be appropriate in an initial replication to validate the results of a GWAS, in terms of ruling out systematic bias, as there is a consensus that the family-wise error (i.e. probability of at least one false positive result) should be kept low in a GWAS. In subsequent replications, however, one might regard each SNP as an individual hypothesis with some prior support, independently of how many other SNPs are replicated at the same time. In that case, $p < 0.05$ for each SNP can be justified. Given the variability in how replication studies

are conceived and designed, no single rule can be appropriate for all studies, but it is common to see p -values of 10^{-3} and 10^{-4} cited as convincing evidence of replication.

22.2.5 A Key Challenge: Heterogeneity

The success or interpretation of a replication study is influenced by heterogeneity of effect sizes, that is, differences in the true effect of a SNP between studies. Various sources of heterogeneity are possible in genetic association studies. One occurs when different patterns of LD exist between the genotyped SNP and untyped causal alleles in different studies. Patterns of LD may vary significantly across different ethnicities at a large fraction of the genome as a result of historical events, in particular genetic drift, inbreeding, natural selection or admixture. Extreme differences in the frequency of the causal alleles and local LD patterns can arise not only among different ethnicities, but also in populations of similar ethnicity which have experienced significant drift or inbreeding in their past. While the effect of a causal SNP may be homogeneous, there will be heterogeneity in the effect of a marker SNP in LD, which (absent knowledge of which SNP is causal) is often the target of replication.

Heterogeneity can also arise from different case ascertainment strategies, such as family history, age of onset or distribution of risk factors. Stronger genetic effects can be expected in cases with a family history. For example, for a polygenic model of breast cancer the sample size required to detect a common risk allele could be reduced by more than twofold if cases with an affected first-degree relative are selected, and by more than fourfold if cases with two affected first-degree relatives are used (Antoniou and Easton, 2003). Compared to a study of unselected cases, heterogeneity in effect sizes may then be observed. However, in practice empirical data have shown little difference in effects between cases with a family history compared to unselected cases (Dudbridge *et al.*, 2012). This can be explained by incomplete LD between marker SNPs and the causal variant, which attenuates both effects towards the null. As genotyping density increases towards whole genome sequences, greater heterogeneity is expected to be revealed between studies with varying ascertainment criteria.

Non-additive interactions with other genetic variants or environmental exposures also create heterogeneity. If these interactions are not modelled explicitly, but are absorbed into the marginal effect of a tested variant, then differences in the distribution of other risk factors between populations will alter the marginal effect size, creating heterogeneity in the estimated effect. For example, Timpson *et al.* (2009) examined how differences in the BMI distribution of type 2 diabetes subjects affected estimates of SNP effects. After stratifying case subjects into above and below the median BMI (30.2 kg/m^2), they showed reproducible heterogeneity in odds ratios for two loci, *FTO* and *TCF7L2*. In particular, the *FTO* association was undetectable in the low-BMI cases. While *FTO* is known to exert its risk of type 2 diabetes via adiposity, this result suggests that the effect is confined to overweight individuals, representing an interaction effect. Heterogeneity in effects will be observed whenever studies differ in their proportions of overweight cases.

22.3 Winner's Curse

22.3.1 Description of the Problem

It is usually found that the effect size of a SNP is lower in replication studies than in the GWAS that discovered the association. Therefore, the power of replication studies is lower than is apparent in the discovery study, and sample size calculations should be adjusted upwards accordingly. This effect is known as the winner's curse.

In its basic form, the winner's curse is a fairly simple concept. Suppose an item is for sale, and we define its value to be the average of the prices that each potential buyer is prepared to pay for it. If the item is then sold to the highest bidder, then the winner will, by definition, pay more for the item than its true value. Thus arises the winner's curse.

This phenomenon occurs in many settings. In ranking sports players, say according to how many points one scores during a season, the top ranking players will tend to be those that had an above-average season, whereas across many seasons those players may not be outstanding. Similarly, when identifying accident black spots, the apparently most dangerous intersections will tend to be those that happened to have an above-average rate during the observation period. And in the context of GWASs, the 'winners' are the SNPs whose effect sizes (usually odds ratios) are stochastically higher in the study to hand than their true values.

To give a more precise description, suppose that the observation for any item consists of two components: its *true effect*, which may vary across items but is constant over time, and some *noise*, which also may vary across items but also varies with time. Then in any period of study the items with the largest observations will tend to be those with both the largest true values and the largest noise. However, in subsequent observations these items will have the same true values but will typically have less noise, leading to less extreme observations. This phenomenon is known as regression to the mean (Figure 22.1), and was first observed by Galton, who noted that tall parents tend to have children who, while also tall, are shorter on average than their parents. In general, a repeat observation of some quantity will tend to be nearer the average value than the initial observation. Note that this is only a tendency, and in some cases the subsequent observations will be more extreme owing to randomly greater noise.

We can distinguish between two sources of winner's curse in GWASs. First, when considering all SNPs within a single GWAS, the most significant markers will tend to have estimated effects higher than their true values. After ranking the SNPs by their significance, the bias will be strongest for the top ranking markers, and so this effect can be described as *ranking bias* (Jeffries, 2009). Secondly, if a SNP is of interest by virtue of achieving statistical significance, then its estimated effect will be higher than its true value within the studies in which it is significant. This effect can be described as *significance bias* (Garner, 2007).

Note that ranking bias may occur without significance bias, and vice versa. Even if no SNPs are statistically significant, those that are closest to significance will exhibit ranking bias. And if

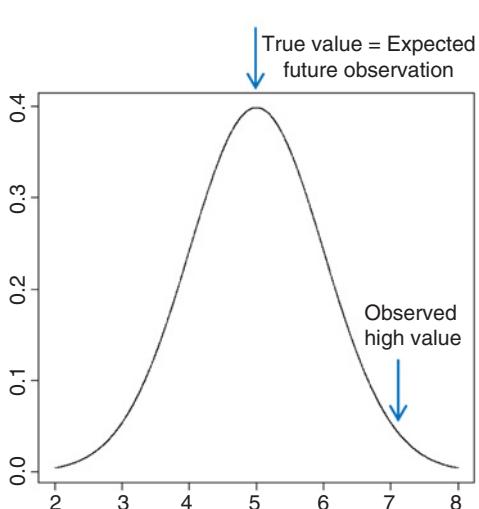


Figure 22.1 Regression to the mean. The most extreme observations tend to be those with the greatest noise. In future observations, the noise will on average be less, so that a future observation is likely to be closer to the true value (the mean) than the initial extreme observation.

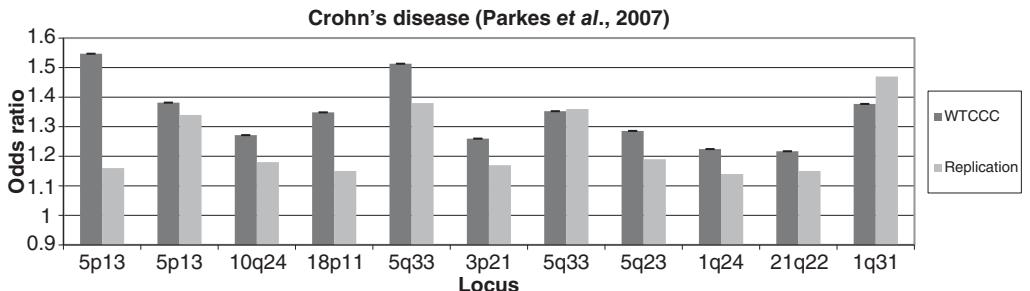


Figure 22.2 Estimated odds ratios for 11 SNPs identified in the Wellcome Trust Case Control Consortium (2007) study of Crohn's disease and subsequently replicated in independent data (Parkes *et al.*, 2007). The most significant associations in the WTCCC study are shown to the left.

only one SNP is studied, its effect will be biased when it is statistically significant, even though there is no ranking. In practice, both types of bias are present in GWASs. The nature of the exercise creates intrinsic ranking bias, and SNPs are followed up when they achieve some level of significance. Since statistical significance has a strong bearing on whether a study is published, the bias is stronger within published studies, and publication bias has been shown to be greatest in higher-impact journals (Palmer and Pe'er, 2017).

An example of both kinds of bias is given in Figure 22.2, which shows odds ratios for Crohn's disease for 11 SNPs identified in the Wellcome Trust Case Control Consortium (2007) study and subsequently replicated in an independent study (Parkes *et al.*, 2007). The replication study provides unbiased estimates of the odds ratios since there was no further selection of SNPs. It is clear that the odds ratios are systematically higher in the discovery study, owing to ranking and significance bias, and the largest biases are seen for the most significant SNPs, owing to ranking bias.

22.3.2 Methods for Correcting for Winner's Curse

Several methods have been proposed to correct for winner's curse in GWASs, each taking slightly different perspectives. In general, a direct replication study would give an unbiased estimate of a SNP effect, although this too may be subject to significance (or publication) bias, or even ranking bias if several SNPs are replicated in parallel. However, direct replication data may not always be to hand, and even if it were, it entails an inefficient use of data since the discovery GWAS is not further used for estimation.

Several authors have provided methods to correct for significance bias (Ghosh *et al.*, 2008; Zhong and Prentice, 2008). Their approaches are essentially to define a likelihood for the effect size, conditional on significance, of the form

$$L(\beta|\hat{\beta}) = \frac{\Pr(\hat{\beta}|\beta)}{\Pr(|(\hat{\beta}-\beta)/\sigma| > c|\beta)} = \frac{\phi((\hat{\beta}-\beta)/\sigma)}{\Phi\left(\frac{-c-\beta}{\sigma}\right) + \Phi\left(\frac{-c+\beta}{\sigma}\right)},$$

where β is the true effect size of the SNP, $\hat{\beta}$ is its estimate in the GWAS, σ is the standard error of $\hat{\beta}$, c is the critical value for declaring significance, and ϕ and Φ are the probability density and distribution functions of the standard normal distribution, respectively. Maximum likelihood or other approaches may be used to obtain corrected estimates of $\hat{\beta}$. However, by ignoring ranking bias these methods implicitly assume that SNPs are of interest regardless of their

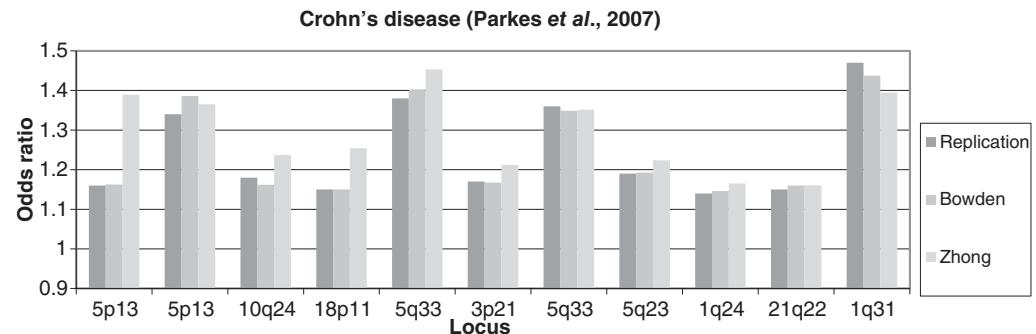


Figure 22.3 Estimated odds ratios for 11 SNPs identified in the WTCCC study of Crohn's disease. Unbiased estimate from replication data compared to bias-corrected estimators by Bowden and Dudbridge (2009) and Zhong and Prentice (2008).

ranking in a GWAS; this assumption does not hold in general, and in practice these methods are found to give estimates with the bias reduced but not eliminated (Figure 22.3).

Both ranking and significance bias may be reduced by a bootstrap procedure (Faye *et al.*, 2011). Here, repeated bootstrap samples are drawn from the data and in each sample the entire analysis and selection of SNPs is repeated. For each bootstrap sample, the subjects not included are taken as a replication sample in which the SNP effects are estimated for the selected markers from the bootstrap analysis. In this way, an estimate of the winner's curse bias is obtained at each rank and used to obtain an adjusted estimate of the form

$$\hat{\beta}_{\text{boot}(k)} = \hat{\beta}_{N(k)} - \frac{1}{N(k)} \sum_{i=1}^{N(k)} (\hat{\beta}_{D_i(k)} - \hat{\beta}_{E_i(k)}),$$

where $\hat{\beta}_{\text{boot}(k)}$ is the adjusted estimate for the k th ranked SNP, $\hat{\beta}_{N(k)}$ is the naive estimate from the discovery data, $N(k)$ is the number of bootstrap samples with at least k significant SNPs, and $\hat{\beta}_{D_i(k)}$, $\hat{\beta}_{E_i(k)}$ are the within- and out-of-bootstrap sample estimates for the k th ranked SNP in sample i . In practice some further adjustments are necessary to allow for unequal allele frequencies across SNPs, and correlation between $\hat{\beta}_{D_i(k)}$ and $\hat{\beta}_{E_i(k)}$; see Faye *et al.* (2011) for details. Of note, different SNPs may be selected at each rank in each bootstrap sample. Ranking bias is considered to apply to a rank rather than a SNP, and the SNP that occupies each particular rank is regarded as random, as is its effect size, while the degree of bias is considered fixed. This procedure has been shown to be slightly conservative in the sense that it over-corrects the bias; this is not a major concern, but as with any resampling approach it can be rather time-consuming, requiring a full reanalysis of the data in each sample.

When independent replication data are available, an exactly unbiased estimator can be obtained by combining the discovery and replication data, adjusting for both ranking and selection bias (Bowden and Dudbridge, 2009). Again the bias is associated with the rank, but here the correction depends on the estimated effect size of the corresponding SNP, as well as those of the SNPs at adjacent ranks. This approach has the advantage of using all the available data, and indeed has the minimum variance among all unbiased estimators using both data sets, conditional on the SNP ranking. Its main disadvantage is the lack of analytic confidence intervals, which must instead be estimated from bootstrap samples, a difficulty shared by the bootstrap bias correction. Furthermore, in practice only small gains in precision are obtained over the replication estimate, unless the discovery study is very much larger than the replication.

An alternative approach to correct for winner's curse is to regard all the true SNP effects as coming from a common probability distribution, which is taken as a prior distribution for the SNP effects. Each estimated effect is combined with this prior using Bayes' theorem to yield a posterior expectation that serves as a bias-corrected estimate. When the prior is correctly specified, this approach yields the property

$$E(\beta|\hat{\beta}) = \hat{\beta},$$

which is complementary to the classical unbiasedness property

$$E(\hat{\beta}|\beta) = \beta.$$

Under classical unbiasedness, the estimator will over repeated samples have expectation equal to the truth. In contrast, the complementary property states that over many SNPs with the same estimated effect, the mean true effect size equals this estimate. Furthermore, over many SNPs with varying estimates, the mean true effect size equals the mean estimate. An advantage of this viewpoint is that, because the conditioning is on the value of the estimator, selection by rank or significance is automatically accounted for.

This approach is arguably the most appropriate within a GWAS since we observe the estimated effects for many SNPs and want an idea of the corresponding true effects (Goddard *et al.*, 2009). One problem is the appropriate choice of prior. A normal distribution has traditionally been used in livestock genetics, whereas more flexible distributions allowing a mass at zero have been explored for human GWASs (Xu *et al.*, 2011). A very useful model is a nonparametric distribution, making minimal assumptions about the distribution of effects, for which the adjusted estimates can be computed very easily (Ferguson *et al.*, 2013):

$$\hat{\beta}_{EB} = \hat{\beta}_N + \sigma \left[\frac{d}{dz} \log f(z) \right]_{\frac{\hat{\beta}_N}{\sigma}}.$$

Here, $\hat{\beta}_{EB}$ is an empirical Bayes adjusted estimate, $\hat{\beta}_N$ is the naive estimate in the discovery GWAS, and $f(z)$ is the probability density of the standardised effects $\hat{\beta}/\sigma$. Conveniently, $\log f(z)$ can be approximated by nonparametric Poisson regression of a histogram of $\hat{\beta}/\sigma$. Full details are provided elsewhere (Ferguson *et al.*, 2013).

Although this empirical Bayes approach gives estimates that, in the sense described above, equal the expected truth, the estimates are biased in the classical sense and may therefore appear inconsistent with replication studies of individual SNPs. This approach is therefore most appropriate within a GWAS, for ensuring an average accuracy across all SNPs, rather than across many studies of the same SNPs of interest. As with other corrections for winner's curse, some care is needed in the interpretation of the bias-corrected estimate.

22.3.3 Applicability of These Methods

One of the original concerns about winner's curse was that replication studies would be designed according to over-optimistic estimates of effect and sample size. However, in practice replication studies have generally been performed in convenience samples not collected specifically to replicate particular SNPs, and typically also, many SNPs are replicated in parallel. As the size of replication cohorts has continued to grow, allowing accurate unbiased estimates of SNP effects, the need for adjusted estimates from discovery studies has waned. However, emerging applications such as polygenic risk scores will benefit from bias-reduced estimates in discovery studies. A polygenic risk score is constructed as the sum of risk-increasing alleles across a large number of SNPs, weighted by their effect sizes. They have been used in many applications

such as risk prediction and Mendelian randomisation (Dudbridge, 2016). Typically, estimates of the effect sizes are taken naively from GWAS data, but this will impart winner's curse on each weight. Correcting for the selection effect can increase the accuracy of polygenic risk scores (Shi *et al.*, 2016; Lall *et al.*, 2017), and awareness of the problem of winner's curse remains a key element in discussing and interpreting the results of GWASs.

22.4 Meta-analysis

22.4.1 Motivation

When several studies have been completed on the same SNPs and traits, it is natural to ask what conclusions are indicated, as a whole, by all the studies together. Meta-analysis ('analysis of analyses') is a set of techniques for combining results across studies. The aims of a meta-analysis are to obtain more precise estimates of the effects of SNPs, to resolve discrepancies between results of different studies, and to determine whether, and to what degree, genetic effects differ between studies.

Genetic meta-analysis has been important since the candidate gene era, when individual studies often gave conflicting results. A dedicated body, the Human Genome Epidemiology Network (HuGENet) was set up in 1998 with the specific aim of collating and reviewing evidence across multiple genetic studies. More recently, meta-analysis has become a core activity in consortium studies combining multiple GWASs of the same conditions. The increased sample size ensures the discovery of more associated SNPs, although replication remains as critical as in individual GWASs.

22.4.2 An Illustrative Example

Standard epidemiological methods of meta-analysis are based on summary estimates of effect sizes, such as odds ratios, and their standard errors. Figure 22.4 shows the results of a meta-analysis of two variants, D9N and HindIII, in the lipoprotein lipase (*LPL*) gene and their potential association with coronary heart disease (CHD). The studies shown are a subset of a large meta-analysis of 89 studies (Sagoo *et al.*, 2008), chosen here to illustrate the main points. The figure shows a *forest plot*, which is a standard way of presenting a meta-analysis. Each row shows the odds ratio from one study, with the width of the line showing its 95% confidence interval and the size of the square reflecting the relative size of the study. The meta-analytic estimates are shown in the final two lines, with the width of each of the diamonds reflecting the 95% confidence interval.

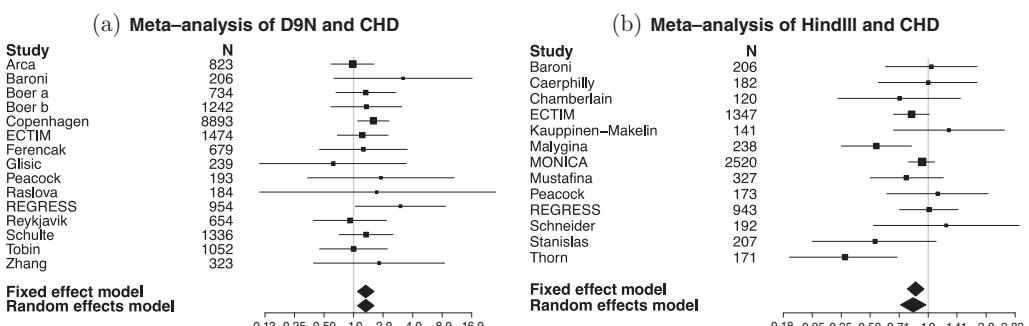


Figure 22.4 Meta-analysis of D9N and HindIII with coronary heart disease.

The following subsections describe different approaches to obtaining such meta-analytic estimates, and discuss other related methods. These approaches are applicable to studies of binary, quantitative or any other traits for which consistent estimates of effect sizes and their standard errors are available.

22.4.3 Fixed Effect Meta-analysis

In order to combine effect size estimates across the studies in Figure 22.4, we might start by taking a weighted mean. This technique is described as *fixed effect meta-analysis*, as we assume all studies contribute data on a single, identical, underlying genotype/phenotype association parameter; that is, the underlying effect is fixed across populations. A sensible weighting is the inverse variance of each study estimate. This ensures that less certain estimates make smaller contributions to the overall mean, and results in a meta-analysis estimate with the minimum variance among all possible weighted means. Denoting the estimate from study i as $\hat{\beta}_i$ and its standard error as σ_i , the inverse variance weighted mean for N studies is given by

$$\bar{\beta} = \frac{\sum_{i=1}^N \hat{\beta}_i \sigma_i^{-2}}{\sum_{i=1}^N \sigma_i^{-2}}.$$

Considering the $\hat{\beta}_i$ as independent random variables, it is straightforward to show that the standard error of the inverse variance weighted mean is

$$SE(\bar{\beta}) = \sqrt{\frac{1}{\sum_{i=1}^N \sigma_i^{-2}}}.$$

For the example in Figure 22.4, we set $\hat{\beta}_i$ equal to the study-reported log odds ratios, since $SE(\hat{\beta}_i)$ is then readily defined and normal distributions for $\hat{\beta}_i$ and $\bar{\beta}$ are justified by theory. Exponentiation of the inverse variance weighted log odds ratio gives meta-analysis odds ratio estimates of 1.33 (95% CI: (1.10, 1.62), $p=0.0038$) for D9N and 0.86 (95% CI: (0.78, 0.95), $p=0.0032$) for HindIII, strong overall evidence of effects at both variants. These are displayed in Figure 22.4 with diamonds denoted by ‘Fixed effect model’. The analysis was conducted and plots generated using the R package meta. For large-scale studies such as meta-analysis of several GWASs, an efficient implementation of fixed effects analysis is provided in the METAL software (Willer *et al.* 2010).

22.4.4 Chi-Square Test for Heterogeneity in Effect

As discussed previously, genetic variants may have different effect sizes in different studies, breaking the assumption of identical underlying effect across studies required for fixed effect meta-analysis. In particular, heterogeneity may arise from differences in LD structure, and from interactions with environmental or other genetic exposures at different frequencies.

A number of techniques are available to investigate the degree of heterogeneity in underlying effect across populations. Naturally we do not expect the summary estimates to be identical, but we are interested in whether the differences are entirely explained by sampling variation. More formally, denoting the true underlying study effects by β_1, \dots, β_N we wish to test the hypothesis that $\beta_1 = \beta_2 = \dots = \beta_N$. The test statistic

$$Q = \sum_{i=1}^N (\hat{\beta}_i - \bar{\beta})^2 \sigma_i^{-2},$$

proposed by Cochran (1954), is often used, and may be compared to a χ^2 distribution with $N - 1$ degrees of freedom. Since this test usually has low power (there are only as many observations as studies), it has become customary to use a significance threshold of 0.1, rather than 0.05 (Fleiss, 1986). Applying this test to the D9N and HindIII data, we obtain $p = 0.882$ and $p = 0.087$, respectively. Therefore there is no reason to believe in underlying differences in effect of D9N on CHD, and as such a fixed effect meta-analysis is reasonable, but there is some evidence of heterogeneity for HindIII.

22.4.5 Random Effects Meta-analysis

When heterogeneity in population effects truly exists, effect estimates across studies have two sources of variability: sampling variation *within* each study, and heterogeneity in underlying effect *between* studies. Fixed effect meta-analysis accounts for the first source of variability (that due to within-study sampling error), but ignores any variability between studies. This would lead to an artificially precise confidence interval, overestimating the strength of evidence. Random effects meta-analysis provides a more flexible framework, in which the effects underlying each study are allowed to vary. First described by DerSimonian and Laird (1986), the underlying study-specific effects are considered to be randomly distributed around a fixed 'global' mean. Typically a normal distribution is used:

$$\beta_i \sim N(\mu, \tau^2),$$

where μ denotes the global mean effect, and τ^2 the variance in true effect between studies. The variance of the individual study estimate $\hat{\beta}_i$ is then $\sigma_i^2 + \tau^2$.

Note that to justify this model, there must be no reason *a priori* to believe any *specific* population has a stronger or weaker underlying effect than another. The corresponding assumption that the study-specific effects are randomly distributed is known as the exchangeability assumption. Linking the study-specific effects via a common distribution makes the assumption that they arise from the same random process. For a particular genetic variant, or group of variants from the same gene, this assumption seems perfectly reasonable.

Estimating μ and τ^2 from the study summary estimates $\hat{\beta}_i$ and standard errors σ_i is not trivial, but good procedures are implemented in all the major statistical packages (e.g. package meta for R, command metan for STATA). Mathematical details on these procedures are omitted, though interested readers are referred to DerSimonian and Laird's (1986) original paper for both approximate formulae (*weighted least squares*) and a method of numerical approximation. As well as estimating μ and τ^2 , both procedures produce 95% confidence intervals for μ under an assumption of normality. Usually the simpler weighted least squares approach is used.

Random effects models are also naturally fitted within the Bayesian framework. The model typically mirrors that described above, the key differences being that priors must be specified on μ and τ^2 , and that a posterior distribution is obtained for the mean effect μ as opposed to a point estimate and standard error. An advantage of the Bayesian approach is that external information may be incorporated via an informative prior on μ . A disadvantage, however, is that prior choices are usually subjective, and so sensitivity analysis is required to assess the dependence of results on the choice of prior. We focus on frequentist inference in this chapter, and refer readers elsewhere for accessible descriptions of Bayesian meta-analysis (Sutton and Abrams, 2001), examples in genetics (Shah *et al.*, 2010) and discussion of possible pitfalls (Gelman, 2006).

If it is unclear whether the degree of heterogeneity warrants a departure from a fixed effect meta-analysis, carrying out a random effects meta-analysis actually helps to answer the question, since estimation of τ^2 allows us to examine how much of the total variability is attributable

to underlying differences between studies. The relative contribution of the between-study variability to the total variability is quantified by

$$I^2 = 100\% \frac{Q - df}{Q},$$

where Q is Cochran's Q and $df = N - 1$ is the number of degrees of freedom in Cochran's Q test for heterogeneity in effect (Higgins and Thompson, 2002). Negative values of I^2 are set to 0, and larger values reflect greater amounts of heterogeneity. Similar in concept to an intra-class correlation coefficient, this ultimately determines the impact of ignoring between-study heterogeneity, and is often used to decide whether a random effects meta-analysis is worthwhile.

The I^2 estimate for the D9N data was less than 0.01% which, consistent with Cochran's Q test, provides no reason to believe in underlying differences between studies. A random effects meta-analysis leads to identical results to the fixed effect meta-analysis, which is unsurprising given the lack of evidence for heterogeneity across studies.

On the other hand, a random effects meta-analysis of the HindIII data gives a global odds ratio estimate of 0.84, with a notably wider confidence interval than the fixed effect analysis at (0.72, 0.97), shown in Figure 22.4, with $p = 0.021$ thus leading to a reduction in evidence. By ignoring variability *between* studies, indicated by an I^2 of 37.0% to account for nearly half the total variance in effect estimates, the fixed effect confidence interval was over-precise, therefore overestimating the strength of evidence. This is consistent with the significant test for heterogeneity using the Q test that we saw earlier.

22.4.6 Interpretation and Significance Testing of Meta-analysis Estimates

A fixed effects meta-analysis assumes that there is a common effect size in all the studies included, and obtains an estimate of that effect size; however, it makes no claim about other studies of the same risk factor and outcome. On the other hand, a random effects meta-analysis assumes that the true effect sizes are drawn from a probability distribution, and while this is an appropriate way to allow for heterogeneity across studies, the resulting estimate is not necessarily the true effect size in any study. Instead, it is the effect size we would expect, on average, when performing a future study of the same risk factor and outcome. Such a quantity may not be particularly relevant in genetic studies, in which the aim is usually to discover SNPs affecting the risk of disease and to quantify their effect in particular populations of interest.

If a SNP has no effect on the studied trait, then its effect size would be zero in *any* study, and there would be no heterogeneity. Therefore, under the null hypothesis of no effect, the fixed effects model is appropriate and may be used for testing the presence of association. Because the fixed effects estimate allows for fewer sources of variation, its standard error is lower than the random effects estimate and so, unless there is high heterogeneity, the fixed effects model has higher power to detect a non-null effect than the random effects model. For these reasons, the fixed effects model is most commonly used in genome-wide meta-analysis studies, with the random effects model usually reserved to examine SNPs that have significant fixed effect associations.

If heterogeneity exists, but the effect sizes are symmetrical around zero, then both fixed and random effect models would yield an estimate of zero, even for a SNP with true effects. In such a case, the variance of the random effects distribution would be non-zero, and a test of $\tau^2 = 0$ would detect the association. In cases of low heterogeneity, and effects that are not symmetric around zero, a joint test of $\mu = \tau^2 = 0$ has been shown to have greater power than both fixed and random effects meta-analyses (Han and Eskin, 2011). Essentially, it is a test of $\mu = 0$ under

a random effects model with the constraint that $\tau^2 = 0$ when all true effects $\beta_i = 0$. The test statistic

$$S = \sum_{i=1}^N \log \left(\frac{\sigma_i^2}{\sigma_i^2 + \hat{\tau}^2} \right) + \frac{\hat{\beta}_i^2}{\sigma_i^2} + \frac{(\hat{\beta}_i - \hat{\mu})^2}{\sigma_i^2 + \hat{\tau}^2},$$

where $\hat{\mu}$ and $\hat{\tau}^2$ are maximum likelihood estimates of μ and τ^2 respectively, is distributed as an equal mixture of $\chi^2_{(1)}$ and $\chi^2_{(2)}$ when $\mu = \sigma^2 = 0$, for large N .

A further refinement is to assume that the effects are present in only a subset of studies, so that $\beta_i = 0$ for some values of i . In this case, the fixed effects test can be modified to (Han and Eskin, 2012)

$$S_m = \frac{\sum_{i=1}^N m_i \hat{\beta}_i \sigma_i^{-2}}{\sqrt{\sum_{i=1}^N m_i^2 \sigma_i^{-2}}},$$

where m_i are posterior probabilities that $\beta_i \neq 0$. Methods for estimating m_i have been described (Han and Eskin, 2012).

In the context of genome-wide meta-analysis, then, when estimation of the effect size is secondary to detection of the effect, the fixed effect model is preferable to the random effects model but can be further improved by methods that allow for some heterogeneity.

22.4.7 Using Funnel Plots to Investigate Small Study Bias in Meta-analysis

Meta-analysis can help to identify the presence of *publication bias*, a particular form of which is *small-study bias*. This problem arises because studies reporting significant results are more likely to be published than those without; the result is that small studies are only likely to be published if their estimated effect sizes are unusually large, reflecting the winner's curse phenomenon discussed earlier. This leads to an over-representation in the literature of small studies with significant effects. *Funnel plots* (Egger *et al.*, 1997) can aid the detection of small-study bias. The effect estimates from each study are plotted on the x -axis against their standard errors on the y -axis. Due to larger random error, effect estimates from smaller studies are more widely distributed around the overall mean effect. The degree of spread should steadily decrease as study size increases, and therefore the points on the plot should roughly fall into a funnel shape. If the likelihood of publication is greater for studies with smaller p -values (or larger effects) this plot may become skewed.

Figure 22.5 shows funnel plots for the D9N and HindIII data, with lines indicating 95% confidence intervals according to the fixed effect mean. While there is no obvious reason for concern for D9N, there is a clear suggestion of publication bias among the HindIII studies. Two of the smallest studies (Malygina and Thorn) report the largest effects and consequently lie outside the funnel – these effects are surprisingly large given the rest of the data.

Upon suspicion of publication bias a typical approach is to check the sensitivity of results by repeating the meta-analysis including larger studies only, for which publication bias is much less likely. Table 22.1 shows Hind III/CHD association results from meta-analyses excluding Malygina and Thorn, and of the largest studies only ($n > 500$). Under both analyses the evidence of effect is consistently and substantially diminished, losing significance at the 5% level. Therefore, publication bias does appear to be an issue in the meta-analysis of these data; any interpretation should take this sensitivity analysis into account.

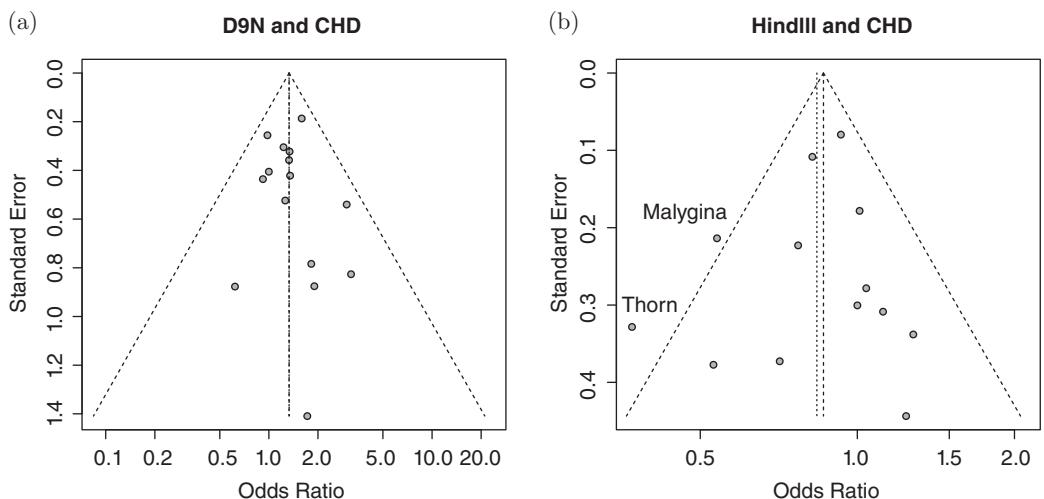


Figure 22.5 Funnel plots for D9N and HindIII meta-analyses.

Upon exclusion of the Malygina and Thorn studies, notice the similarity between fixed and random effects estimates, which is supported by small I^2 estimates. It is clear that these two outlying studies accounted for most of the heterogeneity in reported estimates, and upon exclusion a random effects analysis is no longer necessary.

Finally, we note that while funnel plots are a useful graphical tool it is important to keep in mind that asymmetry does not necessarily equate to publication bias. There are a number of other possible explanations, such as genuine heterogeneity, selective reporting, or poor methodological quality leading to spuriously inflated effects in smaller studies. For more detail see Sterne *et al.*'s (2011) comprehensive discussion of funnel plots and the potential pitfalls in their interpretation.

22.4.8 Improving Analyses via Meta-analysis Consortia and Publicly Available Data

Most meta-analyses are *retrospective*, in that they review and summarise the results of studies that have already been completed and published. While these analyses are of course very effective, they cannot control all aspects of study design and analysis. In contrast, a *prospective meta-analysis* allows control over many key design aspects from the outset. In these analyses, eligible studies are identified and their agreement to participate is confirmed before their results are known or published. Often the studies will form a consortium with central coordination and analysis teams. This model has become common in genetic epidemiology, because individual

Table 22.1 Sensitivity analysis for the meta-analysis of HindIII

Data	No. studies	Model	Mean	95% CI	p	I^2
All studies	13	Fixed effects	0.86	(0.78, 0.95)	0.003	
		Random effects	0.84	(0.72, 0.97)	0.021	0.37
Excluding Malygina and Thorn	11	Fixed effects	0.91	(0.82, 1.00)	0.062	
		Random effects	0.91	(0.82, 1.00)	0.062	< 0.001
Large studies only	3	Fixed effects	0.90	(0.80, 1.02)	0.090	
		Random effects	0.90	(0.80, 1.02)	0.090	< 0.001

studies have limited power to detect the small effects typical of genetic variants. Indeed, for studies of modest size the consortium is often the most realistic way to contribute to a significant discovery. Even for well-powered studies, it has become clear that larger samples lead to more discoveries (Visscher *et al.*, 2012), so that consortium work is of value to almost all research groups. Consortia are now particularly common for GWASs.

22.4.8.1 Better Quality Control

A major advantage of the prospective approach is that more control is possible over data quality, analysis and reporting. In particular, quality control is a key aspect of genome-wide association analysis, as even small errors in genotyping can lead to inflated false-positive rates. In a consortium it is possible to harmonise quality control criteria, or at least to ensure that participating studies have sufficiently rigorous control. Because quality control is closely related to laboratory genotyping protocols, it is often reasonable to allow participating studies to apply their local quality controls, but then to apply relatively liberal quality control to the pooled data at the meta-analysis stage, in order to ensure a baseline level across all studies.

Quality control metrics can then be compared between the participating studies to establish whether the data quality is reasonably uniform. For example, the overall genotyping call rate can be compared between studies, as can the distribution of allele frequencies. These measures can also be compared between groups of studies that use different genotyping platforms, or between case-control and family-based studies, and so on. Such comparisons are useful for ensuring a comparable level of data quality across all studies, or for identifying problematic studies that might bias the meta-analysis.

22.4.8.2 Imputation Using Publicly Available Reference Panels

Often, different studies have genotyped different sets of markers in the same gene or across the whole genome. Imputation is a widely used method which allows such studies to be analysed together (Marchini and Howie, 2010). In this method, a *reference panel* is available which contains genotypes from all the markers used in each study, genotyped on a small number of individuals. The 1000 Genomes Project has proved a very useful resource for this purpose (1000 Genomes Project Consortium, 2015); recently, the Haplotype Reference Consortium has released a more extensive and accurate reference panel for European populations (McCarthy *et al.*, 2016). Using the reference panel, a statistical model is fitted which predicts the genotype at one SNP given the genotypes at others (see Chapter 3). Thus, the SNPs present on one genotyping panel can be used to predict genotypes for SNPs present on another, so that a combined genotyping panel is effectively created for each study. In a consortium, the imputation can be performed by each participating study, which then provides the association data on each imputed SNP as well as those directly genotyped. Meta-analysis is then performed on every SNP genotyped in any study.

22.4.8.3 Mega-analysis

When studies have agreed to collaborate on prospective meta-analysis, it is possible to pool the individual-level data rather than just summary statistics as in standard meta-analysis. This approach, sometimes called ‘mega-analysis’, allows standard methods of analysis to be applied to the pooled data. It is also easy to perform subgroup analyses or to adjust for important covariates. Furthermore, the problem of population stratification can be addressed more directly. Usually, each study will have performed an adjustment for population stratification, using methods such as principal components analysis. However, these adjustments are often incomplete, and the remaining bias becomes magnified as studies are combined in meta-analysis. A common approach to reducing the residual bias is to add a covariate indexing each study to the

meta-analysis. With individual-level data, however, principal components adjustment can be applied to the entire combined data, giving a more direct and complete correction.

Sometimes an individual-level analysis is not possible. This may happen if the export of individual genotype data is excluded by the terms of a study, such as the consent of its subjects or the regulations of the country in which it is based. Combinations of family-based and population-based studies may also present practical obstacles that are most easily overcome by meta-analysis of summary data. Fortunately, when the sample size is large there is no loss of power in performing a fixed effects meta-analysis compared to an individual-level analysis (Lin and Zeng, 2010). The standard meta-analysis could therefore also be used as a first step even when individual data are available, prior to more accurate analyses as described above.

22.4.8.4 Sharing of Expertise

Many large consortia have been formed with the aim of improving the power of genetic association studies. Some of the largest are the Genetic Investigation of Anthropometric Traits (Locke *et al.*, 2015), Psychiatric Genomics Consortium (Psychiatric GWAS Consortium Steering Committee, 2009), Breast Cancer Association Consortium (Michailidou *et al.*, 2017) and CardioGramPlusC4D (Nikpay *et al.*, 2015). Apart from the improved statistical power offered, these consortia also allow researchers to share expertise and develop new methods more rapidly, and as many researchers are members of several consortia the development of best practice across genetic epidemiology is considerably accelerated.

22.5 Summary

Replication and meta-analysis are two important aspects in following up GWASs. Replication is a mandatory step in confirming the validity of associations, regardless of statistical significance in GWASs. In order to eliminate any biases in the GWAS analysis, replication in independent data sets is preferred, and may take the form of technical, direct or indirect replication. Technical replication verifies the integrity of the original data, while direct replication reproduces the association as closely as possible in independent data. Indirect replication, while not providing full support for the original association, provides evidence for its generalisability. Winner's curse, the upward bias of effect sizes in a discovery study, affects the power of replication studies and explains the reduced effects usually seen in replication studies and meta-analysis.

Meta-analysis is an increasingly prominent aspect of genetic epidemiology, as individual studies have limited power to detect genetic associations. Standard epidemiological methods of fixed and random effects meta-analysis can be used. Heterogeneity of effects may arise for various reasons including variation in LD patterns, case ascertainment and interaction effects. More powerful methods are available for detecting effects when heterogeneity is present. Imputation allows the combination of association studies that have genotyped different markers. Increasingly, consortia are being formed in advance of initial data analysis, allowing harmonisation of quality control criteria and sharing of individual-level data.

References

- 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**, 68–74.
- Antoniou, A.C. and Easton, D.F. (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. *Genetic Epidemiology* **25**, 190–202.

- Bowden, J. and Dudbridge, F. (2009). Unbiased estimation of odds ratios: Combining genomewide association scans with replication studies. *Genetic Epidemiology* **33**, 406–418.
- Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics* **10**, 101–129.
- DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.
- Dudbridge, F. (2016). Polygenic epidemiology. *Genetic Epidemiology* **40**, 268–272.
- Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology* **32**, 227–234.
- Dudbridge, F., Fletcher, O., Walker, K., et al. (2012). Estimating causal effects of genetic risk variants for breast cancer using marker data from bilateral and familial cases. *Cancer Epidemiology, Biomarkers & Prevention* **21**(2), 262–272.
- Easton, D.F., Pooley, K.A., Dunning, A.M., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093.
- Egger, M., Davey Smith, G., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.
- Erdmann, J., Grosshennig, A., Braund, P.S., et al. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nature Genetics* **41**, 280–282.
- Faye, L.L., Sun, L., Dimitromanolakis, A. and Bull, S.B. (2011). A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. *Statistics in Medicine* **30**, 1898–1912.
- Ferguson, J.P., Cho, J.H., Yang, C. and Zhao, H. Empirical Bayes correction for the winner's curse in genetic association studies. *Genetic Epidemiology* **37**, 60–68.
- Fleiss, J.L. (1986). Analysis of data from multiclinic trials. *Controlled Clinical Trials* **7**, 267–275.
- Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology* **31**, 288–295.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on an article by Browne and Draper). *Bayesian Analysis* **1**, 515–534.
- Goddard, M.E., Wray, N.R., Verbyla, K. and Visscher, P.M. (2009). Estimating effects and making predictions from genome-wide marker data. *Statistical Science* **24**, 517–529.
- Ghosh, A., Zou, F. and Wright, F.A. (2008). Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *American Journal of Human Genetics* **82**, 1064–1074.
- Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *American Journal of Human Genetics* **88**, 586–598.
- Han, B. and Eskin, E. (2012). Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics* **8**, e1002555.
- Higgins, J.P. and Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.
- Hirschhorn, J.N., Lohmueller, K., Byrne, E. and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine* **4**, 45–61.
- Jeffries, N.O. (2009). Ranking bias in association studies. *Human Heredity* **67**, 267–275.
- Lall, K., Magi, R., Morris, A., Metspalu, A. and Fischer, K. (2017). Personalized risk prediction for type 2 diabetes, the potential of genetic risk scores. *Genetics in Medicine* **19**, 322–329.
- Lin, D.Y. and Zeng, D. (2010). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic Epidemiology* **34**, 60–66.
- Locke, A.E., Kahali, B., Berndt, S.I., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206.

- Loos, R.J. and Yeo, G.S. (2014). The bigger picture of FTO: The first GWAS-identified obesity gene. *Nature Reviews Endocrinology* **10**, 51–61.
- Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511.
- McCarthy, S., Das, S., Kretzschmar, W., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279–1283.
- Michailidou, K., Lindstrom, S., Dennis, J., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94.
- NCI-NHGRI Working Group on Replication in Association Studies (2007). Replicating genotype-phenotype associations. *Nature* **447**, 655–660.
- Nikpay, M., Goel, A., Won, H.H., et al. (2015). A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130.
- Palmer, C. and Pe'er, I. (2017). Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genetics* **13**, e1006916.
- Parkes, M., Barrett, J.C., Prescott, N.J., et al. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics* **39**, 830–832.
- Prentice, R.L. and Qi, L. (2006). Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics* **7**, 339–354.
- Psychiatric GWAS Consortium Steering Committee (2009). A framework for interpreting genome-wide association studies of psychiatric disorders. *Molecular Psychiatry* **14**, 10–17.
- Sagoo, G.S., Tatt, I., Salanti, G., et al. (2008). Seven lipoprotein lipase gene polymorphisms, lipid fractions, and coronary disease: A HuGE association review and meta-analysis. *American Journal of Epidemiology* **168**, 1233–1246.
- Shah, T., Newcombe, P., Smeeth, L., et al. (2010). Ancestry as a determinant of mean population C-reactive protein values: Implications for cardiovascular risk prediction. *Circulation: Cardiovascular Genetics* **3**, 436–444.
- Shi, J., Park, J.H., Duan, J., et al. (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genetics* **12**, e1006493.
- Skol, A.D., Scott, L.J., Abecasis, G.R. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics* **38**, 209–213.
- Stacey, S.N., Manolescu, A., Sulem, P., et al. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature Genetics* **39**, 865–869.
- Sterne, J.A., Sutton, A.J., Ioannidis, J.P., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal* **343**, d4002.
- Sutton, A.J. and Abrams, K.R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10**, 277–303.
- Thomas, D., Xie, R. and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genetic Epidemiology* **27**, 401–414.
- Timpson, N.J., Lindgren, C.M., Weedon, M.N., et al. (2009). Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes 2009*; **58**, 505–510.
- Todd, J.A., Walker, N.M., Cooper, J.D., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics* **39**, 857–864.

- Turnbull, C., Ahmed, S., Morrison, J., *et al.* (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics* **42**, 504–507.
- Vimaleswaran, K.S., Tachmazidou, I., Zhao, J.H., Hirschhorn, J.N., Dudbridge, F. and Loos, R.J. (2012). Candidate genes for obesity-susceptibility show enriched association within a large genome-wide association study for BMI. *Human Molecular Genetics* **21**, 4537–4542.
- Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics* **90**, 7–24.
- Wang, H., Thomas, D.C., Pe'er, I. and Stram, D.O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology* **30**, 356–368.
- Wason, J.M. and Dudbridge, F. (2012). A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *American Journal of Human Genetics* **90**, 760–773.
- Willer, C.J., Li, Y. and Abecasis, G.R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191.
- Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
- Xu, L.Z., Craiu, R.V. and Sun, L. (2011). Bayesian methods to overcome the winner's curse in genetic studies. *Annals of Applied Statistics* **5**, 201–231.
- Zhong, H. and Prentice, R.L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634.

23

Inferring Causal Relationships between Risk Factors and Outcomes Using Genetic Variation

Stephen Burgess,^{1,2} Christopher N. Foley,¹ and Verena Zuber^{1,3}

¹ MRC Biostatistics Unit, University of Cambridge

² Cardiovascular Epidemiology Unit, University of Cambridge

³ School of Public Health, Imperial College London

Abstract

An observational correlation between a suspected risk factor and an outcome does not necessarily imply that interventions on levels of the risk factor will have a causal impact on the outcome (“correlation is not causation”). If genetic variants associated with the risk factor are also associated with the outcome, this increases the plausibility that the risk factor is a causal determinant of the outcome. However, if the genetic variants in the analysis do not have a specific biological link to the risk factor, then causal claims can be spurious. We introduce the Mendelian randomization paradigm for making causal inferences using genetic variants. We consider monogenic analysis, in which genetic variants are taken from a single gene region, and polygenic analysis, which include variants from multiple regions. We focus on answering two questions: when can Mendelian randomization be used to make reliable causal inferences, and when can it make relevant causal inferences.

23.1 Background

In this chapter we discuss how genetic variation can be used to make inferences about whether epidemiological relationships between modifiable risk factors and outcome variables (usually disease status) are causal or not. A risk factor is causal if intervening on the value of the risk factor leads to changes in the average value of the outcome (or the risk of disease for a disease outcome). Questions about causal relationships are fundamental to epidemiological research, such as to determine disease aetiology, to predict the impact of a medical or public health intervention, to guide drug development, to advise clinical practice, or to counsel on the impact of lifestyle choices.

23.1.1 Correlation and Causation

Observational correlations between risk factors and outcomes can arise for many reasons (Hernán and Robins, 2018). For example, individuals who drink more red wine tend to have lower incidence of coronary heart disease (CHD) risk than non-drinkers. It may well be that drinking red wine improves coronary health in a causal way; that is, an intervention to increase

red wine consumption in the population would lower CHD risk. However, it may also be that the relationship has no causal basis. An observational correlation can arise due to confounding (Greenland and Robins, 1986). Individuals in the population who drink red wine may be different than those who do not drink red wine in many ways that have nothing to do with the consumption of wine – for instance, they may be richer and take better care of their health. A confounder is a variable that is a common cause of the risk factor and the outcome. The existence of such variables will lead to correlations between risk factors and outcomes in the absence of a causal relationship. Observational correlations can also arise due to reverse causation. It may not be that increased red wine consumption leads to improved coronary health, but rather that individuals with pre-existing subclinical CHD are advised to reduce their consumption of alcohol due to related health problems.

Questions about causation in epidemiology are ideally addressed by randomized controlled trials (RCTs) (Rubin *et al.*, 1974). By randomizing individuals, we create two groups in the population that are exchangeable – that is, all variables are identically distributed on average between the two groups. A difference between the average outcome in the group allocated to the intervention and in the group allocated to the control can only arise due to the causal effect of the intervention. However, not all variables can be randomly allocated for practical or ethical reasons; it would not be feasible to allocate people at random to either consume or abstain from drinking red wine for the next 20 years to observe whether there was a difference in CHD rates between the groups. Hence, non-experimental or quasi-experimental approaches are required to judge the causal status of relationships. While there are statistical approaches for dealing with confounding, such as adjusting for confounders in a multivariable regression model, these methods typically make the assumption that all confounders are known and measured without error. Both the failure to adjust for a confounder, and inappropriate adjustment for a non-confounder, can lead to bias and Type I error inflation (i.e. a greater than expected rate of false positive findings) (Robins *et al.*, 1989; Christenfeld *et al.*, 2004).

23.1.2 Chapter Outline

The main approach that we discuss in this chapter is that of Mendelian randomization. Mendelian randomization is the use of genetic variants to address questions about the causal status of risk factors by making the assumption that the variants are instrumental variables (IVs) (Davey Smith and Ebrahim, 2003; Burgess and Thompson, 2015). While Mendelian randomization is not the only paradigm for assessing causal relationships using genetic information, it is the most direct assessment of causality. We will also discuss colocalization and linkage disequilibrium (LD) score regression, and how these approaches can help to address causal questions. Software code corresponding to the various methods discussed in this chapter can be found in the appendix to a recent review paper (Burgess *et al.*, 2018).

23.2 Introduction to Mendelian Randomization and Motivating Example

We introduce the technique of Mendelian randomization and the assumptions required using an example. C-reactive protein (CRP) is an acute phase reactant and part of the body's inflammatory response mechanism. Observationally, CRP is correlated with a multitude of traits, including cardiovascular risk factors and CHD risk (Emerging Risk Factors Collaboration, 2010). However, these observational associations do not imply that CRP is necessarily a causal risk factor for CHD. It could be that there are confounders – common determinants of the risk factor and the outcome – that give rise to the association. Alternatively, it could be that CRP

levels increase in response to subclinical CHD, an example of reverse causation. Before spending millions of dollars developing and then validating a drug that decreases CRP levels, it would be prudent to first ensure that decreasing CRP levels will lead to reductions in CHD risk.

Genetic variants in the *CRP* gene region are associated with CRP levels (italics indicate the gene name, non-italics indicate the biomarker). We consider rs1205, one particular single nucleotide polymorphism (SNP) variant in this region with alleles C and T. For each additional copy of the T allele for rs1205, individuals have a CRP level that is lower by an average of 0.169 standard deviations (95% confidence interval (CI) 0.153–0.185) (CRP CHD Genetics Collaboration, 2011). However, the rs1205 variant is not associated with CHD risk, with an odds ratio of 0.999 (95% CI 0.980–1.019) per additional copy of the T allele (CARDIoGRAMplusC4D Consortium, 2015). Subgroups in the population with different numbers of T alleles and therefore different average levels of CRP do not have different prevalences of CHD risk. This suggests that decreasing average CRP levels in the population would not lead to reductions in CHD risk, and that CRP is not a cause of CHD risk. By contrast, genetic associations with CHD risk for variants in the *IL6R* gene region (IL6R Genetics Consortium and Emerging Risk Factors Collaboration, 2012; Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium, 2012) and the *IL1RN* gene region (Interleukin-1 Genetics Consortium, 2015) provide evidence that interleukin-6 and interleukin-1 pathways respectively are implicated in CHD aetiology, and that these pathways are worthwhile therapeutic targets (as has now been demonstrated clinically for interleukin-1 (Ridker *et al.*, 2017)).

23.2.1 Instrumental Variable Assumptions

The risk factors CRP, interleukin-6 and interleukin-1 are excellent candidates for Mendelian randomization investigations because there are gene regions having strong and plausibly specific links with each risk factor: the coding regions for CRP (*CRP*), interleukin-6 receptor (*IL6R*), and interleukin-1 receptor antagonist (*IL1RN*), respectively. For a genetic variant to provide a reliable test of the causal status of a risk factor, it should be associated with the risk factor, but not associated with any confounders of the risk factor–outcome association, nor should it directly influence the outcome (Lawlor *et al.*, 2008; Didelez and Sheehan, 2007). This means that subgroups of the population defined by the genetic variant differ on average systematically only with respect to the risk factor and any downstream consequences of the risk factor.

The assumptions for a genetic variant to be an IV are as follows (Greenland, 2000):

- IV1 (Relevance). The average value of the risk factor depends on the genetic variant.
- IV2 (No association with confounders). The genetic variant is distributed independently of all confounders of the risk factor–outcome association.
- IV3 (No conditional association with outcome). The genetic variant is distributed independently of the outcome conditional on the confounders and the risk factor.

These assumptions imply the exclusion restriction condition – if the risk factor and confounders are kept constant, then intervening on the genetic variant will have no impact on the outcome (Angrist *et al.*, 1996). If the IV assumptions are satisfied, then the genetic variant can be treated like random allocation to a treatment group in an RCT (Hingorani and Humphries, 2005). This is known as a natural experiment, as allocation to ‘treatment’ or to ‘control’ is performed by nature (Davey Smith, 2006). Any association between the genetic variant and the outcome implies that the risk factor is a cause of the outcome in the same way as an intention-to-treat effect implies efficacy of the treatment in a randomized trial (Didelez and Sheehan, 2007). Under further parametric assumptions (sufficient conditions are linearity of relationships between variables and no effect modification (Hernán and Robins, 2006)), a causal effect of the risk factor on the outcome can be estimated.

23.2.2 Assessing the Instrumental Variable Assumptions

The IV assumptions are unlikely to be true for all genetic variants, and require a genetic variant to act like an unconfounded proxy measure of intervention on the risk factor. There are several reasons why this may not be the case (Davey Smith and Ebrahim, 2004; Glymour *et al.*, 2012; VanderWeele *et al.*, 2014). For example, the genetic variant may be pleiotropic – that is, it affects multiple risk factors on different causal pathways. Such a variant would not provide any specific information about intervening on the risk factor under analysis. Alternatively, a genetic variant may not directly affect the risk factor, but instead affect a precursor of the risk factor. If there is only one causal pathway from the genetic variant to the outcome and if this passes via the risk factor of interest, then the genetic variant would still be a valid IV (Burgess *et al.*, 2016). However, if there is an alternative pathway from the precursor of the risk factor to the outcome that does not pass via the risk factor, then the IV assumptions would be violated and the causal status of the risk factor cannot be reliably assessed. Other ways in which the IV assumptions may be violated have been well documented and include population stratification (the population under investigation consists of multiple subpopulations, such as different ethnic groups, meaning that genetic associations may reflect differences in allele frequencies between the subpopulations and not true biological relationships) and LD (the genetic variant is correlated with another variant that influences a competing risk factor) (Lawlor *et al.*, 2008).

The most straightforward way of assessing the IV assumptions is to test the association of the genetic variant with a range of potential confounders (Burgess *et al.*, 2015). However, this approach is far from foolproof. Firstly, not all relevant confounders may be known or measured. Secondly, a genetic association may reflect a downstream consequence of intervention on the risk factor itself, and not a pleiotropic effect (Figure 23.1). For example, variants in the *IL1RN* gene region linked with interleukin-1 are also associated with CRP and interleukin-6 (Interleukin-1 Genetics Consortium, 2015). However, pharmacological intervention on the interleukin-1 receptor antagonist pathway via the drug anakinra also leads to elevated CRP and interleukin-6 levels, suggesting that these associations may reflect mediation along a single causal pathway (also known as vertical pleiotropy), rather than effects on multiple causal pathways and hence pleiotropy (or horizontal pleiotropy). Thirdly, there is a multiple testing problem: one would not want to be overly conservative with testing for violations of the IV assumptions; however, if large numbers of potential confounders have been measured then some associations with the genetic variant may arise due to chance alone.

Alternative ways of assessing the IV assumptions are: (1) to consider positive and negative control variables (Lipsitch *et al.*, 2010; Burgess and Davey Smith, 2017) (as discussed above, CRP and interleukin-6 may be considered as positive controls for the intervention on interleukin-1); (2) to subset the population, particularly if there are subsets in the population that may have different distributions of the risk factor (see Section 23.5.7 for a fuller discussion); and (3) cross-ethnic analyses – if a risk factor is a cause of an outcome, then a causal effect should be evidenced across different populations, whereas patterns of LD and population stratification will differ between populations.



Figure 23.1 (Left) Diagram illustrating pleiotropy (horizontal pleiotropy): genetic variant is separately associated with the risk factor and covariate via different causal pathways. (Right) Diagram illustrating mediation (vertical pleiotropy): genetic variant is directly associated with the risk factor, and the association with the covariate is a downstream consequence of the risk factor.

23.2.3 Two-Sample Mendelian Randomization and Summarized Data

Two ongoing developments that are changing the way that Mendelian randomization analyses are performed are the increasing availability of summarized data from large consortia (Burgess *et al.*, 2015), and the development of large-scale population-based biobanks, such as UK Biobank (Sudlow *et al.*, 2015). Summarized data comprise beta coefficients and standard errors taken from regression (generally linear for continuous variables, logistic for binary variables) of the variable on each genetic variant in turn (Burgess *et al.*, 2016). Beta coefficients represent the average change in the trait per additional copy of the effect allele. Typically (and ideally for Mendelian randomization), these regression models additionally adjust for genetic principal components of ancestry, and sometimes for age and sex, but not for further variables (Bowden *et al.*, 2017). Several consortia, such as the Global Lipids Genetics Consortium for lipid fractions (Global Lipids Genetics Consortium, 2013) and the Coronary Artery Disease Genomewide Replication and Meta-analysis plus The Coronary Artery Disease (CARDIoGRAMplusC4D) consortium for CHD (Mahajan *et al.*, 2014), have made these association estimates publicly available for download; a collated queriable database of these associations can be found at www.phenoscanner.medschl.cam.ac.uk (Staley *et al.*, 2016). This has facilitated the implementation of Mendelian randomization analyses, which can be conducted using these summarized data, and in particular two-sample Mendelian randomization, in which genetic associations with the risk factor and with the outcome are obtained from separate data sources (Pierce and Burgess, 2013).

Two-sample Mendelian randomization is attractive as genetic associations with risk factors should be estimated in cohort or cross-sectional studies of healthy individuals (Bowden and Vansteelandt, 2011), whereas genetic associations with disease outcomes are best estimated in case–control studies. It also sidesteps some concerns about weak instrument bias, a version of winner’s curse in IV analysis (Burgess and Thompson, 2011). In a two-sample analysis, weak instrument bias is in the direction of the null and does not lead to inflated Type I error rates, as opposed to a one-sample analysis, in which bias is in the direction of the confounded association (Burgess *et al.*, 2016).

23.3 Monogenic Mendelian Randomization Analyses: The Easy Case

We divide Mendelian randomization analyses into two categories: monogenic analyses and polygenic analyses. By monogenic analyses, we mean investigations into the causal nature of risk factors using genetic variants from a single gene region. This contrasts with polygenic analyses, which include genetic variants from multiple gene regions.

Among all Mendelian randomization analyses, monogenic analyses (such as assessing the causal effect of CRP using variants in the *CRP* gene region) include the most reliable assessments of causal relationships. Particularly for protein risk factors, there is often a gene region that encodes either the risk factor itself or else a biologically relevant factor in the causal pathway relating to the risk factor (such as the interleukin-1 receptor antagonist for interleukin-1). These gene regions have the greatest plausibility for having specific associations with these risk factors, and for being reliable proxies to assess the effect of intervening on the same pathway by which the variant acts. Such investigations are also of most value to developers of pharmaceutical agents, as the genetic variant often highlights a pathway that can be targeted by pharmacological intervention (Plenge *et al.*, 2013).

However, there is a sense in which performing a monogenic Mendelian randomization analysis (which is essentially a candidate gene study) is to put all one’s eggs into a single basket. Of course, it is necessary to put the eggs into some basket, and in many cases it is better to

put them into a single reliable basket rather than several unreliable baskets. For example, multiple gene regions have been discovered that are associated with CRP at a genome-wide level of statistical significance. It would be possible to conduct Mendelian randomization analyses using each of these genetic variants. However, for some of these gene regions (such as *IL6R*), the allele that is associated with increased CRP is associated with increased CHD risk, and for other gene regions (such as *APOC1*, *LEPR* and *HNF1A*), the CRP-increasing allele is associated with decreased CHD risk (Burgess *et al.*, 2017). Clearly, CRP cannot simultaneously be both protective and harmful for CHD risk. In truth, it is difficult to justify looking beyond the *CRP* gene region to assess the specific causal role of CRP. However, in other cases, combining data from multiple genetic variants associated with a single risk factor may provide convincing evidence that the risk factor is causal even when the IV assumptions cannot be clearly justified for any one variant.

23.4 Polygenic Mendelian Randomization Analyses: The Difficult Case

Polygenic risk factors – risk factors that have multiple genetic determinants – include biomarkers that are more complex than proteins, such as low-density lipoprotein (LDL) cholesterol, exogenous biomarkers, such as calcium, and more complex multifactorial measures, such as blood pressure and body mass index (BMI).

23.4.1 Example: Low-Density Lipoprotein Cholesterol and Coronary Heart Disease Risk

The clearest example of a plausible Mendelian randomization analysis for a polygenic risk factor is for LDL cholesterol and CHD risk. There are several gene regions that contain candidate IVs: each either encodes a biologically relevant compound to LDL cholesterol, or is a proxy for an existing or proposed LDL cholesterol lowering drug. Figure 23.2 (left) shows genetic

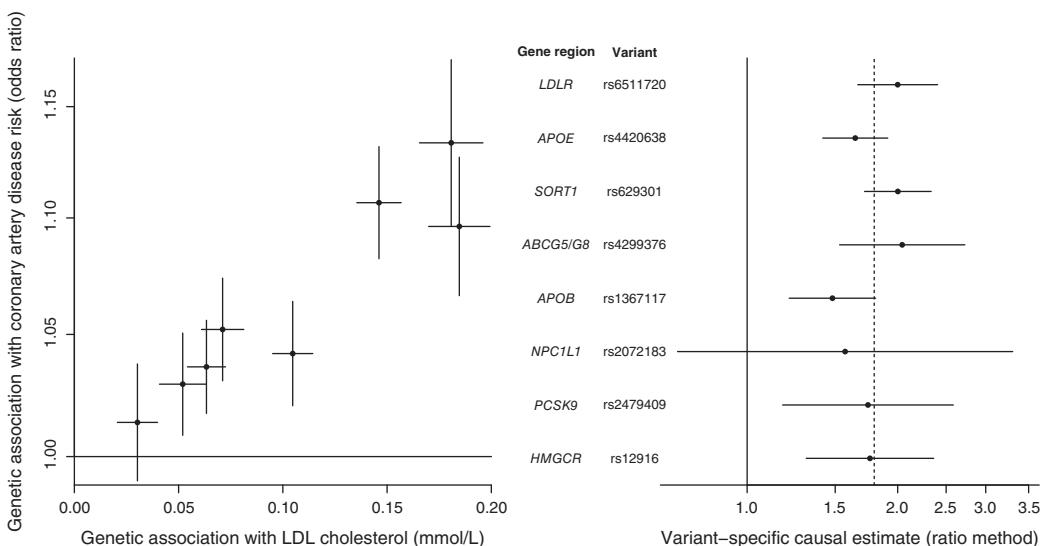


Figure 23.2 (Left) Genetic associations with risk factor (LDL cholesterol) and outcome (CHD risk) for eight genetic variants having biological links to LDL cholesterol. Lines are 95% confidence intervals. (Right) Variant-specific causal estimates (odds ratio for CHD per 1 mmol/L increase in LDL cholesterol) from ratio method for eight variants in turn. Solid lines are 95% confidence intervals, dashed vertical line is the inverse-variance weighted estimate.

associations with CHD risk (odds ratio per LDL-cholesterol-increasing allele) taken from the CARDIoGRAMplusC4D Consortium (CARDIoGRAMplusC4D Consortium, 2015) plotted against genetic associations with LDL cholesterol (per allele changes in LDL cholesterol) taken from the Global Lipids Genetics Consortium (Global Lipids Genetics Consortium, 2013) for variants in eight gene regions. These gene regions are: *HMGCR* (proxy for statin treatment), *PCSK9* (proxy for PCSK9 inhibition), *NPC1L1* (proxy for ezetimibe), *APOB* (encodes biologically relevant apolipoprotein B), *ABCG5/G8* (bile acid sequestrant), *SORT1* (antisense oligonucleotide RNA inhibitor targeting this pathway currently under development), *APOE* (encodes biologically relevant apolipoprotein E), and *LDLR* (encodes biologically relevant LDL receptor). The gradient of the line from the origin to each point is equal to the causal estimate (assumed to be linear on the log odds ratio scale) based on the corresponding genetic variant; this is known as the ratio estimate. It represents the change in the outcome per unit change in the risk factor.

We see not only that LDL-cholesterol-increasing alleles for each of the gene regions are all concordantly associated with increased CHD risk, but also that there is remarkable consistency in the causal estimates (Figure 23.2, right). Even though each variant affects LDL cholesterol via a different biological mechanism, the increase in CHD risk seems to depend only on the magnitude of change in LDL cholesterol. This consistency in the causal estimate has been observed to extend even for rare variants having much larger genetic associations with LDL cholesterol (Ference *et al.*, 2012).

Generally speaking, if multiple genetic variants in different gene regions all show the same direction of association with the outcome, then a causal relationship is plausible, particularly if the causal estimates based on the individual variants are all similar. As a corollary, if the causal estimates of several genetic variants are all similar, but there is one variant whose estimate differs sharply, then this genetic variant may be pleiotropic. However, it would seem unwise to make judgements about the validity of a genetic variant as an IV solely based on its associations with no reference to biological knowledge about the function of the variant; it is possible that the homogeneous variants are the invalid ones, or that the outlying variant provides important information relating to the question of causation.

23.4.2 More Complex Examples

As an example in which causal inferences are less clear, we consider associations of the same eight LDL-cholesterol-related genetic variants with risk of type 2 diabetes taken from the DIAGRAM consortium (Mahajan *et al.*, 2014). For seven of the eight variants, the LDL-cholesterol-increasing allele is associated with a decrease in the risk of type 2 diabetes (Figure 23.3, left). The exception is the variant in the *APOB* locus. Even among those variants suggesting a protective effect of increased LDL cholesterol, the variant-specific causal estimates were not consistent in magnitude (Figure 23.3, right), suggesting either mechanism-specific effects or that the effect of LDL cholesterol on type 2 diabetes risk also depends on particle size or some other aspect of lipid-related biology (Lotta *et al.*, 2016).

As examples in which causal inferences are even less clear, we consider Mendelian randomization analyses for high-density lipoprotein (HDL) cholesterol and CHD risk (Figure 23.4, left), and LDL cholesterol and Alzheimer's disease risk (Figure 23.4, right). Both of the analyses include all variants previously associated with the lipid fraction risk factor at a genome-wide level of statistical significance ($p < 5 \times 10^{-8}$) in the Global Lipids Genetic Consortium (Global Lipids Genetics Consortium, 2013). For HDL cholesterol, this is because there are few if any variants that have specific associations with HDL cholesterol and are not also associated with other lipid fractions. In the first case, while genetic variants having less strong associations with HDL cholesterol do seem to have protective associations on average with CHD risk, variants

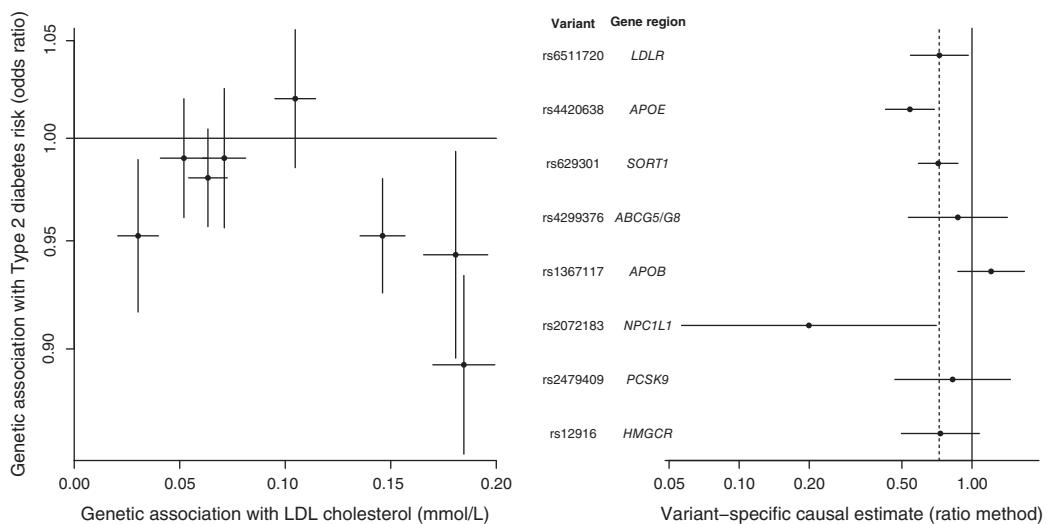


Figure 23.3 (Left) Genetic associations with risk factor (LDL cholesterol) and outcome (type 2 diabetes risk) for eight genetic variants having biological links to LDL cholesterol. Lines are 95% confidence intervals. (Right) Variant-specific causal estimates (odds ratio for type 2 diabetes per 1 mmol/L increase in LDL cholesterol) from ratio method for 8 variants in turn. Solid lines are 95% confidence intervals, dashed vertical line is the inverse-variance weighted estimate.

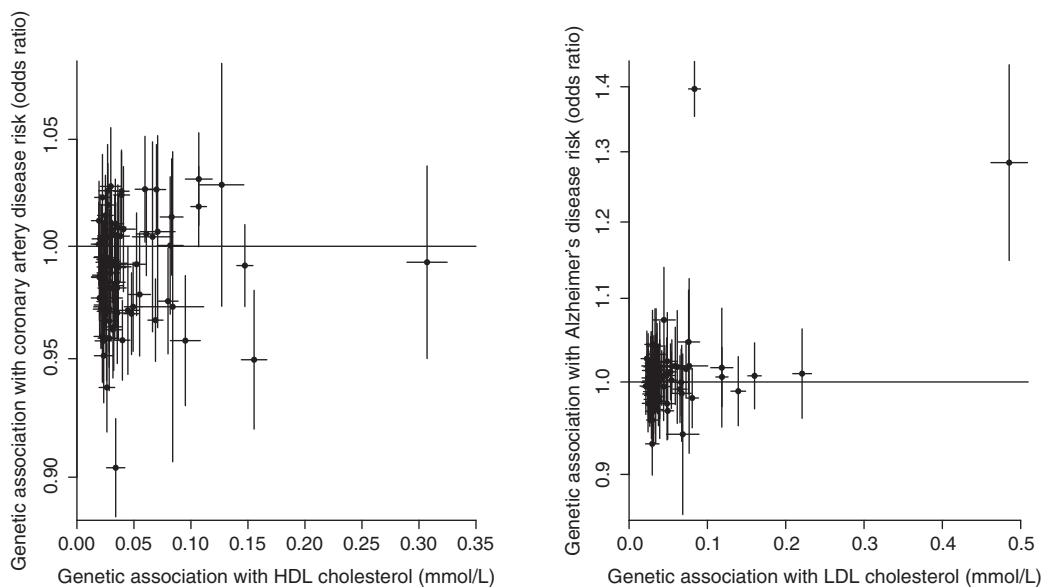


Figure 23.4 (Left) Genetic associations with risk factor (HDL cholesterol) and outcome (coronary heart disease risk) for 86 genetic variants. The lines are 95% confidence intervals for the genetic associations (all associations are orientated to the HDL-cholesterol-increasing allele). (Right) Genetic associations with risk factor (LDL cholesterol) and outcome (Alzheimer's disease risk) for 76 genetic variants. The lines are 95% confidence intervals for the genetic associations (all associations are orientated to the LDL-cholesterol-increasing allele).

having the strongest associations with HDL cholesterol do not also have the strongest associations with CHD risk. In the second case, while two of the genetic variants that are strongly associated with LDL cholesterol do have strong associations with Alzheimer's disease, none of the other variants are strongly associated with Alzheimer's disease, and the association estimates seem to be symmetrically distributed about the null. We return to these examples as we introduce various analysis methods.

23.4.3 Two-Stage Least Squares and Inverse-Variance Weighted Methods

To quantitatively combine evidence on a causal relationship from multiple genetic variants, we estimate the causal effect of the risk factor on the outcome. (The interpretation of this estimate is discussed in Section 23.7.) With a single genetic variant, the causal estimate (known as the ratio estimate) is obtained by dividing the beta coefficient for the association of the variant with the outcome ($\hat{\beta}_Y$) by the beta coefficient for the association of the variant with the risk factor ($\hat{\beta}_X$).

The most efficient assessment of causation when there are several IVs in terms of statistical power is the two-stage least squares estimate (Wooldridge, 2009). When only summarized data on the per allele genetic associations with risk factor and outcome are available, the same estimate can be obtained by the inverse-variance weighted method, which combines the association estimates (beta coefficients and standard errors) into a single estimate of the causal effect (Burgess *et al.*, 2016). This can be obtained from inverse-variance weighted meta-analysis of the ratio estimates from each genetic variant (Burgess *et al.*, 2013). Alternatively, the estimate can be calculated by weighted linear regression of the genetic associations with the outcome ($\hat{\beta}_{Yj}$ for genetic variant j) on the genetic associations with the risk factor ($\hat{\beta}_{Xj}$), using the reciprocals of the variances of the genetic associations with the outcome ($\text{se}(\hat{\beta}_{Xj})^{-2}$) as weights:

$$\hat{\beta}_{Yj} = \theta \hat{\beta}_{Xj} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, \text{se}(\hat{\beta}_{Yj})^2), \quad (23.1)$$

where θ is the causal effect parameter (Thompson and Sharp, 1999). If the genetic variants are correlated in their distributions, then these correlations should be accounted for in the regression model using generalized weighted least squares regression (Burgess *et al.*, 2016).

The estimate from the two-stage least squares/inverse-variance weighted method is a weighted mean of the ratio estimates based on the individual genetic variants, each of which is a consistent estimate of the causal effect when the genetic variant is a valid IV. As such, the inverse-variance weighted method generally only provides a consistent estimator of the causal effect of the risk factor on the outcome if all the genetic variants are valid IVs, and is a biased estimate if one or more of the variants is invalid.

We next consider in turn different methods that can make reliable causal inferences under weaker assumptions. We assume for simplicity of discussion that all the genetic variants are uncorrelated in their distributions; this is usually achieved by including one genetic variant from each gene region in the analysis. If two variants are perfectly correlated, then including both in a Mendelian randomization analysis will not provide additional information. However, if two partially correlated variants each explain independent variance in the risk factor, then precision can be increased by including both variants; extensions to several of these methods to allow for correlated variants have been developed (Burgess *et al.*, 2016; Rees *et al.*, 2017). We also make linearity and homogeneity assumptions as discussed elsewhere (Bowden *et al.*, 2017), in particular assuming that all valid IVs estimate the same causal parameter (see Section 23.7.2 for further discussion of this point).

23.5 Robust Approaches for Polygenic Mendelian Randomization Analyses

A robust method for Mendelian randomization is one that does not require all genetic variants to be valid IVs to give consistent estimates of a causal parameter (Burgess *et al.*, 2017).

23.5.1 Median Estimation Methods

A conceptually straightforward method that provides consistent estimates under the assumption that at least 50% of the genetic variants are valid instruments is the simple median method (Bowden *et al.*, 2016). The median estimate is obtained by first calculating the ratio estimates based on each genetic variant, and then taking the median of these estimates. With an infinite sample size, the genetic associations for each valid instrument should lie on a straight line through the origin. The gradient of this line will be the true causal effect, and will correspond to the median estimate provided that at least half of the genetic variants are valid instruments (Han, 2008). With a finite sample size, the simple median estimate still provides a causal estimate that is supported by the majority of the genetic variants. The median estimate is also not sensitive to genetic variants with outlying estimates which may be pleiotropic – a variant contributes in the same way to an analysis if its estimate is slightly above the median or a long way above. A weighted median method has also been proposed that assigns more weight in the analysis to genetic variants having more precise ratio estimates (Bowden *et al.*, 2016).

23.5.2 Modal Estimation Methods

A related concept is the modal estimation method. With an infinite sample size, in the case where exactly 50% of genetic variants estimate one value and 50% of the genetic variants estimate another value, it would not be possible to determine which of those values is the true causal effect. However, if the invalid genetic variants all estimated different values, then the true causal effect could be identified even when less than 50% of the genetic variants are valid instruments (Guo *et al.*, 2018). If the largest group of genetic variants having the same causal estimate are the valid instruments (referred to by Hartwig *et al.* (2017) as the zero modal pleiotropy assumption), then with an infinite sample size, the modal variant-specific ratio estimate would be the true causal effect.

With finite data, no two estimates would be exactly the same, so a modal estimate cannot be considered directly. Hartwig *et al.* (2017) consider a kernel-density smoothed estimate of the distribution of the variant-specific causal estimates (the ratio estimates), and take the maximum of this distribution as their modal causal estimate. Burgess *et al.* (2018) consider a model averaging procedure giving a mixture distribution based on the causal estimates from each subset of variants, and taking the maximum of this distribution as the modal causal estimate. The model averaging procedure has a number of technical advantages over the kernel-smoothed approach in that it is asymptotically efficient, it does not require the specification of a bandwidth parameter, and it allows for uncertainty in determining which peak is the global maximum in its confidence interval. As a consequence, the method is able to identify the presence of multiple subsets of variants with similar causal estimates, suggesting the existence of multiple causal mechanisms by which the risk factor influences the outcome.

23.5.3 Regularization Methods

A further class of methods invoking a similar assumption for consistent estimation (that is, some variants are valid IVs but we do not know which) uses regularization to identify valid IVs.

Regularization is the use of an external condition to fit a statistical model when there are more parameters than datapoints. Kang *et al.* (2016) consider a statistical model in which genetic variants influence an outcome variable via their associations with a risk factor, but also via direct (pleiotropic) effects on the outcome. These effects cannot all be identified, as the number of parameters is equal to the number of genetic variants plus one (one pleiotropic effect parameter for each variant plus the causal effect parameter). To provide identification, they use L1-penalization (also known as the lasso) with a tuning parameter to control the total of the absolute values of all the pleiotropic effects. A related method using adaptive lasso is considered by Windmeijer *et al.* (2016). We describe a simple implementation of this approach with summarized data (Burgess *et al.*, 2016). The inverse-variance weighted method (equation (23.1)) minimizes the weighted sum of squares,

$$\hat{\theta}_{IVW} = \arg \min_{\theta} \sum_j se(\hat{\beta}_{Yj})^{-2} (\hat{\beta}_{Yj} - \theta \hat{\beta}_{Xj})^2. \quad (23.2)$$

We propose adding a separate term θ_0 for each genetic variant that can be interpreted as the pleiotropic effect of variant j . If the term is equal to zero, this implies that the variant is a valid IV. For identification, we add an L1-penalty term:

$$\hat{\theta}_{LASSO,\lambda} = \arg \min_{\theta_L} \left(\sum_j se(\hat{\beta}_{Yj})^{-2} (\hat{\beta}_{Yj} - \theta_{0j} - \theta_L \hat{\beta}_{Xj})^2 + \lambda \sum_j |\theta_{0j}| \right), \quad (23.3)$$

where λ is a tuning parameter and minimization is across both θ_0 and θ_L . When λ is set to zero, the parameters are not identified; when λ tends to infinity, the inverse-variance weighted estimate is recovered. By considering estimates of the causal effect $\hat{\theta}_{LASSO,\lambda}$ across a range of values for λ , we can compare causal inferences based on different numbers of genetic variants, with other variants being allowed to have pleiotropic effects.

Alternatively, Bayesian methods to analyse the same statistical model have been developed. These approaches use a prior for formal identification of the statistical model, and to force the pleiotropic effects θ_{0j} to take values close to zero unless there is strong evidence of pleiotropy. Methods have been proposed using a horseshoe prior (Berzuini *et al.*, 2018) and a spike-and-slab prior (Li, 2017).

23.5.4 Other Outlier-Robust Methods

Another way of downweighting the contribution to the analysis of genetic variants that are heterogeneous from the other variants in terms of their causal estimate is the use of outlier-robust methods. These include simple approaches, such as leave-one-out estimation, and more sophisticated approaches, such as the use of robust regression rather than standard ordinary least squares regression. The use of MM estimation combined with Tukey's bisquare objective function provides inference that is robust to outliers and influential points (Mosteller and Tukey, 1977; Huber and Ronchetti, 2011), and has been shown to reduce Type I error rates in simulations with invalid IVs (Burgess *et al.*, 2016).

23.5.5 MR-Egger Method

The MR-Egger method (Bowden *et al.*, 2015) was the first robust method for Mendelian randomization to become widely used, and has become the *de facto* sensitivity analysis of choice in many people's minds. This is somewhat unfortunate, as the assumption required for the method

to give consistent estimates and valid causal inferences is not always plausible, and the MR-Egger method is quite fragile to departures from this assumption, as well as being sensitive to influential points in the regression model (Burgess and Thompson, 2017), having low power in many cases (Bowden *et al.*, 2016), and being subject to more severe weak instrument bias than other IV methods in a one-sample setting (Hartwig and Davies, 2016). However, in its defence, the MR-Egger method allows all genetic variants to be invalid IVs provided that they satisfy a different untestable assumption. Also, it is valuable to consider alternative assumptions for causal inference rather than multiple assumptions that are different flavours of ‘some genetic variants are valid instruments’.

In the MR-Egger method, we fit a regression model similar to that for the inverse-variance weighted method, except with an intercept term θ_0 :

$$\hat{\beta}_{Yj} = \theta_0 + \theta_E \hat{\beta}_{Xj} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, se(\hat{\beta}_{Yj})^2). \quad (23.4)$$

The intercept term is similar to the pleiotropic effect parameters in equation (23.3), except that there is a single pleiotropy term θ_0 , rather than a separate term for each genetic variant. This term represents the average pleiotropic effect of a genetic variant (the expected association of a variant with the outcome if its association with the exposure is zero) (Bowden *et al.*, 2016). If all the genetic variants are valid instruments, then this term should tend to zero asymptotically. Indeed, rejection of the statistical test that this term equals zero (known as ‘directional pleiotropy’) implies that the inverse-variance weighted estimate is biased (Burgess and Thompson, 2017). This is the MR-Egger intercept test. It provides a valuable test of the validity of the set of genetic variants as IVs.

Additionally, if the genetic variants satisfy a condition known as the instrument strength independent of direct effect (InSIDE) assumption, then the MR-Egger slope estimate is a consistent estimate of the causal effect even if there is directional pleiotropy (Bowden *et al.*, 2015). The genetic associations with the outcome can be decomposed into an indirect component via the risk factor and a direct (pleiotropic) component:

$$\beta_{Yj} = \theta \beta_{Xj} + \alpha_j. \quad (23.5)$$

The InSIDE assumption states that the pleiotropic effects of genetic variants (α_j) are distributed independently of the genetic associations with the risk factor (β_{Xj}). This assumption would be plausible if pleiotropic effects of genetic variants were to act directly on the outcome via mechanisms unassociated with the risk factor or confounders of the risk factor–outcome association, but is less plausible when pleiotropic effects act via related pathways or via confounders (Burgess and Thompson, 2017). The InSIDE assumption is not testable, and violation of the assumption can lead to anomalous results – for example, the estimated average pleiotropic effect can be larger than the associations of all variants with the outcome (Lee *et al.*, 2016), an implausible situation. Hence, while the MR-Egger method is a useful one, particularly for detecting pleiotropy via the intercept test, it should not be relied on as a primary analysis method, and its findings should be weighed carefully (and particularly when they differ sharply from estimates from other methods), for example by judging whether they are particularly influenced by a single observation.

For the example of HDL cholesterol and CHD risk (Figure 23.4), the straightforward inverse-variance method that assumes all genetic variants are valid instruments suggests that HDL cholesterol is causally protective of CHD, with an odds ratio estimate of 0.85 (95% CI 0.76–0.95) per standard deviation increase in HDL cholesterol. In contrast, the weighted median method gives OR = 0.95 (95% CI 0.87–1.05), and the MR-Egger method gives OR = 1.10 (95% CI 0.93–1.31) with an intercept term that differs from zero ($p = 0.0004$), suggesting no evidence for a

causal relationship. In contrast, all three methods agree that LDL cholesterol is a harmful risk factor for CHD risk (recall Figure 23.2), even when all its genome-wide significant predictors are used in the analysis (Burgess and Davey Smith, 2017).

23.5.6 Multivariable Methods

Multivariable Mendelian randomization (Burgess and Thompson, 2015) is an alternative approach for causal inference that can be used when it is difficult to find genetic variants specifically and uniquely associated with particular risk factors, but it is possible to find genetic variants specifically associated with a set of risk factors. For example, it is difficult to find genetic variants that are associated with HDL cholesterol but not also associated with LDL cholesterol and triglycerides (Burgess *et al.*, 2014). In multivariable Mendelian randomization, genetic variants are allowed to be associated with multiple measured risk factors, so long as they are not associated with confounders of any of the risk factor–outcome associations and they do not directly affect the outcome – any genetic association with the outcome is mediated via one or more of the risk factors.

We assume that summarized data are available on the associations of genetic variants with each risk factor in turn ($\hat{\beta}_{Xj1}, \hat{\beta}_{Xj2}, \dots, \hat{\beta}_{XjK}$ for each variant $j = 1, 2, \dots, J$ with each risk factor $k = 1, 2, \dots, K$). The inverse-variance weighted method can be extended to a multivariable weighted regression model:

$$\hat{\beta}_{Yj} = \theta_1 \hat{\beta}_{Xj1} + \theta_2 \hat{\beta}_{Xj2} + \dots + \theta_K \hat{\beta}_{XjK} + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, \text{se}(\hat{\beta}_{Yj})^2). \quad (23.6)$$

The parameters $\theta_1, \theta_2, \dots, \theta_K$ represent the direct causal effects of each risk factor in turn on the outcome (the effect of varying the corresponding risk factor while keeping all other risk factors fixed). By contrast, the MR-Egger method allows for unmeasured pleiotropic effects, and multivariable Mendelian randomization assumes that all pleiotropic effects are measured and accounted for. These methods can be combined into a multivariable MR-Egger method that accounts for both measured and unmeasured pleiotropy (under the InSIDE assumption) by including an intercept term in equation (23.6) (Rees *et al.*, 2017).

Multivariable Mendelian randomization can also be used to consider problems of mediation to unravel complex aetiological pathways (Burgess *et al.*, 2017). It has been used to demonstrate that the causal effect of age at menarche on breast cancer risk can be decomposed into a harmful indirect effect via decreased BMI and a protective direct effect not via BMI (Day *et al.*, 2017). Another area of application in which this method can be used is when there are multiple biomarkers influenced by several genetic variants in a single region. For example, multivariable Mendelian randomization suggested that genetic associations with atopic dermatitis of variants in the *IL1RL1*–*IL18R1* locus on chromosome 2 are driven by *IL18R1* and *IL1RL2* rather than by *IL1RL1* (Sun *et al.*, 2018).

23.5.7 Interactions and Subsetting

Another approach for assessing the validity of causal inferences that could also be used with a single genetic variant is exploiting subsets within the data set. For example, in East Asian societies, women tend not to drink alcohol for cultural reasons. Hence, if alcohol is a cause of a disease outcome, then genetic variants that influence alcohol metabolism should be associated with the outcome in men, but not in women, as there is no mechanism by which the genetic variants would affect the outcome in individuals who have zero exposure to alcohol (Cho *et al.*, 2015). If the same genetic associations were present in both men and women, this would suggest

that they are driven not by alcohol consumption, but rather by a pleiotropic mechanism (van Kippersluis and Rietveld, 2017).

Additionally, stronger genetic associations with oesophageal cancer risk have been observed in populations that drink heavily, and no association has been observed in men who abstain from alcohol (Lewis and Davey Smith, 2005). Some care is needed here, as the decision to abstain from alcohol is self-determined, and hence conditioning on abstinence may lead to selection bias (Hernán *et al.*, 2004). However, in this case the differences between the genetic associations in drinkers and non-drinkers are so extreme that it is highly unlikely that they are explained by selection bias alone. Conditions required for the subsets to satisfy have been provided; ideally the genetic associations with the risk factor should be present in one subset, but not in the other, but any pleiotropic effects of the genetic variants should be identical across subsets (van Kippersluis and Rietveld, 2017).

This idea has been extended to interactions with a continuous variable under the name Slichter regression (Spiller *et al.*, 2018). Similar conditions are required – genetic associations with the risk factor should depend on the interacting variable, but any pleiotropic effects should not be subject to modification by the interacting variable.

23.5.8 Practical Advice

In practice, we would encourage investigators to plot the genetic associations with the risk factor against those with the outcome (as in Figure 23.2, left) as a routine part of polygenic Mendelian randomization investigations. This provides a visual check for heterogeneity, and enables both investigators and readers to judge whether there is general consistency in genetic associations, and also whether any claimed causal effect is evidenced by all of the genetic variants, or just a subset of genetic variants. Formal tests for heterogeneity are also available (Greco *et al.*, 2015; Bowden *et al.*, 2018). Other plots that have been suggested for investigating pleiotropy include forest plots (Lewis and Clarke, 2001) and funnel plots (Sterne *et al.*, 2011), taken from the meta-analysis literature.

While we have presented a long list of robust methods for Mendelian randomization, we would not expect every Mendelian randomization investigation to include all of these analyses. Indeed, we would encourage investigators to think carefully about which sensitivity analyses would be most appropriate given their particular investigation. For example, if there are genetic variants that are likely to be outliers or influential points (such as variants in the *APOE* gene region for Alzheimer's disease, or in the *FTO* gene region for BMI), then an outlier-robust approach may be more appropriate than the MR-Egger method.

As an example of how robust methods can reveal inconsistencies in a Mendelian randomization investigation, we consider the investigation into the causal effect of LDL cholesterol on Alzheimer's disease using 380 genetic variants reported by Benn *et al.* (2017). Causal estimates represent odds ratios per 1 mmol/L decrease in LDL cholesterol. The authors initially reported a highly significant association based on the inverse-variance weighted method (OR 0.83, 95% CI 0.75–0.92) and the MR-Egger method (OR 0.64, 95% CI 0.52–0.79). However, the weighted median method gave a much different estimate (OR 0.97, 95% CI 0.91–1.02). Further investigation showed that the variants suggesting a positive effect were all in and around the *APOE* gene region, a known risk factor for Alzheimer's disease (note that the two variants strongly associated with Alzheimer's disease in Figure 23.4 are both in the *APOE* gene region). Analyses excluding these variants suggested no causal effect in either the inverse-variance weighted method (OR 0.98, 95% CI: 0.93–1.03) or the MR-Egger method (OR 0.98, 95% CI 0.87–1.09) (Benn *et al.*, 2017; see the Rapid Response written by the authors).

23.6 Alternative Approaches for Causal Inference with Genetic Data

Mendelian randomization is not the only methodological approach for considering causal relationships between risk factors and outcomes, although it is the most direct way of assessing the causal effect of a risk factor on an outcome. Alternative approaches include colocalization and LD score regression.

23.6.1 Fine-Mapping and Colocalization

Fine-mapping is a statistical approach to find genetic variants that causally affect a trait (Benner *et al.*, 2016). We note that a causal genetic variant is not required in a Mendelian randomization investigation – the first IV assumption only requires a genetic variant to divide the population into groups with different average levels of the risk factor (Hernán and Robins, 2006).

Colocalization methods take a single region of the chromosome that is associated with two traits and ask whether the same genetic variants are driving the associations with both traits (Solovieff *et al.*, 2013). As such, they are an extension of fine-mapping to multiple traits. If the same genetic variants (or variant) that give rise to the association with trait A also give rise to the association with trait B, then the same biological mechanism is likely to be responsible for both associations. This does not necessarily mean that trait A causes trait B (or vice versa), as it may be that a trait C causes both A and B. However, even in this case, colocalization implies a shared aetiology between traits A and B.

As an example, colocalization showed that a genome-wide association study ‘hit’ in the *CTSH* gene region drove genetic associations with both type 1 diabetes and narcolepsy (Guo *et al.*, 2015). The same variant was also associated with gene expression, but localized to monocytes and not to B cells. These results demonstrate the utility of genetic association data for highlighting causal cell types and potential druggable mechanisms.

Differences between the assumptions made in Mendelian randomization and in colocalization are illustrated in Figure 23.5. In some cases, the statistical methodology used in the two approaches will be identical (Wallace, 2013); differences in the conclusions (for example, concluding that a risk factor is a cause of the outcome) are conceptual and arise solely due to different prior assumptions.

Colocalization methods are useful when a particular section of the chromosome has multiple genes in close proximity (Hormozdiari *et al.*, 2014; Giambartolomei *et al.*, 2014). It may be that genetic variants that are associated with a risk factor also show an association with a disease outcome. This would lead to a positive finding in a Mendelian randomization analysis. However, if the colocalization analysis showed that the peaks of the associations were being driven by different variants, then this would suggest that the risk factor is not the relevant causal agent. This could result from LD between two functional variants. It could also be that the relevant measure of the risk factor has not been identified (for example, suppose that the risk factor is not LDL cholesterol concentration but rather a measure of LDL particle size).

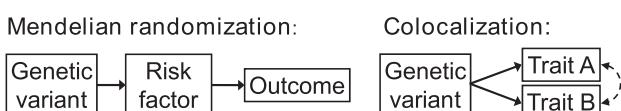


Figure 23.5 Schematic diagram illustrating different assumptions made in Mendelian randomization (in which the genetic variant is assumed to associate directly with the risk factor and with the outcome only via the risk factor) and in colocalization (in which the genetic variant is allowed to associate with both traits directly, and causal effects may occur between the traits in either direction or not at all).

Overall, colocalization focuses on a different causal question than does Mendelian randomization, one that focuses on a single gene region rather than considering the totality of evidence for a relationship between a risk factor and an outcome across multiple gene regions. However, the more detailed interrogation of a single gene region performed in the colocalization method can provide additional evidence to link a risk factor to an outcome. Colocalization can also be performed between two disease outcomes, to assess whether the diseases have a common aetiological link even in the absence of a hypothesized mechanism or risk factor for that link.

23.6.2 LD Score Regression

There are two forms of LD score regression: with a single trait (Bulik-Sullivan *et al.*, 2015), and with a pair of traits (Bulik-Sullivan *et al.*, 2015). The aim of the method is to provide estimates of heritability (shared heritability for two traits) that are unaffected by population stratification. Both approaches make use of LD scores. These are obtained for each variant in turn by taking a window about the target variant, calculating the squared correlation for each variant in this window with the target variant, and summing these squared correlations. A genetic variant with a high LD score is a variant with a high degree of ‘friendliness’ – it has many nearby variants in high LD with it. If biologically relevant genetic variants are uniformly distributed across the whole genome, then a genetic variant with a high LD score is more likely to be correlated with a biologically relevant variant, and hence more likely to be associated with phenotypic traits, including risk factors and outcomes. In contrast, genetic variants that are subject to population stratification would not be more or less likely to have high LD scores, so any genetic associations due to population stratification would be independent of LD score.

With a single trait, LD score regression is implemented by regression of the chi-squared statistics for the genetic associations with the trait against the LD scores for genetic variants across the whole genome (Bulik-Sullivan *et al.*, 2015). The intercept term is interpreted as ‘confounding’ (although by ‘confounding’ the authors are referring to population stratification and cryptic relatedness between individuals rather than conventional epidemiological confounding), and the slope coefficient is interpreted as a revised estimate of heritability (‘narrow-sense heritability’) that excludes population stratification. Weights are used in the regression model to correct for correlations between variants and for the varying precisions of the chi-squared statistics. With two traits, ‘cross-trait’ LD score regression is implemented by multiplying the z -statistic for the genetic association with trait 1 by the z -statistic for the genetic association with trait 2, and regressing this product of statistics against the LD scores (Bulik-Sullivan *et al.*, 2015). The slope coefficient (representing shared heritability or genetic correlation) is large when the same genetic variants predict both of the traits. Regression on the LD score should ensure that estimates of heritability are driven by biologically relevant variants rather than variants subject to population stratification. However, an implicit assumption in the method that variants contribute equally to heritability estimates may be violated in realistic settings, leading to bias and incorrect standard errors (Speed *et al.*, 2017).

LD score regression differs from Mendelian randomization in several ways. Firstly, LD score regression is not just a polygenic method, it is a massively polygenic method, using genetic variants from the whole genome. Secondly, LD score regression is symmetric in its two traits, whereas Mendelian randomization assesses the effect of the risk factor on the outcome. It does not seek to model one trait as a function of the other, but rather uses the LD score as the dependent variable in the regression model.

However, there are also connections between the approaches. In the MR-Egger method, the intercept term in the regression model is viewed as a nuisance parameter relating to the degree of pleiotropy, and the slope parameter is an estimate of the causal effect. In LD score regression,

the intercept term is again a nuisance parameter, and the slope parameter is of interest. This suggests that a version of the InSIDE assumption is required by the LD score regression paper (no correlation between the residual terms in the LD score regression equation and the LD scores). The problem of low power in the MR-Egger method is often sidestepped in cross-trait LD score regression by specifying rather than estimating the intercept term.

A criticism of LD score regression is that every analysis for each pair of traits uses the same LD scores as the dependent variable in the regression model (and as LD scores have been pre-computed by its proponents, literally the same LD scores are used in the majority of applied analyses). This means that any influential points in the regression will affect not only one LD score regression analysis, but all such analyses. LD scores are also likely to be a ‘weak instrument’ in the language of Mendelian randomization, as they will only explain a small proportion of variance in the dependent variable. Additionally, due to the scale of the data, it is not possible to provide a visual representation of an LD score regression analysis. Standard regression diagnostics are rarely, if ever, performed. Finally, results from LD score regression are not always consistent with known causal relationships; for example, the method did not find evidence for a genetic correlation between LDL cholesterol and CHD risk that survived a multiple-testing correction (Bulik-Sullivan *et al.*, 2015). The method has utility in mapping the genetic distance between related phenotypes, such as determining how closely related different psychiatric disorders are in terms of their genetic predictors (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). However, the reliance of the method on numerous linearity and independence assumptions, incorrect weighting in the linear regression model (correct weights would require computation of the Cholesky decomposition of a matrix with dimension equal to the number of genetic variants in the model – misspecified weights are recommended for use in practice), and lack of validation against known causal relationships mean that results from the method should not be treated too seriously as an assessment of causality.

23.7 Causal Estimation in Mendelian Randomization

A Mendelian randomization investigation has two related aims: firstly, to assess whether a risk factor has a causal effect on an outcome; and secondly, to estimate the magnitude of that causal effect. Usually, the first aim is the primary one, with estimation being a secondary aim. As we discuss below, the causal estimate obtained from Mendelian randomization is not always a reliable guide as to the expected impact of intervening on the same risk factor in practice. We discuss various issues relating to the interpretation of causal estimates from Mendelian randomization.

Aside from violations of the IV assumptions, several factors can lead to bias in Mendelian randomization estimates, particularly in a two-sample setting, as estimates of genetic associations may differ between the two data sets (Burgess *et al.*, 2016; Bowden *et al.*, 2017). However, not all biases are equally important – biases that affect estimates but without affecting the validity of causal inferences are less important, as causal estimates from Mendelian randomization are not always reliable even in the absence of bias. Whereas biases that lead to Type I error rate inflation are more important, as they will lead to a greater than expected rate of false positive findings from applied investigations.

23.7.1 Relevance of Causal Estimate

While Mendelian randomization aims to address whether a risk factor is a cause of an outcome, in reality this is not a well-defined question. For example, in answering the question ‘would

lowering BMI reduce cardiovascular mortality?', several further questions arise. How is BMI proposed to be lowered? – by increasing metabolism? by suppressing appetite? How long is BMI proposed to be lowered for? To whom will the intervention be applied? To answer the question 'is high BMI a cause of increased cardiovascular mortality?', one first has to pose the question in a precise way.

A Mendelian randomization investigation compares subgroups in the population defined by their genetic variants. As such, it is analogous to an RCT, but one in which the allocation to a 'treatment group' is determined at conception. In a randomized trial, typically the treatment that is assessed in the trial is also the one that is offered to patients. However, in Mendelian randomization, the 'intervention' that is assessed is changing an individual's genotype at conception, whereas the proposed treatment is to change the value of the risk factor by a different mechanism (that is, not by altering their genes) and at a later timepoint (usually in adulthood) (Swanson *et al.*, 2017). As such, Mendelian randomization estimates (and, equivalently, assessments of causal relationships) typically represent the impact of a long-term (or even a life-long) intervention in the risk factor by a small amount in a primary prevention context (Burgess *et al.*, 2012). In many cases, the Mendelian randomization investigation does address a relevant causal question. For example, one may be interested in whether a CRP-reducing agent should be included in a polypill that is taken every day to reduce mortality risk in healthy individuals. However, a Mendelian randomization investigation would be less relevant for judging whether or not short-term highly elevated CRP levels should be targeted for intervention in individuals displaying acute coronary symptoms.

23.7.2 Heterogeneity and Pleiotropy

For a complex multifactorial risk factor such as BMI, there are many mechanisms by which the risk factor could be lowered, and many genetic variants associated with BMI that influence the risk factor via different biological pathways. It is therefore entirely possible that some genetic variants associated with the risk factor are also associated with the outcome, but not all. Or that all variants are associated with the outcome, but that the causal effects estimated by different variants differ. This is both a blessing and a curse for Mendelian randomization. On the positive side, it can be used to identify which aspects of a complex risk factor influence an outcome, and hence may help identify effective mechanisms for reducing disease risk (Walter *et al.*, 2015). On the negative side, the methods above rely on homogeneity of the causal estimates based on different genetic variants, and interpret deviation from this as pleiotropy.

Even if there is heterogeneity in causal estimates, the ratio estimate based on each genetic variant is still a valid test of the causal null hypothesis (that is, under the IV assumptions it will only differ from zero if the risk factor is a cause of the outcome) (Burgess *et al.*, 2016). Hence the inverse-variance weighted method still provides a valid test of the causal null hypothesis, as do the median-based and mode-based methods. However, other methods such as MR-Egger are more sensitive to these parametric assumptions. One practical solution is the recommendation to use random effects models in estimation, as these models translate heterogeneity between variant-specific causal estimates into wider confidence intervals (Bowden *et al.*, 2017). Heterogeneous points should be investigated carefully to assess whether there is any difference in how the genetic variant influences the risk factor that may give rise to the heterogeneity, or whether it is likely to be a result of pleiotropy.

23.7.3 Weak Instrument Bias and Sample Overlap

If the genetic associations with the risk factor and those with the outcome are estimated in separate samples of the population, then the estimates will be independent. However, in practice,

many large international genetics consortia include the same studies, and hence have participants in common (Lin and Sullivan, 2009). This means that if genetic associations with the risk factor are overestimated, then associations with the outcome will also tend to be overestimated (assuming that the risk factor and outcome are positively correlated). This leads to weak instrument bias, a version of winner's curse in IV analysis (Burgess and Thompson, 2011). In a two-sample analysis, independence between the association estimates means that weak instrument bias is in the direction of the null and does not lead to inflated Type I error rates. In contrast, in a one-sample analysis, bias is in the direction of the confounded association, and can lead to false positive findings (Pierce and Burgess, 2013). When there is some overlap between the two data sets, bias is proportional to the degree of overlap (Burgess *et al.*, 2016). A weakness of summarized data is that it is often not possible to determine the degree of sample overlap, or to correct for sample overlap by omitting individuals from the calculation of genetic association estimates or by cross-validation (Burgess and Thompson, 2013).

23.7.4 Time-Dependent Causal Effects

Some methodological development has been undertaken to try to perform IV analysis with a time-to-event outcome (Tchetgen Tchetgen *et al.*, 2015; Martinussen *et al.*, 2017). While it is possible to assess the causal null hypothesis with a time-to-event outcome (by assessing the association between the genetic variant(s) and outcome), any causal estimates are unlikely to be reliable. Indeed, it would seem impossible to make any detailed judgement about the timing of causal effects using genetic variants, unless one had specific temporal knowledge about the action of the genetic variants on the risk factor. For example, if increased BMI leads to increased cardiovascular mortality, associations of genetic variants would be identical if increased childhood BMI led to increased cardiovascular mortality, or if the causal risk factor were instead increased adolescent BMI or increased BMI in early adulthood – assuming that the genetic variant is associated with increased BMI for the whole life course. These scenarios could only be distinguished by identifying particular genetic variants that increase BMI in childhood, those that increase BMI in adolescence, and so on.

Additionally, it is not possible (without additional information) to judge the 'viscosity' of a risk factor. By viscosity we mean the rate at which changes in the risk factor will influence the outcome. For instance, LDL cholesterol is likely to have high viscosity for CHD risk, as atherosclerosis is a chronic condition resulting from long-term exposure to LDL cholesterol. Short-term interventions in LDL cholesterol would not be expected to lead to immediate reduction in CHD risk. Indeed, it has been shown that increased time of exposure to LDL-cholesterol-lowering drugs is associated with a greater reduction in CHD risk per 1 mmol/L change in LDL cholesterol (Taylor *et al.*, 2013), with Mendelian randomization estimates around two three times greater in magnitude than even estimates of risk reduction from randomized trials with 5-year median follow-up (Burgess *et al.*, 2012). In contrast, blood pressure may have lower viscosity as a risk factor, as although long-term exposure to high blood pressure does lead to arterial stenosis, current blood pressure is likely to be an important risk factor for influencing CHD risk.

This life-course perspective for genetic associations has positive consequences for Mendelian randomization, as it means that genetic associations with disease outcomes may be greater than one would expect from observational research, and hence power to detect a causal relationship may be greater than expected. However, it also means that causal estimates from Mendelian randomization are not realistic guides to the impact of intervening on a risk factor in practice, and may be overly optimistic. Further research is needed to judge the similarity of observational epidemiological, clinical trial, and Mendelian randomization estimates in cases where all are available.

23.7.5 Collider Bias

A collider is a variable that is a common effect of two variables (that is, it is causally downstream of the two variables) (Cole *et al.*, 2010). Even if the two variables are unrelated (they are marginally independent), they will typically be related when conditioning on the collider (conditionally dependent). For example, suppose that ear infections and throat infections are independent events. However, conditional on an individual attending an ear, nose and throat clinic, ear infections and throat infections are conditionally dependent events – if an individual does not have an ear infection, but attends such a clinic, then it is more likely that the individual has a throat infection (Pearl, 2000). Conditioning on a collider between an instrument variable and confounder can induce an association between the two even if they are marginally independent. This can lead to a valid IV becoming invalid, and hence to an association between the IV and the outcome in the absence of a causal effect of the risk factor on the outcome. Intuitively speaking, even if a genetic variant can be treated as if it were randomly distributed in the general population, it may not be randomly distributed in a subset of the general population (in particular, in the sample population under investigation).

Collider bias can occur due to differential selection into the sample population (Hernán *et al.*, 2004). For example, if the sample is preferentially selected with respect to the risk factor or the outcome, then selection bias (a form of collider bias) would occur (both the risk factor and the outcome are common effects of the genetic variant and confounders, and hence both are colliders for these two variables). Collider bias may also occur due to differential survival. If the risk factor affects survival, then selection bias would also occur. This is particularly likely for Mendelian randomization investigations of diseases of old age. It is also relevant for investigations of disease progression, as in order to be included in an analysis of disease progression, one has to have the disease in the first place. A further situation where collider bias may occur is when the population under investigation is stratified, in particular if the stratifying variable is downstream of the risk factor or outcome. In a Mendelian randomization analysis of the causal effect of BMI on breast cancer progression, it was shown that bias would occur if BMI were a cause of breast cancer risk (Guo *et al.*, 2016). However, the magnitude of the selection effect required to lead to substantial Type I error inflation was greater than would be plausibly expected, indicating that findings were likely to be robust to collider bias. Further research is needed to ascertain the extent to which collider bias is likely to be a practically relevant issue in Mendelian randomization investigations.

23.8 Conclusion

This chapter has focused on answering two questions: when can genetic variants be used to make reliable causal inferences about epidemiological relationships between risk factors and outcomes, and when can genetic variants be used to make relevant causal inferences? As for the first question, the most reliable causal inferences from Mendelian randomization will always come from monogenic analyses ('candidate gene Mendelian randomization'), where the genetic variant has a plausible biological link with the risk factor of interest. There are some issues with these analyses, such as the problem of multiple testing. Additionally, the lessons of the candidate gene era – the need for replication and stringent p -value thresholds – must be learnt. While polygenic Mendelian randomization analyses have inherent weaknesses, as the instrumental variable assumptions are unlikely to hold for all genetic predictors of a given risk factor, several methods are available for interrogating the robustness of findings. Although the inverse-variance weighted method will generally be the primary analysis method, all polygenic Mendelian randomization investigations should at minimum assess the heterogeneity between

causal estimates from different variants, including a visual assessment via a scatter plot. Median-based and mode-based methods are useful for determining whether a causal effect is evidenced by the majority of genetic variants and for detecting multiple causal mechanisms within the genetic variants. In some cases, despite its decreased power, the MR-Egger method will also be relevant for detecting pleiotropy (via the intercept test), and for providing another estimate of the causal effect under a different set of assumptions. In other scenarios, such as when there are influential points in the regression model or when pleiotropic effects of multiple variants are likely to act via the same confounder, the MR-Egger method is less appropriate. Other approaches, such as multivariable Mendelian randomization, and looking for interactions in subsets, will be useful in some situations. Cases in which multiple methods making different assumptions lead to the same causal conclusion should be given more evidential weight in favour for a causal relationship.

As for the second question, while Mendelian randomization addresses a relevant causal question (that of the long-term effect of elevated levels of the risk factor on the outcome), it does not answer all relevant causal questions. Additionally, and particularly for highly viscous risk factors, causal estimates and even causal inferences may be misleading guides as to the impact of interventions on the same risk factor in applied practice. While there is still much to learn from the biobank era of data sets whose main distinguishing feature is their sheer size, there are also many biologically relevant questions about timings and mechanisms of causal effects that will require clever epidemiological designs as well as clever statistical analyses – not all causal questions will be answerable simply by increasing the sample size of cross-sectional data sets.

Future methodological questions to be determined include evaluating and comparing the various robust methods for Mendelian randomization, both those included in this review and further proposed methods (Tchetgen Tchetgen *et al.*, 2017; Jiang *et al.*, 2019). A related question is how to take advantage of genome-wide genetic data, and even whether extensive genetic data will ever be useful for making reliable causal inferences. Additionally, how large-scale data on multiple biomarkers and multiple layers of -omics data can be used in Mendelian randomization investigations, particularly to learn about complex aetiological networks and causal mechanisms.

In conclusion, Mendelian randomization is an important and valuable tool for learning about causal relationships using genetic data, but it is also a fallible one. While statistical approaches can go some way in allowing the assumptions on which the approach stands to be assessed, fundamentally the approach depends on being able to find genetic variants that are plausible proxies for intervention on the risk factor. We look forward to the further development of genetic knowledge, statistical methodology, and epidemiological data sets – active communication and close collaboration between these fields has been a defining feature of Mendelian randomization research so far. We hope it will continue.

References

- Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**(434), 444–455.
- Benn, M., Nordestgaard, B.G., Frikke-Schmidt, R. and Tybjærg-Hansen, A. (2017). Low LDL cholesterol, PCSK9 and HMGCR genetic variation, and risk of Alzheimer's disease and Parkinson's disease: Mendelian randomisation study. *British Medical Journal* **357**, j1648.
- Benner, C., Spencer, C.C., Havulinna, A.S., Salomaa, V., Ripatti, S. and Pirinen, M. (2016). FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**(10), 1493–1501.

- Berzuini, C., Guo, H., Burgess, S. and Bernardinelli, L. (2018). A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*. doi: 10.1093/biostatistics/kxy027.
- Bowden, J., Davey Smith, G. and Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**(2), 512–525.
- Bowden, J., Davey Smith, G., Haycock, P.C. and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40**(4), 304–314.
- Bowden, J., Del Greco, F., Minelli, C., Davey Smith, G., Sheehan, N.A. and Thompson, J.R. (2016). Assessing the suitability of summary data for Mendelian randomization analyses using MR-Egger regression: The role of the I^2 statistic. *International Journal of Epidemiology* **45**(6), 1961–1974.
- Bowden, J., Del Greco, F., Minelli, C., Lawlor, D., Sheehan, N., Thompson, J. and Davey Smith, G. (2018). Improving the accuracy of two-sample summary data Mendelian randomization: Moving beyond the NOME assumption. *International Journal of Epidemiology*. doi: 10.1093/ije/dyy258.
- Bowden, J., Del Greco, M.F., Minelli, C., Davey Smith, G., Sheehan, N. and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* **36**(11), 1783–1802.
- Bowden, J. and Vansteelandt, S. (2011). Mendelian randomisation analysis of case-control data using structural mean models. *Statistics in Medicine* **30**(6), 678–694.
- Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., ReproGen Consortium, Psychiatric Genomics Consortium, Anorexia Nervosa Genetic Consortium for the Wellcome Trust Case Control Consortium, Duncan, L., Perry, J.R., Patterson, N., Robinson, E., Daly, M.J., Price, A.L. and Neale, B.M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241.
- Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**(3), 291–295.
- Burgess, S., Bowden, J., Dudbridge, F. and Thompson, S.G. (2016). Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. Preprint, arXiv:1606.03729.
- Burgess, S., Bowden, J., Fall, T., Ingelsson, E. and Thompson, S. (2017). Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28**(1), 30–42.
- Burgess, S., Butterworth, A., Malarstig, A. and Thompson, S. (2012). Use of Mendelian randomisation to assess potential benefit of clinical intervention. *British Medical Journal* **345**, e7325.
- Burgess, S., Butterworth, A.S. and Thompson, J.R. (2016). Beyond Mendelian randomization: How to interpret evidence of shared genetic predictors. *Journal of Clinical Epidemiology* **69**, 208–216.
- Burgess, S., Butterworth, A.S. and Thompson, S.G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37**(7), 658–665.
- Burgess, S. and Davey Smith, G. (2017). Mendelian randomization implicates high-density lipoprotein cholesterol-associated mechanisms in etiology of age-related macular degeneration. *Ophthalmology* **124**(8), 1165–1174.
- Burgess, S., Davies, N.M. and Thompson, S.G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology* **40**(7), 597–608.
- Burgess, S., Dudbridge, F. and Thompson, S.G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Statistics in Medicine* **35**(11), 1880–1906.

- Burgess, S., Foley, C.N. and Zuber, V. (2018). Inferring causal relationships between risk factors and outcomes from genome-wide association study data. *Annual Review of Genomics and Human Genetics* **19**, 303–327.
- Burgess, S., Freitag, D., Khan, H., Gorman, D. and Thompson, S. (2014). Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS ONE* **9**(10), e108 891.
- Burgess, S., Scott, R.A., Timpson, N., Davey Smith, G., Thompson, S.G. and EPIC-InterAct Consortium (2015). Using published data in Mendelian randomization: A blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology* **30**(7), 543–552.
- Burgess, S., Thompson, D.J., Rees, J.M.B., Day, F.R., Perry, J.R. and Ong, K.K. (2017). Dissecting causal pathways using Mendelian randomization with summarized genetic data: Application to age at menarche and risk of breast cancer. *Genetics* **207**, 481–487.
- Burgess, S. and Thompson, S.G. (2011). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine* **30**(11), 1312–1323.
- Burgess, S. and Thompson, S.G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* **42**(4), 1134–1144.
- Burgess, S. and Thompson, S.G. (2015). *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*, Chapman & Hall, Boca Raton, FL.
- Burgess, S. and Thompson, S.G. (2015). Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology* **181**(4), 251–260.
- Burgess, S. and Thompson, S.G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology* **32**(5), 377–389.
- Burgess, S., Zuber, V., Gkatzionis, A. and Foley, C.N. (2018). Modal-based estimation via heterogeneity-penalized weighting: Model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *International Journal of Epidemiology* **47**(4), 1242–1254.
- CARDIoGRAMplusC4D Consortium (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130.
- Cho, Y., Shin, S.Y., Won, S., Relton, C.L., Smith, G.D. and Shin, M.J. (2015). Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study. *Scientific Reports* **5**, 18 422.
- Christenfeld, N., Sloan, R., Carroll, D. and Greenland, S. (2004). Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine* **66**(6), 868–875.
- Cole, S.R., Platt, R.W., Schisterman, E.F., Chu, H., Westreich, D., Richardson, D. and Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* **39**(2), 417–420.
- Cross-Disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *Lancet* **381**, 1371–1379.
- CRP CHD Genetics Collaboration (2011). Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *British Medical Journal* **342**, d548.
- Davey Smith, G. (2006). Randomised by (your) god: Robust inference from an observational study design. *Journal of Epidemiology and Community Health* **60**(5), 382–388.
- Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**(1), 1–22.
- Davey Smith, G. and Ebrahim, S. (2004). Mendelian randomization: Prospects, potentials, and limitations. *International Journal of Epidemiology* **33**(1), 30–42.

- Day, F., Thompson, D., Helgason, H., Chasman, D., Finucane, H., Sulem, P., Ruth, K., Whalen, S., Sarkar, A., Albrecht, E. et al. (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nature Genetics* **49**, 834–841.
- Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**(4), 309–330.
- Emerging Risk Factors Collaboration (2010). C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: An individual participant meta-analysis. *Lancet* **375**(9709), 132–140.
- Ference, B.A., Yoo, W., Alesh, I., Mahajan, N., Mirowska, K.K., Mewada, A., Kahn, J., Afonso, L., Williams, K.A. and Flack, J.M. (2012). Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: A Mendelian randomization analysis. *Journal of the American College of Cardiology* **60**(25), 2631–2639.
- Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**(5), e1004383.
- Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1283.
- Glymour, M., Tchetgen Tchetgen, E. and Robins, J. (2012). Credible Mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology* **175**(4), 332–339.
- Greco, M., Minelli, C., Sheehan, N.A. and Thompson, J.R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine* **34**(21), 2926–2940.
- Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**(4), 722–729.
- Greenland, S. and Robins, J. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* **15**(3), 413–419.
- Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J.A. and Wallace, C. (2015). Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics* **24**(12), 3305–3313.
- Guo, Y., Andersen, S.W., Shu, X.O., Michailidou, K., Bolla, M.K., Wang, Q., Garcia-Closas, M., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Dunning, A., Bojesen, S.E., Ahsan, H., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Arndt, V., Beckmann, M.W., Beeghly-Fadiel, A., Benitez, J., Bogdanova, N.V., Bonanni, B., Børresen-Dale, A.L., Brand, J., Brauch, H., Brenner, H., Brüning, T., Burwinkel, B., Casey, G., Chenevix-Trench, G., Couch, F.J., Cox, A., Cross, S.S., Czene, K., Devilee, P., Dörk, T., Dumont, M., Fasching, P.A., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Fostira, F., Gammon, M., Giles, G.G., Guénél, P., Haiman, C.A., Hamann, U., Hooning, M.J., Hopper, J.L., Jakubowska, A., Jasmine, F., Jenkins, M., John, E.M., Johnson, N., Jones, M.E., Kabisch, M., Kibriya, M., Knight, J.A., Koppert, L.B., Kosma, V.M., Kristensen, V., Marchand, L.L., Lee, E., Li, J., Lindblom, A., Luben, R., Lubinski, J., Malone, K.E., Mannermaa, A., Margolin, S., Marme, F., McLean, C., Meijers-Heijboer, H., Meindl, A., Neuhausen, S.L., Nevanlinna, H., Neven, P., Olson, J.E., Perez, J.I.A., Perkins, B., Peterlongo, P., Phillips, K.A., Pylkäs, K., Rudolph, A., Santella, R., Sawyer, E.J., Schmutzler, R.K., Seynaeve, C., Shah, M., Shrubsole, M.J., Southey, M.C., Swerdlow, A.J., Toland, A.E., Tomlinson, I., Torres, D., Truong, T., Ursin, G., Luijt, R.B.V.D., Verhoef, S., Whittemore, A.S., Winquist, R., Zhao, H., Zhao, S., Hall, P., Simard, J., Kraft, P., Pharoah, P., Hunter, D., Easton, D.F. and Zheng, W. (2016). Genetically predicted body

- mass index and breast cancer risk: Mendelian randomization analyses of data from 145,000 women of European descent. *PLoS Medicine* **13**(8), e1002105.
- Guo, Z., Kang, H., Cai, T.T. and Small, D.S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B*. doi: 10.1111/rssb.12275.
- Han, C. (2008). Detecting invalid instruments using L1-GMM. *Economics Letters* **101**, 285–287.
- Hartwig, F.P., Davey Smith, G. and Bowden, J. (2017). Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46**(6), 1985–1998.
- Hartwig, F.P. and Davies, N.M. (2016). Why internal weights should be avoided (not only) in MR-Egger regression. *International Journal of Epidemiology* **45**, 1676–1678.
- Hernán, M., Hernández-Díaz, S. and Robins, J. (2004). A structural approach to selection bias. *Epidemiology* **15**(5), 615–625.
- Hernán, M. and Robins, J. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17**(4), 360–372.
- Hernán, M. and Robins, J. (2018). *Causal Inference*, Chapman & Hall/CRC Press, Boca Raton, FL. Available at <http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/>.
- Hingorani, A. and Humphries, S. (2005). Nature's randomised trials. *Lancet* **366**(9501), 1906–1908.
- Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**(2), 497–508.
- Huber, P.J. and Ronchetti, E.M. (2011). *Robust Statistics*, Wiley, Hoboken, NJ.
- IL6R Genetics Consortium and Emerging Risk Factors Collaboration (2012). Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *Lancet* **379**(9822), 1205–1213.
- Interleukin-1 Genetics Consortium (2015). Cardiometabolic consequences of genetic up-regulation of the interleukin-1 receptor antagonist: Mendelian randomisation analysis of more than one million individuals. *Lancet: Diabetes and Endocrinology* **3**(4), 243–253.
- Interleukin-6 Receptor Mendelian Randomisation Analysis Consortium (2012). The interleukin-6 receptor as a target for prevention of coronary heart disease: a Mendelian randomisation analysis. *Lancet* **379**(9822), 1214–1224.
- Jiang, L., Oualkacha, K., Didelez, V., Ciampi, A., Rosa, P., Benedet, A.L., Mathotaarachchi, S.S., Richards, B. and Greenwood, C.M.T. (2019). Constrained instruments and their application to Mendelian randomization with pleiotropy. *Genetic Epidemiology*. doi: 10.1002/gepi.22184.
- Kang, H., Zhang, A., Cai, T. and Small, D. (2016). Instrumental variables estimation with some invalid instruments, and its application to Mendelian randomisation. *Journal of the American Statistical Association* **111**(513), 132–144.
- Lawlor, D., Harbord, R., Sterne, J., Timpson, N. and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**(8), 1133–1163.
- Lee, Y.S., Cho, Y., Burgess, S., Davey Smith, G., Relton, C.L., Shin, S.Y. and Shin, M.J. (2016). Serum gamma-glutamyl transferase and risk of type 2 diabetes in the general Korean population: A Mendelian randomization study. *Human Molecular Genetics* **25**(17), 3877–3886.
- Lewis, S. and Clarke, M. (2001). Forest plots: Trying to see the wood and the trees. *British Medical Journal* **322**(7300), 1479–1480.
- Lewis, S. and Davey Smith, G. (2005). Alcohol, ALDH2, and esophageal cancer: A meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology Biomarkers & Prevention* **14**(8), 1967–1971.
- Li, S. (2017). Mendelian randomization when many instruments are invalid: hierarchical empirical Bayes estimation. Preprint, arXiv:1706.01389.

- Lin, D.Y. and Sullivan, P.F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *American Journal of Human Genetics* **85**(6), 862–872.
- Lipsitch, M., Tchetgen Tchetgen, E. and Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21**(3), 383–388.
- Lotta, L.A., Sharp, S.J., Burgess, S., Perry, J.R., Stewart, I.D., Willems, S.M., Luan, J., Ardanaz, E., Arriola, L., Balkau, B. et al. (2016). Association between low-density lipoprotein cholesterol-lowering genetic variants and risk of type 2 diabetes: A meta-analysis. *Journal of the American Medical Association* **316**(13), 1383–1391.
- Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C., Prokopenko, I. et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* **46**(3), 234–244.
- Martinussen, T., Vansteelandt, S., Tchetgen Tchetgen, E.J. and Zucker, D.M. (2017). Instrumental variables estimation of exposure effects on a time-to-event response using structural cumulative survival models. *Biometrics* **73**(4), 1140–1149.
- Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, MA.
- Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.
- Pierce, B. and Burgess, S. (2013). Efficient design for Mendelian randomization studies: Subsample and two-sample instrumental variable estimators. *American Journal of Epidemiology* **178**(7), 1177–1184.
- Plenge, R., Scolnick, E. and Altshuler, D. (2013). Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery* **12**(8), 581–594.
- Rees, J., Wood, A. and Burgess, S. (2017). Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in Medicine* **36**(29), 4705–4718.
- Ridker, P.M., Everett, B.M., Thuren, T., MacFadyen, J.G., Chang, W.H., Ballantyne, C., Fonseca, F., Nicolau, J., Koenig, W., Anker, S.D. et al. (2017). Antiinflammatory therapy with canakinumab for atherosclerotic disease. *New England Journal of Medicine* **377**(12), 1119–1131.
- Robins, J. (1989). The control of confounding by intermediate variables. *Statistics in Medicine* **8**(6), 679–701.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(5), 688–701.
- Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. and Smoller, J.W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics* **14**(7), 483–495.
- Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J. and UCLB Consortium (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**, 986–992.
- Spiller, W., Slichter, D., Bowden, J. and Davey Smith, G. (2018). Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions. *International Journal of Epidemiology*. doi: 10.1093/ije/dyy204.
- Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J., Young, R. and Butterworth, A.S. (2016). PhenoScanner: A database of human genotype-phenotype associations. *Bioinformatics* **32**(20), 3207–3209.
- Sterne, J.A., Sutton, A.J., Ioannidis, J., Terrin, N., Jones, D.R., Lau, J., Carpenter, J., Rücker, G., Harbord, R.M., Schmid, C.H. et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal* **343**, d4002.

- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. et al. (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* **12**(3), e1001779.
- Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M.A., Prins, B.P., Wilcox, S.K., Zimmerman, E.S., Chi, A., Bansal, N., Spain, S.L., Wood, A.M., Morrell, N.W., Bradley, J.R., Janjic, N., Roberts, D.J., Ouwehand, W.H., Todd, J.A., Soranzo, N., Suhre, K., Paul, D.S., Fox, C.S., Plenge, R.M., Danesh, J., Runz, H. and Butterworth, A.S. (2018). Consequences of natural perturbations in the human plasma proteome. *Nature* **558**, 73–79.
- Swanson, S.A., Tiemeier, H., Ikram, M.A. and Hernán, M.A. (2017). Nature as a trialist? Deconstructing the analogy between Mendelian randomization and randomized trials. *Epidemiology* **28**(5), 653–659.
- Taylor, F., Ward, K., Moore, T., Burke, M., Davey Smith, G., Casas, J. and Ebrahim, S. (2013). Statins for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* **1**, CD004816.
- Tchetgen Tchetgen, E., Walter, S., Vansteelandt, S., Martinussen, T. and Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26**(3), 402–410.
- Tchetgen Tchetgen, E.J., Sun, B. and Walter, S. (2017). The GENIUS approach to robust Mendelian randomization inference. Preprint, arXiv:1709.07779.
- Thompson, S. and Sharp, S. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine* **18**(20), 2693–2708.
- van Kippersluis, H. and Rietveld, C.A. (2017). Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology* **47**(4), 1279–1288.
- VanderWeele, T., Tchetgen Tchetgen, E., Cornelis, M. and Kraft, P. (2014). Methodological challenges in Mendelian randomization. *Epidemiology* **25**(3), 427–435.
- Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. *Genetic Epidemiology* **37**(8), 802–813.
- Walter, S., Kubzansky, L.D., Koenen, K.C., Liang, L., Tchetgen Tchetgen, E.J., Cornelis, M.C., Chang, S.C., Rimm, E., Kawachi, I. and Glymour, M.M. (2015). Revisiting mendelian randomization studies of the effect of body mass index on depression. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics* **168**(2), 108–115.
- Windmeijer, F., Farbmacher, H., Davies, N. and Davey Smith, G. (2016). On the use of the lasso for instrumental variables estimation with some invalid instruments. Discussion Paper 16/1674, Department of Economics, University of Bristol.
- Wooldridge, J. (2009). Instrumental variables estimation and two stage least squares, in *Introductory Econometrics: A Modern Approach*, South-Western, Nashville, TN.

24

Improving Genetic Association Analysis through Integration of Functional Annotations of the Human Genome

Qiongshi Lu¹ and Hongyu Zhao²

¹ Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison WI, USA

² Department of Biostatistics, Yale University, New Haven, CT, USA

Abstract

The genome-wide association study (GWAS) has enjoyed success for many complex diseases and traits, identifying tens of thousands of genetic associations. However, our understanding for most diseases and traits is far from complete. Integrative analysis of GWAS data, especially summary association statistics, and functional annotations of the human genome have made some progress in improving both the statistical power of association analysis and the interpretation of GWAS findings. We introduce several types of functional annotation data with direct applications in GWAS and post-GWAS analyses, discuss statistical and computational methods to synthesize these annotation data into concise and interpretable metrics, and summarize several applications of these functional annotations that have attracted much interest in recent years.

24.1 Introduction

Genome-wide association studies (GWASs) have identified tens of thousands of single nucleotide polymorphisms (SNPs) associated with a variety of complex human diseases and traits. These findings have improved our understanding of the genetic architecture of complex traits. Two observations (somewhat unexpected) were made in the early GWAS days: significantly associated loci identified in GWASs only have moderate effect sizes and explain a small fraction of phenotypic variability (Manolio *et al.*, 2009), and most associated SNPs are located in non-coding regions of the human genome (Hindorff *et al.*, 2009). These two observations have been seen across a variety of traits in the past decade. It is becoming increasingly clear that most complex traits are weakly associated with a large number of genetic loci, many of which are non-coding. Boyle *et al.* (2017) recently summarized these observations using a unified, ‘omni-genic’ model. Consequently, as sample sizes for GWASs continue to grow, the number of identified associations will keep increasing but the effect sizes of newly identified loci will most likely be very weak (Visscher *et al.*, 2017). For many complex traits, the focus in GWAS and post-GWAS analyses has gradually shifted from identifying trait-associated SNPs to interpreting the findings.

Integrative analysis of GWAS data and functional annotations of the human genome has proven to be a fruitful approach. Over the past decade, large consortia such as the Encyclopedia of DNA Elements (ENCODE: ENCODE Project Consortium, 2012), the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium, 2015), and the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015, 2017) have generated vast amounts of transcriptomic and epigenetic annotation data. Many computational and statistical frameworks have been developed to synthesize these data into concise and interpretable annotations. Statistical methods that integrate these functional annotations with GWAS data, especially the widely accessible GWAS summary statistics (Pasaniuc and Price, 2017), have provided novel biological insights into the genetic basis of many complex traits.

In this chapter, we begin by introducing different types of functional annotation data that have applications in GWAS analyses. Then we summarize recent efforts to use computational and statistical methods to synthesize these data. Finally, we introduce several applications of integrating functional annotations with GWAS summary statistics that have achieved great success in recent years.

24.2 Types of Functional Annotation Data in GWAS Applications

Numerous studies have suggested that the non-coding genome harbors the majority of risk variants for complex diseases and traits (Hindorff *et al.*, 2009; Schaub *et al.*, 2012; Maurano *et al.*, 2012). Many different assays have been developed to identify non-coding functionality in the human genome (Ward and Kellis, 2012a). Here, we mainly focus on the advances in recent years that have direct and rich applications in GWAS and post-GWAS analyses.

24.2.1 Transcriptomic Annotation Data

Transcription can be used to annotate activity in the genome. Numerous consortia/projects – among them GTEx (GTEx Consortium, 2015, 2017), CommonMind (Frommer *et al.*, 2016), and STARNET (Franzén *et al.*, 2016) – have generated a large amount of gene expression data for various cell lines as well as human tissues (Table 24.1). Among studies aiming to characterize transcriptional activities, GTEx stands out because it provides gene expression data for dozens of tissue types from hundreds of healthy donors (GTEx Consortium, 2015, 2017). Additionally, genome-wide genotype data in GTEx make it possible to study the regulatory effect of individual SNPs on gene expression across human tissues. GTEx data are useful in GWAS analyses for several reasons. First, regions that are transcribed in disease-related tissues are more likely

Table 24.1 Major data portals providing rich functional annotation data

Project	Description	URL
GTEx	gene expression, genotype	https://www.gtexportal.org/home/
CommonMind	gene expression, genotype	https://www.synapse.org/#!Synapse:syn2759792/
ENCODE	epigenetic marks	https://www.encodeproject.org
PsychENCODE	gene expression, epigenetic marks	https://www.synapse.org//#!Synapse:syn4921369/
Roadmap	epigenetic marks	http://www.roadmapepigenomics.org

to harbor risk variants for that disease (Finucane *et al.*, 2018). In addition, expression quantitative trait loci (eQTL), that is, SNPs associated with gene expression, are also known to strongly enrich for GWAS associations (Nicolae *et al.*, 2010). These annotations can be used to prioritize GWAS associations and fine-map functional candidates among SNPs in linkage disequilibrium (LD). Further, since the majority of GWAS associations are located in the non-coding genome, it remains challenging to infer candidate genes using GWAS results. Expression quantitative trait loci are a unique functional annotation that may suggest genes regulated by SNPs identified in GWASs.

24.2.2 Epigenetic Annotation Data

Epigenetic functional annotation is another major resource that has recently attracted widespread interest. Several large consortia, including ENCODE (ENCODE Project Consortium, 2012; PsychENCODE Consortium, 2015), Roadmap Epigenomics Project (Roadmap Epigenomics Consortium, 2015), and BLUEPRINT project (Stunnenberg *et al.*, 2016), have generated rich, high-throughput data for a variety of epigenetic marks, including histone modifications, DNA methylation, transcription factor binding activities, and markers for chromatin accessibility. Each of these epigenetic marks is known to be associated with various non-coding regulatory activities (ENCODE Project Consortium, 2012; Trynka *et al.*, 2013; Lawrence *et al.*, 2016). They can be used to test specific hypotheses in GWAS downstream analyses (e.g. whether an identified locus shows enhancer activities in a disease-related cell type). Additionally, these annotations are the building blocks for designing machine learning algorithms to distinguish functional variants from neutral DNA elements. Almost all computational annotation methods use data from ENCODE and Roadmap as predictive features. Of note, similar to gene expression data, epigenetic annotations are tissue-specific. Both ENCODE and Roadmap provide data for a large number of cell lines and human tissues. While cell line data usually provide the most complete collection of assays, data generated for primary human tissues may lead to more interpretable results in post-GWAS analyses. On the other hand, the ensemble profiles of primary tissue samples may be insensitive to cell type heterogeneity. That is, when the relevant cell type is only present in a small fraction of sampled cells, interpretation of data remains challenging.

24.2.3 DNA Conservation

DNA conservation metrics based on DNA alignment across multiple species are another type of commonly used functional annotation in human genetics research. High conservation of a DNA segment across species suggests strong purifying selection and therefore hints at functionality. Based on this idea, numerous methods have been developed to quantify conservation, such as GERP (Cooper *et al.*, 2005; Davydov *et al.*, 2010), PhyloP (Pollard *et al.*, 2010), and PhastCons (Siepel *et al.*, 2005). Conserved DNA regions are strongly enriched for heritability of complex diseases (Finucane *et al.*, 2015), suggesting the importance of considering this information in GWAS analyses. However, only 4.5% of the human genome is conserved across mammals (Lindblad-Toh *et al.*, 2011), which is a substantially smaller fraction compared to the coverage of transcriptomic and epigenetic annotations. Additionally, conservation metrics are solely based on DNA sequences and therefore do not provide tissue-specific insights. Finally, although it may be reasonable to use conservation to annotate DNA regions that are critical for fitness, it is unclear if conservation will remain useful when studying the genetics of late-onset diseases (e.g. Alzheimer's disease: Lambert *et al.*, 2013) or human-specific traits (e.g. substance dependence: Tobacco and Genetics Consortium, 2010).

24.3 Methods to Synthesize Annotation Data

Although a variety of functional annotation data have been generated, raw annotation information is not easy to use in practice. The thousands of data sets present in ENCODE can be overwhelming, especially when it is unclear what assay and cell type may be suitable for a particular study. Therefore, it is of practical interest to synthesize various functional annotation data into more concise and informative metrics. To this end, computational methods predicting deleterious variants in protein coding genes have been successful. Moreover, many machine learning methods have been developed and are widely used in genetic studies (Ng and Henikoff, 2001; Adzhubei *et al.*, 2010; Schwarz *et al.*, 2010; Dong *et al.*, 2014). However, these tools are not suitable for GWAS applications due to their inability to annotate non-coding variants. In the past few years, several statistical and computational methods have been developed to predict non-coding functionality in the human genome (Table 24.2).

24.3.1 Genome Browsers and Annotator Software

Several annotation browsers have been developed (Boyle *et al.*, 2012; Ward and Kellis, 2012b; Kent *et al.*, 2002; Flieck *et al.*, 2012). These servers do not provide sophisticated metrics to weigh different annotation tracks or predict functionality. Instead, they provide well-designed user interfaces to facilitate visual presentation of vast amount of annotation information. These browsers can be very useful for exploratory analysis but are not suitable for quantitative analyses. Annotator software is another type of useful tool (Wang *et al.*, 2010; Cingolani *et al.*, 2012). These toolboxes have comprehensive built-in databases and annotate variants and genome regions based on user-selected functional annotations. Similarly to genome browsers, these methods are not suitable for quantitative analyses, but they provide a fast, one-stop solution for generating a large number of functional annotations for users' data.

24.3.2 Supervised Learning Methods

Supervised learning methods that predict deleteriousness of variants in protein-coding genes have been successful, therefore it is natural to expand these methods to annotate variants in the non-coding regions. Standard supervised learning algorithms such as random forests or support vector machines can be applied to annotate non-coding variants. Additionally, transcriptional activities, epigenetic marks, and DNA conservation can effectively distinguish functional and neutral DNA elements. However, the lack of gold-standard training data makes it challenging to train an accurate variant classifier. Nevertheless, a few methods have been developed to tackle this difficult problem. CADD (Kircher *et al.*, 2014) and GWAVA (Ritchie *et al.*, 2014) are two popular methods among the first efforts. CADD contrasted human-derived variants that are nearly fixed in the human population with simulated *de novo* mutations. A support vector machine was trained to classify these two sets of variants. Since deleterious variants are depleted in fixed human-derived variants but not in randomly simulated mutations, CADD score can be used to quantify deleteriousness. GWAVA used a different strategy and defined regulatory variants from the Human Gene Mutation Database (HGMD) as the positive set in its training data. Three sets of random variants with different levels of proximity with functional variants (i.e. unmatched, matched to the closest transcription start site, and matched to the same 1 kb region) were used as the negative set, respectively. A modified random forest algorithm was trained to distinguish functional variants from random SNPs. Of note, although all three models showed reasonable predictive power, prediction accuracy was sensitive to the

Table 24.2 Browsers, software, and composite scores for non-coding annotation

Tools	Description	Ref.	URL
UCSC genome browser	annotation browser	Kent <i>et al.</i> (2002)	https://genome.ucsc.edu/cgi-bin/hgGateway
Ensembl genome browser	annotation browser	Flicek <i>et al.</i> (2012)	http://useast.ensembl.org/index.html
GTEx data browser	annotation browser	GTEx Consortium (2015)	https://www.gtexportal.org/home/
Roadmap data browser	annotation browser	Roadmap Epigenomics Consortium (2015)	http://egg2.wustl.edu/roadmap/web_portal
Haploreg	annotation browser	Ward and Kellis (2012b)	http://archive.broadinstitute.org/mammals/haploreg
RegulomeDB	annotation browser	Boyle <i>et al.</i> (2012)	http://www.regulomedb.org
ANNOVAR	annotation software	Wang <i>et al.</i> (2010)	http://annovar.openbioinformatics.org
SnpEff	annotation software	Cingolani <i>et al.</i> (2012)	http://snpeff.sourceforge.net
ChromHMM	chromatin states	Ernst and Kellis (2012)	http://compbio.mit.edu/ChromHMM/
Segway2.0	chromatin states	Chan <i>et al.</i> (2018)	https://segway.hoffmanlab.org
CADD	composite score	Kircher <i>et al.</i> (2014)	http://cadd.gs.washington.edu
GWAVA	composite score	Ritchie <i>et al.</i> (2014)	http://www.sanger.ac.uk/science/tools/gwava
DANN	composite score	Quang <i>et al.</i> (2014)	https://cbcl.ics.uci.edu/public_data/DANN/
FATHMM-XF	composite score	Rogers <i>et al.</i> (2018)	http://fathmm.biocompute.org.uk/fathmm-xf/
GenoCanyon	composite score	Lu <i>et al.</i> (2015)	http://genocanyon.med.yale.edu
GenoSkyline	composite score	Lu <i>et al.</i> (2016a)	http://genocanyon.med.yale.edu/GenoSkyline
EIGEN	composite score	Ionita-Laza <i>et al.</i> (2016)	http://www.columbia.edu/~i2135/eigen.html
FUN-LDA	composite score	Backenroth <i>et al.</i> (2018)	http://www.columbia.edu/~i2135/funlda.html
LINSIGHT	composite score	Huang <i>et al.</i> (2017)	https://github.com/CshlSiepelLab/LINSIGHT
DIVAN	composite score	Chen <i>et al.</i> (2016)	https://sites.google.com/site/emorydivan/
PINES	composite score	Bodea <i>et al.</i> (2016)	http://genetics.bwh.harvard.edu/pines/
Funseq2	composite score	Fu <i>et al.</i> (2014)	http://funseq2.gersteinlab.org

choice of negative set and substantially deteriorated when using random SNPs matched to the same region as the negative training set (Ritchie *et al.*, 2014).

To overcome the issues introduced by insufficient and potentially biased training data, two methods, DeltaSVM (Lee *et al.*, 2015) and DeepSea (Zhou and Troyanskaya, 2015), adopted an innovative strategy. Instead of directly predicting deleteriousness/functionality of variants, these methods predict regulatory activities (e.g. transcription factor binding) using short segments of DNA sequences as predictive features. Rich training data for these biochemical activities are available in ENCODE, therefore reliable classifiers could be built. Then, for each variant of interest, predicted regulatory activities based on different alleles of the variant and the surrounding DNA sequence were compared. The difference in the prediction scores was used to quantify the impact of a DNA variant. A major advantage of these methods is that they do not rely on labeled training data for functional and non-functional non-coding variants. However, a predefined regulatory activity in the genome was required. Prediction results based on different epigenetic marks (e.g. binding sites of different transcription factors) could be substantially different (Zhou and Troyanskaya, 2015). Additionally, since DNA sequence was the only predictor in the model, these methods cannot be used to predict a SNP's impact on molecular activities that are not associated with a pattern in the DNA sequence. Of note, DeepSea is one of the earliest successes in deep learning application in non-coding functional annotation. Although deep learning has also been directly applied to distinguish deleterious and neutral variants (Quang *et al.*, 2014), limited sample size in training data remains a critical issue.

24.3.3 Unsupervised Learning Methods

Unsupervised methods provide another solution to potential biases in labeled training data. These unsupervised learning methods use all data as algorithm input and cluster data points into different categories based on patterns learned from predictive features. However, interpretation of data clusters can be challenging. Based on the idea that recurrent combinatorial patterns in epigenetic marks (e.g. histone modifications) may be associated with chromatin states representing different regulatory functions, several methods have been developed. ChromHMM (Ernst and Kellis, 2010, 2012) and Segway (Hoffman *et al.*, 2012a; Chan *et al.*, 2018) are two methods that introduced and implemented this concept. ChromHMM was based on a hidden Markov model, while Segway was trained using a dynamic Bayesian network. Trained on data from ENCODE, both methods identified multiple chromatin states associated with various genomic features (e.g. transcription and heterochromatin) (Hoffman *et al.*, 2012b).

When trained on different data sets, the number of chromatin states predicted by ChromHMM and Segway may differ. In addition, although some chromatin states are associated with known genomic features, not all states are easy to interpret. Several recent methods simplified the model by allowing only two latent states. Such a strategy improved the robustness of model and made results easier to interpret. However, by using a 'functional' state to summarize diverse regulatory machinery in the non-coding genome, these methods may not be able to distinguish important biological signatures. Among these methods, GenoCanyon (Lu *et al.*, 2015) integrated multiple conservation metrics and dozens of epigenetic marks from ENCODE via a naive Bayes mixture model. The expectation-maximization algorithm was used to estimate parameters in the model. Similar to ChromHMM and Segway, since all included annotations were at the nucleotide level, GenoCanyon quantifies the functional potential of each base in the genome and does not specifically focus on DNA variants. Based on a similar idea, EIGEN (Ionita-Laza *et al.*, 2016) adopted a spectral method (Parisi *et al.*, 2014) that allows correlations among annotations conditioning on the functional state. Since variant-specific annotations such as allele frequencies were used in the model, EIGEN annotates DNA variants instead of genomic regions.

24.3.4 Improving Specificity of Computational Annotations

Due to advances in both genotyping technology and imputation methods, newly conducted GWASs typically include more than 10 million SNPs, most of which are non-coding. It is of practical interest to improve the specificity of functional annotation tools so that they can effectively filter/prioritize variants in genetic studies. The tissue-dependent nature of transcriptomic and epigenetic annotations makes it possible to expand some of the introduced methods to predict functional regions or variants for specific tissue types. However, there is extremely scarce information on tissue-specific functional and non-functional variants in non-coding regions. Therefore, unsupervised learning methods may be the only option at present. ChromHMM and Segway were trained using data from particular cell lines in ENCODE when they were first introduced, so naturally these annotations will be tissue-specific when data from a given tissue is used as input. Although GenoCanyon and EIGEN used data across cell lines as predictive features and were consequently non-tissue-specific, both methods have been expanded to predict tissue and cell type-specific non-coding functionality. GenoSkyline (Lu *et al.*, 2016a), the tissue-specific expansion of GenoCanyon, was initially available for seven broadly defined tissues and has been expanded to the complete set of 127 tissue and cell types in the Roadmap Epigenomics Project in its most recent update (Lu *et al.*, 2017b). Similarly, FUN-LDA (Backenroth *et al.*, 2018), a tissue-specific expansion of EIGEN, has also been implemented for all cell types in the Roadmap project.

Another direction to improve the specificity of annotations is to incorporate disease-specific information. Several methods, such as Phevor (Singleton *et al.*, 2014) and Phen-Gen (Javed *et al.*, 2014), integrated variant information with biomedical ontologies that summarize prior knowledge of diseases and achieved some success in protein-coding genes. Similar ideas have been recently applied to annotate non-coding variants. DIVAN (Chen *et al.*, 2016) and PINES (Bodea *et al.*, 2016) used SNPs associated with the disease of interest as part of the training data, thereby achieving disease-specific functional prediction. However, since GWAS results have been used during model training, it is unclear if such annotations can still be used in post-GWAS analyses. Finally, methods specifically targeting non-coding variants in cancer have also been developed (Khurana *et al.*, 2013; Fu *et al.*, 2014).

24.4 Methods to Integrate Functional Annotations in Genetic Association Analysis

GWASs are simple and effective; with sufficient sample size, they are able to identify robust associations for complex traits. Nevertheless, findings from a GWAS can be difficult to interpret. Due to the wide accessibility of annotation data and GWAS results, integrative analysis of GWAS summary statistics and functional annotations has been successful. Numerous statistical methods have been developed for various applications in post-GWAS analyses and have provided insights into the genetic basis of many complex diseases and traits. Here, we summarize two major advances that have gained great interest in recent years and briefly discuss other applications.

24.4.1 Partitioning Heritability and Genetic Covariance

Missing heritability (i.e. the observation that GWAS associations only explained a small proportion of phenotypic variability) was a mystery that intrigued human geneticists following publication of the first series of GWASs (Manolio *et al.*, 2009). In their seminal paper, Yang *et al.* (2010) demonstrated that the gap of missing heritability can be largely filled when effects

of all SNPs are taken into account via a linear mixed model and GREML algorithm. Interestingly, heritability estimates based on all SNPs remain consistent even when only a small fraction of SNPs are causal, that is, have non-zero contribution to heritability (Jiang *et al.*, 2016). Therefore, given the total heritability estimate, it is biologically interesting to investigate what SNPs contribute to this. Using a generalized version of GREML, Yang *et al.* (2011) demonstrated that the contribution to the heritability of human height from each chromosome is proportional to chromosome length, suggesting a polygenic genetic architecture for height. Further, via a similar approach, it was shown that heritability for height and body mass index is enriched in variants with lower minor allele frequencies, hinting at selection effects (Yang *et al.*, 2015). These results were very interesting, but since GREML required individual-level genotype and phenotype data as input, it was not easy to apply GREML to other GWASs, especially when the GWAS was a meta-analysis of results from multiple groups.

LD score regression was introduced in 2014 and has quickly gained popularity. First introduced as a method to distinguish polygenicity from unadjusted confounding in GWASs (Bulik-Sullivan *et al.*, 2015b), LD score regression only requires GWAS summary statistics and externally estimated LD as inputs. It is based on the same model that was previously used in GREML:

$$Y = X\beta + \varepsilon, \quad \beta \sim N\left(0, \frac{h^2}{m}I\right), \quad \varepsilon \sim N(0, (1 - h^2)I),$$

where X is the standardized genotype matrix; Y is the standardized trait value; β and ε denote genetic and non-genetic effects, respectively; h^2 denotes heritability; and m is the number of SNPs. When there is no unadjusted confounding in the model, it can be shown that

$$E(z_j^2) = \frac{nh^2}{m}l_j + 1,$$

where z_j is the z -score of the j th SNP in the GWAS; n denotes the GWAS sample size; and l_j , the LD score for the j th SNP, is defined as the sum of LD between the j th SNP and all SNPs in the data:

$$l_j = \sum_{k=1}^m r_{jk}^2.$$

Then, when regressing z_j^2 on LD scores l_j , the weighted least squares estimator for regression coefficient can be used to estimate heritability h^2 . More interestingly, LD score regression can be extended to estimate annotation-dependent heritability (Finucan *et al.*, 2015). When K functional annotations are present, the model is generalized as follows:

$$Y = \sum_{i=1}^K X_i \beta_i + \varepsilon, \quad \beta_i \sim N\left(0, \frac{h_i^2}{m_i}I\right), \quad \varepsilon \sim N\left(0, \left(1 - \sum_{i=1}^K h_i^2\right)I\right),$$

where X_i is the genotype matrix for m_i SNPs in the i th functional annotation category, and h_i^2 is the fraction of phenotypic variance that can be explained by SNPs in the i th annotation. Based on this model, LD score regression becomes a multiple regression problem,

$$E(z_j^2) = \sum_{i=1}^K \frac{nh_i^2}{m_i} l_j^{(i)} + 1.$$

Here, the annotation stratified LD score $l_j^{(i)}$ is defined as the sum of LD between the j th SNP and all SNPs in the i th functional annotation:

$$l_j^{(i)} = \sum_{k \in A_i} r_{jk}^2.$$

Similar to the univariate case, regression coefficient estimates can be used to estimate heritability parameters.

Heritability enrichment analysis based on LD score regression has become routine in GWASs. Using the ratio between the proportion of heritability explained by the SNPs in a functional annotation and the proportion of genome covered by the annotation to quantify enrichment for GWAS associations, LD score regression can identify functional annotations that are relevant to the disease of interest and help generate novel hypotheses about disease etiology. Even more interestingly, when tissue-specific functional annotations are used, this type of enrichment analysis can robustly identify disease-related tissue and cell types (Finucan *et al.*, 2015, 2018; Lu *et al.*, 2016a, 2017b). Despite its success, LD score regression has limitations. First, no constraint is applied to heritability estimates, therefore it is possible to get non-interpretable estimates that are below 0 or above 1. Second, variance of estimators is calculated using a resampling method, the blockwise jackknife. For moderate GWAS sample size, this approach often leads to very wide confidence intervals for enrichment, thereby reducing statistical power. Finally, z -scores for SNPs in LD may be strongly correlated. LD score regression reduces the impact of dependence among data points by using specially designed weights in weighted least squares estimation and only including HapMap SNPs in the model. Still, it remains unclear if such empirical approaches are sufficient to remove bias. Since LD score regression was introduced, a few methods have been proposed to fix some of these issues (Zhou, 2017; Speed *et al.*, 2017; Bulik-Sullivan, 2015; Hecker *et al.*, 2018; H. Shi *et al.*, 2016).

After its success in heritability estimation, GREML was generalized to study shared genetics across multiple complex traits (Lee *et al.*, 2012; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). LD score regression can be extended in a similar way (Bulik-Sullivan *et al.*, 2015a). Cross-trait LD score regression is based on the following model:

$$\begin{aligned} Y_1 &= X\beta + \varepsilon, \quad \beta \sim N\left(0, \frac{h_1^2}{m}I\right), \quad \varepsilon \sim N\left(0, (1 - h_1^2)I\right), \\ Y_2 &= Z\gamma + \delta, \quad \gamma \sim N\left(0, \frac{h_2^2}{m}I\right), \quad \delta \sim N\left(0, (1 - h_2^2)I\right). \end{aligned}$$

SNPs' effects on two different traits are assumed to be correlated:

$$E(\beta\gamma^T) = \frac{\rho_g}{m}I,$$

where ρ_g is the genetic covariance parameter that quantifies the genetic sharing between traits Y_1 and Y_2 . Importantly, LD score regression allows two GWASs to share a fraction of samples. Without loss of generality, assume the first n_s samples (i.e. rows) in X and Z are identical and the non-genetic effects for shared samples are correlated:

$$E(\varepsilon_i\delta_i) = \rho_e, \quad i = 1, \dots, n_s$$

Then, the cross-trait LD score regression formula is

$$E((z_1)_j(z_2)_j) = \frac{\sqrt{n_1 n_2} \rho_g}{m} l_j + \frac{(\rho_g + \rho_e) n_s}{\sqrt{n_1 n_2}}$$

Similar to single-trait analysis, regression coefficients can be used to estimate genetic covariance, or a close-related but more interpretable metric – genetic correlation:

$$\text{corr} = \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}$$

A recent method, GNOVA (Lu *et al.*, 2017a), has extended cross-tissue LD score regression to model annotation-stratified genetic covariance. When K functional annotations are present in the model,

$$\begin{aligned} Y_1 &= \sum_{i=1}^K X_i \beta_i + \varepsilon, \\ Y_2 &= \sum_{i=1}^K Z_i \gamma_i + \chi, \\ E(\beta_i \gamma_i^T) &= \frac{\rho_i}{m_i} I. \end{aligned}$$

Parameters $\rho_i (i = 1, \dots, K)$ quantify the genetic covariance components for each functional annotation. GNOVA used an estimator based on the method of moments to estimate genetic covariance:

$$\begin{pmatrix} \hat{\rho}_1 \\ \vdots \\ \hat{\rho}_K \end{pmatrix} = \frac{1}{\sqrt{n_1 n_2}} \begin{pmatrix} \frac{1}{m_1 m_1} l_{11} & \cdots & \frac{1}{m_K m_1} l_{K1} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1 m_K} l_{1K} & \cdots & \frac{1}{m_K m_K} l_{KK} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{m_1} (z_1)_1^T (z_2)_1 \\ \vdots \\ \frac{1}{m_K} (z_1)_K^T (z_2)_K \end{pmatrix},$$

where l_{ij} denotes the sum of all pairwise LD between SNPs in the i th and j th functional annotations; $(z_j)_i$ denotes the vector of z -scores from the j th study for SNPs in the i th functional annotation. To date, LD score regression is not able to estimate annotation-stratified genetic covariance. GNOVA also shows higher statistical power than LD score regression in both simulations and real data (Lu *et al.*, 2017a).

Since their introduction, these summary statistics-based methods have been applied to dissect the complex relationship among many complex diseases and traits (Anttila *et al.*, 2016; Tylee *et al.*, 2018) and are being used in most recently-conducted GWASs. A few new methods have been recently developed to estimate local genetic correlation for specific risk loci (Shi *et al.*, 2017), perform trans-ethnic genetic correlation estimation (Brown *et al.*, 2016), and improve genetic correlation estimates for case–control studies (Weissbrod *et al.*, 2018). Further, online servers have been developed for researchers to estimate genetic correlation between their data and hundreds of traits with publicly accessible summary statistics (Zheng *et al.*, 2017).

24.4.2 Imputation-Based Gene-Level Association Analysis

GWASs are effective in identifying SNP–trait associations. However, due to LD at the associated loci, it is not always clear what genes may be involved in disease etiology. Methods have

been developed to convert SNP-level *p*-values obtained in GWASs into association *p*-values for genes. Early methods, such as VEGAS (Liu *et al.*, 2010) often tested if at least one SNP in a predefined window surrounding the candidate gene is associated to disease. These methods have evolved over the years, and recent advances have shown much improved statistical power and computational speed (Barnett *et al.*, 2017; Sun and Lin, 2017). However, these methods do not take into account functional annotations and the results depend on the width of predefined windows. If the windows are too wide, many genes in the region may show similar *p*-values; if the windows are too narrow, users will only be able to identify genes that are close to significant SNPs, which may not be truly causal, as demonstrated in many examples (Claussnitzer *et al.*, 2015; Whalen *et al.*, 2016; GTEx Consortium, 2017).

Expression quantitative trait loci are a type of functional annotation that naturally connects SNPs, genes, and traits. Co-localization methods that aim to identify target genes by aligning GWAS associations with eQTL association statistics have been developed (Nicolae *et al.*, 2010; Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016). However, recent advances in eQTL analysis have revealed that a large fraction of common SNPs in the genome are associated with expression of at least one gene in at least one tissue (GTEx Consortium, 2017), which casts a shadow on co-localization analyses due to potential false positives. Two recent methods, PrediXcan (Gamazon *et al.*, 2015) and TWAS (Gusev *et al.*, 2016), introduced an innovative approach to integrating eQTL with GWAS data. First, an imputation model is trained on data from a reference panel (e.g. GTEx) to predict gene expression in a given tissue (e.g. whole blood) using SNPs. After the imputation model is trained, genotype data in GWASs are used to impute the expression of the given gene. Finally, an association test is performed between the disease/trait of interest and the imputed gene expression values via a simple regression analysis:

$$Y = \alpha + \hat{G}\gamma + \delta,$$

where Y denotes trait values; \hat{G} is the vector of imputed gene expression; and δ is a term for random noise. Interestingly, if \hat{G} is imputed from the genotype matrix X via a linear function

$$\hat{G} = XW,$$

then it can be shown that the test statistic (i.e. *z*-score) for γ is a function of GWAS summary statistics (Gusev *et al.*, 2016; Barbeira *et al.*, 2018)

$$Z = \frac{\hat{\gamma}}{se(\hat{\gamma})} \approx W^T \begin{pmatrix} \frac{\sigma_1}{\eta} \\ \ddots \\ \frac{\sigma_m}{\eta} \end{pmatrix} \tilde{Z}.$$

Here, \tilde{Z} denotes the vector of SNP-level *z*-scores in the GWAS; σ_j denotes the standard deviation of the j th SNP and η denotes the standard deviation of imputed expression values, both of which can be estimated using a reference genotype panel (e.g. 1000 Genomes Project Consortium, 2012). Therefore, gene-level association tests can be performed without having access to individual-level genotype and phenotype data.

Compared to classic gene-level association tests, PrediXcan and TWAS effectively integrate eQTL annotations with GWAS summary statistics and identify biologically interpretable gene targets. Compared to directly testing for differential expression between disease cases and healthy controls, these methods can reduce reverse causality because gene expression imputation models are trained in an independent, healthy panel. Therefore, significant genes identified by PrediXcan and TWAS are more likely to be upstream of disease. However, it has been

pointed out that due to complex LD, co-regulation, and coexpression patterns in the genome and transcriptome, PrediXcan and TWAS do not guarantee causal effects between significant genes and disease (Wainberg *et al.*, 2017). Therefore, association results still need to be interpreted with caution.

These imputation-based gene-level association methods have gained much attention in the past two years and have been applied to a large number of complex diseases and traits (Mancuso *et al.*, 2017a). A few methods have been developed to further improve their performance. Xu *et al.* (2017) proposed an adaptive test that generalized PrediXcan, TWAS, and VEGAS into a more powerful metric. Recently, a cross-tissue TWAS framework, UTMOST, was introduced (Hu *et al.*, 2019). UTMOST imputes gene expression in 44 tissues simultaneously via a multivariate penalized regression model and combines single-tissue association statistics into a powerful metric via a generalized Berk–Jones test framework. There have also been efforts to expand these methods to incorporate other types of annotations such as splicing quantitative trait loci (Raj *et al.*, 2018). Finally, fine-mapping methods have been proposed to identify biologically relevant genes among co-regulated gene candidates (Mancuso *et al.*, 2017a).

24.4.3 Other Applications and Future Directions

Here we have mainly focused on two applications of functional annotations in GWAS analyses: partitioning heritability and genetic covariance, and imputation-based transcriptome-wide association analysis. The former utilizes genome annotations at the base level and helps dissect the genetic architecture of complex traits, while the latter mainly uses eQTL annotations to identify target genes associated to disease. We note that this is a very active field and applications of functional annotations in GWASs are not limited to what we introduced above. Numerous methods have been developed to integrate functional annotations with GWAS data, especially GWAS summary statistics, to fine-map functional SNPs (Pickrell, 2014; Lu *et al.*, 2016b; Li and Kellis, 2016; Kichaev *et al.*, 2014; Kichaev and Pasaniuc, 2015), to increase statistical power in association tests (Kichaev *et al.*, 2019; Eskin, 2008; Sveinbjornsson *et al.*, 2016; He *et al.*, 2017), and to improve risk prediction accuracy (Speed and Balding, 2014; Hu *et al.*, 2017; J. Shi *et al.*, 2016). We also note that more types of functional annotations, for example, ATAC-seq (Buenrostro *et al.*, 2015), Hi-C (Lieberman-Aiden *et al.*, 2009), and single-cell assays (Rotem *et al.*, 2015; Svensson *et al.*, 2018), are being generated and curated. It remains an open question how to utilize these new annotations to benefit genetic association studies. Further, many functional annotation tools have been developed and it is challenging to decide which tools may be suitable for a specific study. It is of practical interest to benchmark different types of functional annotations in terms of their enrichment for GWAS associations and performance in various GWAS applications (Li *et al.*, 2017). Finally, it remains to be investigated if methods developed for GWASs can be directly applied to sequencing-based studies in which rare variants play a significant role. Nevertheless, the increasingly accessible association summary statistics with large sample sizes, coupled with rich annotation data generated by large consortia, provide great opportunities for methodology development and shed light on the future of complex trait research.

Acknowledgements

This study was supported in part by the National Institutes of Health grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of

Research and Development. Dr. Lu's work was supported by the Clinical and Translational Science Award (CTSA) program through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427.

References

- 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., et al. (2010). A method and server for predicting damaging missense mutations. *Nature Methods* **7**(4), 248–249.
- Anttila, V., Bulik-Sullivan, B., Finucane, H.K., et al. (2016). Analysis of shared heritability in common disorders of the brain. Preprint, bioRxiv 048991.
- Backenroth, D., He, Z., Kiryluk, K. et al. (2018) FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: Methods and application. *American Journal of Human Genetics* **102**(5), 920–942.
- Barbeira, A.N., Dickinson, S.P., Bonazzola, R., et al. (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1825.
- Barnett, I., Mukherjee, R. and Lin, X. (2017) The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association* **112**(517), 64–76.
- Bodea, C.A., Mitchell, A.A., Day-Williams, A.G., et al. (2016). Phenotype-specific information improves prediction of functional impact for noncoding variants. Preprint, bioRxiv 083642.
- Boyle, A.P., Hong, E.L., Hariharan, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22**(9), 1790–1797.
- Boyle, E.A., Y.I. Li., and J.K. Pritchard, (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169**(7), 1177–1186.
- Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C.J., et al. (2016). Transethnic genetic-correlation estimates from summary statistics. *American Journal of Human Genetics* **99**(1), 76–88.
- Buenrostro, J.D., Wu, B., Chang, H.Y. and Greenleaf, W.J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology* **109**, 21.29.1–9.
- Bulik-Sullivan, B. (2015) Relationship between LD score and Haseman-Elston regression. Preprint, bioRxiv 018283.
- Bulik-Sullivan, B., Finucane, H.K., Anttila, V., et al., (2015a). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**(11), 1236–1241.
- Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., et al., (2015b). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**(3), 291–295.
- Chan, R.C., Libbrecht, M.W., Roberts, E.G., et al. (2018). Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics* **34**(4), 669–671.
- Chen, L., Jin, P. and Qin, Z.S. (2016). DIVAN: Accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biology* **17**(1), 252.
- Cingolani, P., Platts, A., Wang, L.L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w¹¹⁸; iso-2; iso-3*. *Fly* **6**(2), 80–92.
- Claussnitzer, M., Dankel, S.N., Kim, K.-H. et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine* **373**(10), 895–907.

- Cooper, G.M., Stone, E.A., Asimenos, G., *et al.* (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**(7), 901–913.
- Cross-Disease Group of the Psychiatric Genomics Consortium (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics* **45**(9), 984–994.
- Davydov, E.V., Goode, D.L., Sirota, M., *et al.* (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**(12), e1001025.
- Dong, C., Wei P., Jian X. *et al.* (2014). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**(8), 2125–2137.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74.
- Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28**(8), 817–825.
- Ernst, J. and Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods* **9**(3), 215.
- Eskin, E. (2008). Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Research* **18**(4), 653–660.
- Finucane, H.K., Bulik-Sullivan, B., Gusev, A., *et al.* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**(11), 1228–1235.
- Finucane, H.K., Reshef, Y.A., Anttila, V., *et al.* (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics* **50**(4), 621–629.
- Flicek, P., Amode, M.R., Barrell, D., *et al.* (2012). Ensembl 2012. *Nucleic Acids Research* **40**(D1), D84–D90.
- Franzén, O., Ermel, R., Cohain, A., *et al.* (2016). Cardiometabolic risk loci share downstream *cis*- and *trans*-gene regulation across tissues and diseases. *Science* **353**(6301), 827–830.
- Fromer, M., Roussos, P., Sieberts, S.K., *et al.* (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19**(11), 1442.
- Fu, Y., Liu, Z., Lou, S., *et al.* (2014). FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology* **15**(10), 480.
- Gamazon, E.R., Wheeler, H.E., Shah, K.P., *et al.* (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**(9), 1091–1098.
- Giambartolomei, C., Vukcevic, D., Schadt, E.E., *et al.* (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**(5), e1004383.
- GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**(6235), 648–660.
- GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* **550**(7675), 204–213.
- Gusev, A., Ko, A., Bhatia, G., *et al.*, (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**(3), 245–252.
- He, Z., Xu, B., Lee, S. *et al.* (2017). Unified Sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *American Journal of Human Genetics* **101**(3), 340–352.

- Hecker, J., Prokopenko, D., Lange, C., et al. (2018). PolyGEE: A generalized estimating equation approach to the efficient and robust estimation of polygenic effects in large-scale association studies. *Biostatistics* **19**(3), 295–306.
- Hindorff, L.A., Sethupathy, P., Junkins, H.A., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**(23), 9362–9367.
- Hoffman, M.M., Buske, O.J., Wang, J., et al. (2012a). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**(5), 473.
- Hoffman, M.M., Ernst, J., Wilder, S.P., et al. (2012b). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research* **41**(2), 827–841.
- Hormozdiari, F., van de Bunt, M., Segre, A.V., et al. (2016). Colocalization of GWAS and eQTL signals detects target genes. *American Journal of Human Genetics* **99**(6), 1245–1260.
- Hu, Y., Lu, Q., Powles, R., et al. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Computational Biology* **13**(6), e1005589.
- Hu, Y., Li, M., Lu, Q., et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics* **51**(3), 568–576.
- Huang, Y.-F., Gukko, B. and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics* **49**(4), 618.
- Ionita-Laza, I., McCallum, K., Xu, B., et al. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics* **48**(2), 214–220.
- Javed, A., Agrawal, S. and Ng, P.C. (2014). Phen-Gen: Combining phenotype and genotype to analyze rare disorders. *Nature Methods* **11**(9), 935–937.
- Jiang, J., Li, C., Paul, D., et al. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *Annals of Statistics* **44**(5), 2127–2160.
- Kent, W.J., Sugnet, C.W., Furey, T.S., et al. (2002). The human genome browser at UCSC. *Genome Research* **12**(6), 996–1006.
- Khurana, E., Fu, Y., Colonna, V., et al. (2013). Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science* **342**(6154), 1235587.
- Kichaev, G. and Pasaniuc, B. (2015). Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *American Journal of Human Genetics* **97**(2), 260–271.
- Kichaev, G., Wang, W.Y., Lindstrom, S., et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics* **10**(10), e1004722.
- Kichaev, G., Bhatia, G., Loh, P.-R., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *American Journal of Human Genetics* **104**(1), 65–75.
- Kircher, M., Witten, D.M., Jain, P., et al. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**(3), 310–315.
- Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**(12), 1452–1458.
- Lawrence, M., Daujat, S. and Schneider, R. (2016). Lateral thinking: How histone modifications regulate gene expression. *Trends in Genetics* **32**(1), 42–56.
- Lee, D., Gorkin, D.U., Baker, M., et al. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics* **47**(8), 955.
- Lee, S.H., Yang, J., Goddard, M.E., et al. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**(19), 2540–2542.
- Li, B., Lu, Q. and Zhao, H. (2017). An evaluation of noncoding genome annotation tools through enrichment analysis of 15 genome-wide association studies. *Briefings in Bioinformatics*. doi: 10.1093/bib/bbx131.

- Li, Y. and Kellis, M. (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research* **44**(18), e144.
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950), 289–293.
- Lindblad-Toh, K., Garber, M., Zuk, O., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**(7370), 476–482.
- Liu, J.Z., McRae, A.F., Nyholt, D.R. et al. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics* **87**(1), 139–145.
- Lu, Q., Hu, Y., Sun, J. et al. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific Reports* **5**, 10576.
- Lu, Q., Powles, R.L., Wang, Q., et al., (2016a). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genetics* **12**(4), e1005947.
- Lu, Q., Yao, X., Hu, Y., et al. (2016b). GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics* **32**(4), 542–548.
- Lu, Q., Li, B., Ou, D., et al. (2017a). A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *American Journal of Human Genetics* **101**(6), 939–964.
- Lu, Q., Powles, R.L., Abdallah, S., et al. (2017b). Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genetics* **13**(7), e1006933.
- Mancuso, N., Kichaev, G., Shi, H., et al. (2017a). Probabilistic fine-mapping of transcriptome-wide association studies. Preprint, bioRxiv 236869.
- Mancuso, N., Shi, H., Goddard, P., et al. (2017b). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *American Journal of Human Genetics* **100**(3), 473–487.
- Manolio, T.A., Collins, F.S., Cox, N.J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753.
- Maurano, M.T., Humbert, R., Ynes, E., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**(6099), 1190–1195.
- Ng, P.C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Research* **11**(5), 863–874.
- Nicolae, D.L., Gamazon, E., Zhang, W., et al., (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* **6**(4), e1000888.
- Parisi, F., Strino, F., Nadler, B., et al. (2014). Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences of the United States of America* **111**(4), 1253–1258.
- Pasaniuc, B. and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18**(2), 117–127.
- Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics* **94**(4), 559–573.
- Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., et al. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**(1), 110–121.
- PsychENCODE Consortium (2015). The PsychENCODE project. *Nature Neuroscience* **18**(12), 1707–1712.
- Quang, D., Chen, Y. and Xie, X. (2014). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**(5), 761–763.

- Raj, T., Li, Y.I., Humphrey, J. *et al.* (2018). Integrative analyses of splicing in the aging brain: Role in susceptibility to Alzheimer's disease. *Nature Genetics* **50**(11), 1584–1592.
- Ritchie, G.R., Dunham I., Zeggini E., *et al.* (2014). Functional annotation of noncoding sequence variants. *Nature Methods* **11**(3), 294–296.
- Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518**(7539), 317–330.
- Rogers, M.F., Shihab, H.A., Mort, M., *et al.* (2018). FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**(3), 511–513.
- Rotem, A., Ram, O., Shores, N., *et al.* (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology* **33**(11), 1165.
- Schaub, M.A., Boyle, A.P., Kundaje, A., *et al.* (2012). Linking disease associations with regulatory information in the human genome. *Genome Research* **22**(9), 1748–1759.
- Schwarz, J.M., Rodelsperger, C., Schuelke, M., *et al.* (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**(8), 575–576.
- Shi, H., Kichaev, G. and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *American Journal of Human Genetics* **99**(1), 139–153.
- Shi, H., Mancuso, N., Spendlove, S., *et al.* (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *American Journal of Human Genetics* **101**(5), 737–751.
- Shi, J., Park, J.H., Duan, J., *et al.* (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genetics* **12**(12), e1006493.
- Siepel, A., Bejerano, G., Pedersen, J.S., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**(8), 1034–1050.
- Singleton, M.V., Guthrey, S.L., Voelkerding, K.V., *et al.* (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics* **94**(4), 599–610.
- Speed, D. and Balding, D.J. (2014). MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Research* **24**(9), 1550–1557.
- Speed, D., Cai, N., UCLEB Consortium, *et al.* (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**(7), 986.
- Stunnenberg, H.G., International Human Epigenome Consortium and Hirst, M. (2016). The International Human Epigenome Consortium: A blueprint for scientific collaboration and discovery. *Cell* **167**(5), 1145–1149.
- Sun, R. and Lin, X. (2017). Set-based tests for genetic association using the generalized Berk-Jones statistic. Preprint, arXiv:1710.02469.
- Sveinbjornsson, G., Albrechtsen, A., Zink, F., *et al.* (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nature Genetics* **48**(3), 314.
- Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**(4), 599.
- Tobacco and Genetics Consortium (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics* **42**(5), 441–447.
- Trynka, G., Sandor, C., Han, B., *et al.* (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45**(2), 124–130.
- Tylee, D.S., Sun, J., Hess, J.L., *et al.* (2018). Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *American Journal of Medical Genetics B* **177**(7), 641–657.

- Visscher, P.M., Wray, N.R., Zhang, Q., *et al.* (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics* **101**(1), 5–22.
- Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., *et al.* (2017). Vulnerabilities of transcriptome-wide association studies. Preprint, bioRxiv 206961.
- Wang, K., Li, M. and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**(16), e164.
- Ward, L.D. and Kellis, M. (2012a). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* **30**(11), 1095–1106.
- Ward, L.D. and Kellis, M. (2012b). HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40**(Database issue), D930–D934.
- Weissbrod, O., Flint, J. and Rosset, S. (2018). Estimating heritability and genetic correlation in case control studies directly and with summary statistics. Preprint, bioRxiv 256388.
- Whalen, S., ruty, R.M.T. and Pollard, K.S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48**(5), 488.
- Xu, Z., Wu, C., Wei, P., *et al.* (2017). A powerful framework for integrating eQTL and GWAS summary data. *Genetics* **207**(3), 893–902.
- Yang, J., Benyamin, B., McEvoy, B.P., *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**(7), 565–569.
- Yang, J., Manolio, T.A., Pasquale, L.R., *et al.* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**(6), 519–525.
- Yang, J., Bakshi, A., Zhu, Z., *et al.* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**(10), 1114.
- Zheng, J., Mesut Erzurumluoglu, A., Elsworth, B.L., *et al.* (2017). LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for S2.NP heritability and genetic correlation analysis. *Bioinformatics* **33**(2), 272–279.
- Zhou, J. and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**(10), 931.
- Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Annals of Applied Statistics* **11**(4), 2027–2051.

25

Inferring Causal Associations between Genes and Disease via the Mapping of Expression Quantitative Trait Loci

Solveig K. Sieberts¹ and Eric E. Schadt^{2,3}

¹Sage Bionetworks, Seattle, WA, USA

²Sema4, Stamford, CT, USA

³Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Abstract

The ability to monitor molecular traits across populations and over time in a comprehensive fashion allows for a more general characterization of gene networks and their relationship to disease and other complex physiological processes. Information on how variations in DNA impact complex physiological processes flows through molecular networks. Therefore, integrating DNA variation, molecular phenotypes such as gene or protein expression, and phenotypic data has the potential to enhance identification of the associations between DNA variation and disease, as well as characterize those parts of the molecular networks that drive disease. Towards that end, we discuss mapping expression quantitative trait loci (eQTL) for gene expression traits and then detail a method for integrating eQTL, expression and clinical data to infer causal relationships among gene expression traits and between expression and clinical traits, with natural generalizations to other molecular traits such as protein expression, metabolite levels, and epigenomic changes. We further describe methods to integrate these data in a more comprehensive manner by constructing interaction-based networks, such as weighted gene coexpression networks, which leverage gene interaction data to represent more general relationships. To infer gene networks that capture causal information we describe a Bayesian algorithm that further integrates eQTL, expression and clinical phenotype data to reconstruct whole gene networks capable of representing causal relationships among genes and traits in the network. The flexibility of the Bayesian framework allows for the incorporation of more general structure via appropriate priors, such as DNA–protein binding, protein–protein interactions, protein–small molecule interactions and so on.

25.1 Introduction

DNA microarrays and now RNA-sequencing (RNA-seq) technologies have revolutionized the way we study genes and the role they play in everything from the regulation of normal cellular processes to complex diseases such as obesity and cancer. These high-throughput technologies are capable of simultaneously providing quantitative measures of RNA present in single cells or complex mixtures of cells from tissues and organs, enabling the profiling of tens of

thousands of protein-coding genes, in addition to what is now a great diversity of non-coding transcribed sequences. Newer RNA-seq technologies can also allow for the mapping of the expressed RNA to specific isoforms, RNA editing, and allelic imbalance. The quantitative measures of transcript abundances in cells are often referred to as gene expression traits. In their typical use, gene expression assays allow researchers to screen thousands of genes for differences in expression levels or differences in how genes are connected in the network (Califano *et al.*, 2012; Zhang *et al.*, 2013; Franzen *et al.*, 2016; Peters *et al.*, 2017) between experimental conditions of interest (see **Chapters 26, 30 and 31**; Huber *et al.*, 2007; Pounds *et al.*, 2007, Sections 6.4 and 6.5; Lewin and Richardson, 2007). These data are often used to discover genes that differ between normal and diseased tissue, to model and predict continuous or binary measures, to predict patient survival, and to classify disease or tumor sub-types. Because gene expression levels in a given sample are measured simultaneously, researchers are able to identify genes whose expression levels are correlated, implying an association under specific conditions or more generally. Of course, gene expression is just one of many types of molecular phenotypes: metabolomic, proteomic, and epigenomic measurements also provide useful insights into cellular and higher-order system functioning. However, the focus of this chapter is on gene expression data – analogous approaches can generally be applied for these alternative omics measurements.

Causal associations among genes or between genes and traits have also been investigated using time series experiments, gene knockouts or transgenics that over-express a gene of interest, RNAi-based knockdown or viral-mediated over-expression of genes of interest, and chemical activation or inhibition of genes of interest. However, a number of studies have demonstrated that naturally occurring DNA polymorphism can be used to help establish causal associations, since gene expression and other molecular phenotypes in a number of species have been shown to be significantly heritable and at least partially under the control of specific genetic loci (Schadt *et al.*, 2003, 2005b, 2008; Doss *et al.*, 2005; Zhu *et al.*, 2008; Greenawalt *et al.*, 2011; Zhu *et al.*, 2012; Zhang *et al.*, 2013; Franzen *et al.*, 2016; Peters *et al.*, 2017). By examining the effects that naturally occurring variations in DNA have on variations in gene expression traits in human or experimental populations, other phenotypes (including disease) can then be examined with respect to these same DNA variants and ultimately ordered with respect to genes to infer causal control (Figure 25.1). The power of this integrative genomics strategy rests in the molecular processes that transcribe DNA into RNA and then RNA into protein, such that information on how variations in DNA impact complex physiological processes often flows directly through transcriptional networks. As a result, integrating DNA variation, expression, and phenotypic data has the potential to enhance identification of the associations between DNA variation and disease, as well as to characterize those parts of the molecular networks that drive disease.

A number of groups have published approaches for identifying key drivers of complex traits by examining genes located in regions of the genome genetically linked to a complex phenotype of interest, and then looking for colocalization of *cis*-acting expression quantitative trait loci (eQTL) for those genes residing in a genomic region linked to the phenotype (Jansen and Nap, 2001; Brem *et al.*, 2002; Schadt *et al.*, 2003, 2005a; Monks *et al.*, 2004; Morley *et al.*, 2004; Alberts *et al.*, 2005; Chesler *et al.*, 2005; Cheung *et al.*, 2005; Petretto *et al.*, 2006; He *et al.*, 2013; Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2014; Chun *et al.*, 2017). Those genes with (1) expression values that are significantly correlated with the complex phenotype of interest (including disease), (2) gene expression abundances controlled by QTL that colocalize with the phenotype QTL, and (3) physical locations supported by the phenotype and expression QTL, are natural causal candidates for the complex phenotype of interest. Since DNA variation leads to changes in transcription and other molecular trait activity, it can be used to partition the thousands of gene expression traits that may be correlated with a given phenotype into sets

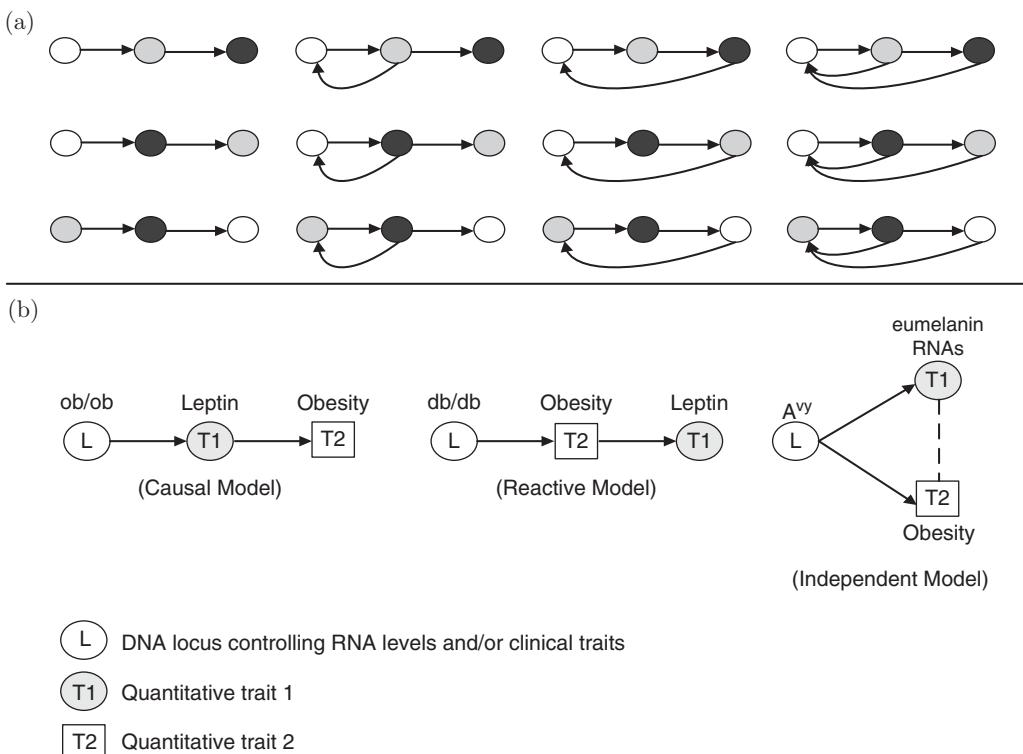


Figure 25.1 Possible relationships between phenotypes with and without genetic information. Edges between nodes in each of the graphs represent an association between the nodes. A directed edge indicates a causal association between the nodes. (a) A subset of the number of possible relationships between three variables. (b) The set of possible relationships between two traits and a controlling genetic locus, when feedback mechanisms are ignored.

of genes that are supported as causal for, reacting to, or independent of the given phenotype. The key to the success of this approach is the unambiguous flow of information from changes in DNA to changes in RNA and protein function (Figure 25.1). That is, given that two traits are linked to the same DNA locus, and a few simplifying assumptions discussed below, there are a limited number of ways in which such traits can be related with respect to that locus (Schadt, 2005, 2006; Schadt *et al.*, 2005a), whereas in the absence of such genetic information many indistinguishable relationships would be possible, so that additional information would be required to establish the correct relationships.

Here we discuss general approaches to causal modeling of biological processes. We then focus on methods that incorporate genetic information to infer causal directions, starting with general principles for mapping eQTL for gene expression traits, and detail a method for integrating eQTL, expression and clinical data to infer causal relationships among gene expression traits and between expression and clinical traits. We further describe methods to integrate these data in a more comprehensive manner by constructing coexpression gene networks, which leverage pairwise gene interaction data to represent more general relationships. This type of network provides a useful construct for characterizing the topological properties of biological networks and for partitioning such networks into functional units (modules) that underlie complex phenotypes such as disease. However, these networks are, by design, undirected and so do not capture causal relationships among genes. To infer gene networks that capture causal information we describe a Bayesian algorithm that, like the methods operating on only two or three

expression traits and/or clinical traits mentioned above, integrates eQTL, expression and clinical phenotype data to reconstruct whole gene networks capable of representing direction along the edges of the network. Here, directionality among the edges corresponds to causal relationships among genes and between genes and clinical phenotypes. These high-dimensional data analysis approaches, which integrate large-scale data from multiple sources, have contributed to statistical genetics moving away from considering one trait (or even locus) at a time and towards operating in a network context.

25.1.1 An Overview of Transcription as a Complex Process

Messenger RNA (mRNA) is an intermediate product of a process that generates protein starting with DNA, where DNA is transcribed to produce RNA, and then RNA is translated to produce protein. The nucleus is where DNA is transcribed into its RNA analog, and in the case where RNA corresponds to a protein-coding gene, the RNA is processed into a mature mRNA form. The mRNA then migrates to the cytoplasm where it is translated, via transfer RNA, into a protein, which is typically considered the functional product of the gene. Thus, gene expression measures corresponding to protein-coding genes are considered as surrogates for either the state of a protein or the amount of protein that is being produced by the corresponding gene. The amount of mRNA detected in the cytoplasm of a cell is itself the result of a number of different molecular processes, including the process of transcription itself, the rate of RNA degradation (which is sequence specific and often sensitive to different functional forms of a protein), transport from the nucleus, and alternative splicing.

The overall regulation of transcription is a complex process, especially in eukaryotic cells. Transcription begins when RNA polymerase binds to specific DNA sequences known as promoters, which are generally located close to the 5' end of the gene. In order to prevent the binding of the polymerase, and thus regulate the otherwise constant transcription of a given gene sequence, many genes have negative control elements called operators, which when bound by specific proteins called repressor proteins prevent the binding of the polymerase to the promoter region. Using these simple positive and negative controllers, cells can regulate the transcription of DNA sequences in response to specific conditions or, as appropriate, for a given tissue type. In eukaryotes, promoters and operators are usually complemented by other regulatory sequences that can operate at much longer distances, often tens of kilobases (kb) or even longer. Some elements, such as enhancers, encourage and speed transcription, while others, such as silencers, discourage and slow transcription. While promoters and operators cause binary (on/off) regulation, enhancers and silencers can have much more subtle effects on transcription. Nevertheless, mutations in any of these sequence elements can alter gene expression at the basal level or affect a gene's ability to react under varied cellular conditions. These regulatory elements, which occur in or near the gene transcription region, are referred to as *cis*-acting elements (Figure 25.2).

Complementary to the sequence-based regulatory elements are the protein elements that interact with the sequence elements to regulate transcription. Repressor proteins are one example of regulatory proteins that directly bind to DNA sequences. Other proteins may bind to the promoter region to increase the affinity of RNA polymerase to the region and, thus, are positive controllers of transcription. In eukaryotes, in order for polymerase to bind the DNA and begin transcription, a number of additional molecules must first bind the promoter region. These transcription factor binding-associated proteins (TAFs) form transcription factor complexes that help position the polymerase in the promoter region. This complex alone provides basal regulation of transcription but may also be enhanced by additional ancillary proteins called activators, which bind to enhancer sequence elements to enhance the rate of transcription.

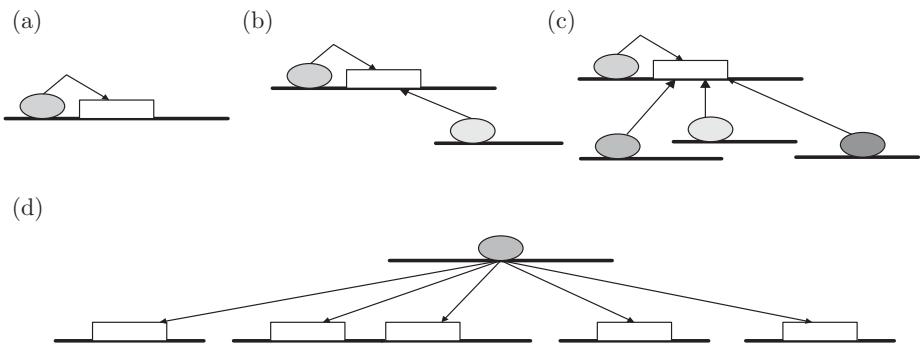


Figure 25.2 Mapping proximal and distal eQTL for gene expression traits. The white rectangles represent genes that are controlled by transcriptional units. The ellipses represent the transcriptional control units, which could be transcription regulatory sites, other genes that control the expression of the indicated gene, and so on. (a) *Cis*-acting control unit acting on a gene. DNA variations in this control unit that affected the gene's expression would lead to a *cis*-acting (proximal) eQTL. (b) *Cis* and *trans* control units regulating the indicated gene. DNA variations in these control units that affected the gene's expression would lead to proximal and distal eQTL. (c) *Cis* control unit and multiple *trans* control units regulating the indicated genes. DNA variations in these control units would lead to a complex eQTL signature for the gene. (d) A single control unit regulating multiple genes. DNA variations in this single control unit could lead to a cluster of distal eQTL (an eQTL hot spot).

Repressor proteins may bind to silencer sequence elements in the DNA to slow transcription. Of course, these regulatory proteins may themselves be regulated, where feedback control or other such processes are known to regulate the quantities of these proteins in the cell. Other regulatory proteins may require that allosteric effectors be in their active form or that they become inactivated by effector proteins. Other proteins may inhibit the formation of the transcription factor complexes. These elements, which are usually encoded far from the genes they regulate, are called *trans*-acting elements (Figure 25.2).

An additional mechanism affecting gene expression is epigenetics. The most commonly studied epigenetic modification is DNA methylation, in which a methyl group can be added to the cytosine (or in specific instances adenine) DNA molecule, which in turn can interfere with gene expression, especially when occurring in a promoter region. This is especially so in G/C-rich DNA sequences (see Chapter 33). Methylation activity can affect expression in *cis* when DNA sequence variants affect the density of methylation in an upstream regulatory region, but it can also act in *trans* when there is variation that changes the efficiency of the machinery responsible for methylation. Similarly, splicing effectors can act in *cis*- or *trans*-.

The view of transcription just discussed, although necessarily simplified for this presentation, highlights that it is indeed a complex process, and with the discovery of large numbers of non-coding RNAs (ncRNAs), many of which have already been shown to regulate transcription, the complexity of how cells regulate transcription will no doubt be shown to be more complex than we can appreciate at this time. Variations in transcript abundances can be caused in a number of ways, many of which are ultimately due to variation in genetic sequences, where altering the sequence in the regulatory region or altering the proteins or quantity of the proteins that bind to these regions results in variations in transcript abundances. Additionally, the rate of transcription and degradation of gene transcripts as well as functions such as alternative splicing can also be affected by ncRNAs, including microRNAs, short interfering RNAs, and long non-coding RNAs.

Importantly, recent advances in sequencing technologies, combined with specific approaches for cleaving and selecting DNA or converting (e.g. when bound by protein) allow us to

quantitatively assay these varied molecular layers. Among many different assays, we can explore the binding (and thus activity) of specific transcription factors and histone modifications across the genome (ChIP-seq), uncover segments of DNA unbound by histone and therefore open for transcription (ATAC-seq), methylation (bisulfite sequencing) and long-range interactions between DNA sequences (Hi-C). Moreover, adaptations of existing RNA-seq protocols also allow the expression of ncRNAs to be quantified. Each of these data types can contribute to the understanding of how DNA variants affect RNA abundance and phenotypes, and the models presented here can be adapted to include these data. Further discussion of these data can be found in **Chapter 33**. For the purposes of this chapter, however, we focus on the simplest case, linking DNA modifications, gene expression levels, and phenotypes.

25.1.2 Modeling Approaches for Biological Processes

A true understanding of complex systems and the complex behaviors they exhibit can only be achieved if we understand the causal relationships among the hierarchy of constituent components comprising the system. However, inferring causality between variables, especially recovering causal networks from observational data, is a particularly challenging task. Given the complexity of biological data and the complexity of methods that can be applied to deriving meaning from such data, an awareness of the different classes of models is warranted, even though in this chapter we focus primarily on probabilistic causal reasoning. The different types of modeling that can be applied to biological data can be broken down into a number of different classes, with the selection of modeling approach depending upon a number of factors such as the extent of prior knowledge, the dimensionality of the data to be modeled, the scale of data available to model, and of course what you hope to derive from the data and the model (Figure 25.3).

Type of Model		Characteristics of Model						
Bottom-up Modeling	Kinetic	Limited to small #	Very large # of data points needed to fit model	Extensive prior knowledge required	Can reveal strong mechanistic insights	Prior Knowledge Dependence	Novel Mechanistic Insights	
	Fuzzy Logic	Limited to small to moderate #						
Top-down Modeling	Boolean Network	Can have moderate #	Larger # of data points needed to fit model	Strong prior knowledge required	Can reveal strong mechanistic insights	Prior Knowledge Dependence	Novel Mechanistic Insights	
	Bayesian Network	Can have moderate to large #	Larger # of data points needed to fit model	Less prior knowledge required	Potential to provide mechanistic insights			
Correlation-based Modeling	PLS Regression	Can have large #	Moderate to large # of data points to fit model	Prior knowledge not required, but can be leveraged	Can learn novel causal relationships	Prior Knowledge Dependence	Novel Mechanistic Insights	
	PCA Multi-Regression & WGCNA	Can have very large #	Small to moderate # of data points to fit model	Prior knowledge not required, some ability to model prior data	Does not implicitly infer causality but informs on relationships			
				Small # of data points to fit model	Prior knowledge not required, limited ability to incorporate prior knowledge			

Figure 25.3 Modeling biological data using different classes of mathematical modeling approaches. The primary aim of these different approaches is uncovering relationships in the data that may help predict phenotypes of interest, elucidate causal relationships among traits and biological processes, and derive mechanistic insights into the causes of disease, wellness, drug response, and other phenotypes of interest. A more detailed description of the different modeling approaches is given in the main text. WGCNA is weighted correlation network analysis.

Across the spectrum of modeling classes ranging from those assuming the most complete knowledge of pathways and networks to those assuming no knowledge (preferring instead to learn the network structures directly from the data), kinetic models are at the most extreme end of the distribution with respect to requiring extensive prior knowledge. Kinetic models are typically represented as systems of ordinary differential equations (ODEs), which require extensive prior knowledge as the ODEs fix the connectivity structure among the variables being modeled (e.g. the pathway is assumed to be known). The model is thus defined by a series of parameters that are fitted from the data, and with these parameter estimates the behavior of the system can be directly explored via simulations run on the model. Via these simulations, kinetic models provide mechanistic insights. These models can also be fitted from smaller, more focused data sets, although typically this modeling approach is restricted to smaller network structures and the models can be difficult to calibrate (Azeloglu and Iyengar, 2015). Modeling of the dynamics of physiologic glucose-insulin levels, metabolic flux, and drug response are just a few of many examples that have been effectively modeled using this approach.

Logic models represent another class of models that require significant prior knowledge. However, they also have an adaptive component that can be learned from the data, which reduces dependence on the prior knowledge required to model the biological system of interest. Logic models also maintain a simple and intuitive framework for understanding complex signaling networks (Morris *et al.*, 2010). In addition, this type of modeling approach still allows direct mechanistic insights to be derived through appropriate simulations. In sum, kinetic and logic models are representatives of bottom-up modeling approaches that begin with strong prior knowledge regarding how pathways are put together, but then estimate the kinetic parameters to describe the flow of information through the system.

Boolean network modeling represents another class of approaches, where a flexible framework for modeling biomolecules as binary variables directly links relevant state information to downstream biological processes. However, in contrast to kinetic models, the regulation of the different states represented is described in a parameter-free way, thus enabling a more exploratory characterization of the dynamics of a complex system (Albert and Thakar, 2014). While these types of models can represent many more variables than kinetic models, they provide less mechanistic insight.

Bayesian network models, the approach discussed in more detail later in this chapter, are an even more flexible framework for modeling complex biological processes, requiring no prior knowledge. However, they do this while still providing a natural and mathematically elegant way to incorporate prior knowledge if available. Bayesian networks provide a way to learn regulatory relationships directly from the data. With the use of heuristic searching, networks comprised of many thousands of variables can be constructed, although equally large sets of data are required to effectively construct this type of model. The causal relationships represented in these models are statistically inferred and so deriving mechanistic insights is more difficult. The Boolean and Bayesian network modeling approaches are examples of what we refer to as top-down modeling approaches that seek to learn relationships directly from the data (otherwise known as structure-based learning).

The final classes of modeling approaches are correlation-based and more exploratory in nature. They seek to elucidate correlation structures, potentially using many data sets, in order to begin to understand relationships that may be well reflected in them, and that may aid in understanding key processes involved in complex processes associated with phenotypes of interest such as disease. Partial least squares (PLS) regression and principal components analysis (PCA) multi-regression are examples of two classes of such modeling approaches. No prior knowledge is required to fit the models: they can operate on extremely large data sets; scale to any number of variables; give rise to very large-scale networks; and they are easy to calibrate. However, such models do not explicitly infer causality but rather reflect connections and

influences on those connections, a first step for learning important relationships that are involved in complex processes such as disease.

In this broad spectrum of methods, Bayesian networks strike a nice balance between resolving mechanisms and structure and more broadly reflecting connections and their influences, thereby providing an efficient path for understanding information flow. Whereas ODEs are hypothesis-driven, where the relationships among variables is assumed known, Bayesian methods operate in a hypothesis-free context in which we attempt to infer the relationships among variables given the data. As a result, Bayesian networks have emerged as a state-of-the-art approach for understanding complex systems in which the relationships among the constituent components of the system are not generally known. Equally important, they can seamlessly incorporate existing knowledge as structural and parameter priors and then infer directed relationships among the nodes in the network using conditional dependency arguments (Zhu *et al.*, 2008, 2012; Chang *et al.*, 2015). However, there are also limitations with this modeling approach that relate to the ability of Bayesian networks to distinguish causal structures that have equivalent joint probability and conditional independence structures (Markov equivalence). The severity of this problem cannot be understated, given that statistically indistinguishable structures may reflect completely contradictory causal relationships. We will explore below how appropriate prior information can be incorporated to help resolve these and related issues.

In this chapter, we focus on approaches that leverage DNA sequence or genotype information to help infer causal relationships with respect to gene expression and clinical phenotypes, starting with eQTL identification and continuing with simple causal inference and larger-scale network methods. Alternate approaches to causal inference are treated in **Chapter 23**.

25.1.3 Human versus Experimental Models

The types of analyses described in this chapter can be performed using either designed crosses of experimental organisms or natural segregating populations, such as humans, using methods appropriate for the study design (**Chapters 17, 20 and 21**). Mapping of eQTL has been successful both in animal crosses and in humans using both linkage and association (**Chapters 17, 20 and 21**). The choice of animal model versus human should depend on many factors that include the goal of the study and the availability of tissue samples in the chosen population.

In humans, there are two primary difficulties that make these studies less common. The first is the availability of tissue samples. Preliminary studies in humans were done using lymphoblastoid cell lines (Schadt *et al.*, 2003; Morley *et al.*, 2004; Montgomery *et al.*, 2010; Pickrell *et al.*, 2010), and, with the exception of cell types that can be extracted from blood, the procurement of most tissue types of interest is relatively invasive and often prohibitive. With the exception of blood, most eQTL studies are performed in post-mortem tissue samples, though data and results from many studies, comprising many different tissues, are available for access by researchers (Westra *et al.*, 2013; GTEx Consortium, 2015, 2017; Fromer *et al.*, 2016; Miller *et al.*, 2016; Joehanes *et al.*, 2017). Another barrier to performing these types of studies in humans is that of power, which is always an issue when choosing between human and designed animal experiments, due to the lack of full informativeness and typically low variant effect size for most complex traits in human data. In the context of eQTL mapping, the issue of power is exacerbated by the need to correct for multiple testing when examining thousands of expression traits rather than one or few physical phenotypes. Studies must be appropriately powered to distinguish signal from noise, given both the dimensionality of the expression data and the size of the genome. To be well powered, human eQTL association studies require sample sizes in the hundreds. While this is an order of magnitude or less than that required to be well

powered for genome-wide association studies (GWASs; see **Chapter 21**), it also means that typical eQTL cohorts are not sufficiently powered to also assess phenotype associations.

Crosses of inbred lines in animal models are more informative in terms of identifying the grandparental source of alleles, so fewer individuals are required to detect QTL. Additionally, tissue samples of all types are more readily available in animal experiments than they are from human. Still, questions exist about the validity of extrapolating results from animal to human. Additional experiments are always required to validate findings in human. Crosses of inbred lines are also limited by the genetic differences between the two parental lines. Because all inbred strains are related in some manner, between any two strains there will always be some portion of the genome that is non-varying (identical by descent) between the two, and thus, identifiable QTL are limited to those at loci with segregating polymorphisms. These are likely to be fewer in number than in a natural population. As a result, no one experiment can fully identify all QTL and findings from one cross may not replicate in another, though analysis of recombinant inbred lines such as those generated through the Collaborative Cross (Threadgill and Churchill, 2012) effort is one way to ameliorate this concern.

25.2 Modeling for eQTL Detection and Causal Inference

Gene expression traits are quantitative measures that can be analyzed like any other quantitative trait, via linkage or association methods, though special care should be taken in the quality control and normalization of the expression data to fit the distributional assumptions of the method being used. In the case of RNA-seq data, this typically means transformation via log or inverse normal to induce normality. Data should be adjusted for measured confounders such as known processing batches, and RNA integrity number (RIN), and methods to adjust for hidden confounders such as SVA (Leek and Storey, 2007) or PEER (Stegle *et al.*, 2012) are often employed prior to eQTL analysis. Note that these approaches remove association between gene expression traits and therefore should be used with extreme caution or excluded prior to network analysis. Application of these methods prior to the discovery of *trans* eQTL may also be problematic as they can obscure loci affecting a large number of genes (e.g. as might be observed when a mutation affects the expression or binding of a transcription factor) (GTEx Consortium, 2017). A full discussion of gene expression normalization is outside the scope of this chapter.

25.2.1 Heritability of Expression Traits

As a quantitative measure, gene expression can be treated as any other quantitative trait in a genetic analysis. A variety of linkage and association methods for the analysis of quantitative traits are described in Höschle (2007) and Gianola (2007), which are just as useful for identifying genetic controllers of gene expression as they are for, say, regulation of plasma low-density lipoprotein cholesterol levels, though special care is required to appropriately model or transform RNA-seq data, which in canonical form is a counting process. This premise, of course, is only valid if expression is, indeed, a trait under genetic control. Thus, characterizing the proportion of the trait variance that is attributed to genetics (i.e. the heritability, see **Sections 17.2 and 24.4.1**) over a large set of gene expression traits provides important information about the landscape of genetic control, and ultimately determines how useful eQTL mapping strategies will be. For instance, if the genetic component of expression control is vastly outweighed by environmental factors and measurement error, then strategies to map genetic contributions and ultimately correlate these contributions with complex disease will fail.

Equally important is identifying the types of expression traits that are heritable. Many studies have identified gene transcripts that (1) are associated with complex disease phenotypes (Karp *et al.*, 2000; Schadt *et al.*, 2003), (2) are alternatively spliced (Johnson *et al.*, 2003), (3) elucidate novel gene structures (Shoemaker *et al.*, 2001; Mural *et al.*, 2002; Schadt *et al.*, 2004), (4) can serve as biomarkers of disease or drug response (DePrimo *et al.*, 2003), (5) lead to the identification of disease subtypes (van 't Veer *et al.*, 2002; Mootha *et al.*, 2003; Schadt *et al.*, 2003), and (6) elucidate mechanisms of drug toxicity (Waring *et al.*, 2001). However, identifying the heritable traits and the extent of their genetic variability provides insight about the evolutionary forces contributing to the changes in expression associated with biological processes that underlie complex traits such as disease, beyond what can be gained by looking at the transcript abundance data alone.

While genetic heritability of gene expression traits, at this point, is well established (Price *et al.*, 2011; Wright *et al.*, 2014; Yang *et al.*, 2014; Wheeler *et al.*, 2016; Lloyd-Jones *et al.*, 2017), the extent of heritability of a given gene may vary by tissue due to tissue or condition (e.g. developmental timepoint) specific regulation. In the case of inbred crosses, the heritability observed is relative to the genetic background of the parents, so cannot be extrapolated from one experiment to another. An additional consideration for this specific application is that partitioning heritability into components for both proximal single nucleotide polymorphisms (SNPs, near the gene) and distal SNPs (distant or unlinked from the gene) can allow us to estimate the proportion of the variance explained by *cis* elements versus that explained by *trans* elements, which is useful for understanding the sources of regulatory variation for a given gene (see **Section 24.4.1** for more detail).

25.2.2 Single-Trait eQTL Mapping

While gene expression traits are typically analyzed like any other quantitative trait (**Chapters 17, 20 and 21**), the difficulty in analysis and interpretation comes with the large number of traits examined. Whatever method is chosen must be computationally tractable enough to be performed hundreds of thousands of times, though a variety of computationally efficient tools are available specifically for this purpose (Shabalin, 2012; Ongen *et al.*, 2016). In addition, significance thresholds must be adjusted for multiple testing. Multiple testing issues relate not only to the number of transcripts tested, but also to the number of markers or proportion of the genome tested.

In human association studies, standard algorithms for the estimation of the false discovery rate (FDR) such as the Benjamini–Hochberg (Benjamini and Hochberg, 1995) and Storey's method (Storey, 2002) are typically sufficient as long as (1) a minor allele frequency threshold is employed which is appropriate for the sample size, (2) outliers are removed from gene expression traits prior to eQTL analysis, and (3) appropriate steps to account for population stratification have been employed. Permutation can be used to ensure *p*-values do not show inflation of Type I error, and special care may be necessary to control the FDR for low-frequency variants when sample sizes are small (Huang *et al.*, 2017). Power to identify *cis* eQTL can be maximized by assessing the FDR separately in proximal (often defined as ± 1 Mb around the gene) and distal regions. Approaches to estimate the number of genes under genetic control ('eGenes') require empirical *p*-value estimation via permutation, since the best association per gene will not follow a U[0,1] distribution (GTEx Consortium, 2015, 2017). Further approaches to eQTL mapping are discussed in **Chapter 30**.

25.2.3 Joint eQTL Mapping

Single-trait analyses lose information by ignoring correlation among associated expression traits. Appropriately chosen joint mapping methods can leverage trait correlation and more

elegantly handle multiple testing. Typical approaches to the joint analysis of genetic traits involve mapping each gene expression trait individually and inferring the genetic correlation between pairs or sets of expression traits based on pairwise Pearson correlation, eQTL overlaps, and/or tests for pleiotropy. Using a family-based sample, Monks *et al.* (2004) estimated the genetic correlation between pairs of traits using a bivariate variance-component-based segregation analysis, and showed that the genetic correlation was better able to distinguish clusters of genes in pathways than correlations based on the observed expression traits. This method is similar to single-trait variance-component-based segregation analysis. For two trait vectors \mathbf{v} and \mathbf{y} , let \mathbf{t} denote the bivariate trait vector

$$\mathbf{t} = \begin{bmatrix} \mathbf{v} \\ \mathbf{y} \end{bmatrix},$$

with piecewise mean vector

$$\mu_t = \begin{bmatrix} \mu_v \\ \mu_y \end{bmatrix}.$$

The partitioned covariance matrix can be written as

$$\mathbf{V}_t = \begin{bmatrix} \mathbf{V}_v & \mathbf{V}_{vy} \\ \mathbf{V}_{vy} & \mathbf{V}_y \end{bmatrix},$$

where \mathbf{V}_v and \mathbf{V}_y are the univariate covariance matrices for traits v and y , respectively. The trait covariance matrix \mathbf{V}_{vy} is modeled as

$$\mathbf{V}_{vy} = \mathbf{A}\sigma_{uvy}^2 + \mathbf{I}\sigma_{evy}^2,$$

where \mathbf{A} is twice the kinship matrix, \mathbf{I} is the identity matrix, σ_{uvy}^2 is the genetic covariance between the two traits, and σ_{evy}^2 is the non-genetic covariance. The genetic and non-genetic covariances can be expressed in terms of the genetic and non-genetic correlation as

$$\sigma_{uvy}^2 = \sigma_{uv}\sigma_{uy}\rho_{uvy}$$

and

$$\sigma_{evy}^2 = \sigma_{ev}\sigma_{ey}\rho_{evy},$$

where σ_{uv} and σ_{uy} are the square root of the genetic variances for traits v and y , σ_{ev} and σ_{ey} are the square root of the environmental variances, and ρ_{uvy} and ρ_{evy} are the genetic and non-genetic correlations, respectively.

These methods can be extended to perform bivariate and multivariate linkage analysis, which can be more highly powered to detect linkage when traits are correlated. Clusters of correlated gene expression traits can often contain hundreds or thousands of genes, which would be computationally prohibitive in a joint analysis. Kendziorski *et al.* (2006) approached this problem in a different way by employing a Bayesian mixture model to exploit the increased information from the joint mapping of correlated gene expression traits, which is computationally tractable for large sets of genes. Instead of performing a linkage scan by computing LOD scores at positions along the genome, Kendziorski *et al.* (2006) compute the posterior probability that a particular gene expression trait maps to marker m for each marker, as well as the posterior probability that the trait maps nowhere in the genome. More specifically, for a particular gene expression trait, k , from a correlated gene cluster of interest, the marginal distribution of the data, \mathbf{y}_k , is

$$p_0 f_0(\mathbf{y}_k) + \sum_m p_m f_m(\mathbf{y}_k),$$

where p_m is the probability that the transcript maps to marker m and f_m is the distribution of the data if it does, p_0 is the probability that the transcript does not map to the genome and f_0 is the data distribution in this case. Given appropriate choices for the distributions, model parameters can be estimated via the EM algorithm, and posterior estimates of $\{p_m\}$ can be obtained. Non-linkage is declared for a transcript if the posterior probability of non-linkage exceeds a threshold that bounds the posterior expected FDR. One benefit to this approach is that it controls false discovery for the number of expression traits being tested, whereas assessing the appropriate significance cutoffs in single-transcript linkage analysis often requires data permutation analyses. The drawback is that this method assumes that linkage occurs at either one or none of the markers tested, and lacks a well-defined method for the case when multiple eQTL control an expression trait. Bayesian approaches to joint eQTL mapping are one way to control error in this case (Chapter 30). More recently, computational advances have allowed genome-wide multi-trait eQTL analysis through the application of machine learning algorithms (Hore *et al.*, 2016).

25.2.4 eQTL and Clinical Trait Linkage Mapping to Infer Causal Associations

While understanding the mechanisms of RNA expression is in itself important for understanding biological processes, the ultimate use of this information is identifying the relationship between variation in expression levels and disease phenotypes in an organism of interest. Microarray or RNA-seq experiments are commonly used to explore differential expression between disease and normal tissue samples or between samples from different disease sub-types (see Chapter 30). These studies are designed to detect association between gene expression and disease associated traits, which in turn can lead to the identification of biomarkers of disease or disease sub-types. However, in the absence of supporting experimental data, these data alone are not able to distinguish genes that drive disease from those that respond. As discussed above, eQTL mapping can aid traditional clinical trait QTL (cQTL) mapping by narrowing the set of candidate genes underlying a given cQTL peak and by identifying expression traits that are causally associated with the clinical traits.

Expression traits detected as significantly correlated with a clinical phenotype may reflect a causal relationship between the traits, either because the expression trait contributes to, or is causal for, the clinical phenotype, or because the expression trait is reactive to, or a marker of, the clinical phenotype. However, correlation may also exist in cases when the two traits are not causally associated. Two traits may appear correlated due to confounding factors such as tight linkage of causal mutations (Schadt *et al.*, 2005a) or may arise independently from a common genetic source. The A^y mouse provides an example of correlations between eumelanin RNA levels and obesity phenotypes induced by an allele that acts independently on these different traits, causing both decreased levels of eumelanin RNA and an obesity phenotype (Figure 25.1(b)). More generally, a clinical and expression trait for a particular gene may depend on the activity of a second gene, in such a manner that, conditional on the second gene, the clinical and expression traits are independent.

Correlation data alone cannot indicate which of the possible relationships between gene expression traits and a clinical trait are true. For example, given two expression traits and a clinical trait detected as correlated in a population of interest, there are 112 ways to order the traits with respect to one another. That is, for each pair of nodes there are five possible ways the nodes can be connected: (1) connected by an undirected edge; (2) connected by a directed edge moving from left to right; (3) connected by a directed edge moving from right to left; (4) connected by a directed edge moving from right to left and a directed edge moving from left to right; (5) not connected by an edge. Since there are three pairs of nodes, there are

$5 \times 5 \times 5 = 125$ possible graphs. However, since we start with the assumption that all traits are correlated, we exclude 12 of the 125 graphs in which one node is not connected to either of the other two nodes, and we exclude the graph in which none of the nodes are connected, leaving us with 112 possible graphs, some of which are illustrated in Figure 25.1(a). The joint trait distributions induced by these different graphs are often statistically indistinguishable from one another (i.e. they are Markov equivalent, so that their distributions are identical), making it nearly impossible in most cases to infer the true relationship. On the other hand, when the two traits are at least partially controlled by the same genetic locus and when more complicated methods of control (e.g. feedback loops) are ignored, the number of relationships between the QTL and the two traits of interest can be reduced to the three models illustrated graphically in Figure 25.1(b). The dramatic reduction in the number of possible graphs to consider (from 112 to 3) is mainly driven by the fact that changes in DNA drive changes in phenotypes and not vice versa (it is extremely unlikely that changes in RNA or protein lead to changes in DNA at a high enough frequency to detect associations between germ-line transmitted polymorphisms and phenotype).

It is important to note here that when we use the term ‘causality’, it is perhaps meant in a more non-standard sense than most researchers in the life sciences may be accustomed to. In the molecular biology or biochemistry setting, claiming a causal relationship between, say, two proteins usually means that one protein has been determined experimentally to physically interact with or to induce processes that directly affect another protein, and that in turn leads to a phenotypic change of interest. In such instances, an understanding of the causal factors relevant to this activity are known, and careful experimental manipulation of these factors subsequently allows for the identification of genuine causal relationships. However, in the present setting, the term ‘causal’ is used from the standpoint of statistical inference, where statistical associations between changes in DNA, changes in expression (or other molecular phenotypes), and changes in complex phenotypes such as disease are examined for patterns of statistical dependency among these variables that allows directionality to be inferred among them, where the directionality then provides the source of causal information (highlighting putative regulatory control as opposed to physical interaction). The graphical models (networks) described here, therefore, are necessarily probabilistic structures that use the available data to infer the correct structure of relationships among genes, and between genes and clinical phenotypes. In a single experiment, with one timepoint measurement, these methods cannot easily model more complex regulatory structures that are known to exist, such as negative feedback control. However, the methods can be useful in providing a broad picture of correlation and causative relationships, and while the more complex structures may not be explicitly represented in this setting, they are captured nevertheless, given that they represent observed states that are reached as a result of more complicated processes such as feedback control. A mathematical theory of causal inferences from observed dependency patterns from raw data has been established, and Judea Pearl, a pioneer of mathematical and computational methods for this purpose, provides an excellent description and treatment of this underlying theory (Pearl, 1988, 2000).

While in principle the techniques we describe can be applied to any pair of traits, continuous or binary, clinical or molecular (e.g., RNA, protein, or metabolite), we focus on the problem of inferring causality among gene expression traits and between gene expression and continuous clinical phenotypes. We let T_1 and T_2 denote RNA expression traits or an expression trait and a clinical trait, and L the locus genotype. In the first model illustrated in Figure 25.1(b), the genetic locus is causal for T_2 only through T_1 (causal model). In the second model, T_1 is a reaction to T_2 (reactive model). In the final model, the genetic locus affects both T_1 and T_2 independently (independence model). With the inclusion of genetic marker information, the data models are distinguishable due to the conditional independence structure of the variables.

25.2.4.1 A simple model for inferring causal relationships

Assuming the conditional independence structure implied in the graphical models in Figure 25.1(b), the following simplifications can be made for the joint probability distributions for the causal, reactive and independence models, respectively:

$$\begin{aligned} P_1(L, T_1, T_2) &= P_{\theta_L}(L)P_{\theta_{T_1 L}}(T_1|L)P_{\theta_{T_2 T_1}}(T_2|T_1), \\ P_2(L, T_1, T_2) &= P_{\theta_L}(L)P_{\theta_{T_2 L}}(T_2|L)P_{\theta_{T_1 T_2}}(T_1|T_2), \\ P_3(L, T_1, T_2) &= P_{\theta_L}(L)P_{\theta_{T_1 L}}(T_1|L)P_{\theta_{T_2 L}^*}(T_2|L, T_1) \\ &= P_{\theta_L}(L)P_{\theta_{T_2 L}}(T_2|L)P_{\theta_{T_1 L}^*}(T_1|L, T_2). \end{aligned} \quad (25.1)$$

In the causal model depicted in Figure 25.1(b), the clinical trait (T_2) is independent of the genetic locus conditional on the gene expression trait (T_1). In other words, the locus genotype lends no additional information about the clinical phenotype when the gene expression measurement is known (P_1 above). Similarly, in the reactive model (P_2), the expression trait is independent of the underlying genetics conditional on the clinical trait. In the independence model (P_3), the gene expression and clinical trait are not assumed to be conditionally independent, allowing for correlation due to other shared genetic and environmental influences. Failure to account for this correlation, when it is of moderate size, can result in falsely choosing one of the two other models because the two traits contain information, in addition to that provided by the genetic locus, about the other, due to unmeasured common influences.

The modeling framework in equation (25.1) is quite general and can accommodate a wide range of genetic and trait dependence models. For continuous traits, a variety of trait models have been developed for the purposes of testing for linkage of a QTL to a single locus. These same models can be used to model an expression or clinical trait conditional on the genetic locus.

Given appropriate choices for the conditional distributions in the three models, $P(T_1|L)$, $P(T_2|L)$, $P(T_1|T_2)$, $P(T_2|T_1)$ and $P(T_1|L, T_2)$, likelihoods for the three different models can be maximized with respect to the model parameters and the likelihoods can be subsequently compared. Note that the log likelihood of each model is the sum of log likelihoods for each of the three variables with no common parameters. For example,

$$\begin{aligned} \log P_1(L, T_1, T_2) &= \log P_{\theta_L}(L) + \log P_{\theta_{T_1 L}}(T_1|L) + \log P_{\theta_{T_2 T_1}}(T_2|T_1) \\ &= \log L(\theta_L|L) + \log L(\theta_{T_1 L}|T_1, L) + \log L(\theta_{T_2 T_1}|T_2, T_1) \\ &= \log L_i(\theta_i|L, T_1, T_2). \end{aligned}$$

Thus maximum likelihood estimates (MLEs) for θ_L , $\theta_{T_1 L}$, and $\theta_{T_2 T_1}$ can be obtained by separately maximizing each corresponding term. The likelihoods are then compared among the different models in order to find the most likely of the three. When the number of model parameters among the models differs, a penalized function of the likelihood is used to avoid the bias against parsimony: the model with the smallest value of the penalized statistic

$$-2 \log L_i(\hat{\theta}_i|L, T_1, T_2) + k \times p_i$$

is chosen. Here, $L_i(\hat{\theta}_i|L, T_1, T_2)$ is the MLE for the i th model, p_i is the number of parameters in the i th model, and k is a constant. Common choices for k include $k = 2$, in which case the statistic is known as the Akaike information criterion (AIC), and $k = \log(n)$, where n is the number of observations, which is called the Bayesian information criterion (BIC).

An alternate approach, which is often taken in causal inference, is to use significance tests to identify a model consistent with the data such as the causal inference test (CIT; Millstein *et al.*, 2009). For example, in the causal model depicted in Figure 25.1(b), the clinical trait is

conditionally independent of the genetic locus, L . Thus, a rejected test of conditional independence between T_2 and L implies that the data are not consistent with the causal model, but the conditions required to infer the causal model are: (1) L and T_2 are associated; (2) L is associated with $T_1|T_2$; (3) T_1 is associated with $T_2|L$; and (4) L is independent of $T_2|T_1$.

In order to infer among the three models, the following regressions are performed to assess the previously mentioned conditions:

$$\begin{aligned} T_2 &= \alpha_1 + \beta_1 L + \varepsilon_1, \\ T_1 &= \alpha_2 + \beta_2 T_2 + \beta_3 L + \varepsilon_2, \\ T_2 &= \alpha_3 + \beta_4 T_1 + \beta_5 L + \varepsilon_3, \end{aligned}$$

where ε_i is a random variable representing independent noise. Then the previous conditions are assessed via the following hypothesis tests:

1. $H_0: \beta_1 = 0$,
2. $H_0: \beta_3 = 0$,
3. $H_0: \beta_4 = 0$,
4. $H_0: \beta_5 = 0$.

Note that the reactive model can be tested using the same three regressions by swapping β_2 for β_4 in hypothesis (3) and β_3 for β_5 in hypothesis (4). While testing hypotheses (1)–(3) is a straightforward t -test (or F -test allowing for codominance), hypothesis (4) requires a formulated equivalence test because failure to reject $\beta_5 = 0$ is not equivalent to rejecting hypothesis (4). To appropriately test this hypothesis, the CIT employs a simulation strategy to compute the test statistic under conditional independence between T_2 and $T_1|L$ by permuting the residuals of $T_1 = \alpha + \beta L + \varepsilon$. Finally, an omnibus test for causality corresponds to the supremum of the p -values of each of the four component tests.

25.2.4.2 Distinguishing proximal eQTL effects from distal

As discussed in previous sections, all genes expressed in living systems are *cis*-regulated at some level and so are under the control of various *cis*-acting elements such as promoters and TATA boxes. In this context, expression as a quantitative trait for eQTL mapping presents a unique situation in quantitative trait genetics because the expression trait corresponds to a physical location in the genome (the structural gene that is transcribed, giving rise to the expression trait). The transcription process operates on the structural gene, and so DNA variations in the structural gene that affect transcription will be identified as eQTL in the mapping process. In such cases we would identify eQTL as *cis*-acting, given that the most reasonable explanation for seeing an eQTL coincident with the physical location of the gene will be that variations within the gene region itself give rise to variations in its expression (Doss *et al.*, 2005). However, because we cannot guarantee that the eQTL is truly *cis*-acting (i.e. it could arise from variation in a gene that is closely linked to the gene expression trait in question), it is more accurate to refer to such eQTL as “proximal”, given they are close to the gene corresponding to the expression trait. Because the *cis*-regulated components of expression traits are among the most proximal traits in a biological system with respect to the DNA, we might expect that true *cis*-acting genetic variance components of expression traits are among the easiest components to detect via QTL analysis if they exist. This indeed has been observed in a number of studies, where proximal (presumably *cis*-acting) eQTL have been identified that explain unprecedented proportions of a trait’s overall variance (several published studies highlight examples where more than 90% of the overall variation was explained by a single *cis*-acting eQTL) (Brem *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Cervino *et al.*, 2005; Cheung *et al.*, 2005; Lum *et al.*, 2006). In recent human association studies, eQTL have been shown to be most enriched

within 100 kb of the transcription start site, and the strongest associations tend to be enriched in tissue-specific regulatory elements as estimated from ChIP-seq data (Fromer *et al.*, 2016). *Cis*-acting associations in or very near the gene can be detected via allele-specific expression methods when expression is assessed via RNA-seq (e.g. Pickrell *et al.*, 2010). Recent approaches have aimed to disentangle eQTL heterogeneity due to cellular composition in primary tissue or other sources (Westra *et al.*, 2015; Jansen *et al.*, 2017; Dobbyn *et al.*, 2018). These methods are not addressed in this chapter.

Variations in expression levels induced by DNA variations in or near the gene itself may in turn induce changes in the expression levels of other genes. Each of these genes in a population of interest may not harbor any DNA variation in their structural gene, so that they do not give rise to true *cis*-acting eQTL, but they would give rise to a so-called *trans* eQTL (Figure 25.2). In reality, variation in expression traits may be due to variation in *cis*-acting elements and/or one or multiple *trans*-acting elements. Additionally, master regulators of transcription, which affect the expression of many traits in *trans*, may exist, though the evidence on this is mixed (Figure 25.2).

Because in many cases it is not possible to infer the true regulatory effects (i.e. *cis* versus *trans*) of an eQTL without complex bioinformatics study and experimental validation, we instead categorize eQTL into proximal and distal types based on the distance between the eQTL and the location of the structural genes. If these are on different chromosomes the eQTL is obviously distal, but if they fall on the same chromosome then we require the distance between the structural gene and the eQTL to not exceed some threshold. The exact threshold will be a function of the number of meioses and extent of recombination in a given population data set. In a completely outbred population where association mapping has been used to discover the eQTL, it is reasonable to require the distance between the proximal eQTL and structural gene to be less than 1 Mb (Cheung *et al.*, 2005), and strategies that assess FDR separately in proximal and distal regions can maximize discovery of eQTL due to the enrichment of eQTL near the gene. However, in an F2 intercross population constructed from two inbred lines of mice, the extent of linkage disequilibrium (LD) will be extreme given that all animals are descended from a single F1 founder, with only two meiotic events separating any two mice in the population. In such cases, the resolution of linkage peaks is quite low, requiring the threshold of peak-to-physical gene distance to be more relaxed, so that eQTL that are within 20 or 30 Mb could be considered proximal (Schadt *et al.*, 2003; Doss *et al.*, 2005). While the proximal eQTL provide an easy path to making causal inference, given that the larger effect sizes commonly associated with proximal eQTL make them easier to detect (Brem *et al.*, 2002; Schadt *et al.*, 2003; Monks *et al.*, 2004; Cervino *et al.*, 2005; Cheung *et al.*, 2005; Lum *et al.*, 2006), the methods discussed above work for distal as well as proximal eQTL.

25.2.4.3 Trait-Gene inference from GWAS

Using similar logic to the CIT, Mendelian randomization (MR), a version of instrumental variable analysis, is used to perform causal inference from GWAS data (see **Chapter 23**). As with the CIT, this inference requires that both disease outcome and gene expression data be assessed in the same samples. However, to be powered to detect disease associations, cohorts of thousands to tens of thousands are required. Assessing gene expression (at least genome-wide) at this scale tends to be cost-prohibitive. This, combined with challenges in obtaining disease relevant tissues, means that the ability to perform causal inference in GWAS cohorts is rare. Given the relatively ready abundance of GWAS summary statistics, one emerging approach attempts to perform mediation analysis logic using inference from expression data and GWAS summary statistics (Park *et al.*, 2017a); however, given the infancy of the field, it is not treated in detail here.

Given the public availability of summary statistics from large GWAS meta-analyses, combined with the public availability of eQTL results and/or data (Westra *et al.*, 2013; Fromer *et al.*, 2016; Miller *et al.*, 2016; Joehanes *et al.*, 2017) including the large-scale effort to catalogue many tissues (GTEx Consortium, 2015, 2017), attempts to integrate eQTL with GWASs tends to focus on methods to identify genes whose eQTL association signatures colocalize with GWAS associations. While so-called colocalization analyses fall short of causal inference, they can be useful in identifying candidate genes driving the GWAS signal, especially when the associations fall in intergenic regions. As with causal inference, failure to identify eQTL genes colocalizing with GWAS signal may result from using the wrong tissue or developmental timepoint, or that the disease risk is not mediated by gene expression level. Because these methods effectively compare the patterns of association between clinical and expression traits, it is important that the cohorts used to infer associations be comparable with respect to population background.

Colocalization is an emerging field with several competing approaches currently available. Most similar to MR, SMR (Zhu *et al.*, 2016) computes the unconfounded effect of expression on the trait as the ratio of the effect sizes obtained as summary statistics; however, it cannot distinguish between causal and pleiotropic effects due to LD. In order to assess this likelihood, it also implements a test of the null hypothesis that there is a single causal eQTL in the region (HEIDI), with failure to reject this hypothesis potentially providing support for a causal model; however, this approach falls short of causality testing. The RTC (Nica *et al.*, 2010; Ongen *et al.*, 2017) approach computes *cis* eQTL statistics conditional on the top GWAS-associated SNP to determine if doing so removes the eQTL association at that locus, and thus requires availability of data (not just summary statistics) in the eQTL cohort. JLIM (Chun *et al.*, 2017) and eCaviar (Hormozdiari *et al.*, 2014) both capitalize on methods that attempt to identify causal GWAS variants and infer when those inferred causal variants correspond between GWAS and eQTL. Sherlock (He *et al.*, 2013) and coloc (Giambartolomei *et al.*, 2014) both employ a Bayesian framework to assess colocalization. Sherlock is the only method that includes distal eQTL, and therefore does not penalize for instances in which there is an eQTL signal but no GWAS signal, though separate priors for proximal and distal eQTL allow one to downweight distal eQTL. Coloc applies a principled approach to compute the posterior probability of five hypotheses: H_0 : no association with either trait; H_1 : association with trait 1, not with trait 2; H_2 : association with trait 2, not with trait 1; H_3 : association with trait 1 and trait 2, two independent SNPs; and H_4 : association with trait 1 and trait 2, one shared SNP. A recent extension, MOLOC, jointly models three traits, which can be useful in identifying epigenetic changes mediating the gene expression variation (Giambartolomei *et al.*, 2018).

In practice, most of these methods perform best when there is one shared causal SNP, and may suffer when multiple SNPs show similarly strong association. Some methods may also report significant associations when both gene expression and the trait show neighboring associations, but the strongest eQTL are not the strongest GWAS associations (coloc's H_4). A simple scatterplot of $-\log_{10}(p\text{-value})$ for eQTL and GWAS can be used to rule out these spurious findings.

Another approach to identifying candidate GWAS genes involves imputing gene expression from SNP data and testing whether the imputed gene expression is significantly associated with the trait of interest in a GWAS cohort. This approach, dubbed transcript-wide association analysis (TWAS), is essentially a dimensionality reduction method to test whether weighted sums of expression-associated SNPs also show association with disease. Instead of the millions of association tests in a GWAS, the number of tests is reduced to thousands or tens of thousands. The first approach to transcript-wide association analysis allowed expression to be assayed in a cohort separate from the GWAS study, with the predictive model being built in the eQTL cohort, and the predictions made using genotype data in the GWAS cohort (Gamazon

et al., 2015). In other words, access to GWAS data was required. Because of the wide availability of GWAS summary statistics and limited availability of GWAS cohorts, later approaches (Barbeira *et al.*, 2017; Park *et al.*, 2017b) were adapted to allow prediction from summary statistics combined with SNP variance and covariance (LD) information which is readily available. Transcript-wide association analysis may also be prone to spuriously infer associations due to LD when both the expression trait and clinical trait show association in the region, but the strongest associations do not colocalize, so further analysis is necessary to rule out this situation. As with colocalization, this field is rapidly evolving.

25.3 Inferring Gene Regulatory Networks

25.3.1 From Assessing Causal Relationships among Trait Pairs to Predictive Gene Networks

Leveraging DNA variation as a systematic perturbation source to resolve the causal relationships among traits is necessary but not sufficient for understanding the complexity of living systems. Cells are comprised of many tens of thousands of proteins, metabolites, RNA, and DNA, all interacting in complex ways. Modeling the extent of such relationships between molecular entities, between cells, and between organ systems is a daunting task. Networks are a convenient framework for representing the relationships among these different variables. In the context of biological systems, a network can be viewed as a graphical model that represents relationships among DNA, RNA, protein, metabolite, and higher-order phenotypes such as disease state. In this way, networks provide a way to represent extremely large-scale and complex relationships among molecular and higher-order phenotypes such as disease in any given context.

25.3.2 Building from the Bottom Up or Top Down?

Two fundamental approaches to the reconstruction of molecular networks dominate computational biology today. The first is what is referred to as the bottom-up approach in which fundamental relationships between small sets of genes that may comprise a given pathway are established, thus providing the fundamental building blocks of higher-order processes that are then constructed from the bottom up. This approach typically assumes that we have more complete knowledge regarding the fundamental topology (connectivity structure) of pathways, and, given this knowledge, models are constructed that precisely detail how changes to any component of the pathway affect other components as well as the known functions carried out by the pathway (i.e. bottom-up approaches are hypothesis-driven). The second approach is referred to as a top-down approach in which we take into account all data and our existing understanding of systems and construct a model that reflects whole-system behavior and from there tease apart the fundamental components from the top down. This approach typically assumes that our understanding of how the network is actually wired is sufficiently incomplete, that our knowledge is sufficiently incomplete, and that we must objectively infer the relationships by considering large-scale, high-dimensional data that inform on all relationships of interest (i.e. top-down approaches are data-driven).

Given our incomplete understanding of more general networks and pathways in living systems, in this chapter we focus on a top-down approach to reconstructing predictive networks, given that this type of structure learning from data is critical to derive hypotheses that cannot otherwise be efficiently proposed in the context of what is known (from the literature, pathway databases, or other such sources).

In the context of integrating genetic, molecular profiling and higher-order phenotypic data, biological networks are comprised of nodes that represent molecular entities that are observed to vary in a given population under study (e.g. DNA variations, RNA levels, protein states, or metabolite levels). Edges between the nodes represent relationships between the molecular entities, and these edges can be either directed, indicating a cause–effect relationship, or undirected, indicating an association or interaction. For example, a DNA node in the network representing a given locus that varies in a population of interest may be connected to a transcript abundance trait, indicating that changes at the particular DNA locus induce changes in the levels of the transcript. The potentially millions of such relationships represented in a network define the overall connectivity structure of the network, otherwise known as the topology of the network. Any realistic network topology will be necessarily complicated and nonlinear from the standpoint of the more classic biochemical pathway diagrams represented in textbooks and pathway databases like KEGG (Kanehisa *et al.*, 2016). The more classic pathway view represents molecular processes on an individual level, while networks represent global (population-level) metrics that describe variation between individuals in a population of interest, which, in turn, define coherent biological processes in the tissue or cells associated with the network. One way to manage the complexity of network structures that can be obtained is to impose constraints on network structures to make them more computationally tractable. For example, it is common when learning network structures to disallow loops or cycles in the network structure, in which cases we refer to the network as acyclic.

The neurosciences have a rich history of employing network-based approaches to understand the complexity of the human brain and the causes of psychiatric illnesses. Resources like the Allen Brain Atlas (<http://portal.brain-map.org/#>) provide an anatomically comprehensive map of gene expression of the human brain that can facilitate network-based analyses (Ding *et al.*, 2016). Others have employed techniques developed for constructing gene coexpression networks to construct interaction networks on fMRI data (Mumford *et al.*, 2010), and others still have generated protein interaction networks to reflect features of the network architecture in brains of those with illnesses such as Huntington's disease (Shirasaki *et al.*, 2012). Larger-scale efforts have also been undertaken to integrate larger-scale transcriptomic data in the context of diseases such as autism to understand how changes in these networks may give rise to autism or reflect the types of pathways or biological processes involved in such a disease (Voineagu *et al.*, 2011). These efforts are important not only for better understanding psychiatric diseases, but also for elucidating novel drug targets or biomarkers that better assess disease risk or severity. However, most of these current efforts do not lead to predictive models of disease, but rather provide a descriptive framework within which to uncover associations between a myriad of molecular, cellular, imaging, and clinical traits and disease.

25.3.3 Using eQTL Data to Reconstruct Coexpression Networks

Networks provide a convenient framework for representing high-dimensional data in which relationships among the many variables making up such data are the key to understanding the properties that emerge from the complex systems they represent. Networks are simply graphical models comprised of nodes and edges. For gene networks associated with biological systems, the nodes in the network typically represent genes, and edges (links) between any two nodes indicate a relationship between the two corresponding genes. For example, an edge between two genes may indicate that the corresponding expression traits are correlated in a given population of interest (Zhu *et al.*, 2004), that the corresponding proteins interact (Kim *et al.*, 2005), or that changes in the activity of one gene lead to changes in the activity of the other gene (Schadt *et al.*, 2005a). Interaction or association networks have recently gained more

widespread use in the biological community, where networks are formed by considering only pairwise relationships between genes, including protein interaction relationships (Han *et al.*, 2004), coexpression relationships (Gargalovic *et al.*, 2006; Ghazalpour *et al.*, 2006), as well as other straightforward measures that may indicate association between two genes.

Forming association networks from expression data based purely on correlations between genes in a set of experiments of interest can give rise to links in the network driven by correlated noise structures between array-based experiments or other such artifacts. The eQTL data can be simply leveraged in this case by filtering out gene–gene correlations in which the expression traits are not at least partially explained by common genetic effects. For example, we may connect two genes with an edge in a coexpression network if the *p*-value for the Pearson correlation coefficient between the two genes was less than some pre-specified threshold, and if the two genes had at least one eQTL in common. Note that this approach works best in the experimental model case, when *trans* eQTL analyses are well powered, and may be too stringent in the case of population association studies. One intuitive way to establish whether two genes share at least one eQTL is to carry out single-trait eQTL mapping for each expression trait and then consider eQTL for each trait overlapping if the corresponding LODs for the eQTL are above some threshold and if the eQTL are in close proximity to one another.

The *p*-value threshold for considering two genes linked in the coexpression networks is chosen such that the resulting network exhibits the scale-free property (Barabasi and Albert, 1999; Ghazalpour *et al.*, 2006; Lum *et al.*, 2006) and the FDR for the gene–gene pairs represented in the network is constrained. Filtering correlations based on eQTL overlap can be seen to improve the quality of the coexpression networks by examining whether networks reconstructed without the eQTL data are more coherent than networks reconstructed with the eQTL data with respect to Gene Ontology (GO) biological process categories. We say one network is more coherent than another with respect to GO biological process categories according to its percentage of gene–gene pairs sharing a common GO biological process category. The pathways represented in GO are independently determined and so provide an independent source of information in testing how much of the known information is captured in the network.

25.3.3.1 More Formally Assessing eQTL Overlaps in Reconstructing Coexpression Networks

While testing whether gene–gene pairs that are significantly correlated are partially controlled by common genetic loci using the overlap method discussed above is intuitively appealing, it fails to make full use of the data to infer whether overlapping eQTL are really the same locus or closely linked eQTL, and such an approach does not lend itself to statistically robust hypothesis testing. One way to test more formally whether two overlapping eQTL represent a single eQTL or closely linked eQTL is to employ a pleiotropy effects test (PET) based on a pleiotropy model initially described by Jiang and Zeng (Jiang and Zeng, 1995; Zeng *et al.*, 2000). While we discuss this method for considering only two traits simultaneously, the method can be easily extended to consider more traits. The statistical model for the PET is an extension of the single-trait model as defined by

$$\begin{pmatrix} y_{11} & \cdots & y_{n1} \\ y_{12} & \cdots & y_{n2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} (x_1 \cdots x_n) + \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} (z_1 \cdots z_n) + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

where y_i is the vector of trait values for individual i ($i = 1, \dots, n$), a_j and d_j are the additive and dominance effects for trait j ($j = 1, 2$), x_i is as defined above, and e_j is the residual effect for trait j . In this case, we are assuming an F2 intercross population constructed from two inbred lines of mice, but the model is easily generalizable to other experimental cross populations.

From this statistical model, a series of tests of hypotheses can be performed to test whether the two traits are supported as being driven by a single QTL at a given position. The first test involves testing whether a given region is linked to the joint trait vector for the traits under study:

$$\begin{aligned} H_0: \alpha_1 = 0, d_1 = 0, \alpha_2 = 0, d_2 = 0, \\ H_1: \text{at least one of the above terms is not } 0. \end{aligned}$$

To test the above null hypothesis of no linkage against the alternative linkage hypothesis, likelihoods associated with the null and alternative hypothesis are maximized with respect to the model parameters. From the maximum likelihoods, the log likelihood ratio statistic is formed and used to test whether the alternative hypothesis (H_1) is supported by the data. With this model, the log likelihood ratio statistic under the null hypothesis is chi-square distributed with 4 degrees of freedom. If the null hypothesis (H_0) is rejected, the implication is that trait 1 and/or trait 2 have a QTL at the given test locus.

Subsequent to the test just described resulting in a rejection of the null hypothesis, second and third tests of hypotheses can be performed to establish whether the detected QTL affects both traits. For a given QTL test position,

$$\begin{aligned} H_{10}: \alpha_1 = 0, d_1 = 0, \alpha_2 \neq 0, d_2 \neq 0, \\ H_{11}: \alpha_1 \neq 0, d_1 \neq 0, \alpha_2 \neq 0, d_2 \neq 0 \end{aligned}$$

assesses whether the first trait has a QTL at the test position, and

$$\begin{aligned} H_{20}: \alpha_1 \neq 0, d_1 \neq 0, \alpha_2 = 0, d_2 = 0, \\ H_{21}: \alpha_1 \neq 0, d_1 \neq 0, \alpha_2 \neq 0, d_2 \neq 0 \end{aligned}$$

assesses whether the second trait has a QTL at the test position. As above, the log likelihood ratio statistics are formed for each of these tests, where under the null hypotheses these statistics are chi-square distributed with 2 degrees of freedom. If both null hypotheses H_{10} and H_{20} are rejected, the QTL is supported as having pleiotropic effects on the two traits under study.

At the 2.8 LOD score threshold indicating suggestive linkage in an F2 intercross population, the expected number of QTL detected is 1 for the conditional linkage tests just described (Lander and Kruglyak, 1995). Therefore, an intuitive way to compute the probability that both of these tests for independent traits give rise to a QTL at a given location (so rejection of the null hypothesis for both tests at the same chromosome position), is to consider the probability that the two QTL expected by chance to be identified for each test happen to be detected at the same chromosome location. This probability can be approximated by the fraction 1/655, where 655 represents the effective number of tests carried out in searching the entire genome at the 2.8 LOD threshold (Lander and Kruglyak, 1995). However, because such an estimate is based on theoretical arguments relying on assumptions that may not hold exactly, and because in practice the traits under consideration will not be independent (especially not in the coexpression network setting where we are interested in highly interconnected sets of genes), permutation testing can also be used to assess the significance of both tests leading to a rejection of the null hypothesis at the 2.8 LOD threshold. In the adipose BXH cross expression data described above, the significance level associated with the 2.8 LOD score threshold was estimated to be 0.004 using permutation methods, only slightly larger than the theoretical estimate of 0.002 given above. Therefore, using the suggestive 2.8 LOD score threshold established for single traits seems reasonable in this situation as it corresponds to a genome-wide significance level of 0.004 when considering two traits at a time.

25.3.3.2 Identifying Modules of Highly Interconnected Genes in Coexpression Networks

Given the scale-free and hierarchical nature of co-expression networks (Barabasi and Oltvai, 2004; Ghazalpour *et al.*, 2006; Lum *et al.*, 2006), one of the problems is to identify the key network modules, representing those hub nodes (nodes that are significantly correlated with many other nodes) that are highly interconnected with one another, but that are not as highly connected with other hub nodes. Ravasz *et al.* (2002) used a manually selected height cutoff to separate tree branches after hierarchical clustering, in contrast to Lee *et al.* (2004) who formed maximally coherent gene modules with respect to GO functional categories. Another strategy is to employ a measure similar to that used by Lee *et al.* (2004), but without the dependence on the GO functional annotations, given it is of interest to determine independently whether coexpression modules are enriched for GO functional annotations. A gene module in the co-expression network is defined as a maximum set of interconnected genes. A coherence measure for a given gene module can be defined as

$$\text{Coherence} = \frac{GP_{\text{obs}}}{GP_{\text{tot}}},$$

where GP_{obs} is the number of gene pairs that are connected, and GP_{tot} is the total number of possible gene pairs in the module. The efficiency of a gene module can then be defined as

$$\text{Efficiency} = \frac{\text{Coherence} \times G_{\text{mod}}}{G_{\text{net}}},$$

where G_{mod} is the number of genes in the module, and G_{net} is the number of genes in the network. Given these definitions, we define a process to iteratively construct gene modules by the following steps:

1. Order genes in the gene–gene connectivity matrix according to an agglomerative hierarchical clustering algorithm as previously described (Hughes *et al.*, 2000).
2. Calculate the efficiency $e_{i,j}$ for every possible module, including genes from i to j as given in the ordered connectivity matrix, where $j \geq i + 9$ (i.e., minimum module size is 10), using a dynamic programming algorithm.
3. Determine the maximum $e_{i,j}$.
4. Set $e_{i\dots j, 1\dots G_{\text{net}}} = 0$ and $e_{1\dots G_{\text{net}}, i\dots j} = 0$.
5. Go to step 3 until no additional modules can be found.

The modules identified in this way are informative for identifying the functional components of the network that may underlie complex traits like disease (Lum *et al.*, 2006). It has been widely demonstrated that coexpression modules are often enriched for known biological pathways, for genes that are associated with disease traits, and for genes that are linked to common genetic loci (Ghazalpour *et al.*, 2006; Lum *et al.*, 2006; Schadt *et al.*, 2008; Zhu *et al.*, 2008; Greenawalt *et al.*, 2011; Zhang *et al.*, 2013; Franzen *et al.*, 2016; Peters *et al.*, 2017).

25.3.4 An Integrative Genomics Approach to Constructing Predictive Network Models

Systematically integrating different types of data into probabilistic networks using Bayesian networks has been proposed and applied for the purpose of predicting protein–protein interactions (Jansen *et al.*, 2003) and protein function (Lee *et al.*, 2004). However, these Bayesian networks are still based on associations between nodes in the network as opposed to causal relationships. As discussed above for the simple case of two traits, from these types of networks we cannot infer whether a specific perturbation will affect a complex disease trait. To

make such predictions, we need networks capable of representing causal relationships. Probabilistic causal networks are one way to model such relationships from the top down, where causality again in this context reflects a probabilistic belief that one node in the network affects the behavior of another. Bayesian networks (Pearl, 1988) are one type of probabilistic causal network that provides a natural framework for integrating highly dissimilar types of data.

Bayesian networks are directed acyclic graphs in which the edges of the graph are defined by conditional probabilities that characterize the distribution of states of each node given the state of its parents (Pearl, 1988). The network topology defines a partitioned joint probability distribution over all nodes in a network, such that the probability distribution of states of a node depends only on the states of its parent nodes: formally, a joint probability distribution $p(X)$ on a set of nodes X can be decomposed as $p(X) = \prod p(X^i | \text{Pa}(X^i))$, where $\text{Pa}(X^i)$ represents the parent set of X^i . The biological networks of interest we wish to construct are comprised of nodes that represent a quantitative trait such as the transcript abundance of a given gene or levels of a given metabolite. The conditional probabilities reflect not only relationships between genes, but also the stochastic nature of these relationships, as well as noise in the data used to reconstruct the network.

The aim in any network reconstruction such as this is to find the best model, the model that best reflects the relationships between all of the variables under consideration, given a set of data that informs on the variables of interest. In a probabilistic sense, we want to search the space of all possible networks (or models) for the network that gives the highest likelihood of occurring given the data. Bayes' formula allows us to determine the likelihood of a network model M given observed data D as a function of our prior belief that the model is correct and the probability of the observed data given the model, $P(M|D) \propto P(D|M)P(M)$. The number of possible network structures grows superexponentially with the number of nodes, so an exhaustive search of all possible structures to find the one best supported by the data is not feasible, even for a relatively small number of nodes. A number of algorithms exist to find the optimal network without searching exhaustively, such as Monte Carlo Markov Chain (MCMC; Madigan and York, 1995) simulation. With the MCMC algorithm, optimal networks are constructed from a set of starting conditions. This algorithm is run thousands of times to identify different plausible networks, each time beginning with different starting conditions. These most plausible networks can then be combined to obtain a consensus network. For each of the reconstructions using the MCMC algorithm, the starting point is a null network. Small random changes are made to the network by flipping, adding, or deleting individual edges, ultimately accepting those changes that lead to an overall improvement in the fit of the network to the data. To assess whether a change improves the network model or not, information measures such as the BIC (Schwarz, 1978) are employed, reducing overfitting by imposing a cost on the addition of new parameters. This is equivalent to imposing a lower prior probability $P(M)$ on models with larger numbers of parameters.

Even though edges in Bayesian networks are directed, we cannot in general infer causal relationships from the structure directly, similar to the earlier discussion in relation to the causal inference test. For a network with three nodes, X_1 , X_2 , and X_3 , there are multiple groups of structures that are mathematically equivalent. For example, the three models, $M1 : X_1 \rightarrow X_2, X_2 \rightarrow X_3$, $M2 : X_2 \rightarrow X_1, X_2 \rightarrow X_3$, and $M3 : X_2 \rightarrow X_1, X_3 \rightarrow X_2$, are all Markov equivalent, meaning that they all encode for the same conditional independence relationship: $X_1 \perp\!\!\!\perp X_3 | X_2$, X_1 , and X_3 are independent conditional on X_2 . In addition, these models are mathematically equivalent:

$$\begin{aligned} p(X) &= p(M1|D) = p(X_2|X_1)p(X_1)p(X_3|X_2) \\ &= p(M2|D) = p(X_1|X_2)p(X_2)p(X_3|X_2) \\ &= p(M3|D) = p(X_2|X_3)p(X_3)p(X_1|X_2). \end{aligned}$$

Thus, from correlation data alone we cannot infer whether X_1 is causal for X_2 or vice versa from these types of structures. It is worth noting, however, that there is a class of structures, V-shape structures (e.g. $Mv : X_1 \rightarrow X_2, X_3 \rightarrow X_2$), that have no Markov-equivalent structure. Because there are more parameters to estimate in the Mv model than in the $M1$, $M2$, or $M3$ models, there is a large penalty in the BIC score for the Mv model. Therefore, in practice, a large sample size is needed to differentiate the Mv model from the $M1$, $M2$, or $M3$ models.

25.3.5 Integrating Genetic Data as a Structure Prior to Enhance Causal Inference in the Bayesian Network Reconstruction Process

In general, Bayesian networks can only be solved to Markov-equivalent structures, so it is often not possible to determine the causal direction of a link between two nodes even though Bayesian networks are directed graphs. However, the Bayesian network reconstruction algorithm can take advantage of genetic data to break the symmetry among nodes in the network that lead to Markov-equivalent structures, thereby providing a way to infer causal directions in the network in an unambiguous fashion (Zhu *et al.*, 2004). The reconstruction algorithm can be modified to incorporate genetic data as prior evidence that two quantitative traits may be causally related based on a causality test (Zhu *et al.*, 2004). The genetic priors can be constructed from three basic sources. First, gene expression traits associated with DNA variants that are coincident with the gene's physical location (*cis* eQTL) (Doss *et al.*, 2005) are allowed to be parent nodes of genes with coincident *trans* eQTL, $p(cis \rightarrow trans) = 1$, but genes with *trans* eQTL are not allowed to be parents of genes with *cis* eQTL, $p(trans \rightarrow cis) = 0$. Second, after identifying all associations between different genetic loci and expression traits at some reasonable significance threshold, genes from this analysis with *cis* or *trans* eQTL can be tested individually for pleiotropic effects at each of their eQTL to determine whether any other genes in the set are driven by common eQTL (Lum *et al.*, 2006). If such pleiotropic effects are detected, the corresponding gene pair and locus giving rise to the pleiotropic effect can then be used to infer a causal/reactive or independent relationship based on the causality test described above. If an independent relationship is inferred, then the prior probability that gene A is a parent of gene B can be scaled as

$$p(A \rightarrow B) = 1 - \frac{\sum_i p(A \perp B | A, B, l_i)}{\sum_i 1},$$

where the sums are taken over all loci used to infer the relationship. If a causal or reactive relationship is inferred, then the prior probability is scaled as

$$p(A \rightarrow B) = \frac{2 \sum_i (A \rightarrow B | A, B, l_i)}{\sum_i (A \rightarrow B | A, B, l_i) + p(B \rightarrow A | A, B, l_i)}.$$

Finally, if the causal/reactive relationship between genes A and B cannot be determined from the first two sources, the complexity of the eQTL signature for each gene can be taken into consideration. Genes with a simpler, albeit stronger, eQTL signature (i.e. a small number of eQTL that explain the genetic variance component for the gene, with a significant proportion of the overall variance explained by the genetic effects) can be considered as more likely to be causal compared with genes with more complex and possibly weaker eQTL signatures (i.e. a larger number of eQTL explaining the genetic variance component for the gene, with less of

the overall variance explained by the genetic effects). The structure prior that gene A is a parent of gene B can then be taken to be

$$p(A \rightarrow B) = 2 \frac{1 + n(B)}{2 + n(A) + n(B)},$$

where $n(A)$ and $n(B)$ are the number of eQTL at some predetermined significance level for genes A and B , respectively.

25.3.6 Incorporating Other Omics Data as Network Priors in the Bayesian Network Reconstruction Process

Just as genetic data can be incorporated as a network prior in the Bayesian network reconstruction algorithm, so can other types of data such as transcription factor binding site (TFBS) data, protein–protein interaction (PPI) data, and protein–small molecule interaction data. PPI data can be used to infer protein complexes to enhance the set of manually curated protein complexes (Guldener *et al.*, 2006). PPI-inferred protein complexes can be combined with manually curated sets, and each protein complex can then be examined for common transcription factor binding sites at the corresponding genes. If some proportion of the genes in a protein complex (e.g. half) carry a given TFBS, then all genes in the complex can be included in the TFBS gene set as being under the control of the corresponding transcription factor.

Given that the scale-free property is a general property of biological networks (i.e. most nodes in the network are linked to a small number of nodes, whereas a smaller number of nodes are linked to many nodes) (Albert *et al.*, 2000), inferred and experimentally determined TFBS data can be incorporated into the network reconstruction process by constructing scale-free priors, in a manner similar to the scale-free priors others have constructed to integrate expression and genetic data (Lee *et al.*, 2006). Given a transcription factor, T , and a set of genes, G , that contain the binding site of T , the transcription factor (TF) prior, p_{tf} , can be defined so that it is proportional to the number of expression traits correlated with the TF expression levels, for genes carrying the corresponding TFBS:

$$\log(p_{tf}(T \rightarrow g)) \propto \log \left(\sum_{g_i \in G} p_{qtl}(T \rightarrow g_i) \delta \right),$$

where $p_{qtl}(T \rightarrow g)$ is the prior for the QTL and

$$\delta = \begin{cases} 1, & \text{if } \text{corr}(T, g_i) \geq r_{\text{cutoff}}, \\ 0, & \text{if } \text{corr}(T, g_i) < r_{\text{cutoff}}. \end{cases}$$

The correlation cutoff r_{cutoff} can be determined by permuting the data and then selecting the maximum correlation values in the permuted data sets (corresponding to some predetermined, reasonable FDR). This form of the structure prior favors TFs that have a large number of correlated responding genes. From the set of priors computed from the inferred and experimentally determined TFBS set, only non-negative priors should be used to reconstruct the Bayesian network. For those protein complexes that could not be integrated into the network reconstruction process using scale-free priors, uniform priors were used for pairs of genes in these complexes (i.e. $p_{pc}(g_i \rightarrow g_j) = p_{pc}(g_j \rightarrow g_i) = c$).

Small molecule–protein interactions can also be incorporated into the Bayesian network reconstruction process. Chemical reactions reflected in biochemical pathways and the associated catalyzing enzymes can be identified as metabolite–enzyme pairs from existing pathway databases such as KEGG. These relationships can then be stored in an adjacency matrix in which a 1 in a cell represents a direct connection between the metabolite and the enzyme. The shortest distance $d_{m,e}$ from an enzyme e to a metabolite m can then be calculated using the repeated matrix multiplication algorithm. The structure prior for the gene expression of an enzyme e affecting the metabolite concentration is related to their shortest distance $d_{m,e}$ as $p(m \rightarrow e) \propto e^{-\lambda d_{m,e}}$. The shorter the distance, the stronger the prior.

25.3.7 Illustrating the Construction of Predictive Bayesian Networks with an Example

The Bayesian network reconstruction algorithm can be employed to elucidate the rich correlation structure reflected in the module connectivity of association-based structures such as coexpression networks. Because reconstruction of Bayesian networks is an NP-hard problem (Garey and Johnson, 1979), the number of nodes that can be considered in the network and the extent of connections (edges) among these nodes must be reduced (over what can be considered in reconstructing coexpression networks) in order to make the problem tractable, thereby making such networks sparser than coexpression networks. Toward this end, Figure 25.4 shows the result of the Bayesian network reconstruction algorithm discussed above applied to an inflammatory bowel disease (IBD) associated coexpression network we identified as conserved across multiple IBD disease cohorts (Peters *et al.*, 2017). The network was constructed from gene expression data generated from intestinal tissues isolated from the IBD cohort, in addition to eQTL data identified from this gene expression and genome-wide genotype data from this same cohort. The nodes depicted in Figure 25.4(a) represent gene expression traits and clinical features measured in the IBD cohort. The edges represent the causal inferences made from the Bayesian network reconstruction process. The colored nodes represent a subnetwork identified from a coexpression network module generated on this same IBD cohort that was identified as the most enriched for the IBD disease gene expression signature and for genes harboring genetic variants associated with IBD (Peters *et al.*, 2017).

With such a probabilistic causal network structure, we can carry out key driver analysis to identify those nodes in a given subnetwork that are predicted to modulate the regulatory state of the subnetwork (Zhang *et al.*, 2013; Peters *et al.*, 2017). The large diamond nodes in the colored subnetwork of Figure 25.4(a) represent the top five key driver genes for the subnetwork that have not previously been causally associated with IBD. The different colors represent the different local network neighborhoods for each of the five key driver genes. Each of these colored subnetworks generates two core hypotheses relating to IBD: (1) variations in the state of these local subnetworks will alter IBD-related phenotypes; and (2) changing the state of the key driver gene will alter the states of the gene expression traits predicted by the network to change in response to changes in the indicated key driver gene. An example of this is indicated in Figure 25.4(b) for the key driver gene *GPSM3*. Not only did perturbations to *GPSM3* in an animal model for IBD increase the severity of disease in the animal model (knocking down *GPSM3* in the animal model resulted in a more severe phenotype), but also the changes in gene expression in the intestinal biopsies isolated between the control and knockout *GPSM3* animal models were significantly predicted by the network model (the network was enriched roughly fivefold for genes that were differentially expressed in the *GPSM3* knockout model), directly molecularly validating the accuracy of the network model.

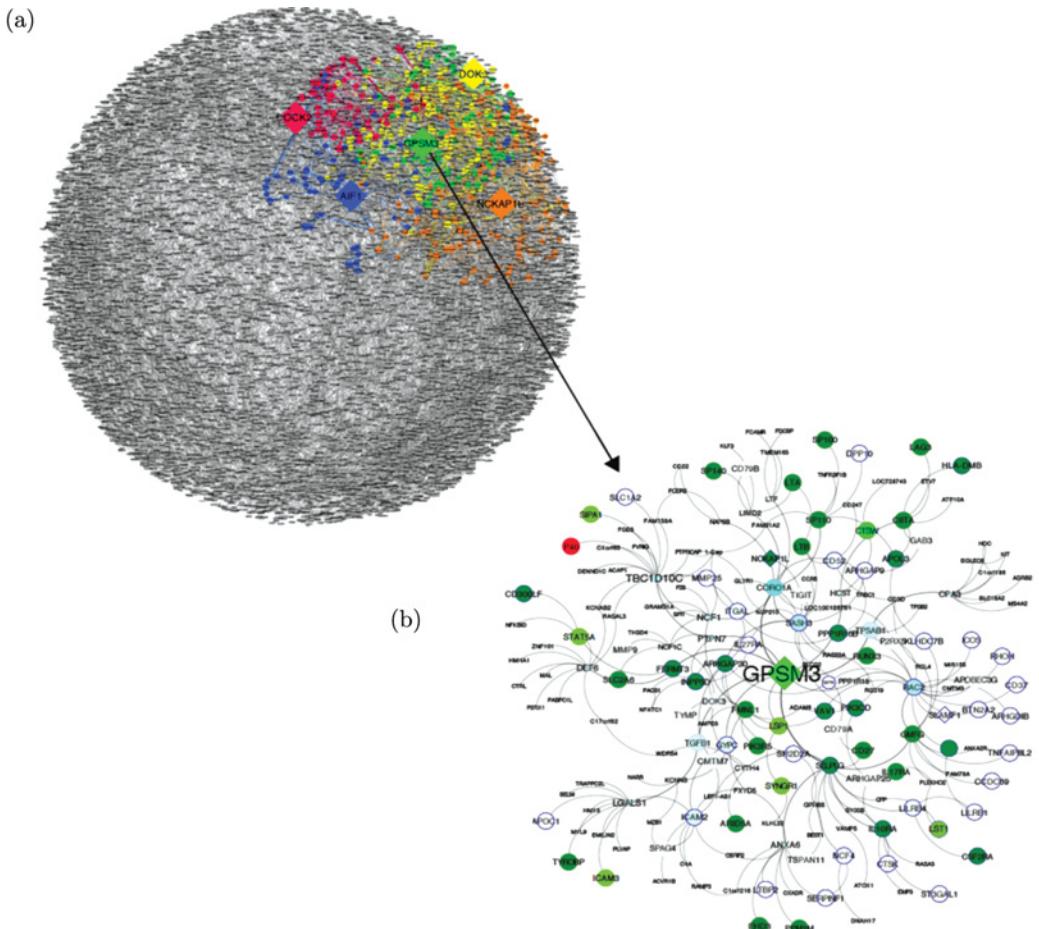


Figure 25.4 Bayesian network constructed from an inflammatory bowel disease cohort to identify novel key driver genes for IBD. (a) Bayesian network constructed from intestinal gene expression data generated on tissues from an IBD cohort. The colored nodes and diamond nodes are described in the main text. (b) The key driver gene, *GPSM3*, subnetwork is identified as a causal network for IBD that is enriched for *cis* eQTL that are also associated with IBD in a genome-wide significant fashion in large-scale genome-wide association studies for IBD. *GPSM3* was identified as a novel causal gene for IBD and experimentally validated at the physiologic and molecular levels. The gene expression signature derived from a *Gpsm3* knockout signature in an IBD mouse model was enriched nearly fivefold for genes in this *GPSM3* subnetwork, directly experimentally validating the causal structure of this subnetwork.

25.4 Conclusions

The eQTL mapping methods and network reconstruction methods, including the causality test, discussed here provide a convenient framework for moving beyond examining genes one at a time to understand complex phenotypes such as common human diseases. Whether considering the relationship between two traits with respect to common QTL driving each of the traits, interaction networks, Bayesian networks, or other types of networks reconstructed by integrating genetic and molecular profiling data, the advantage the network view affords is the ability to consider more of the raw data informing on complex phenotypes simultaneously, taking into

account dependency structures between all of the different fundamental components of the system, and providing a framework to integrate a diversity of data types (something Bayesian network reconstruction methods are particularly good at). Researchers in the life and biomedical sciences will have few other options available to them in considering the vast amounts of data being generated to elucidate how the hundreds of thousands or even millions of fundamental components within the cell interact to give rise to complex phenotypes. Networks provide one of the only frameworks of which we are aware to systematically and simultaneously take all of the fundamental components into account. Statistical inferences on networks will come to define much of the future research needed in this field to adequately leverage what network models can provide.

Of course, the types of approaches reviewed here represent only the first steps being taken to reconstruct meaningful gene networks, and even in this chapter we have largely restricted attention to genetic, gene expression, and disease phenotype data. Ultimately it will be necessary to integrate many different lines of experimental data simultaneously. Protein–protein interactions, protein–DNA interactions, protein–RNA interactions, RNA–RNA interactions, protein state, methylation state and especially differential methylation states that have now been shown to act transgenerationally (Anway *et al.*, 2005), epigenomic state, and interactions with metabolites, among other interactions, are important components that define complex phenotypes that emerge in living systems. What a given protein and RNA does will give way to what a network of protein, RNA, DNA, and metabolite interactions do, where such networks of interaction are defined by the context in which they operate, with environment playing a critical role. A particular network state that drives disease (or other complex phenotypes that define living systems) will require not only knowledge of DNA and environmental variation and the changes these variation components induce in the network, but also information on the previous states of the network that led to the current state, where environmental stresses interacting in complex ways with genetic background not only influence the current state of the network, but also can lead to longer-lasting effects on the network that act transgenerationally.

While this more comprehensive reconstruction of biological networks is still outside the scope of what is presently doable, the types of approaches discussed here represent useful steps toward this ultimate goal. Even though the number of networks that can be reconstructed from the fundamental components of living systems is truly daunting, as work progresses in this area we will learn the rules that necessarily constrain the possible ranges of molecular interactions, and as a result begin to capture the more conserved network motifs that form the framework upon which all other interactions are based. The complexity revealed by a systems-biology-motivated approach to elucidating complex phenotypes such as disease should be embraced, given the potential to develop a better understanding of the true diversity of disease and the constellation of genes that need to be targeted to effectively treat disease.

25.5 Software

qvalue: <https://github.com/StoreyLab/qvalue>

SVA: <https://bioconductor.org/packages/release/bioc/html/sva.html>

PEER: <https://github.com/PMBio/peer>

Matrix eQTL: http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

fastQTL: <http://fastqtl.sourceforge.net/>

Kendziora's mixture of markers method: <http://www.biostat.wisc.edu/~kendzior/MOM/>

CIT: <https://cran.r-project.org/web/packages/cit/index.html>

SMR: <http://cnsgenomics.com/software/smr>
 RTC: <https://qtltools.github.io/qtltools/>
 JLIM: <http://genetics.bwh.harvard.edu/wiki/sunyaevlab/jlim>
 eCaviar: <http://genetics.cs.ucla.edu/caviar/>
 Sherlock: <http://sherlock.ucsf.edu/>
 COLOC: <https://CRAN.R-project.org/package=coloc>
 MOLOC: <https://github.com/clagiamba/moloc>
 PrediXcan: <https://github.com/hakyim/PrediXcan>
 MetaXcan: <https://github.com/hakyimlab/MetaXcan>
 fQTL: <https://github.com/YPARK/fqtl>

References

- Albert, R. and Thakar, J. (2014). Boolean modeling: A logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **6**(5), 353–369.
- Albert, R., Jeong, H. and Barabasi, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* **406**(6794), 378–382.
- Alberts, R., Terpstra, P., Bystrykh, L.V., de Haan, G. and Jansen, R.C. (2005). A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171**(3), 1437–1439.
- Anway, M.D., Cupp, A.S., Uzumcu, M. and Skinner, M.K. (2005). Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* **308**(5727), 1466–1469.
- Azeloglu, E.U. and Iyengar, R. (2015). Good practices for building dynamical models in systems biology. *Science Signaling* **8**(371), fs8.
- Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**(5439), 509–512.
- Barabasi, A.-L. and Oltvai, Z.N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* **5**(2), 101–113.
- Barbeira, A.N., Dickinson, S.P., Torres, J.M., Bonazzola, R., Zheng, J., Torstenson, E.S., Wheeler, H.E., Shah, K.P., Edwards, T., Garcia, T., GTEx Consortium, Nicolae, D., Cox, N.J. and Im, H.K. (2017). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Preprint, bioRxiv 045260.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**(1), 289–300.
- Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**(5568), 752–755.
- Califano, A., Butte, A.J., Friend, S., Ideker, T. and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genetics* **44**(8), 841–847.
- Cervino, A.C., Li, G., Edwards, S., Zhu, J., Laurie, C., Tokiwa, G., Lum, P.Y., Wang, S., Castellini, L.W., Lusis, A.J., Carlson, S., Sachs, A.B. and Schadt, E.E. (2005). Integrating QTL and high-density SNP analyses in mice to identify *Insig2* as a susceptibility gene for plasma cholesterol levels. *Genomics* **86**(5), 505–517.
- Chang, R., Karr, J.R. and Schadt, E.E. (2015). Causal inference in biology networks with integrated belief propagation. In *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 359–370.

- Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., Threadgill, D.W., Manly, K.F. and Williams, R.W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics* **37**(3), 233–242.
- Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**(7063), 1365–1369.
- Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R. and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics* **49**(4), 600–605.
- DePrimo, S.E., Wong, L.M., Khatri, D.B., Nicholas, S.L., Manning, W.C., Smolich, B.D., O'Farrell, A.M. and Cherrington, J.M. (2003). Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: A novel strategy for biomarker identification. *BMC Cancer* **3**, 3.
- Ding, S.L., Royall, J.J., Sunkin, S.M., Ng, L., Facer, B.A., Lesnar, P., Guillozet-Bongaarts, A., McMurray, B., Szafer, A., Dolbeare, T.A., Stevens, A., Tirrell, L., Benner, T., Caldejon, S., Dalley, R.A., Dee, N., Lau, C., Nyhus, J., Reding, M., Riley, Z.L., Sandman, D., Shen, E., van der Kouwe, A., Varjabedian, A., Write, M., Zollei, L., Dang, C., Knowles, J.A., Koch, C., Phillips, J.W., Sestan, N., Wohnoutka, P., Zielke, H.R., Hohmann, J.G., Jones, A.R., Bernard, A., Hawrylycz, M.J., Hof, P.R., Fischl, B. and Lein, E.S. (2016). Comprehensive cellular-resolution atlas of the adult human brain. *Journal of Comparative Neurology* **524**(16), 3127–3481.
- Dobbyn, A., Huckins, L.M., Boocock, J., Sloofman, L.G., Glicksberg, B.S., Giambartolomei, C., Hoffman, G.E., Perumal, T.M., Girdhar, K., Jiang, Y., Raj, T., Ruderfer, D.M., Kramer, R.S., Pinto, D., CommonMind, C., Akbarian, S., Roussos, P., Domenici, E., Devlin, B., Sklar, P., Stahl, E.A. and Sieberts, S.K. (2018). Landscape of conditional eQTL in dorsolateral prefrontal cortex and co-localization with schizophrenia GWAS. *American Journal of Human Genetics* **102**(6), 1169–1184.
- Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research* **15**(5), 681–691.
- Franzen, O., Ermel, R., Cohain, A., Akers, N.K., Di Narzo, A., Talukdar, H.A., Foroughi-Asl, H., Giambartolomei, C., Fullard, J.F., Sukhavasi, K., Koks, S., Gan, L.M., Giannarelli, C., Kovacic, J.C., Betsholtz, C., Losic, B., Michoel, T., Hao, K., Roussos, P., Skogsberg, J., Ruusalepp, A., Schadt, E.E. and Bjorkegren, J.L. (2016). Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* **353**(6301), 827–830.
- Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., Klei, L.L., Kramer, R., Pinto, D., Gumus, Z.H., Cicek, A.E., Dang, K.K., Browne, A., Lu, C., Xie, L., Readhead, B., Stahl, E.A., Xiao, J., Parvizi, M., Hamamsy, T., Fullard, J.F., Wang, Y.C., Mahajan, M.C., Derry, J.M., Dudley, J.T., Hemby, S.E., Logsdon, B.A., Talbot, K., Raj, T., Bennett, D.A., De Jager, P.L., Zhu, J., Zhang, B., Sullivan, P.F., Chess, A., Purcell, S.M., Shinobu, L.A., Mangravite, L.M., Toyoshiba, H., Gur, R.E., Hahn, C.G., Lewis, D.A., Haroutunian, V., Peters, M.A., Lipska, B.K., Buxbaum, J.D., Schadt, E.E., Hirai, K., Roeder, K., Brennan, K.J., Katsanis, N., Domenici, E., Devlin, B. and Sklar, P. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience* **19**(11), 1442–1453.
- Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J. and Im, H.K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47**(9), 1091–1098.

- Garey, M.R. and Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, W. H. Freeman.
- Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., Clark, M.J., Pagnon, J., Yang, W.P., He, A., Truong, A., Patel, S., Nelson, S.F., Horvath, S., Berliner, J.A., Kirchgessner, T.G. and Lusis, A.J. (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences of the United States of America* **103**(34), 12741–12746.
- Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J. and Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* **2**(8).
- Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics* **10**(5), e1004383.
- Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A.E., CommonMind, C., Pasaniuc, B. and Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**(15), 2538–2545.
- Gianola, D. (2007). Inferences from mixed models in quantitative genetics. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edition. John Wiley & Sons, Chichester, pp. 678–717.
- Greenawalt, D.M., Dobrin, R., Chudin, E., Hatoum, I.J., Suver, C., Beaulaurier, J., Zhang, B., Castro, V., Zhu, J., Sieberts, S.K., Wang, S., Molony, C., Heymsfield, S.B., Kemp, D.M., Reitman, M.L., Lum, P.Y., Schadt, E.E. and Kaplan, L.M. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Research* **21**(7), 1008–1016.
- GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**(6235), 648–660.
- GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* **550**(7675), 204–213.
- Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W. and Stumpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Research* **34**, D436–441.
- Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**(6995), 88–93.
- He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X. and Li, H. (2013). Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics* **92**(5), 667–680.
- Hore, V., Vinuela, A., Buil, A., Knight, J., McCarthy, M.I., Small, K. and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* **48**(9), 1094–1100.
- Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics* **198**(2), 497–508.
- Höschele, I. (2007). Mapping quantitative trait loci in outbred pedigrees. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edition. John Wiley & Sons, Chichester, pp. 623–677.
- Huang, Q.Q., Ritchie, S.C., Brozynska, M. and Inouye, M. (2017). Power, false discovery rate and winner's curse in eQTL studies. Preprint, bioRxiv 209171.
- Huber, W., von Heydebreck, A. and Vingron, M. (2007). Analysis of microarray gene expression data. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edition. John Wiley & Sons, Chichester, pp. 203–230.

- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M. and Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**(1), 109–126.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**(5644), 449–453.
- Jansen, R., Hottenga, J.J., Nivard, M.G., Abdellaoui, A., Laport, B., de Geus, E.J., Wright, F.A., Penninx, B. and Boomsma, D.I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics* **26**(8), 1444–1451.
- Jansen, R.C. and Nap, J.P. (2001). Genetical genomics: The added value from segregation. *Trends in Genetics* **17**(7), 388–391.
- Jiang, C. and Zeng, Z.B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**(3), 1111–1127.
- Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.X., Nguyen, Q.T., Demirkale, C.Y., Feolo, M.L., Sharopova, N.R., Sturcke, A., Schaffer, A.A., Heard-Costa, N., Chen, H., Liu, P.C., Wang, R., Woodhouse, K.A., Tanriverdi, K., Freedman, J.E., Raghavachari, N., Dupuis, J., Johnson, A.D., O'Donnell, C.J., Levy, D. and Munson, P.J. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biology* **18**(1), 16.
- Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653), 2141–2144.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**(D1), D457–462.
- Karp, C.L., Grupe, A., Schadt, E., Ewart, S.L., Keane-Moore, M., Cuomo, P.J., Kohl, J., Wahl, L., Kuperman, D., Germer, S., Aud, D., Peltz, G. and Wills-Karp, M. (2000). Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nature Immunology* **1**(3), 221–226.
- Kendziorski, C.M., Chen, M., Yuan, M., Lan, H. and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**(1), 19–27.
- Kim, J.K., Gabel, H.W., Kamath, R.S., Tewari, M., Pasquinelli, A., Rual, J.F., Kennedy, S., Dybbs, M., Bertin, N., Kaplan, J.M., Vidal, M. and Ruvkun, G. (2005). Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**(5725), 1164–1167.
- Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**(3), 241–247.
- Lee, I., Date, S.V., Adai, A.T. and Marcotte, E.M. (2004). A probabilistic functional network of yeast genes. *Science* **306**(5701), 1555–1558.
- Lee, S.I., Pe'er, D., Dudley, A.M., Church, G.M. and Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proceedings of the National Academy of Sciences of the United States of America* **103**(38), 14062–14067.
- Leek, J.T. and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**(9), 1724–1735.
- Lewin, A. and Richardson, S. (2007). Bayesian methods for microarray data. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edition. John Wiley & Sons, Chichester, pp. 267–295.
- Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M., Gibson, G., Spector, T., Montgomery, G., Esko, T., Visscher, P.M.

- and Powell, J.E. (2017). The genetic architecture of gene expression in peripheral blood. *American Journal of Human Genetics* **100**(2), 228–237.
- Lum, P.Y., Chen, Y., Zhu, J., Lamb, J., Melmed, S., Wang, S., Drake, T.A., Lusis, A.J. and Schadt, E.E. (2006). Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *Journal of Neurochemistry* **97**(Suppl. 1), 50–62.
- Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63**(2), 215–232.
- Miller, C.L., Pjanic, M., Wang, T., Nguyen, T., Cohain, A., Lee, J.D., Perisic, L., Hedin, U., Kundu, R.K., Majmudar, D., Kim, J.B., Wang, O., Betsholtz, C., Ruusalepp, A., Franzen, O., Assimes, T.L., Montgomery, S.B., Schadt, E.E., Bjorkegren, J.L. and Quertermous, T. (2016). Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nature Communications* **7**, 12092.
- Millstein, J., Zhang, B., Zhu, J. and Schadt, E.E. (2009). Disentangling molecular relationships with a causal inference test. *BMC Genetics* **10**, 23.
- Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A. and Schadt, E.E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics* **75**(6), 1094–1105.
- Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R. and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**(7289), 773–777.
- Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**(3), 267–273.
- Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**(7001), 743–747.
- Morris, M.K., Saez-Rodriguez, J., Sorger, P.K. and Lauffenburger, D.A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry* **49**(15), 3216–3224.
- Mumford, J.A., Horvath, S., Oldham, M.C., Langfelder, P., Geschwind, D.H. and Poldrack, R.A. (2010). Detecting network modules in fMRI time series: A weighted network analysis approach. *Neuroimage* **52**(4), 1465–1476.
- Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., Salzberg, S.L., Holt, R.A., Kodira, C.D., Lu, F., Chen, L., Deng, Z., Evangelista, C.C., Gan, W., Heiman, T.J., Li, J., Li, Z., Merkulov, G.V., Milshina, N.V., Naik, A.K., Qi, R., Shue, B.C., Wang, A., Wang, J., Wang, X., Yan, X., Ye, J., Yooseph, S., Zhao, Q., Zheng, L., Zhu, S.C., Biddick, K., Bolanos, R., Delcher, A.L., Dew, I.M., Fasulo, D., Flanigan, M.J., Huson, D.H., Kravitz, S.A., Miller, J.R., Mobarry, C.M., Reinert, K., Remington, K.A., Zhang, Q., Zheng, X.H., Nusskern, D.R., Lai, Z., Lei, Y., Zhong, W., Yao, A., Guan, P., Ji, R.R., Gu, Z., Wang, Z.Y., Zhong, F., Xiao, C., Chiang, C.C., Yandell, M., Wortman, J.R., Amanatides, P.G., Hladun, S.L., Pratts, E.C., Johnson, J.E., Dodson, K.L., Woodford, K.J., Evans, C.A., Gropman, B., Rusch, D.B., Venter, E., Wang, M., Smith, T.J., Houck, J.T., Tompkins, D.E., Haynes, C., Jacob, D., Chin, S.H., Allen, D.R., Dahlke, C.E., Sanders, R., Li, K., Liu, X., Levitsky, A.A., Majoros, W.H., Chen, Q., Xia, A.C., Lopez, J.R., Donnelly, M.T., Newman, M.H., Glodek, A., Kraft, C.L., Nodell, M., Ali, F., An, H.J., Baldwin-Pitts, D., Beeson, K.Y., Cai, S., Carnes, M., Carver, A., Caulk, P.M., Center, A., Chen, Y.H., Cheng, M.L., Coyne, M.D., Crowder, M., Danaher, S., Davenport, L.B., Desilets, R., Dietz, S.M., Doucet, L., Dullaghan, P., Ferriera, S., Fosler, C.R., Gire, H.C., Gluecksmann, A.,

- Gocayne, J.D., Gray, J., Hart, B., Haynes, J., Hoover, J., Howland, T., Ibegwam, C., Jalali, M., Johns, D., Kline, L., Ma, D.S., MacCawley, S., Magooon, A., Mann, F., May, D., McIntosh, T.C., Mehta, S., Moy, L., Moy, M.C., Murphy, B.J., Murphy, S.D., Nelson, K.A., Nuri, Z., Parker, K.A., Prudhomme, A.C., Puri, V.N., Qureshi, H., Raley, J.C., Reardon, M.S., Regier, M.A., Rogers, Y.H., Romblad, D.L., Schutz, J., Scott, J.L., Scott, R., Sitter, C.D., Smallwood, M., Sprague, A.C., Stewart, E., Strong, R.V., Suh, E., Sylvester, K., Thomas, R., Tint, N.N., Tsionis, C., Wang, G., Wang, G., Williams, M.S., Williams, S.M., Windsor, S.M., Wolfe, K., Wu, M.M., Zaveri, J., Chaturvedi, K., Gabrielian, A.E., Ke, Z., Sun, J., Subramanian, G., Venter, J.C., Pfannkoch, C.M., Barnstead, M. and Stephenson, L.D. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**(5573), 1661–1671.
- Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I. and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics* **6**(4), e1000895.
- Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., GTEx Consortium and Dermitzakis, E.T. (2017). Estimating the causal tissues for complex traits and diseases. *Nature Genetics* **49**(12), 1676–1683.
- Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**(10), 1479–1485.
- Park, Y., Sarkar, A., He, L., Davilla-Velderrain, J., Jager, P.L.D. and Kellis, M. (2017a). Causal gene inference by multivariate mediation analysis in Alzheimer's disease. Preprint, bioRxiv 219428.
- Park, Y., Sarkar, A.K., Bhutani, K. and Kellis, M. (2017b). Multi-tissue polygenic models for transcriptome-wide association studies. Preprint, bioRxiv 107623.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.
- Pearl, J. (2000). *Causality*. Cambridge University Press, New York.
- Peters, L.A., Perrigoue, J., Mortha, A., Iuga, A., Song, W.M., Neiman, E.M., Llewellyn, S.R., Di Narzo, A., Kidd, B.A., Telesco, S.E., Zhao, Y., Stojmirovic, A., Sendecki, J., Shameer, K., Miotti, R., Losic, B., Shah, H., Lee, E., Wang, M., Faith, J.J., Kasarskis, A., Brodmerkel, C., Curran, M., Das, A., Friedman, J.R., Fukui, Y., Humphrey, M.B., Iritani, B.M., Sibinga, N., Tarrant, T.K., Argmann, C., Hao, K., Roussos, P., Zhu, J., Zhang, B., Dobrin, R., Mayer, L.F. and Schadt, E.E. (2017). A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature Genetics* **49**(10), 1437–1449.
- Petretto, E., Mangion, J., Dickens, N.J., Cook, S.A., Kumaran, M.K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N. and Aitman, T.J. (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics* **2**(10), e172.
- Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**(7289), 768–772.
- Pounds, S.B., Cheng, C. and Onar, A. (2007). Statistical inference for microarray studies. In D.J. Balding, M. Bishop and C. Cannings (eds), *Handbook of Statistical Genetics*, 3rd edition. John Wiley & Sons, Chichester, pp. 231–266.
- Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A. and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics* **7**(2), e1001317.
- Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**(5586), 1551–1555.
- Schadt, E.E. (2005). Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. *Current Opinion in Biotechnology* **16**(6), 647–654.

- Schadt, E.E. (2006). Novel integrative genomics strategies to identify genes for complex traits. *Animal Genetics* **37**(Suppl. 1), 18–23.
- Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**(6929), 297–302.
- Schadt, E.E., Edwards, S.W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K.W., Russell, A., Li, G., Cavet, G., Castle, J., McDonagh, P., Kan, Z., Chen, R., Kasarskis, A., Margarint, M., Caceres, R.M., Johnson, J.M., Armour, C.D., Garrett-Engele, P.W., Tsinoremas, N.F. and Shoemaker, D.D. (2004). A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology* **5**(10), R73.
- Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., Lum, P.Y., Leonardson, A., Thieringer, R., Metzger, J.M., Yang, L., Castle, J., Zhu, H., Kash, S.F., Drake, T.A., Sachs, A. and Lusis, A.J. (2005a). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**(7), 710–717.
- Schadt, E.E., Sachs, A. and Friend, S. (2005b). Embracing complexity, inching closer to reality. *Science's STKE: Signal Transduction Knowledge Environment* **2005**(295), pe40.
- Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., Guhathakurta, D., Derry, J., Storey, J.D., Avila-Campillo, I., Kruger, M.J., Johnson, J.M., Rohl, C.A., van Nas, A., Mehrabian, M., Drake, T.A., Lusis, A.J., Smith, R.C., Guengerich, F.P., Strom, S.C., Schuetz, E., Rushmore, T.H. and Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6**(5), e107.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461–464.
- Shabalin, A.A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**(10), 1353–1358.
- Shirasaki, D.I., Greiner, E.R., Al-Ramahi, I., Gray, M., Boontheung, P., Geschwind, D.H., Botas, J., Coppola, G., Horvath, S., Loo, J.A. and Yang, X.W. (2012). Network organization of the huntingtin proteomic interactome in mammalian brain. *Neuron* **75**(1), 41–57.
- Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y., Cavet, G., Wu, L.F., Altschuler, S.J., Edwards, S., King, J., Tsang, J.S., Schimmack, G., Schelter, J.M., Koch, J., Ziman, M., Marton, M.J., Li, B., Cundiff, P., Ward, T., Castle, J., Krolewski, M., Meyer, M.R., Mao, M., Burchard, J., Kidd, M.J., Dai, H., Phillips, J.W., Linsley, P.S., Stoughton, R., Scherer, S. and Boguski, M.S. (2001). Experimental annotation of the human genome using microarray technology. *Nature* **409**(6822), 922–927.
- Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* **7**(3), 500–507.
- Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**: 479–498.
- Threadgill, D.W. and Churchill, G.A. (2012). Ten years of the Collaborative Cross. *Genetics* **190**(2), 291–294.
- van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871), 530–536.
- Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J. and Geschwind, D.H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**(7351), 380–384.

- Waring, J.F., Jolly, R.A., Ciurlionis, R., Lum, P.Y., Praestgaard, J.T., Morfitt, D.C., Buratto, B., Roberts, C., Schadt, E. and Ulrich, R.G. (2001). Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicology and Applied Pharmacology* **175**(1), 28–42.
- Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., Hoen, P.A.C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B.J. and Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**(10), 1238–1243.
- Westra, H.J., Arends, D., Esko, T., Peters, M.J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B.P., Andiappan, A.K., Li, Y., Fu, J., Karjalainen, J., Plattee, M., Visschedijk, M., Weersma, R.K., Kasela, S., Milani, L., Tserel, L., Peterson, P., Reinmaa, E., Hofman, A., Uitterlinden, A.G., Rivadeneira, F., Homuth, G., Petersmann, A., Lorbeer, R., Prokisch, H., Meitinger, T., Herder, C., Roden, M., Grallert, H., Ripatti, S., Perola, M., Wood, A.R., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Knight, J.C., Melchiotti, R., Lee, B., Poidinger, M., Zolezzi, F., Larbi, A., Wang de, Y., van den Berg, L.H., Veldink, J.H., Rotzschke, O., Makino, S., Salomaa, V., Strauch, K., Volker, U., van Meurs, J.B., Metspalu, A., Wijmenga, C., Jansen, R.C. and Franke, L. (2015). Cell specific eQTL analysis without sorting cells. *PLoS Genetics* **11**(5), e1005223.
- Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., GTEx Consortium, Cox, N.J., Nicolae, D.L. and Im, H.K. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS Genetics* **12**(11), e1006423.
- Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T.H., D'Ambrosio, D., Gallins, P., Ha, M.J., Hottenga, J.J., Huang, S., Kattenberg, M., Kochhar, J., Middeldorp, C.M., Qu, A., Shabalina, A., Tischfield, J., Todd, L., Tzeng, J.Y., van Grootheest, G., Vink, J.M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J.H., de Geus, E.J., Yin, Z., Penninx, B.W. and Boomsma, D.I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics* **46**(5), 430–437.
- Yang, S., Liu, Y., Jiang, N., Chen, J., Leach, L., Luo, Z. and Wang, M. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics* **15**: 13.
- Zeng, Z.B., Liu, J., Stam, L.F., Kao, C.H., Mercer, J.M. and Laurie, C.C. (2000). Genetic architecture of a morphological shape difference between two *Drosophila* species. *Genetics* **154**(1), 299–310.
- Zhang, B., Gaiteri, C., Bodea, L.G., Wang, Z., McElwee, J., Podtelezhnikov, A.A., Zhang, C., Xie, T., Tran, L., Dobrin, R., Fluder, E., Clurman, B., Melquist, S., Narayanan, M., Suver, C., Shah, H., Mahajan, M., Gillis, T., Mysore, J., MacDonald, M.E., Lamb, J.R., Bennett, D.A., Molony, C., Stone, D.J., Gudnason, V., Myers, A.J., Schadt, E.E., Neumann, H., Zhu, J. and Emilsson, V. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**(3), 707–720.
- Zhu, J., Lum, P.Y., Lamb, J., GuhaThakurta, D., Edwards, S.W., Thieringer, R., Berger, J.P., Wu, M.S., Thompson, J., Sachs, A.B. and Schadt, E.E. (2004). An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and Genome Research* **105**(2–4), 363–374.

- Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E. and Schadt, E.E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40**(7), 854–861.
- Zhu, J., Sova, P., Xu, Q., Dombek, K.M., Xu, E.Y., Vu, H., Tu, Z., Brem, R.B., Bumgarner, R.E. and Schadt, E.E. (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology* **10**(4), e1001301.
- Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**(5), 481–487.

26

Statistical Methods for Single-Cell RNA-Sequencing

Tallulah S. Andrews, Vladimir Yu. Kiselev, and Martin Hemberg

Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

Abstract

Advances in technology have made it possible to profile the transcriptomes of individual cells using the method known as single-cell RNA-sequencing (scRNA-seq). The cellular resolution provided by scRNA-seq makes it possible to examine the heterogeneity of complex tissues. Single-cell RNA-seq has made it possible to address fundamental biological questions as well as providing important insights for clinical applications. However, a typical scRNA-seq experiment generates large amounts of noisy data, and analysis remains challenging. In this chapter we cover some of the most common types of analyses that are required to gain biological insights. We highlight the most challenging aspects from a statistical point of view as well as the strategies that have been used by researchers.

26.1 Introduction

Over the last two decades, technological advances have made it possible to measure transcript levels in individual cells in an unbiased, genome-wide manner. This process of measuring the transcriptome is generally referred to as RNA-sequencing, or RNA-seq for short (Wang *et al.*, 2009). Since the transcriptome is relevant for a wide range of biological processes, RNA-seq has provided a transformative technology and it has allowed for important insights in both basic biology and translational settings.

RNA-seq emerged as a successor to microarrays in the mid-2000s, and since the first protocols required millions of cells, they are typically referred to as bulk RNA-seq. Tang *et al.* (2009) were the first to publish single-cell RNA-seq (scRNA-seq) in 2009. Since then, many different groups have published improved protocols (for an overview, see, for example, Haque *et al.*, 2017; Wu *et al.*, 2017; Kalisky *et al.*, 2017), and several commercial platforms have also become available from companies such as 10X Genomics, Fluidigm and Dolomite. The overall workflow for an scRNA-seq experiment involves several steps: (i) dissociation of individual cells; (ii) cell lysis and extraction of transcripts; (iii) reverse transcription and initial amplification; (iv) library preparation; (v) sequencing. Developing new protocols and technologies to improve the accuracy, throughput and cost of scRNA-seq experiments is currently an active area of research, and although the overall workflow is likely to remain the same, the individual steps will be made more efficient. The improved protocols for cell isolation and library preparation, combined with the drastic reduction in sequencing costs, mean that scientists

can now routinely analyse ~10,000 individual cells in a single experiment. As of early 2018, data sets with more than 1 million cells are publicly available and, based on historical trends, the number of cells in a typical experiment is set to continue to grow rapidly (Svensson *et al.*, 2018).

Although an scRNA-seq experiment typically is more expensive than a bulk RNA-seq from the same sample, it has several advantages. Most importantly, since data are obtained for individual cells, it is possible to characterize the heterogeneity of cellular populations. Consequently, researchers are able to study the cellular composition of complex tissues across space and time. This is particularly important for highly complex tissues such as the central nervous system (Ofengheim *et al.*, 2017), or for situations where one is interested in very rare cell types, as is often the case when studying the immune system (Papalexi and Satija, 2018; Stubbington *et al.*, 2017). Single-cell approaches have also made it possible to study the transcriptome in situations where very small amounts of cells are available, for example in early embryonic development (Deng *et al.*, 2014; Petropoulos *et al.*, 2016). However, the additional steps in scRNA-seq protocols make them more complex than bulk protocols, and this can make them more sensitive. The additional steps of cell handling and dissociation may result in biases due to the additional stress. Moreover, differences in size and shape of cells are important, and some cell types are much more amenable to scRNA-seq experiments than others. Although bulk and single-cell experiments are similar in many ways, there are fundamental differences between the data obtained. Thus, there are novel statistical challenges associated with analysing scRNA-seq data. The goal of this chapter is to provide an overview of some of those challenges and to survey the different approaches that have been used to tackle them. First, we present a more detailed overview of the most widely used experimental workflows for scRNA-seq. We then discuss the various challenges involved in processing the reads coming off the sequencer to produce an expression matrix that summarizes the transcript levels across all of the cells that were profiled in the experiment. The expression matrix is the starting point for many different types of analyses, such as clustering, pseudo-time ordering and network inference, that can help provide novel biological insights.

26.2 Overview of scRNA-Seq Experimental Platforms and Low-Level Analysis

Once a biological sample has been obtained, cells must be separated before they can be further analysed. For some samples (e.g. blood or other liquids), isolating individual cells is straightforward, whereas for others (e.g. brain or heart), it can be challenging to extract intact cells without inducing a stress response. Dissociation protocols are often tissue/organism-specific, and are largely independent of the cell/transcript capture and library preparation technique. Most dissociation protocols will produce a solution of freely flowing individual cells.

26.2.1 Low-Throughput Methods

Many early scRNA-seq experiments relied on capturing single cells using the Fluidigm C1 chip. While this system is able to perform lysis, reverse transcription and amplification of cDNA efficiently and in small reaction volumes, the high cost and small size of the microfluidic chips and potentially high doublet rates make this one of the least cost-effective protocols (Ziegenhain *et al.*, 2017). One advantage of the C1 is that it enables imaging of cells to identify doublets and damaged cells.

A popular method for isolating cells is fluorescent activated cell sorting (FACS), which has the added advantage of being able to estimate cell size and quality prior to capture, as well as enabling different samples to be sorted into specific wells on the same plate to facilitate the removal of batch effects. In addition, FACS enables the selection/enrichment of a specific subset of cells based on protein surface markers which is helpful for characterizing rare cell types. However, this comes at the cost of lower throughput: polymerase chain reaction (PCR) plates typically capture only 96–384 single cells, and involve higher costs per cell due to the larger reaction volumes. Another consideration when using FACS is cell stress and time restrictions, since transporting the sample and the sorting procedure itself take time and may further stress the dissociated cells.

An important advantage of Fluidigm C1 and other plate-sorting based strategies is that they allow for full-length coverage, that is, the distribution of reads is relatively uniform across the length of the transcripts. Smartseq2 (Picelli *et al.*, 2014) is the predominant full-transcript protocol. Sequencing the entire transcript increases the coverage of single nucleotide polymorphisms which can be used to identify different donors, allele-specific expression (Deng *et al.*, 2014), or tumour subclones (Tirosh *et al.*, 2016). However, correcting expression levels for gene-length bias is not trivial in single-cell data due to incomplete and biased coverage (Archer *et al.*, 2016).

26.2.2 High-Throughput Methods

Much higher throughput can be achieved by pooling cells prior to library preparation. The reads can later be assigned to their cell of origin since they are assigned a cell-specific barcode (i.e. a short sequence of nucleotides) prior to mixing. Today, the most popular high-throughput cell capture systems are based on microfluidic droplet systems (Table 26.1). These include the Chromium system by 10X Genomics and the open-source Drop-seq and InDrop systems. All of these methods produce water droplets containing barcoded primers that will hybridize to the mRNA and lysis buffer. The Chromium and InDrop package the barcoded primers into hydrogel ‘beads’ and perform reverse transcription inside the individual droplets. By contrast, Drop-seq uses barcoded primers attached to solid beads which capture the mRNA through hybridization. Droplets are then pooled and the RNA is extracted before performing reverse transcription. It has been observed that Drop-seq generally has lower mRNA capture efficiency than InDrop/Chromium (Svensson *et al.*, 2017). One potential explanation for this is that droplet formation is a partially stochastic process as there is variation in the size of droplets which can have consequences for the library sizes generated from the encapsulated cells. By contrast, cell capture systems with predetermined compartment sizes (e.g. chip- or plate-based methods) will have a more uniform distribution of reaction efficiencies.

Recently a micro-well-based alternative to droplet-based capture known as SeqWell (Gierahn *et al.*, 2017) has been proposed. This method retains the advantages of high-throughput capture of individual cells and small reaction volumes but also enables imaging of captured cells and the application of additional reagents to the cells prior to lysing. SeqWell has similar performance to Drop-seq, which is unsurprising since it collects mRNAs through hybridization to solid beads prior to reverse transcription similar to Drop-seq (Gierahn *et al.*, 2017). Another recently proposed method demonstrated that combinatorial indexing (i.e. iteratively attaching short nucleotide sequences to random subsets of cells to build up a unique barcode for each cell) can be used to generate cell-specific barcoded libraries from dissociated cells without the need to isolate individual cells (Cao *et al.*, 2017). An additional advantage of the combinatorial indexing strategy over droplet-based methods is that it can be applied to

Table 26.1 Methods for single-cell RNA-sequencing

Protocol	Cell capture	Included steps	Unique molecular identifiers	Spike-ins	Transcript coverage	Cell number*
Chromium	droplet	mRNA capture + reverse transcription in droplets	Yes	No	3' or 5'	1,000–10,000 [4000]
Drop-seq	droplet	mRNA capture only	Yes	No	3'	1,000–10,000 [4000/hour]
CEL-seq	plate	Linear amplification (IVT**)	Yes	Yes	3'	96–384/plate
MARS-seq	plate	Linear amplification (IVT**)	Yes	Yes	3'	96–384/plate
SeqWell	microwell	mRNA capture only	Yes	Yes	3'	1,000–80,000 [10,000]
Smartseq	plate	PCR amplification	No	Yes	Full-length	96–384/plate
STRT-seq	plate	PCR amplification	Yes	Yes	5'	96–384/plate
Fluidigm C1	Microfluidic chip	Cell capture, RT, pre-amplification	No	Yes	Full-length	96/chip

*Number in square brackets is the expected number when operating at a 3–5% doublet rate

** *In vitro* transcription

profile nuclei. Single-nuclei sequencing is an important variant of scRNA-seq which is used when studying tissues that are difficult to dissociate (e.g. neurons).

All high-throughput methods discussed above assume that the number of cells in each droplet will follow a Poisson distribution. Thus there is a trade-off between cell capture efficiency and doublet rate. The stochastic capture of cells in droplets creates a novel statistical challenge to correctly identify those droplet barcodes which correspond to droplets that contain a cell as opposed to just background RNA or multiple cells.

Once cells have been captured, RNA must be reverse transcribed into cDNA and then amplified before performing library preparation. High-throughput methods pool cells prior to library preparation, and thus prior to transcript fragmentation. As a consequence of the pooling, one can only retain information about one end of the transcript, which means that it is impossible to obtain coverage of the entire transcriptome using short reads.

26.2.3 Computational Analysis

Once the sequencing reads are obtained, processing the data is straightforward. First, the reads are filtered to remove low-quality specimens. This step is no different from other types of sequencing experiments and it involves inspection of the quality scores associated with each read, as well as evaluation of their diversity. These challenges are not unique to scRNA-seq and they have been reviewed elsewhere (Conesa *et al.*, 2016).

Once a high-quality set of reads have been identified, mapped and compared to a suitable annotation for quantification of transcripts or genes, the output is typically summarized as an expression matrix, or count matrix, X . By convention, each row of the expression matrix represents a gene, while each column represents a cell, although some authors work with the transposed version. Thus, in the following X_{ij} represents the expression level of gene i in cell j , and it is assumed that there are a total of g genes and n cells.

26.3 Novel Statistical Challenges Posed by scRNA-Seq

26.3.1 Estimating Transcript Levels

Gene expression can be viewed as a stochastic process, with the variability stemming from three sources: (i) thermodynamic variability which is inherent due to the low number of molecules involved in some of the key reactions inside the cell (Becskei *et al.*, 2005); (ii) differences in the biological state, for example cell type, chromatin accessibility or cell cycle (Elowitz *et al.*, 2002); and (iii) technical variability due to the experimental procedure (McIntyre *et al.*, 2011; Brennecke *et al.*, 2013). The technical variability stems from the fact that not all transcripts are captured and amplified with the same efficiency. Due to the low amounts of RNA found in an individual cell, carrying out technical replicates in the same way as for bulk RNA-seq (where a sample can be split prior to sequencing) is very challenging. Consequently, we have an incomplete understanding of the variability which is due to experimental factors, and it is challenging to deconvolute the biological and the technical variability. Fortunately, there are experimental tricks that can be employed to estimate or remove the technical variability. The two most popular are spike-ins and unique molecular identifiers (UMIs).

Spike-ins are a set of synthetic RNAs that are added at known concentrations to the sample after lysis and RNA extraction. The concentration of the most popular set of spike-ins generated by the External RNA Control Consortium (2005) spans seven orders of magnitude, which is meant to cover the dynamic range encountered in most eukaryote cells. Technical noise models

can be fitted to spike-ins, but this could be problematic since spike-ins may not experience the same technical noise as endogenous transcripts (Svensson *et al.*, 2017). Another disadvantage of spike-ins is that they cannot be used in droplet-based protocols.

Since the amount of mRNA in an individual cell is very small, it needs to be amplified prior to sequencing. Both the amplification rate and the sequencing efficiency depend on the nucleotide composition of the transcript, and consequently this could lead to additional biases. The idea behind UMIs is to attach a short (4–10 bp) tag to each transcript prior to amplification (Islam *et al.*, 2014). Thus, if two reads from the same mRNA are obtained after sequencing, one can determine if they originate from the same molecule based on their UMI tags. Most platforms will add UMIs to the 3' end of transcripts, which means that only the end of each molecule provides useful information. Thus, UMIs enable noise due to amplification of the low initial quantities of RNA in single cells to be computationally eliminated after sequencing. The pool of UMIs is typically chosen such that all pairs of barcodes have a difference, as quantified by the Hamming distance between cell i and j , d_{ij} , that exceeds a minimal value.

26.3.1.1 How to Handle Unique Molecular Identifiers

Since a typical cell contains $\sim 10^5$ mRNAs, each transcript will not have a unique barcode. Thus, a combination of UMI identity and location in the genome are typically used to identify unique transcripts. Due to the presence of errors, the number of unique transcripts may be either overestimated or underestimated without careful analysis. Depending on the RNA capture efficiency of the library preparation method and UMI length, multiple unique mRNAs derived from the same locus may be tagged with identical UMIs; these are known as ‘collisions’ and lead to an underestimation of the number of molecules present. By contrast, sequencing errors of the UMI will increase the number of distinct barcodes originating from a single mRNA molecule, resulting in an inflated molecule count. Several statistical approaches have been proposed to correct for sequencing errors and collisions to improve the abundance estimates provided by UMIs.

Single base-pair substitutions will generate new UMIs that differ from the original barcode by one or more bases (i.e. Hamming distance 1 or greater). If the pool of UMIs was designed such that no two valid UMIs have Hamming distance less than 2, then it is relatively straightforward to identify and collapse erroneous UMIs. Since the error rate per base is typically $\sim 0.5\%$ (Quail *et al.*, 2012), multiple errors are rare. However, for some protocols UMIs are synthesized at random and the set of possible barcodes contains true UMIs that have a minimal Hamming distance of 1.

Heuristic algorithms for correcting sequencing errors combine both the Hamming distance between UMIs and the relative frequency with which a particular locus–UMI combination was observed (f_i) to collapse false UMIs resulting from sequencing errors. The directed-adjacency algorithm (Smith *et al.*, 2017) creates a directed graph for each set of UMIs that map to the same locus as where pairs of nodes are connected if $d_{ij} = 1$ and $f_i/f_j > 2$. The connected components of this directed graph represent ‘true’ UMIs connected to all erroneous UMIs derived from them. Thus, the number of molecules at each locus is the number of connected components in the directed-adjacency graph.

In addition, heuristic approaches based on Bayesian statistics have been developed within the dropEst package (Petukhov *et al.*, 2018). Here UMI frequency, number of adjacent UMIs, empirical distribution of UMIs, and the location and type of the putative substitution are combined to calculate the posterior probability of a UMI being the result of a sequencing error.

A collision-correction factor can be estimated using a model where all possible UMIs are present in equal abundance during the reverse transcription reaction (Grün *et al.*, 2014). Under this model, the distribution of the number of molecules tagged by a particular UMI, C , is

approximately Poisson with rate parameter m/n , where m is the true number of molecules and n the number of possible UMIs. Thus the probability of seeing a UMI at least once is $1 - \exp(-m/n)$ and the expected number of observed number of barcodes, b , is given by $n[1 - \exp(-m/n)]$. These two relations can be rearranged to estimate the true number of molecules from the observed number of barcodes as $m = -n\log(1 - b/n)$. However, analysis of real data sets reveals that the assumption of equally frequent barcodes is typically invalid (Petukhov *et al.*, 2018). To adjust for collisions with a non-uniform distribution of UMIs, dropEst uses the empirical distribution of UMIs to estimate the collision probability. This bootstrapping approach randomly samples different numbers of UMIs from the empirical distribution and calculates the number of collisions.

26.3.1.2 Cell-Containing Droplet Identification

In addition to the challenge of estimating the true UMI count due to collisions and sequencing errors, for droplet-based scRNA-seq protocols one must also identify which cell barcodes correspond to droplets containing a single, intact cell as opposed to those containing only debris from damaged cells or multiple cells. The simplest method for distinguishing droplets containing cells from the background is to employ a total-UMI count threshold.

CellRanger (Zheng *et al.*, 2017) first corrects for cell barcode sequencing errors by merging barcodes that are found within a Hamming distance of 1 of each other. Then the threshold is calculated as one-tenth of the 99th percentile of the top N barcodes by total UMI counts, where N is the expected number of barcodes. This cut-off is based on the empirical observation that RNA content varies by no more than one order of magnitude among cells of a similar cell type.

dropEst (Petukhov *et al.*, 2018) has a more elaborate procedure as it corrects for cell barcode sequencing errors by comparing the UMI–locus distributions across cells. It is assumed that barcodes obtained from sequencing errors in real cells will have a similar distribution of UMIs across different transcripts, but with fewer reads at each locus. The expected number of common UMI–locus combinations was used to estimate the Poisson parameter for the null hypothesis that two cell barcodes are independent of each other.

If UMIs are not available, but spike-ins are, then a different strategy can be used. Assuming that all cells contain similar amounts of mRNA, one can identify outliers by considering the fraction of reads that are mapped to the spike-ins. This approach has been criticized based on the analysis of bulk RNA-seq data since the addition of spike-ins to a sample is usually a manual procedure and there is a risk of introducing uncontrolled technical biases (Risso *et al.*, 2014). There are also other aspects of the transcriptome that can be used to identify low-quality cells in a relatively straightforward manner. These include relatively simple features such as percentage of mitochondrial reads, number of detected genes or total library size (total number of reads) (Ilicic *et al.*, 2016).

For some protocols, it is possible to obtain an image of the cell before it is lysed. If such information is available, then it is possible to devise classifiers to automate this task. Building on this approach, more sophisticated methods have been developed (Ilicic *et al.*, 2016), where low-quality cells are identified by using a support vector machine which is trained on a curated set of generic features.

26.3.1.3 Normalizing for Library Size

For high-throughput sequencing experiments, the number of reads obtained per cell will fluctuate. Although some of the differences may be explained by differences in RNA content, it is advisable to normalize for sequencing depth to compensate for differences that may arise. Note that such normalization implicitly assumes that relative differences in expression, conditional on total mRNA content, are the feature of primary interest. Normalization is a very important

step of the analysis and a large number of methods have been developed, going back to bulk RNA-seq.

A straightforward method, based on reads per kilobase per million that was first presented for bulk RNA-seq (Mortazavi *et al.*, 2008), is counts per million (CPM). The expression level for gene i in cell j in read counts is divided by the total in the cell then multiplied by 1 million: $y_{ij} = 10^6 x_{ij} / \sum_i x_{ij}$. When applied to UMI counts as opposed to reads, the result is called transcripts per million (TPM).

A method that has become very popular for normalizing bulk RNA-seq data is to use size factors (Anders and Huber, 2010). The expression value of each gene is divided by a size factor which can be calculated by first obtaining the geometric mean of each gene across all cells, $m_i = \text{geomean}(x_{i\cdot})$. The size factor for cell j is then given by the median of the ratio of the expression values and m_i . Due to the presence of a large number of dropouts, size-factors tend to work poorly for scRNA-seq data. Quantile normalization is another popular strategy from bulk RNA-seq that has been adapted for single-cell data. The idea is to divide the counts for each cell by a specific quantile (e.g. 75th) of the expression values within a given cell. Similarly to size factors, quantile normalization may perform poorly due to the large number of zeros, and for shallowly sequenced data sets the 99th percentile may be required to get sensible values (McCarthy *et al.*, 2017). Using trimmed means of the \log_2 fold changes between cells has been shown to work well for bulk data (Robinson and Oshlack, 2010), but the large number of dropouts can again be problematic.

In recent years, normalization methods specifically adapted for scRNA-seq have been developed. In scran (Lun *et al.*, 2016), cells with similar library sizes are pooled together, and then a method similar to CPM is used to calculate a normalization factor for each group of cells. The pooling procedure is then repeated multiple times with different cells in each group and the size factor for each cell can be deconvoluted by solving an overdetermined set of linear equations. SCnorm (Bacher *et al.*, 2017) uses median quantile regression to identify the relationship between sequencing depth and gene expression. Genes with similar expression levels are then grouped together and a scaling factor is calculated to eliminate the dependence on the read depth.

As part of the nonparametric differential expression for single cells (NODES) package (Sengupta *et al.*, 2016), there is a nonparametric normalization strategy based on pseudocounted quantiles (pQ). The pQ strategy is similar to classic quantile normalization, but it also adds a pseudocount to lowly expressed genes. The authors demonstrate that pQ normalization reduces the dependence of the variance in gene expression upon the number of detected genes. The basis of Linnorm (Yip *et al.*, 2017) is a set of genes which are widely expressed and have low variance and skewness. This set of stable genes is identified automatically, which means that the method does not rely on synthetic spike-ins. The stable genes are then used as a basis for a transformation that aims to simultaneously minimize the skewness while making the data more homoscedastic.

Even though normalization for library size may seem like a straightforward and basic task, it is still an area of active research. As of early 2018, there is still no strong consensus in the field about what is the best method, and our understanding of how the different methods impact the downstream analyses remains incomplete. One of the reasons why it is difficult to establish best practice for library size normalization is that the sample is frequently heterogeneous and we do not know how to correct for the composition of cell types. Moreover, it is not clear what is the best way to evaluate whether the normalization method has improved the results. Many publications (e.g. scNorm, NODES and Linnorm) utilize differential expression (DE) to evaluate the quality of the normalization. Some authors have tried to address these issues by carrying out systematic comparisons of normalization methods. The scone package (Cole *et al.*, 2017) scores

normalization methods based on three aspects: how well they separate wanted from unwanted variation; how principal components are related to wanted and unwanted variation; and how distributional properties differ between samples. An important conclusion from scone analysis is that there is no globally optimal normalization method and the best method depends not only on the criterion, but also on the data. NormExpression (Wu *et al.*, 2018) serves a similar purpose as it allows the user to evaluate the consistency of different normalization methods.

26.3.1.4 Correcting for Technical and Biological Factors

A well-known problem in experimental design is that of batch effects, where differences can be attributed solely to the point in time at which the experiment was conducted, rather than any meaningful biological effects. Specifically, when multiple experimental batches are performed there are slight differences in sample handling, reagent quality, reaction rates and so on, which result in systematic differences between batches. Although it is advisable to minimize the number of batches of cells, there are many reason why cell collection and processing would have to occur in multiple experimental batches. A common reason is that technologies are limited in the number of cells which can be captured at once, meaning that multiple runs are required to obtain enough cells. Additionally, samples may not be available at the same place and time, while storage and transport are not feasible.

Fortunately, with proper experimental design, it is possible to remove some batch effects from the data through statistical analysis. If each experimental batch contains the same complement of biological conditions, known as ‘balanced batches’, or partially overlapping biological conditions, known as ‘unbalanced batches’, then a linear model or more sophisticated alternatives can be used to regress out the batch effects. By contrast, if multiple experimental replicates of each biological condition were performed then batch effects can only be removed using genes that are known to be constant, such as spike-ins.

The general framework of batch effect correction is to first identify one or more sets of control genes and/or equivalent cells that are subject only to technical noise. The next step is to learn the structure of the technical noise, and finally remove the learned technical noise from the whole expression matrix. Control genes must be selected based on external information and typically are chosen either as a subset of housekeeping genes or from external spike-ins. One complication stems from the fact that currently available lists of housekeeping genes are based on bulk experiments, and it is not clear to what extent they are valid for an scRNA-seq experiment from a heterogeneous population of cells. Equivalent cells may be provided as metadata, or they may be learned by comparing cell–cell similarities across batches.

The most straightforward approach is to model batch effects directly using a general linear model (GLM). In this framework the expression of each gene j in cell i from batch b is modelled as

$$X_{ijb} = \mu_j + C\beta_j + \gamma_{jb} + \delta_{jb}\epsilon_{ijb},$$

where μ_j is the overall expression level of the gene, $C\beta_j$ is the response to biological conditions provided in design matrix C , γ_{jb} and δ_{jb} are the effect of batch b on the expression level and variance of gene j respectively, and ϵ_{ijb} are normally distributed errors with mean zero and standard deviation σ_j . The corrected expression level is then

$$X_{ijb}^* = (X_{ijb} - \mu_j - C\beta_j - \gamma_{jb}) / (\delta_{jb} + \mu_j + C\beta_j).$$

Combat (Johnson *et al.*, 2007), which was developed for microarray data, uses this model but applies an empirical Bayes fitting procedure which shares information across genes based on the assumption that batch effects affect many genes in similar ways. Combat ‘shrinks’ parameter estimates to improve fitting for data sets with small sample sizes. A simplified version of this

method which excluded the effect of batches on variance (δ_{jb}) was shown by Tung *et al.* (2017) to outperform less explicit methods. However, to use these methods both the biological state (i.e. cell type) and the technical group must be known for every cell *a priori* and must never be confounded. Both assumptions rarely hold for single-cell data sets.

Another popular method for batch correction is the remove unwanted variation (RUv) method which uses singular value decomposition (SVD) to identify variables associated with unwanted variation, as defined by control genes (RUvG) or control cells (RUvS), associated with each cell (Risso *et al.*, 2014). SVD factorizes the transposed expression matrix as the product of three matrices $X^T = U\Sigma V$, where U is a $n \times n$ matrix, Σ is an $n \times g$ diagonal matrix, and V is a $g \times g$ matrix. The factors are used to define the matrix of unwanted variation as $W = U\Sigma_k$, where Σ_k is obtained by retaining only the k largest singular values from Σ . The number of unwanted components (k) must be specified *a priori* but typically any k larger than the number of batches will give consistent results. W is then incorporated into a general linear model,

$$X = W\alpha + C\beta + S.$$

Here C is the matrix of biological covariates of interest (i.e. different cell types/conditions), and a major challenge in applying RUv is that this information is rarely known *a priori*. S are offsets from another normalization method. Both S and C may be set to 0 if they are not known. This model can be used to directly infer DE (β) or to regress out the unwanted variation ($W\alpha$).

It is often relevant to regress out other known biological factors, e.g. cell-cycle or stress signals. To this effect, the single-cell latent variable models (scLVM) method (Buettnner *et al.*, 2015) identifies hidden variables corresponding to these effects, making it possible to separate, and optionally remove, them from the expression matrix. scLVM learns the hidden effects using a Gaussian latent variable model from the cell–cell covariance matrix. For example, technical effects can be learned by supplying control genes, such as spike-ins. Other biological variables that are considered confounding, such as apoptosis, may be learned by supplying a list of relevant genes. The gene-specific parameters for each of the hidden variables are estimated using a linear mixed model,

$$X_j \sim N(\mu_j, \Sigma_h \sigma_{jh}^2 C_h + v_j^2 I),$$

where σ_{jh}^2 is the gene-specific variance attributable to hidden factor h and C_h is the corresponding cell–cell covariance structure, and v_j^2 is the residual biological variation of interest. The fitted σ_{jh}^2 can then be used to calculate the predictive power of hidden effects for each cell and the original expression matrix can be corrected by subtracting the means of this predictive distribution. Although scLVM provides a powerful and general framework for regressing out unwanted effects, an important caveat is that the user must provide a list of genes that are relevant for the effect of interest.

Correcting for batch and other confounding variables remains one of the central statistical challenges for scRNA-seq analysis. As with read-depth normalization, it is not always clear how to best evaluate the different procedures. Many authors evaluate batch removal visually by considering a two-dimensional projection of the data with the goal of having no separation of data points due to batch. To provide a more quantitative approach, the k -nearest-neighbours batch effect test (kBET) (Buttner *et al.*, 2019) has been developed. kBET assumes that the data contain balanced batches, and tests whether the local distribution of batch labels is the same as the global distribution. Specifically, a random subset of cells are chosen and the frequency of batch labels among their k nearest neighbours is tested against the global frequency using a χ^2 test. The overall rejection rate for these tests is a measure of how well mixed batches are in the corrected data.

Table 26.2 Approaches for imputing single-cell RNA-sequencing data

Method	Approach	Local or global	Implementation
BISCUIT	Gaussian mixture model	Global	R
MAGIC	Markov diffusion	Global	Matlab or Python
scImpute	Gamma-Normal mixture	Local	R
SAVER	Bayesian estimation of Poisson model	Global	R
DrImpute	k -means clustering of correlation matrices	Local	R
CIDR	Weighted mean	Global	R

26.3.1.5 Imputation

One of the challenges in analysing dropout events is that we do not know why there is a zero in the expression matrix. On the one hand, the zero could reflect the fact that the gene was not expressed in the cell. On the other hand, the absence could be due to technical noise; a failure in capturing the transcript, in the reverse transcription, during the amplification, or any of the other steps involved in the library preparation. Even if the transcript is present in the library, it may not show up due to insufficient sequencing depth, thus representing missing data.

Imputation of missing data is a common problem in many other areas, such as predicting unobserved alleles based on population structure (Li *et al.*, 2009). The imputation methods developed for scRNA-seq can be classified as either local or global, depending on what type of information they use to infer the values of dropouts (Table 26.2).

Bayesian Inference for Single-Cell clUstering and ImpuTing (BISCUIT; Prakhakaran *et al.*, 2016) is based on the assumption that gene expression values are drawn from a Gaussian mixture model. BISCUIT uses a hierarchical Dirichlet process mixture model as a framework for learning the parameters. One drawback of BISCUIT is that it is computationally intensive since it uses a Monte Carlo method for inferring the parameters. The authors of BISCUIT followed up their work with Markov Affinity-based Graph Imputation of Cells (MAGIC; van Dijk *et al.*, 2018), in which a k -nearest neighbour graph is built from the data and from this graph an affinity matrix is inferred based on a Gaussian kernel of the distance matrix. The normalized affinity matrix, M , is viewed as the transition matrix of a Markov process. M is then treated as a backward diffusion operator, and by applying it t times to the original count matrix D , the counts are influenced by their neighbours, $D_{\text{imputed}} = D_{\text{original}}M^t$. Since MAGIC is based on a diffusion model, the method may alter values that are non-zero as well, thus introducing additional biases and providing undesirable smoothing of the data.

scImpute (Li and Li, 2018) uses a two-stage procedure where in the first step cells are clustered using k -means to identify a set of neighbours for each cell. In the second step a gamma–normal mixture distribution is fitted for each gene using an EM algorithm. Based on the inferred distribution and a user-specified threshold, t , scImpute will infer the expression value of genes that are estimated to have a dropout probability greater than t using non-negative linear regression. A fourth imputation method currently available is Single-cell Analysis Via Expression Recovery (SAVER; Huang *et al.*, 2018). Like MAGIC and scImpute, the starting point is the count matrix, and for a UMI-based experiment it is assumed that the counts are Poisson distributed with $X_{ij} \sim \text{Poi}(s_j \lambda_{ij})$, where s_j is a cell-specific size factor and λ_{ij} represents the true number of molecules. SAVER assumes a gamma distribution as a prior for λ_{ij} , and since it is a Bayesian method it outputs a posterior distribution as well as a point estimate. DrImpute (Kwak *et al.*, 2018) uses a consensus approach similar to the one used by SC3 (Kiselev *et al.*, 2017) for

unsupervised clustering. Each dropout event X_{ij} is given the average of all other cells in the same cluster, and this procedure is repeated for clusterings based on both Pearson and Spearman correlations with different number of eigenvectors used for the k -means algorithm. Clustering through Imputation and Dimensionality Reduction (CIDR; Ntranos *et al.*, 2016) imputes by first identifying a cell-specific threshold, T_i . For all cell pairs where $x_{ij} < T_i$, the imputed value is set to the weighted mean of all other cells that are expressed above the threshold.

One of the challenges with imputation is to determine how to evaluate whether the imputation has led to any improvements. Many authors consider either clustering or pseudotime alignment since those are two of the most common applications. Zhang and Zhang (2018) have provided a thorough evaluation of many of the available imputation methods. They considered the impact of imputation on clustering, pseudotime alignment and DE using both real and synthetic data. Zhang and Zhang report that, in general, imputation improves clustering and differential expression analysis, while the impact on pseudotime alignment appears to be more modest. Given the difficulties of establishing whether or not imputation results in any improvements, there is still a debate on whether or not it is a good idea to use it. The main arguments against imputing are based on the fact that it introduces circularity which may distort the downstream analysis, for example, by introducing false positives or false negatives among the set of differentially expressed genes.

26.3.1.6 Isoform Quantification and Splice Junction Identification

One of the outstanding challenges for RNA-seq analysis relates to isoform quantification and splice junction identification (Finotello and Di Camillo, 2015). Although closely related, the two problems are often treated separately. For isoform quantification, the goal is to find the expression (i.e. number of reads or UMIs) of each isoform represented in the annotation. For splice junction identification the goal is to find out how often each splice junction is used. Splice junction usage is quantified by the percent splicing index (PSI), which is defined as the number of reads that support the inclusion of the exon divided by the sum of the number of reads that do and do not support the inclusion. Since scRNA-seq overcomes the issue of heterogeneous cell populations which can complicate matters in bulk samples, it is conceivable that splice junction identification and isoform quantification will be made easier for protocols that can capture information about the full length of the transcript.

Bayesian Regression for Isoform Estimation (BRIE; Huang and Sanguinetti, 2017) uses a prior based on sequence features and conservation to infer the PSI. The prior allows BRIE to overcome some of the issues related to dropouts and it serves as a means for imputing missing information. By contrast, SingleSplice (Welch *et al.*, 2016b) estimates the expression of groups of exons, or modules, that can be matched to a specific isoform. SingleSplice then uses technical noise estimates based on spike-ins to identify modules that are differentially expressed. The Expedition software package (Song *et al.*, 2017) contains three separate modules – outrigger, anchor and bon voyage – for carrying out *de novo* splice junction detection. Outrigger allows for the *de novo* detection of splicing events consistent with skipped and mutually exclusive exons based on read coverage and the presence of splice sites. Having identified splice junctions and estimated their PSI values, anchor is a Bayesian framework for categorizing them as being included, excluded, bimodal, middle (PSI $\sim 50\%$) or multimodal. Finally, bon voyage is a novel approach based on non-negative matrix factorization for visualizing splicing events across different cell populations.

To the best of our knowledge, no isoform quantification methods have been specifically developed for scRNA-seq data. However, a recent benchmarking study (Westoby *et al.*, 2018) suggests that some of the methods developed for bulk RNA-seq, such as Kallisto (Bray *et al.*, 2016), Salmon (Patro *et al.*, 2017) and RSEM (Li and Dewey, 2011), perform well for single-cell data.

For simulated Smart-seq2 data with relatively deep coverage, most methods show near-perfect recall, suggesting that very few isoforms are missed.

26.3.2 Analysis of the Expression Matrix

scRNA-seq data differs from bulk RNA-seq in two important ways: dropouts and sample size. Full-transcript protocols produce zero-inflated (dropout) gene expression distributions which are not well modelled by bulk RNA-seq methods. Even for deeply sequenced samples (Li *et al.*, 2016; Kolodziejczyk *et al.*, 2015), the dropout rate is ~50%, whereas for sparsely sequenced droplet-based experiments it can exceed 90%. Moreover, scRNA-seq experiments contain hundreds to thousands of samples whereas bulk experiments rarely contain more than a dozen samples. The large number of noisy samples in scRNA-seq facilitates the fitting of parametric distributions but also increases the computation cost. By contrast, bulk RNA-seq methods optimized for small, less noisy samples may have an inflated Type I error rate when applied to scRNA-seq data sets.

Bulk RNA-seq methods, such as edgeR and DESeq, model read counts with a negative binomial model. While this model may be appropriate for UMI counts from scRNA-seq, it is not appropriate for raw read counts from scRNA-seq due to the excess of zeros (Pierson and Yau, 2015). Alternatives which better describe scRNA-seq read counts are the zero-inflated negative binomial distribution and the Poisson-beta distribution. The latter is based on a model of stochastic gene expression which has strong empirical support (Kim and Marioli, 2013), but is challenging to work with analytically.

26.3.2.1 Dimensionality Reduction and Visualization

Due to the high-dimensional nature of scRNA-seq data, dimensionality reduction plays a central role in many analyses. In particular, visualization generally requires projecting data into two-dimensional space. Many popular, general-purpose methods for dimensionality reduction, including principal components analysis (PCA), diffusion maps (Singer *et al.*, 2009) and non-negative matrix factorization (Lee and Seung, 1999) have been used. However, currently the most popular method for visualizing data is tSNE (van der Maaten and Hinton, 2008) since it is efficient for large data sets and produces visually appealing plots with distinct clusters. The main shortcomings of tSNE are its stochasticity, nonlinearity and reliance on a ‘perplexity’ parameter which is difficult to interpret. Taken together, these aspects make it difficult to understand intuitively what tSNE does to the data set and how to interpret distances and angles in the plot.

Zero-inflated factor analysis (ZIFA) can be used to reduce the number of genes by identifying latent factors (Pierson and Yau, 2015). The zero-inflation provides ZIFA with a better fit than PCA-based models. A similar approach is zero-inflated negative binomial-based wanted variation extraction (ZINB-WaVE; Risso *et al.*, 2018), which can be viewed as an extension of the RUV framework. As the name implies, ZINB-WaVE assumes that the gene expression is a mixture of a negative binomial distribution and a Dirac distribution. It has also been demonstrated that neural networks can be helpful for the visualization task (Lin *et al.*, 2017).

26.3.2.2 Feature Selection

Single-cell RNA-seq assays tens of thousands of genes. However, in most situations only a small portion of those genes will be responding to the biological condition of interest (e.g. differences in cell type, drivers of differentiation or response to an environmental stimulus). The large number of genes not affected by the biology will still exhibit technical noise and batch effects which can obscure the biological signal of interest.

Thus, it is often advantageous to perform feature selection to exclude genes which do not contain any biological signal from downstream analysis. Importantly, this will increase the signal to noise ratio, and it will also reduce the amount of data that needs to be processed, thereby reducing the computational complexity of analyses. Supervised feature selection, such as DE, relies on *a priori* information of the structure of the data, which is often unavailable for single-cell experiments. Instead, one is often interested in unsupervised feature selection for which there are two main approaches.

The first approach is to identify genes which are outliers from a null model which describes the technical noise expected in the data set. The most common method for feature selection in scRNA-seq is to identify highly variable genes (Brennecke *et al.*, 2013). Since there is a positive relationship between the mean expression of a gene and the variance, adjustments are required. This can be done by considering the normalized variance (Brennecke *et al.*, 2013), using a moving median/mean (Kolodziejczyk *et al.*, 2015), statistically smoothing (Jiang *et al.*, 2016), or by binning genes by expression level and locally comparing variances (Macosko *et al.*, 2015; Satija *et al.*, 2015). Instead of using variance one can use the dropout rate as it is less sensitive to sampling noise (Andrews and Hemberg, 2016). Since zeros dominate scRNA-seq experiments, dropout-based feature selection has been shown to outperform variance-based feature selection in a variety of situations (Andrews and Hemberg, 2016; Kiselev and Hemberg, 2018).

The other approach to unsupervised feature selection is to identify the major sources of variability in the data using PCA or gene–gene correlations (Macosko *et al.*, 2015; Pollen *et al.*, 2014; Usoskin *et al.*, 2015). These approaches exploit the fact that real biological effects will affect multiple genes in similar ways. Technical noise should affect each gene independently and thus not give rise to gene–gene correlations/covariation; however, this is not the case for batch effects which systematically affect large numbers of genes. Hence, this approach is less commonly used than methods based on outlier detection.

26.3.2.3 Clustering

One of the most important applications of scRNA-seq is to characterize the composition of cell types in a complex tissue. For most tissues, our knowledge of which cell types are present and their relative proportions is incomplete. Consequently, an unsupervised clustering approach to identify cell types is called for (Table 26.3).

In many situations, it is not only the clustering itself that is challenging, but also the problem of determining the number of clusters, k . From a biological point of view, estimating k is

Table 26.3 Strategies for clustering single-cell RNA-sequencing data

Method	Estimate k	Approach	Implementation
SINCERA	Yes	Hierarchical	R
SC3	Yes	PCA + k -means	R
Seurat	Yes	PCA + Louvain	R
SCENIC	Yes	Regulatory modules activity	R
BackSPIN	Yes	Biclustering	Python
SIMLR	Yes	Kernel dim red + k -means	R, Matlab
CIDR	Yes	PCA + hierarchical	R
dropClust	Yes	Locality sensitive hashing + Louvain	R, Python

challenging since, for most samples, there is no unique choice that is unambiguously the best one. Instead, there are frequently multiple hierarchies present and several choices of k are reasonable from a biological point of view.

The starting point for clustering is the cell-by-cell distance matrix. Most methods include either dimensionality reduction or some sort of sparsification applied to the distance matrix prior to clustering. The most popular method for large data sets is Seurat (Satija *et al.*, 2015), which first creates a sparse representation of the distance matrix using a k -nearest-neighbour algorithm. The network is then partitioned into clusters using the Louvain algorithm (Blondel *et al.*, 2008), a very fast method for identifying modules in networks. dropClust (Sinha *et al.*, 2018) is similar to Seurat in that it first builds a k -nearest-neighbour network using locality-sensitive hashing and then uses the Louvain method to identify modules.

Due to the difficulties involved in identifying the best distance metric, researchers have developed more sophisticated strategies based on combining multiple estimates. Single-cell consensus clustering (SC3) (Kiselev *et al.*, 2017) uses a consensus approach whereby many different solutions obtained by applying k -means to different distance metrics are combined. Single-cell Interpretation via Multikernel LeaRning (SIMLR) is also based on k -means clustering but has a more sophisticated strategy since it automatically learns the distance metric that best fits the data (B. Wang *et al.*, 2017).

BackSPIN is a recursive method which splits the cell–cell correlation matrix and outputs a one-dimensional ordering of genes and cells. SINCERA (Guo *et al.*, 2015) is probably the most straightforward method as it centres each cell and normalizes the variance before carrying out hierarchical clustering. Clustering does not need to be done based on gene expression data alone. For example, Single Cell rEgulatory Network Inference and Clustering (SCENIC) (Aibar *et al.*, 2017) analyses regulatory regions for transcription factor binding sites to identify coexpression modules. The activity of the coexpression modules is then used as a basis for unsupervised clustering.

There are only a few samples with relatively small n where the ground truth is known, and this makes validation challenging (Kiselev *et al.*, 2017). Short of evaluating accuracy, one can evaluate consistency. Freytag *et al.* (2018) demonstrated that for a Drop-seq sample, the consistency between different methods was relatively low. A similar study was carried out by Menon (2017) who considered the trade-off between sequencing depth and number of cells. In his benchmarking study the total number of reads or cells were downsampled, and the results show that the Louvain-based Seurat algorithm performs best for samples with low read depth and high cell counts, whereas PCA-based methods perform better for deeply sequenced samples with a small number of cells.

26.3.2.4 Pseudotime

Many biological processes are continuous where cells do not belong to a distinct cell type. The most prominent examples are differentiation processes that take place during development and the cell cycle. In these cases cells change from one cell type or cell state to another over time in a continuous manner. Ideally, one would like to monitor the expression levels of an individual cell over time, but it is not possible with scRNA-seq since the cells are lysed when the RNA is extracted. Therefore, one must sample at multiple time-points and obtain snapshots of the gene expression trajectories. Since some of the cells will proceed faster along the trajectory than others, each snapshot may contain cells at varying states. A crucial task for the analysis of this type of experiment is to order the cells along one or more low-dimensional trajectories which represent the underlying biological process. This ordering is referred to as ‘pseudotime’ (see Table 26.4). Of particular interest are processes where cells can differentiate to distinct fates,

Table 26.4 Methods for ordering single-cell RNA-sequencing data in pseudotime

Method	Branches	Approach	Implementation
Monocle	Yes	Independent component analysis	R
Monocle2	Yes	Reverse graph embedding	R
TSCAN	Yes	Clustering + minimum spanning tree	R
SCUBA	Yes	Clustering + binary tree	R, Matlab
Wanderlust	No	Bootstrap k nearest neighbours + shortest path	Matlab
SLICER	Yes	Locally linear embedding + k nearest neighbours + shortest path	R
Waterfall	No	Clustering + shortest path	R

often represented as branched trajectories. Branched processes bring additional challenges in cell ordering.

The first and the most popular pseudotime method was Monocle (at the time of writing it has been replaced by Monocle 2) (Trapnell *et al.*, 2014; Qiu *et al.*, 2017). Historically, the idea of ordering was not new and the Monocle algorithm was inspired by a method developed for time-ordering of microarray samples (Magwene *et al.*, 2003). Monocle uses independent component analysis to reduce the dimensionality of the data. Then, a weighted complete graph is constructed, where vertices represent cells and edges are weighted by the distance in the independent components space. Next, the algorithm finds the minimum spanning tree (MST) on the graph, which is ultimately used to order the cells. The updated Monocle2 uses a completely different method whereby cells are ordered using a technique called reverse graph embedding (Mao *et al.*, 2016) following dimensionality reduction. Time reconstruction in Single-Cell RNA-seq ANalysis (TSCAN; Ji and Ji, 2016) is also based on MST, but it uses a two-stage approach whereby the cells are first clustered and then an MST is constructed based on the cluster centres. The idea is that clustering cells before MST construction reduces the complexity of the tree space and improves cell ordering. In addition, it also allows utilization of prior knowledge (e.g. known time points) to adjust the ordering.

Numerous other methods are also available for pseudotime ordering such as SCUBA (Marco *et al.*, 2014), Wanderlust (Bendall *et al.*, 2014), SLICER (Welch *et al.*, 2016a) and Waterfall (Shin *et al.*, 2015). All of them follow similar approach to the ones described above whereby the expression matrix is subject to dimensionality reduction before a trajectory is inferred. A more comprehensive overview of these methods can be found in Cannoodt *et al* (2016).

26.3.2.5 Differential Expression and Marker Genes

A common question in scRNA-seq is to identify those genes differentially expressed between cell types or biological conditions. Testing for DE was one of the most common analyses for bulk RNA-seq and microarray experiments. As a result many methods have been created and tailored specifically for those data types. The problem remains important for single-cell analysis, and there are a large number of methods available. A more detailed overview and comparison of available methods has recently been presented by Soneson and Robinson (2018).

The zero-inflated negative binomial model is the most commonly used parametric model of single-cell RNA-seq data, and several DE methods employ this model, including MAST (Finak *et al.*, 2015), scDD (Korthauer *et al.*, 2016), DECENT (Ye *et al.*, 2017) and SCDE (Kharchenko *et al.*, 2014). Both MAST and scDD test for differences in zero-inflation and changes in mean

expression separately, in addition to a combined test. In contrast, SCDE assumes that zero-inflation is simply a function of the mean-expression level of a gene. DECENT statistically infers whether zeros result from low mean expression, or from technical factors and imputes the latter. In addition, DESCEND (J. Wang *et al.*, 2018) considers a family of zero-inflated models when fitting scRNA-seq data for DE testing.

Other DE methods use the log-normal distribution to model normalized counts, such as limma-voom (Costa-Silva *et al.*, 2017), TASC (Jia *et al.*, 2017), and Seurat (Satija *et al.*, 2015), or a standard negative binomial model to model raw read counts, such as edgeR (Robinson *et al.*, 2010), DESeq (Anders and Huber, 2010) and Monocle (Trapnell *et al.*, 2014). BPSC is the only method using the Poisson-beta distribution to model scRNA-seq read counts (Vu *et al.*, 2016).

In addition, scRNA-seq allows parameters other than the mean to be tested for differences across populations. For instance, one may be interested in investigating whether the variability of a gene changes in different biological conditions (Kolodziejczyk *et al.*, 2015). As previously mentioned, MAST and scDD test for differences in zero-inflation rate in addition to differences in mean expression. DESCEND also allows separate testing of differences in zero-inflation and the non-zero mean. BASiCS (Vallejos *et al.*, 2015) models data with a negative binomial model and can test for differences in both the mean and dispersion parameters while adjusting for technical noise.

Parametric methods are popular for scRNA-seq analysis but are limited to modelling the raw counts rather than normalized and/or batch-effect-corrected expression levels. Thus, multiple DE methods based on nonparametric statistics have emerged. Nonparametric statistics are generalizable to data from any underlying distribution, but they have the drawback that they are unable to control for technical covariates within the model. The Cramér–von Mises and Kolmogorov–Smirnov tests are key components of the D3E method (Delmans and Hemberg, 2016). SC3 (Kiselev *et al.*, 2017) uses the Kruskal–Wallis test and the Wilcox rank-sum test to identify DE and marker genes, respectively. NODES (Sengupta *et al.*, 2016) employs a novel variant of the D statistic, a nonparametric test which combines a test of the magnitude of the difference in means and the Wilcox rank-sum test.

The approaches discussed above can be used to compare any two biological groups in a single-cell experiment, and the resulting significant genes are typically referred to as ‘differentially expressed’ genes, whereas if one group is tested against all others the resulting significant genes are typically referred to as ‘marker genes’. Both differentially expressed genes and marker genes may be used to interpret the biological function of identified groups of cells. Marker genes specifically are of interest for designing follow-up experiments to functionally characterize identified groups by enabling sorting and/or labelling of the specific cell population.

26.3.2.6 Network Inference

An important goal in genomics is to understand not only the biological role of individual genes, but also how they interact. Gene regulation is a complex process, with each gene typically being regulated by multiple other genes in a cell-type-specific manner. In theory, scRNA-seq data should be more powerful than bulk RNA-seq data since purer cell populations can be obtained, thereby avoiding confounding effects from a heterogeneous mixture of cells (Table 26.5). A more complete survey of network inference strategies can be found in the recent review by Fiers *et al.* (2018).

Chan *et al.* (2017) present a method for network inference based on mutual information estimates. For each target gene, the proportion of unique contribution, defined as the unique information divided by the mutual information for every other pair of genes, is computed and this quantity is used to determine which edges are kept. An alternative approach is provided by BoolTraineR (BTR; Lim *et al.*, 2016) which is based on Boolean models of regulatory networks

Table 26.5 Network inference models

Method	Time series	Approach	Implementation
PIDC	No	Information theory	Julia
SCODE	Yes	ODE model	R
SINCERITIES	Yes	Ridge regression	R, Matlab
BTR	No	Boolean networks	R

(Ribeiro *et al.*, 2006). BTR assumes an asynchronous update model and uses a scoring function based on the Bayesian information criterion (Schwarz, 1978) to identify a sparse set of update rules that is consistent with the observed expression patterns.

By contrast, SCODE (Matsumoto *et al.*, 2017) is a method for inferring networks from time series data, and it is based on the assumption that the evolution of the gene expression vector, X , can be described by a first-order linear ordinary differential equation (ODE) $dX = AXdt$, where A is a square matrix which encapsulates the regulatory interactions. SCODE uses a linear regression approach combined with dimensionality reduction for the inference. Similarly, SINGLE CELL Regularized Inference using TIme-stamped Expression profileS (SINCERITIES; Papili Gao *et al.*, 2017) also uses a regularized linear regression approach to infer directed regulatory relationships from time series of scRNA-seq data. Rather than using the gene expression values directly when carrying out ridge regression, SINCERITIES uses the Kolmogorov–Smirnov distance between distributions to assign weights to edges. Directions are inferred based on partial Spearman correlations via the Granger causality framework.

Recently, several authors have successfully combined CRISPR-Cas9 knockout screens with scRNA-seq (Datlinger *et al.*, 2017; Adamson *et al.*, 2016; Dixit *et al.*, 2016; Jaitin *et al.*, 2016). This strategy utilizes a library of guide RNAs targeting specific genes, and, thanks to barcodes, it is possible to infer what guide was associated with each cell. Thus, one obtains single-cell readouts of the effect of individual knock-outs. These technologies provide powerful means for probing networks and are likely to be highly informative in our endeavours to better understand gene regulation.

26.3.2.7 Combining Data Sets

As the number of publicly available scRNA-seq data sets is steadily increasing, being able to combine or compare them is becoming increasingly important. As discussed previously, in the case of two data sets coming from different laboratories, batch effects are a major issue and overcoming the technical noise remains a considerable challenge. In the summer of 2017, several methods addressing the difficulties of combining and comparing data sets were presented.

The first method is called MetaNeighbour (Crow *et al.*, 2018) and is based on building a cell–cell Spearman correlation network (which can include multiple data sets). MetaNeighbour then checks whether the cell type labels are consistent across the data sets. By holding out the cell labels of one data set at a time, MetaNeighbour predicts the cell labels of the held-out data set through weighted votes of its nearest neighbours in the cell–cell network. By calculating the area under the receiver operating characteristic curve for the label prediction task, MetaNeighbour can determine the consistency of the labels.

The second method is called mnnCorrect, implemented as a part of the scran Bioconductor R package (Haghverdi *et al.*, 2018). mnnCorrect is a pure batch correction method, but instead of using linear regression, it utilizes the concept of mutual nearest neighbours (MNNs), allowing for unbiased joint analysis of data sets coming from completely different laboratories. MNNs are individual cells that match across experiments, that is, they are mutual k -nearest neighbours

to each other. These cells represent overlapping structure between the experiments, which is then used to learn (using SVD) which dimensions of cell expression correspond to the biological state and which dimensions correspond to batch/experiment effect. mnnCorrect assumes that the batch/experiment effect is orthogonal to the biological subspace, and that batch effects variation is much smaller than the biological effects variation between different cell types. Finally mnnCorrect removes the batch/experiment effects from all experiments and return the corrected expression matrices.

The third method is based on canonical correlation analysis (CCA) and it is implemented as part of the Seurat R package (Butler and Satija, 2018). CCA identifies shared correlation structures across data sets and Seurat then uses this information to quantify how well the shared structures explain each cell's expression profile. By comparing to a normal PCA, cells where the explained variance is reduced in CCA are identified. It is assumed that these cells are subject to variability that is not defined by shared sources of variability and therefore these cells are removed from further analysis. Finally, the data sets are aligned using a 'warping' algorithm resulting in a single integrated low-dimensional space. Importantly, this space is not a corrected expression matrix itself, so further analysis cannot provide specific genes of interest.

Finally, scmap (Kiselev and Hemberg, 2018) allows the user to project cells from a query data set to either cell types or cells of a reference data set. In the former case, scmap represents each cluster (known *a priori*, e.g. a cell type) of the reference data set by its centroid, and measures the similarity between each cell of the query data set and each centroid. This is done by calculating consensus similarity of three different distance measures (Pearson and Spearman correlations and cosine similarity). A notable feature of scmap cluster projection is that it is very fast since the number of clusters in the reference is typically much smaller than the number of cells. To speed up the search when performing cell-to-cell projection, scmap carries out an approximate nearest-neighbour search using a product quantizer (Jégou *et al.*, 2011).

References

- Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., Pak, R.A., Gray, A.N., Gross, C.A., Dixit, A., Parnas, O., Regev, A. and Weissman, J.S. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**(7), 1867–1882.e21.
- Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z.K., Wouters, J. and Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14**(11), 1083–1086.
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**(10), R106.
- Andrews, T.S. and Hemberg, M. (2016). Modelling dropouts for feature selection in scRNASeq experiments. Preprint, bioRxiv 065094.
- Archer, N., Walsh, M.D., Shahrezaei, V. and Hebenstreit, D. (2016). Modeling enzyme processivity reveals that RNA-seq libraries are biased in characteristic and correctable ways. *Cell Systems* **3**(5), 467–479.e12.
- Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M. and Kendziora, C. (2017). SCnorm: Robust normalization of single-cell RNA-seq data. *Nature Methods* **14**(6), 584–586.
- Becskei, A., Kaufmann, B.B. and van Oudenaarden, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genetics* **37**(9), 937–944.

- Bendall, S.C., Davis, K.L., Amir, E.-A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P. and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**(3), 714–725.
- Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008.
- Bray, N.L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**(5), 525–527.
- Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C. and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**(11), 1093–1095.
- Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C. and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**(2), 155–160.
- Butler, A. and Satija, R. (2018). Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *Nature Biotechnology* **36**(5), 411–420.
- Buttner, M., Miao, Z., Wolf, A., Teichmann, S.A. and Theis, F.J. (2019). Assessment of batch-correction methods for scRNA-seq data with a new test metric. *Nature Methods* **16**(1), 43–49.
- Cannoodt, R., Saelens, W. and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology* **46**(11), 2496–2506.
- Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., Adey, A., Waterston, R.H., Trapnell, C. and Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**(6352), 661–667.
- Chan, T.E., Stumpf, M.P.H. and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems* **5**(3), 251–267.e3.
- Cole, M.B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S. and Yosef, N. (2017). Performance assessment and selection of normalization procedures for single-cell RNA-seq. Preprint, bioRxiv 235382.
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczęśniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**(1), 13.
- Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017). RNA-seq differential expression analysis: An extended review and a software tool. *PloS One* **12**(12), e0190152.
- Crow, M., Paul, A., Ballouz, S., Huang, Z.J. and Gillis, J. (2018). Addressing the looming identity crisis in single cell RNA-seq. *Nature Communications* **9**(1), 884.
- Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D. and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* **14**(3), 297–301.
- Delmans, M. and Hemberg, M. (2016). Discrete distributional differential expression (D3E) – a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17**, 110.
- Deng, Q., Ramsköld, D., Reinarius, B. and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**(6167), 193–196.
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T.M., Lander, E.S., Weissman, J.S., Friedman, N. and Regev, A. (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**(7), 1853–1866.e17.
- Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science* **297**(5584), 1183–1186.

- External RNA Controls Consortium (2005). Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6**, 150.
- Fiers, M.W.E.J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z. and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics* **17**(4), 246–254.
- Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S. and Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278.
- Finotello, F. and Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Briefings in Functional Genomics* **14**(2), 130–142.
- Freytag, S., Lonnstedt, I., Ng, M. and Bahlo, M. (2018). Cluster headache: Comparing clustering tools for 10X single cell sequencing data. *F1000Research* **7**, 1297.
- Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C. and Shalek, A.K. (2017). Seq-Well: Portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods* **14**(4), 395–398.
- Grün, D., Kester, L. and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**(6), 637–640.
- Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. and Xu, Y. (2015). SINCERA: A pipeline for single-cell RNA-seq profiling analysis. *PLoS Computational Biology* **11**(11), e1004575.
- Haghverdi, L., Lun, A.T.L., Morgan, M.D. and Marioni, J.C. (2018). Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours. *Nature Biotechnology* **36**(5), 421–427.
- Haque, A., Engel, J., Teichmann, S.A. and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**(1), 75.
- Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J., Raj, A., Li, M. and Zhang, N.R. (2018). Gene expression recovery for single cell RNA sequencing. *Nature Methods* **15**(7), 539–542.
- Huang, Y. and Sanguinetti, G. (2017). BRIE: Transcriptome-wide splicing quantification in single cells. *Genome Biology* **18**(1), 123.
- Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C. and Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biology* **17**, 29.
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11**(2), 163–166.
- Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A. and Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**(7), 1883–1896.e15.
- Jégou, H., Douze, M. and Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(1), 117–128.
- Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research* **44**(13), e117.
- Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M. and Zhang, N.R. (2017). Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Research* **45**(19), 10978–10988.
- Jiang, L., Chen, H., Pinello, L. and Yuan, G.-C. (2016). GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biology* **17**(1), 144.

- Johnson, W.E., Li, C. and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1), 118–127.
- Kalisky, T., Oriel, S., Bar-Lev, T.H., Ben-Haim, N., Trink, A., Wineberg, Y., Kanter, I., Gilad, S. and Pyne, S. (2017). A brief review of single-cell transcriptomic technologies. *Briefings in Functional Genomics* **17**(1), 64–76.
- Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**(7), 740–742.
- Kim, J.K. and Marioni, J.C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology* **14**(1), R7.
- Kiselev, V.Y. and Hemberg, M. (2018). scmap – A tool for unsupervised projection of single cell RNA-seq data. *Nature Methods* **14**(5), 483–486.
- Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R. and Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods* **14**(5), 483–486.
- Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühl, M., Liu, P., Marioni, J.C. and Teichmann, S.A. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**(4), 471–485.
- Korthauer, K.D., Chu, L.-F., Newton, M.A., Li, Y., Thomson, J., Stewart, R. and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* **17**(1), 222.
- Kwak, I.-Y., Gong, W., Koyano-Nakagawa, N. and Garry, D. (2018). DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**(1), 220.
- Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791.
- Li, B. and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.
- Li, C.-L., Li, K.-C., Wu, D., Chen, Y., Luo, H., Zhao, J.-R., Wang, S.-S., Sun, M.-M., Lu, Y.-J., Zhong, Y.-Q., Hu, X.-Y., Hou, R., Zhou, B.-B., Bao, L., Xiao, H.-S. and Zhang, X. (2016). Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Research* **26**(8), 967.
- Li, W.V. and Li, J.J. (2018). scImpute: Accurate and robust imputation for single cell RNA-Seq data. *Nature Communications* **9**(1), 997.
- Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics* **10**, 387–406.
- Lim, C.Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J. and Göttgens, B. (2016). BTR: Training asynchronous Boolean models using single-cell expression data. *BMC Bioinformatics* **17**(1), 355.
- Lin, C., Jain, S., Kim, H. and Bar-Joseph, Z. (2017). Using neural networks for reducing the dimensions of single-cell RNA-seq data. *Nucleic Acids Research* **45**(17), e156.
- Lun, A.T.L., Bach, K. and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17**, 75.
- Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A. and McCarroll, S.A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**(5), 1202–1214.
- Magwene, P.M., Lizardi, P. and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19**(7), 842–850.

- Mao, Q., Wang, L., Tsang, I. and Sun, Y. (2016). Principal graph and structure learning based on reversed graph embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(11), 2227–2241.
- Marco, E., Karp, R.L., Guo, G., Robson, P., Hart, A.H., Trippa, L. and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America* **111**(52), E5643–5650.
- Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T. and Nikaido, I. (2017). SCODE: An efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* **33**(15), 2314–2321.
- McCarthy, D.J., Campbell, K.R., Lun, A.T.L. and Wills, Q.F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**(8), 1179–1186.
- McIntyre, L.M., Lopiano, K.K., Morse, A.M., Amin, V., Oberg, A.L., Young, L.J. and Nuzhdin, S.V. (2011). RNA-seq: Technical variability and sampling. *BMC Genomics* **12**, 293.
- Menon, V. (2017). Clustering single cells: A review of approaches on high-and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics* **17**(4), 240–245.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**(7), 621–628.
- Ntranos, V., Kamath, G.M., Zhang, J.M., Pachter, L. and Tse, D.N. (2016). Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts. *Genome Biology* **17**(1), 112.
- Ofengheim, D., Giagtzoglou, N., Huh, D., Zou, C. and Yuan, J. (2017). Single-cell RNA sequencing: Unraveling the brain one cell at a time. *Trends in Molecular Medicine* **23**(6), 563–576.
- Papalexis, E. and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18**(1), 35–45.
- Papili Gao, N., Ud-Dean, S.M.M., Gandrillon, O. and Gunawan, R. (2017). SINCRETIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**(2), 258–266.
- Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**(4), 417–419.
- Petropoulos, S., Edsgård, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R. and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**(4), 1012–1026.
- Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D., Samsonova, M.G. and Kharchenko, P.V. (2018). Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biology* **19**(1), 78.
- Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S. and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9**(1), 171–181.
- Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16**, 241.
- Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., Ramalingam, N., Sun, G., Thu, M., Norris, M., Lebofsky, R., Toppani, D., Kemp, D.W., Wong, M., Clerkson, B., Jones, B.N. and West, J.A.A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology* **32**(10), 1053–1058.
- Prakhakaran, S., Azizi, E., Carr, A and Pe'er, D. ((2016)). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *Proceedings of Machine Learning Research* **48**, 1070–1079.

- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A. and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**(10), 979–982.
- Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341.
- Ribeiro, A., Zhu, R. and Kauffman, S.A. (2006). A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology* **13**(9), 1630–1639.
- Risso, D., Ngai, J., Speed, T.P. and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* **32**(9), 896–902.
- Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. and Vert, J.-P. (2018). ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9**(1), 284.
- Robinson, M.D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**(3), R25.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140.
- Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**(5), 495–502.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* **6**(2), 461–464.
- Sengupta, D., Rayan, N.A., Lim, M., Lim, B. and Prabhakar, S. (2016). Fast, scalable and accurate differential expression analysis for single cells. Preprint, bioRxiv 049734.
- Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G. and Song, H. (2015). Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17**(3), 360–372.
- Singer, A., Erban, R., Kevrekidis, I.G. and Coifman, R.R. (2009). Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* **106**(38), 16090–16095.
- Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S. and Sengupta, D. (2018). dropClust: Efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Research* **46**(6), e36.
- Smith, T., Heger, A. and Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Research* **27**(3), 491–499.
- Soneson, C. and Robinson, M.D. (2018). Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data. *Nature Methods* **15**(4), 255–261.
- Song, Y., Botvinnik, O.B., Lovci, M.T., Kakaradov, B., Liu, P., Xu, J.L. and Yeo, G.W. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Molecular Cell* **67**(1), 148–161.e5.
- Stubbington, M.J.T., Rozenblatt-Rosen, O., Regev, A. and Teichmann, S.A. (2017). Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**(6359), 58–63.
- Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A. and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Methods* **14**(4), 381–387.
- Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* **13**(4), 599–604.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K. and Surani, M.A. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**(5), 377–382.
- Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G., Neftel, C., Desai, N., Nyman, J., Izar, B., Luo, C.C., Francis, J.M.,

- Patel, A.A., Onozato, M.L., Riggi, N., Livak, K.J. and Suvà, M.L. (2016). Single-cell RNA-seq supports a developmental hierarchy in human oligodendrogloma. *Nature* **539**(7628), 309–313.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**(4), 381–386.
- Tung, P.-Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K. and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* **7**, 39921.
- Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P.V., Linnarsson, S. and Ernfors, P. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience* **18**(1), 145–153.
- Vallejos, C.A., Marioni, J.C. and Richardson, S. (2015). BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Computational Biology* **11**(6), e1004333.
- van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605.
- van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A.J., Moon, K.R., Mazutis, L., Wolf, G., Krishnaswamy, S. and Pe'er, D. (2018). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *Cell* **174**(3), 716–729.
- Vu, T.N., Wills, Q.F., Kalari, K.R., Niu, N., Wang, L., Rantalainen, M. and Pawitan, Y. (2016). Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **32**(14), 2128–2135.
- Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods* **14**(4), 414–416.
- Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M. and Zhang, N.R. (2018). Gene expression distribution deconvolution in single cell RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **115**(28), 6437–6446.
- Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63.
- Welch, J.D., Hartemink, A.J. and Prins, J.F. (2016). SLICER: Inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology* **17**(1), 106.
- Welch, J.D., Hu, Y. and Prins, J.F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research* **44**(8), e73.
- Westoby, J., Sjoberg, M., Ferguson-Smith, A. and Hemberg, M. (2018). Simulation based benchmarking of isoform quantification in single-cell RNA-seq. *Genome Biology* **19**(1), 191.
- Wu, A.R., Wang, J., Streets, A.M. and Huang, Y. (2017). Single-cell transcriptional analysis. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)* **10**(1), 439–462.
- Wu, Z., Liu, W., Ji, H., Yu, D., Wang, H., Liu, L., Ji, S. and Shan, G. (2018). NormExpression: An R package to normalize gene expression data using evaluated methods. Preprint, bioRxiv 251140.
- Ye, C., Speed, T.P. and Salim, A. (2017). DECENT: Differential Expression with Capture Efficiency AdjustmeNT for single-cell RNA-seq data. Preprint, bioRxiv 225177.
- Yip, S.H., Wang, P., Kocher, J.-P.A., Sham, P.C. and Wang, J. (2017). Linnorm: Improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research* **45**(22), e179.
- Zhang, L. and Zhang, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
- Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G.,

- Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R. and Bielas, J.H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049.
- Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. and Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell* **65**(4), 631–643.e4.

Variant Interpretation and Genomic Medicine

K. Carss,¹ D. Goldstein,² V. Aggarwal,² and S. Petrovski¹

¹Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

²Institute for Genomic Medicine, Columbia University Medical Center, New York, USA

Abstract

Despite advances in sequencing technologies and reductions in sequencing costs, linking an individual's genetic variation to their risk of developing disease remains an ongoing clinical medicine challenge. Genetic variants found in an individual are commonly divided into five categories: pathogenic, likely pathogenic, uncertain significance, likely benign, and benign. These classifications aim to reflect the level of confidence supporting an individual variant causing a phenotype in an individual. In the era of precision medicine, improving these classifications is of profound importance for clinical management.

27.1 Introduction and Current Challenges

In this chapter we aim to give the reader a broad overview of the current challenges faced when interpreting individual genomes, the diverse types of evidence that often influence variant classification, some of the tools and resources commonly adopted when interpreting variant pathogenicity, and some common pitfalls to avoid. We provide links to tools for those readers interested in expanding their variant interpretation toolkit, and citations for readers interested in a deeper dive into specific variant interpretation issues. Several reviews and recommendations on aspects of variant interpretation and genomic medicine have been published in recent years (D.G. MacArthur *et al.*, 2014; Eilbeck *et al.*, 2017; Hoskinson *et al.*, 2017; Richards *et al.*, 2015; Guillermin *et al.*, 2018; Goldstein *et al.*, 2013). Given the brisk pace of change in human genomics, including the increasing scale of human reference cohorts and the rapidly falling cost of genome sequencing, it is important to summarise and update the state of the art in variant interpretation.

Genomic medicine is broadly characterised by using an individual's genomic data to make clinical management decisions. These decisions ultimately depend on a correct molecular diagnosis, which is in turn dependent upon robust interpretation of the molecular data, including genomic variants. Although genomics-driven precision medicine is not novel, several key factors have aligned to enable recent successful applications: we have witnessed an accelerating pace of genomic technology development, wider access to genomic technologies, the generation of large collections of human variation data, the development of advanced analytical tools

to handle the complex nature of genomics data, increased therapeutic opportunities emerging from gene editing technologies, and rapidly growing knowledge of human disease and disease-associated genes and variants.

Notwithstanding these advances, interpreting individual genomes remains challenging. In fact, one could argue that the rapid pace and increased access to genomic technology have in some cases negatively affected the literature by permitting variants to be recorded as causally associated, and thus to be annotated as causative in the various gene and variant databases used in routine clinical evaluations without sufficient evidence. Some aspects of variant interpretation have remained subjective, resulting in inconsistencies in interpretation and leading to some incorrect claims of causality. Additionally, consensus standards and robust resources to support the interpretation of individual genomes do not appear to be accelerating at the same pace as technological advances and genomic data generation.

Interpreting an individual's genome involves both understanding which genes are clinically important, and which specific variants in that person influence the observed or latent clinical phenotype(s). In the interpretation of a patient genome, the first focus is on identifying the clinically responsible gene(s) and what is known about the phenotypes they influence and how closely they align to the patient's phenotype(s). There are well-established standards for assessing whether a gene influences a human trait such as a disease, such as those used by ClinGen to curate gene–disease associations (Rehm *et al.*, 2015). Although these standards are not necessarily always strictly followed, false positive claims at the level of the gene are less frequent and often easier to correct than false positive claims at the variant level.

When the genes that could be responsible for an individual's clinical phenotype are well known, the challenge primarily lies in evaluating whether a specific candidate variant in those genes contributes to the phenotype observed in that individual. There is an acute need for this in the genomics community; however, there currently is no unifying statistical framework to quantitatively assess our confidence that a specific variant contributes to disease in an individual patient. This challenge is inflated in the absence of full clinical phenotype at assessment. This raises major challenges, given that the interpretation of patient genomes is already influencing management in some clinical settings.

Early approaches to gene sequencing for the diagnosis of disease were limited to panels of genes with a high suspicion for their potential role in the patient's phenotype. Given the prior disease association for the gene, there was an expectation that any sufficiently rare variants in such genes were likely to be pathogenic. However, this approach suffered due to its lack of appreciation for the scale of human variation, resulting in some genes and variants being falsely implicated in disease (Rehm *et al.*, 2015; Shah *et al.*, 2018; Piton *et al.*, 2013). This is now becoming considerably better appreciated through the generation of large human population sequence reference cohorts, which allow quantification of the search space; for example, one can calculate the probability of finding a variant of similar allele frequency and similar functional prediction to the case-ascertained variant, even if not exactly the same variant. The community is also making progress towards defining the necessary standards on the type, quantity and quality of evidence required to securely implicate a variant as pathogenic for a disease, although there are still opportunities for improvement.

Just as biological understanding is required for appropriate interpretation of variants, particularly in genes not previously associated with disease, an appreciation of statistical confidence is important when interpreting the output of analytical frameworks built to handle data generated by high-throughput sequencing. Thus, to mitigate the contamination of literature records, awareness of statistical as well as biological concepts should be the foundations underpinning evidence used for robust variant interpretation.

Another challenge for the goal of developing a unified statistical framework for variant interpretation is that variability in genetic architecture underpinning different diseases introduces an additional layer of complexity, especially for diseases for which the underlying genetic architecture is poorly understood. The appropriate approach to identifying genes and variants that are associated with human disease depends in part on locus and allelic heterogeneity, but also on the frequency and effect size of the risk variants. Mendelian diseases such as epileptic encephalopathy are rare and devastating, and often caused by *de novo* high-penetrance variants (i.e. a germline mutation not observed in biological parents). Due to their severity and the fact that the clinical consequences severely affect reproductive fitness, *de novo* mutations have a major contribution to the genetic architecture of those Mendelian disorders (Epi4K Consortium *et al.*, 2013; Deciphering Developmental Disorders Study, 2017; de Ligt *et al.*, 2012; Rauch *et al.*, 2012). In these clinical presentations, the knowledge of a patient's variant having arisen *de novo* is often considered among the strongest evidence for pathogenicity. However, while we do not argue that this is generally appropriate weighting, we do want to add the cautionary note that on average all humans are expected to carry approximately one protein-coding *de novo* mutation, and whether that specific variant contributes to disease or not remains a separate line of questioning (Francioli *et al.*, 2015; Rahbari *et al.*, 2015). On the other hand, common diseases such as inflammatory bowel disease tend to have a lesser effect on reproductive fitness. Such common disorders can have multi-factorial causes, including multiple, common, low-effect-size genetic variants combined with environmental and sometimes epigenetic factors (Figure 27.1). Robust statistical methods leveraging multiple information sources must be applied to the process of variant interpretation at every point along the frequency spectrum, although one would suspect that the weight of evidence applied to individual considerations might change along the spectrum.

Historical approaches to identifying Mendelian disease-associated genes relied on methods to narrow the probable cause to a refined set of genes. This came through either linkage analysis where genomic markers could define a limited region of the genome that harboured a causative signal, or the functional disruption of metabolic pathways or other biological processes that

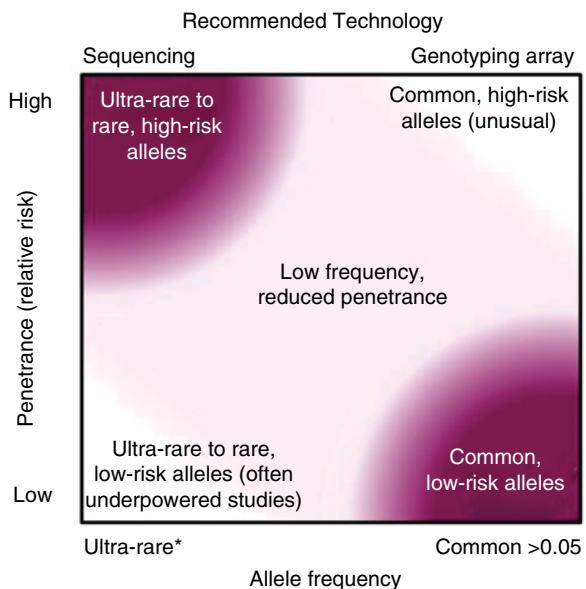


Figure 27.1 The majority of known clinically relevant variants reside closer to the top left-hand corner representing low frequency and higher penetrance, such as Mendelian familial and early-onset disorders like intellectual disabilities caused by *de novo* mutations. At the other end of the spectrum reside variants with increased allele frequency but limited to low effect size, such as those associated with inflammatory bowel disease. Variants associated with age-related macular degeneration and Alzheimer's disease are just a couple of the uncommon genotype–phenotype examples approaching the top right-hand quadrant. *Ultra-rare defined as a variant found in an index sample that is not observed across available population reference cohorts.

could point genomic analysis to a faulty enzyme or other element of a clearly disrupted biological process. This refined set of genes would be Sanger sequenced, and researchers would seek a biologically plausible gene that harboured variant(s) that either segregate within affected pedigrees, or occur in affected members of independent families with the same clinical diagnosis. More recently, with high-throughput sequencing of the exome or genome we get a candidate-free reflection of genetic variation across the adequately sequenced region of an individual.

The increasing awareness of high allelic heterogeneity contributing to the genetic architecture of many diseases has motivated a move from the standard variant-level association statistics to the successful application of locus-based rare-variant collapsing and burden association statistics (Allen *et al.*, 2017; Cirulli *et al.*, 2015; Petrovski *et al.*, 2017). The key assumption that these locus-based analytical approaches make is that there are multiple variants that can disrupt a gene where the end biological result is often similar and thus reflects a similar clinical presentation. Therefore, locus-based analysis frameworks focus on estimating the rates that a predefined class of variation (qualifying variant) occurs in a case collection in comparison to control individuals. These analytical frameworks also provide an opportunity to define the classes of variants where the case signal is enriched, and permit an objective understanding of the genetic architecture contributing to increased disease risk (Petrovski *et al.*, 2017; Allen *et al.*, 2017). These approaches are also valuable for diseases that do not entirely fall into the two extremes of the frequency spectrum, such as familial hypercholesterolaemia, which is a common Mendelian disorder characterised by highly penetrant causal alleles, but also with evidence of both reduced-penetrance variant contribution and polygenic inheritance (Defesche *et al.*, 2017; Futema *et al.*, 2014).

At the common end of the variant frequency spectrum are complex diseases such as inflammatory bowel disease. These associated variants have mostly been identified by microarray-based genome-wide association studies (GWASs) (MacArthur *et al.*, 2017). These experiments are performed on genotyping arrays rather than sequencing so that much larger cohorts can be interrogated in cost-efficient manner. Access to cohort-scale genomics data required development of robust statistical methods for correct interpretation of GWAS results. It is increasingly recognised that rare variants of large effect sizes also have a role in common disease (Zuk *et al.*, 2014; Cirulli and Goldstein, 2010; Cirulli *et al.*, 2015; Allen *et al.*, 2017; Raghavan *et al.*, 2018). To detect these, high-throughput sequencing of large cohorts is required. There were some high-profile retractions in the candidate gene and very early GWAS literature, largely caused by a lack of understanding at the time of the biological and technical sources of confounding, and the lack of a robust statistical framework proposed and adopted community-wide, including a strong emphasis on appropriate multiplicity adjustment through a genome-wide significance threshold (Lambert and Black, 2012). Such incidents have become much less frequent since the publication and widespread adoption of guidelines that address these issues (McCarthy *et al.*, 2008; Pe'er *et al.*, 2008).

In this chapter we will dive deeper into the multiple lines of evidence facilitating robust variant interpretation, all of which must be anchored in sound biological and statistical understanding. These are relevant to every type of variant, and across the allele frequency spectrum; however, in this chapter we primarily focus on the interpretation of highly penetrant, rare, germline protein-coding variants. This chapter is structured as follows. Section 27.2 explores the importance of understanding variant consequence, and introduces various *in silico* prediction tools commonly applied to estimate how likely a specific variant is to disrupt the normal biological function. In Section 27.3 we discuss the value of understanding genomic variation context by leveraging standing variation data from large and diverse human population samples. Section 27.4 discusses how appropriate experimental design around functional assays of genetic variation can assist in defining genotypes with a greater

disease potential, and Section 27.5 shows the value of utilising existing literature knowledge about gene function, including disease-association information from both human and model organism data to further guide variant interpretation. Section 27.6 brings together the multiple sources of data, and illustrates how a holistic approach to variant interpretation can increase confidence in variant classifications. Finally, Section 27.7 discusses some of the outstanding challenges around variant interpretation.

27.2 Understanding the Effect of a Variant

The human genome is a tapestry of different classes of variation, including single nucleotide variants (SNVs) and indels, missense and protein-truncating variants, sequence and structural variants (SVs), coding and non-coding variants, and germline and somatic variants. These different classes present different challenges and often require different bioinformatic, statistical and biological considerations during interpretation. Here we describe why predicting the likely effect(s) of a variant is important. A prerequisite to this is that one must have confidence that the variant is not a technical error. Then one must have confidence that the variant has a functional effect on the gene product. Finally, one must have confidence that such a perturbation of the gene product is sufficient to cause a clinical phenotype.

The initial hurdle is being confident that the variant represents a biological variant, rather than a false positive technical artefact. Variant callers ascribe many confidence metrics to each variant, including read depth, strand bias, mapping quality, genotype quality, to name a few. There are standard protocols for leveraging these metrics to focus on high-confidence variant calls (e.g. GATK best practices; Van der Auwera *et al.*, 2013). This process is often automated, more so than the later stages of variant interpretation. Nevertheless, there are five important pitfalls to be avoided.

1. Variant-level quality control (QC) must be combined with appropriate sample-level QC to identify samples that are likely to have disproportionately high frequency of false positive calls. Contamination, heterozygous–homozygous SNV ratio, transition–transversion ratio, and uniformity of coverage can all be used to inform this decision (Lek *et al.*, 2016).
2. It is valuable, if possible, to consider the variant not in isolation but in the context of the whole cohort distribution (Poplin *et al.*, 2017). A variant/position that fails standard QC confidence metrics in most individuals where it is observed should be flagged as cautionary when it comes to interpretation.
3. The context of the sequence region in which the variant occurs should be considered. There are certain genomic ‘blind spots’, including centromeres, telomeres, GC-rich regions, repetitive regions, and homopolymer stretches, which tend to be poorly represented by high-throughput sequencing. For GC-rich regions, whole-genome sequencing (WGS) generally achieves better coverage than whole-exome sequencing (WES), because library preparation does not require polymerase chain reaction (PCR) amplification (Carss *et al.*, 2017). Also, one must consider absence of variation in a region of interest as either sufficient coverage yet no variant identified or no information in that region of interest due to insufficient read-depth coverage. The nature of the sequence context in simple repeat and similar regions means that they are also more prone to false positive events (Huang *et al.*, 2015). This can be clinically important. For example, variants in *RPGR* can cause X-linked cone-rod dystrophy, and it includes an exon known as *ORF15* that is a mutational hotspot containing pathogenic variants, but it is highly repetitive and generally not well represented by either WES or WGS (Vervoort *et al.*, 2000; Carss *et al.*, 2017).

4. *De novo* mutations are an important class of variation (Deciphering Developmental Disorders Study, 2017), and an estimated 6.5% of *de novo* variants detected in conventional germline variant screens turn out to be mosaic mutation events, that is, to have occurred during early development and thus not to be present in all cells (Acuna-Hidalgo *et al.*, 2015). Such events may be missed if strict allele ratio filters are applied; however, the allele ratio filter is also a useful metric to screen out technical artefacts in sequence data. Fortunately, additional QC metrics such as mapping quality, read-position rank sum, Fisher's strand bias and others can further assist in discriminating putative somatic from likely technical artefacts.
5. Due to regions with high homology and low complexity, there are increased opportunities for misalignment, thus making indel calling marginally less reliable than SNV calling. Currently, however, the experienced human eye can often distinguish edge case real variant calls from false positive calls better than the collection of QC metrics, therefore visualising the region of a variant in a genome browser such as IGV is recommended practice for variants of interest (Thorvaldsdóttir *et al.*, 2013). With the accelerating adoption of artificial intelligence in imaging signal detection, we see an opportunity to improve variant calling accuracy by training on images of the regional alignment around a variant call in combination with conventional variant-calling QC metrics.

The next step on the path towards confidence in the causality between a variant and a disease is amassing sufficient evidence that the variant damages the gene product. The distinction between high-impact 'protein-truncating variants' (PTVs) – which typically includes nonsense SNVs, frameshifting indels, and canonical splice acceptor or donor variants – and moderate-impact missense variants or in-frame indels is one of the most important in variant interpretation. PTVs are often assumed to negate protein function by producing null alleles and therefore as a class of variant are considered damaging, as also reflected in the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) scoring system (Richards *et al.*, 2015). However, there are exceptions to this general rule, so we encourage researchers to have an appreciation of the typical patterns and rates of an observed PTV in a specific gene context when interpreting their findings. The term 'loss-of-function' variant is commonly used to describe PTVs, but this can be misleading because of these exceptions. Therefore, we will use 'PTV' throughout this chapter, except when referring to variants for which loss of function has been functionally confirmed.

First, a PTV, particularly one located close to the 3' end of a gene, might result in a protein that is truncated but still expressed and escapes nonsense-mediated decay either fully or partially (MacArthur *et al.*, 2012). It may retain its normal function and therefore be benign, or counter-intuitively it may have a gain of protein function or toxic effect on the protein, as seen, for example, in floating harbour syndrome and *WASF1*-associated intellectual disability with seizures (Hood *et al.*, 2012; Ito *et al.*, 2018; Hong *et al.*, 2004). Next, there are disease-associated genes for which the mechanism is gain-of-function rather than loss-of-function. In these genes, pathogenic mutations tend to more often be missense changes, and can cluster within regions of a gene. In genes that cause disease through gain of function, PTVs might be benign and occur in the general population, given that the disease mechanism is not haploinsufficiency, for example *KCNT1*-associated epileptic encephalopathy (Milligan *et al.*, 2014). Less frequently, a PTV close to the 5' end of the gene might be rescued by an alternative start site (MacArthur *et al.*, 2012). Finally, PTVs in an exon that is not present in all transcripts of a disease-associated gene might be tolerated if the clinically relevant transcripts are unaffected. Enrichment of PTVs in control populations among exons not present in all transcripts of disease-associated genes further supports this (Lek *et al.*, 2016).

Splice variants are a challenging class of PTVs, as they do not always render the protein non-functional. The general assumption is that a change within the canonical splice site (2 base pair intronic sequence) causes some form of protein truncation. The real picture is often more complex. For example, deep intronic variants can introduce cryptic splice sites (Cummings *et al.*, 2017; Vaz-Drago *et al.*, 2017; Carss *et al.*, 2017), and even some canonical splice site variants do not alter the amount of full-length transcript (Rivas *et al.*, 2015). Thus, the definitive effect of splice-site variants on the protein is often hard to predict. Some *in silico* splice prediction tools can be helpful, such as NNSPLICE, HSF, MaxEntScan, and TraP (Reese *et al.*, 1997; Desmet *et al.*, 2009; Yeo and Burge, 2004; Gelfman *et al.*, 2017). Generally, splice acceptor variants result in exon skipping, which may or may not cause a frameshift. The effect of splice donor variants is less predictable and can result in intron retention, exon skipping, or effects that are harder to predict.

The effects of missense variants are harder to predict and thus interpret than PTVs because their effect on the protein product can range from benign to catastrophic. Commonly used *in silico* tools and scores for variant interpretation are summarised in Table 27.1. Such scores are generally based on evolutionary conservation, with or without additional sources of information such as biochemical properties of the amino acids, and are used to predict whether a SNV and/or indel is likely to have a detrimental effect on protein function. Machine learning approaches have been adopted to integrate multiple information sources into a single ensemble score. These generally improve sensitivity and specificity over scores that use only a single information source, but because they leverage similar information, composite scores have moderate to high correlation, whereas the correlation between classical scores tends to range between low and high (Traynelis *et al.*, 2017; Chun and Fay, 2009).

Some of the more commonly adopted scores include: PolyPhen-2, SIFT, LRT, PhyloP, PhastCons, GERP++, Condel, MutationTaster, eXtasy, CADD, REVEL, VEST, and DANN (Adzhubei *et al.*, 2013; Vasser *et al.*, 2016; Chun and Fay, 2009; Cooper *et al.*, 2005; Siepel *et al.*, 2005; Davydov *et al.*, 2010; González-Pérez and López-Bigas, 2011; Schwarz *et al.*, 2010; Sifrim *et al.*, 2013; Kircher *et al.*, 2014; Ioannidis *et al.*, 2016; Carter *et al.*, 2013; Quang *et al.*, 2015). Also available are *in silico* models of the predicted effect of a variant on the three-dimensional structure of the protein, including cBioPortal and MuPIT, and structural biology-based scores such as mCSM, SDM, and DUET (Cerami *et al.*, 2012; Niknafs *et al.*, 2013; Worth *et al.*, 2011; Pires *et al.*, 2014a,b). These scores and others have been reviewed and discussed in various other publications over recent years (Wu and Jiang, 2013; Niroula and Viñinen, 2016).

There are several limitations to bear in mind when interpreting *in silico* missense scores. One is that some of the prediction scores are constructed to preferentially predict the possible loss of function or hypomorphic effect of missense variants (Flanagan *et al.*, 2010). Another is that high probability of damage to the protein does not necessarily equate to high probability of pathogenicity. Therefore, users must resist the temptation to overinterpret a novel damaging missense variant based on *in silico* predictors alone. Another important caveat when comparing scores is that they are not highly independent of each other. Some are nested within others, and many use the same or similar sources of biological information. This can lead to bias in assessing the tools, and problems due to hierarchical granularity (Popovic *et al.*, 2015). The large number of different scores and the fact that often multiple different scores are required to interpret a variant have led to a need for a compiled database that can annotate variants with many of the scores at once, such as dbNSFP (Liu *et al.*, 2016).

One of the major limitations when adopting *in silico* score as evidence for causality has been the frequent lack of understanding of the gene-specific empirical null distribution of these *in silico* scores. For example, a PolyPhen-2 score of 0.95 becomes less informative if almost all the population control missense variants in that gene also achieved a PolyPhen-2 score of

Table 27.1 Some commonly used software tools for variant interpretation

Class	Name	Full name	Website(s)	Reference
Splice prediction	NNSPLICE	NNSPLICE	http://www.fruitfly.org/seq.tools/splice.html	Reese <i>et al.</i> (1997)
	HSF	Human Splicing Finder	http://www.umd.be/HSF3/	Desmet <i>et al.</i> (2009)
	MaxEntScan	Maximum Entropy Scan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	Yeo and Burge (2004)
	TraP	Transcript-inferred Pathogenicity	http://trap-score.org/	Gelfman <i>et al.</i> (2017)
Function prediction scores	PolyPhen-2	Polymorphism Phenotyping v2	http://genetics.bwh.harvard.edu/pph2/	Adzhubei <i>et al.</i> (2013)
	SIFT	Sorting Intolerant From Tolerant	http://sift.jcvi.org/	Vaser <i>et al.</i> (2016)
	LRT	Likelihood Ratio Test	http://www.genetics.wustl.edu/jflab/lrt_query.html	Chun and Fay (2009)
Scores of interspecies conservation	PhyloP	PhyloP	http://compgen.cshl.edu/phast/background.php	Cooper <i>et al.</i> (2005)
	PhastCons	PhastCons	http://compgen.cshl.edu/phast/background.php	Siepel <i>et al.</i> (2005)
	GERP++	Genomic Evolutionary Rate Profiling	http://mendel.stanford.edu/SidowLab/downloads/gerp/	Davydov <i>et al.</i> (2010)
Ensemble scores	Condel	CONsensus DELetiousness score	http://bbglab.irbbarcelona.org/fannsdb/	González-Pérez and López-Bigas (2011)
	MutationTaster	MutationTaster	http://www.mutationtaster.org/	Schwarz <i>et al.</i> (2010)
	eXtasy	eXtasy	http://extasy.esat.kuleuven.be/	Sifrim <i>et al.</i> (2013)
	CADD	Combined Annotation Dependent Depletion	http://cadd.gs.washington.edu/home	Kircher <i>et al.</i> (2014)
	REVEL	Rare Exome Variant Ensemble Learner	https://sites.google.com/site/revelgenomics/	Ioannidis <i>et al.</i> (2016)
	VEST	Variant Effect Scoring Tool	http://karchinlab.org/apps/appVest.html	Carter <i>et al.</i> (2013)
	DANN	Deleterious Annotation of genetic variants using Neural Networks	https://cbcl.ics.uci.edu/public.data/DANN/	Quang <i>et al.</i> (2015)

Structural effect on 3D protein	cBioPortal MuPIT	cBioPortal Mutation Position Imaging Toolbox	http://www.cbioperl.org/ http://mupit.icmbio.org/	Cerami <i>et al.</i> (2012) Niknafs <i>et al.</i> (2013)
Structural biology based scores	mCSM SDM DUET	Mutation Cutoff Scanning Matrix Site Directed Mutator DUET	http://structure.bioc.cam.ac.uk/mcsm http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php http://structure.bioc.cam.ac.uk/duet	Pires <i>et al.</i> (2014b) Worth <i>et al.</i> (2011) Pires <i>et al.</i> (2014a)
Distinguishes low- and high-confidence protein truncating variants	LOFTEE	Loss-Of-Function Transcript Effect Estimator	https://github.com/konradjk/loftee	–
Genome browser	IGV	Integrative Genomics Viewer	http://software.broadinstitute.org/software/igv/	Thorvaldsdóttir <i>et al.</i> (2013)
Database to annotate variants with multiple scores	dbNSFP	Database of Non-Synonymous Functional Predictions	https://sites.google.com/site/jpopgen/dbNSFP	Liu <i>et al.</i> (2016)

at least 0.95 (Traynelis *et al.*, 2017). Variant interpretation would be more robust if the score's empirical null distribution were considered and described alongside use of a score, especially when used as evidence of the pathogenicity of a variant.

Another important question has been the effect of a variant on the transcriptome, proteome, metabolome, and other omics. Focusing on the transcriptome as this is where knowledge is currently greatest, a variant found in a transcript that is not highly expressed in the relevant tissue or at the relevant developmental stage is unlikely to be pathogenic, even if it is predicted to be damaging. Also, exonic variants can have different consequences on different transcripts that they affect. Typically, only one transcript is analysed because of logistical and interpretive challenges. There are three general approaches to selecting the transcript, each with advantages and disadvantages. The first and most common is to select the 'canonical' transcript, which is defined in databases as the longest coding transcript of a gene. Alternatively, researchers can select the transcript with the highest expression in the tissue(s) of interest. Finally, researchers can focus on the transcript that corresponds with the most damaging variant effect prediction. The first two approaches increase the probability that the annotated transcript affects a biologically relevant isoform; however, both approaches can miss a clinically relevant PTV in an alternative isoform (transcript) of the gene that does not qualify as longest or most expressed transcript. An important counter-argument to selecting the most damaging predicted variant effect is that a damaging variant in a non-canonical transcript, particularly if that transcript is not expressed, may have less clinical relevance, so although it might be the most damaging prediction it may be the least clinically meaningful. Various attempts have been made to address this challenge. One is to rely on the most damaging prediction, but limited to the consensus coding sequence set of transcripts or focused on a subset of transcripts that have been validated as clinically relevant (J.A.L. MacArthur *et al.*, 2014; DiStefano *et al.*, 2018). At present, however, these transcript-level resources do not represent all Mendelian disease-associated genes. Complementary tools such as LOFTEE are available to provide PTV predictions, giving helpful annotations specific for PTVs including the proportion of affected transcripts in which the predicted consequence is protein-truncating, whether the PTV affects an exon that is conserved, affects a non-canonical splice site, and other valuable annotations for interpreting the likely final effect of a PTV on the relevant transcript(s).

Although this chapter is primarily focused on germline variants, we will here briefly discuss some specific implications and limitations when performing somatic variant interpretation. Readers interested in more on this topic should refer to the recently published 'Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer' (Li *et al.*, 2017). The biological importance of distinguishing between germline and somatic variants is illustrated by the fact that some genes can be associated with very different phenotypes depending on whether the variants are germline or somatic (Petrovski *et al.*, 2016). For example, germline PTVs in *ASXL1* cause a severe neurodevelopmental disease, Bohring–Opitz syndrome (Hoischen *et al.*, 2011), whereas the same variants in the same gene can contribute to acute myeloid leukaemia when mitotically mutated (Gelsi-Boyer *et al.*, 2009) due to clonal haematopoiesis, whereby somatic mutations occur in blood cell progenitors and contribute to forming a genetically distinct blood cell subpopulation (Genovese *et al.*, 2014). Somatic variant detection and interpretation, which may be done with or without a matched normal sample, requires special considerations. Additional tools are often adopted to cope with issues such as desired detection of low alternate allele ratios, sub-clonal heterogeneity, and normal tissue contamination (Hansen *et al.*, 2013). When filtering on population allele frequency, somatic-specific databases should be used to give an indication of how mitotically mutable the variant site is, such as the Catalog of Somatic Mutations in Cancer (COSMIC; Forbes *et al.*, 2017). Unlike the well-established germline *de novo* mutation rate, the mitotic mutation rate is less

well understood and is dependent on many factors, collectively making it more difficult to formulate statistical tests for somatic mutation enrichment.

Evidence that a putative variant is real, damaging to the gene product, and affects a clinically relevant transcript are all key components to interpreting a candidate variant. There is a diverse collection of bioinformatic tools that capture information for each component, and these tools should be leveraged to improve variant pathogenicity classification. These include quality scores allocated by variant callers, *in silico* predictors of deleteriousness, databases of clinically relevant transcripts, splice prediction algorithms, and somatic variant-specific tools. The next critical consideration to be discussed is understanding the genomic variation context and the population frequency of the variant as another key source of evidence.

27.3 Understanding Genomic Variation Context through Large Human Reference Cohorts

Another source of evidence to support variant interpretation is accurate assessment of its frequency in the population. To achieve adequate resolution for rare variants, this requires access to large sequenced human cohorts. Moreover, understanding the distribution of functional variation within a gene is important. Prior to the availability of large-scale population reference cohorts, the absence of a candidate variant from a small control collection was commonly adopted to suggest variant pathogenicity, and in some cases new gene–disease associations. Some of these earlier reports have since been deemed unlikely to be pathogenic based on increased information (Shah *et al.*, 2018; Piton *et al.*, 2013; Rehm *et al.*, 2015).

Accessibility to high-throughput sequencing has enabled the collection of large sequenced cohorts (summarised in Table 27.2 and Figure 27.2). The scale of these population cohorts increased initially to thousands, such as the 1000 Genomes Project, NHLBI Exome Sequencing Project and UK10K (Walter *et al.*, 2015; Auton *et al.*, 2015). With reduced sequencing costs and increased data sharing policies, the cohorts have since increased in size to the hundreds of thousands. The larger cohorts include ExAC, gnomAD, Bravo TOPMed, DiscovEHR and HLI (Lek *et al.*, 2016; Telenti *et al.*, 2016). Looking into the future, the sizes of sequenced cohorts are set to increase further, with projects such as Genomics England, AstraZeneca's Genomics Initiative, UK Biobank, and the "All of Us" study of the US NIH Precision Medicine Initiative (Turnbull *et al.*, 2018; Collins and Varmus, 2015; Trehearn, 2016). In addition to the scale, the population diversity of the cohorts has also increased (Table 27.2), which is vital for variant interpretation, as we will discuss later in this section.

One of the most exciting opportunities arising from sequencing large and diverse human populations is that empirical sequence data from large samples of the human population have improved understanding of the 'typical' patterns of genetic variation for any given gene. Access to these empirical patterns provides extremely valuable population genomics information that has not been available historically due to lack of the source data. Knowing what constitutes a 'typical' pattern of variation within a gene makes it possible to identify 'atypical' genotypes. This has motivated the development of several metrics in recent years that solely use large-scale human sequence data collections to quantify how well a given genomic region tolerates functional variation (Petrovski *et al.*, 2013; Samocha *et al.*, 2014). These purely standing variation-based metrics are summarised in Table 27.3. Some quantify variation intolerance specifically to PTVs (e.g., LoF depletion, pLI and LoFtool: Samocha *et al.*, 2014; Lek *et al.*, 2016; Fadista *et al.*, 2017), within gene intolerance (e.g., sub-RVIS, missense tolerance ratio (MTR), regional missense z-score: Gussow *et al.*, 2016; Traynelis *et al.*, 2017; Samocha *et al.*, 2017), and even among the non-coding sequence of the human genome (ncRVIS, Orion and CDTs: Petrovski

Table 27.2 Available large human reference cohorts, in approximate chronological order (UK10K data is available under managed access agreements.)

Name	Lead institute and country	Size	Population diversity	Technology	Phenotypes	Website	Reference
1000 Genomes	Wellcome Genome Campus, UK	2,504	26 populations, 503/2504 (20%) European ancestry	WGS (7×), and WES (66×)	Healthy	http://www.internationalgenome.org/	Auton <i>et al.</i> (2015)
NHLBI Exome Sequencing Project	University of Washington, USA	6,503	2 populations, 4300/6503 (66%) European ancestry and 2203/6503 (34%) African-American	WES	Heart, lung and blood disorders	http://evs.gs.washington.edu/EVS/	–
UK10K	Wellcome Genome Campus, UK	8,963	3781/3781 (100%) healthy cohort European ancestry	WGS (7×), or WES (80×)	Healthy (<i>n</i> = 3781) or rare disorders (<i>n</i> = 5182)	https://www.uk10k.org/	Walter <i>et al.</i> (2015)
ExAC	BROAD, USA	60,706	7 populations, 33,370/60,706 (55%) European ancestry	WES	Excluded as far as possible patients with severe paediatric diseases. Some common disease cohorts (e.g. schizophrenia) included.	http://exac.broadinstitute.org/	Lek <i>et al.</i> (2016)
Bravo TOPMed	University of Michigan, USA	62,784	Various	WGS (38×)	As above	https://bravo.sph.umich.edu/freeze5/hg38/	Taliun <i>et al.</i> (2019)
gnomAD	BROAD, USA	141,456	8 populations, 64,603/141,456 (46%) European ancestry	125,748 WES and 15,708 WGS	As above	http://gnomad.broadinstitute.org/	Karczewski <i>et al.</i> (2019)

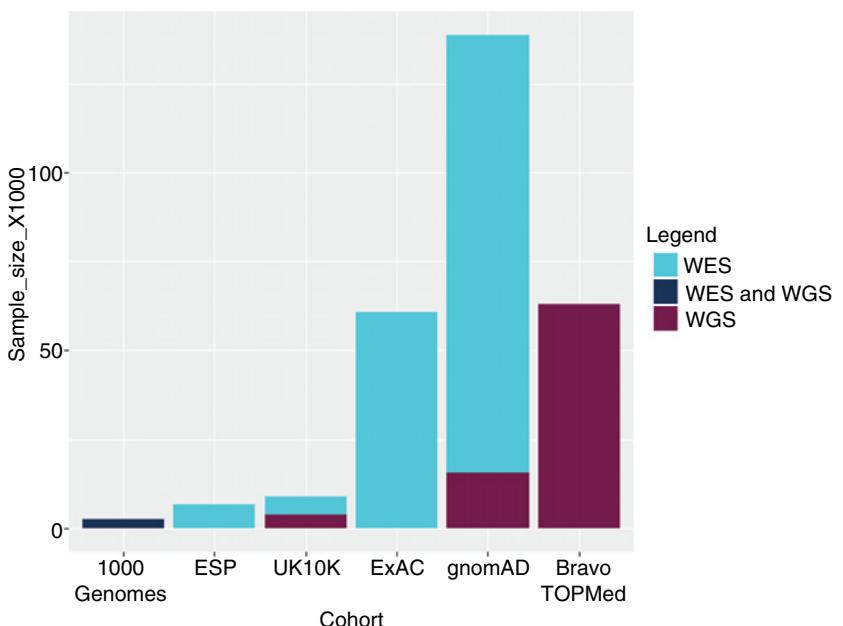


Figure 27.2 Examples of available human reference cohorts and their sequencing data compositions.

et al., 2015; di Julio *et al.*, 2018; Gussow *et al.*, 2017). This knowledge has had significant implications for variant interpretation; for example, genes showing a pattern of having lower tolerance to functional variation are now well described as more likely to be linked to disease (Petrovski *et al.*, 2013; Samocha *et al.*, 2014). Also, given this is a novel source of information, effective interpretational frameworks to leverage these gene-level scores in combination with variant-level scores are also emerging (Petrovski *et al.*, 2013; Zhu *et al.*, 2015; Traynelis *et al.*, 2017; Samocha *et al.*, 2017). Due to their dependency on empirical sequence data from large human samplings, most of these scores could not have been published earlier (Figure 27.3), yet have already been enthusiastically adopted by the community, and the power of such scores to facilitate accurate variant interpretation will only improve over time.

One of the important applications of variant intolerance scores is understanding tolerated alterations to gene dose. For example, if a gain-of-function variant leading to overexpression of a gene is identified as the disease mechanism, then development of an inhibitor is an obvious therapeutic strategy. If that gene is known to tolerate protein-truncating variation in the general population then inhibition is more likely to be considered a safe intervention. An example is *PCSK9* gain-of-function mutations which cause familial hypercholesterolaemia (Di Taranto *et al.*, 2017). *PCSK9* has a LoF depletion false discovery rate (FDR) *p*-value of 0.96 and an ExAC PLI of 0, both indicating that *PCSK9* tolerates reduced dose based on observation that there are many ‘natural inhibitors’ walking among us in the general population. Inhibitors such as evolocumab are effective and well tolerated (Sabatine *et al.*, 2015). If, on the other hand, the gene is identified as being intolerant to PTVs and, more specifically, dosage reduction then complete pharmaceutical inhibition may pose an unacceptable level of risk to patient safety. For example, in *SCN8A*, gain-of-function mutations cause severe and life-threatening epileptic encephalopathy (Veeramah *et al.*, 2012), but *SCN8A* also has a LoF depletion FDR *p*-value of 3.6×10^{-6} and a PLI of 1. Both scores indicate that loss-of-function mutations in *SCN8A* are also not tolerated in the general population, and indeed heterozygous *SCN8A* loss of function has been associated with cognitive impairment (Trudeau *et al.*, 2006).

Table 27.3 Scores that quantify intolerance to different classes of genetic variation and constructed solely relying on human standing variation data

Category	Score	Website	Reference
Gene level	RVIS z-score (constraint)	http://genic-intolerance.org/ http://exac.broadinstitute.org/	Petrovski <i>et al.</i> (2013) Samocha <i>et al.</i> (2014)
Loss-of-function-specific scores	LoF depletion pLI LoFtool	http://genic-intolerance.org/ http://exac.broadinstitute.org/ https://academic.oup.com/bioinformatics/article/33/4/471/2525582	Petrovski <i>et al.</i> (2015) Lek <i>et al.</i> (2016) Fadista <i>et al.</i> (2017)
Non-coding sequence	ncRVIS Orion CDTS	http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005492 http://www.genomic-orion.org/ http://www.hli-opendata.com/noncoding/	Petrovski <i>et al.</i> (2015) Gussow <i>et al.</i> (2017) di Julio <i>et al.</i> (2018)
Regional level (within gene)	subRVIS Missense tolerance ratio (MTR) Regional missense z-score	http://subrvvis.org/ http://mtr-viewer.mdhs.unimelb.edu.au https://www.biorxiv.org/content/early/2017/06/12/148353	Gussow <i>et al.</i> (2016) Traynelis <i>et al.</i> (2017) Samocha <i>et al.</i> (2017)

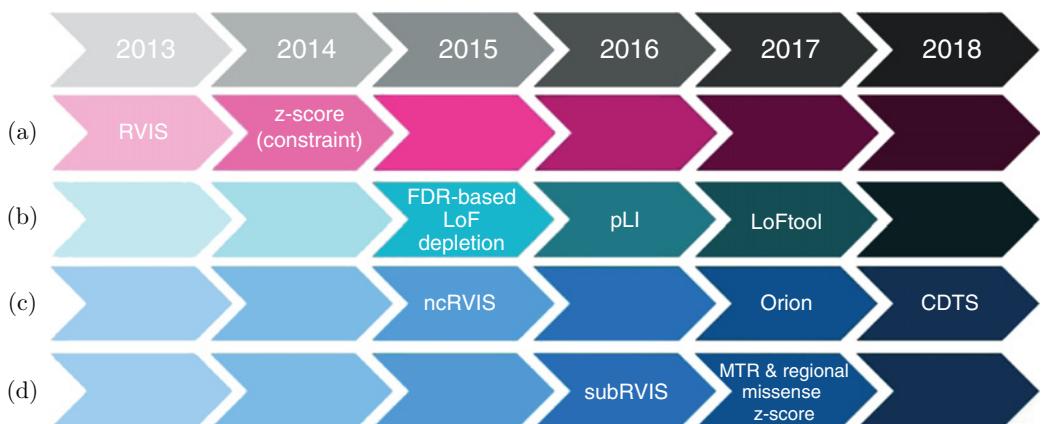


Figure 27.3 Publication timeline for the class of scores that estimate intolerance to genetic variation and are constructed based solely on standing variation genomics data. (a) Gene-level, non-synonymous intolerance. (b) Gene-level, loss-of-function intolerance. (c) Non-coding sequence, regional intolerance. (d) Within-gene, sub-regional protein-coding intolerance.

As discussed in Section 27.2, missense variants are particularly challenging to interpret, particularly when they have not been previously reported. Pathogenic missense mutations preferentially cluster in missense-intolerant regions, demonstrating that as well as gene-level and variant-level evidence, one must consider the regional mutational context (Eilbeck *et al.*, 2017; Traynelis *et al.*, 2017; Gussow *et al.*, 2016; Samocha *et al.*, 2017).

Gene-level metrics perform especially well for dominant variants. Alternative population genetics approaches are more informative for recessive traits. The basic concept for recessive disease can be illustrated with an example. A nullizygous genotype (i.e. losing both copies of the gene) in a patient is often considered a rare and highly interesting observation, yet it is well known that humans can tolerate complete loss of some proteins (Narasimhan *et al.*, 2016; MacArthur *et al.*, 2012). Taking the collection of null alleles found in the ExAC database for *NGLY1* (58 null alleles observed among 60,706 sequenced samples; none observed in a homozygous genotype) and collapsing them across the gene gives a conservative estimate of 2.3×10^{-7} for the rate of non-consanguineous individuals conceived having inherited a loss-of-function null allele from each parent. This translates to approximately one *NGLY1* nullizygous individual expected per 4.4 million non-consanguineous conceptions. While this metric does not directly translate to *NGLY1* nullizygosity being pathogenic, it provides simple population genetics support to appreciate that an *NGLY1* nullizygous genotype is atypical in the general population. In this example, like many others, the *NGLY1* nullizygous genotype is indeed pathogenic despite being discovered in the *n*-of-1 setting (Need *et al.*, 2012; Enns *et al.*, 2014). Such population genetics approaches promise continual improvement as the collection of sequenced samples increases – with attention now being focused on sequencing of underrepresented ethnicities (Petrovski and Goldstein, 2016).

Most of the human reference cohorts predominantly comprise individuals of European ancestry. This has important implications for variant interpretation, particularly in individuals with non-European ancestry. We have an increasing understanding that the minor allele frequency resolution is among the most valuable metrics when interpreting genetic variants, with rarer variants considerably more likely to be clinically relevant for most Mendelian and even some complex disorders. African populations are more genetically diverse than other populations (Pagani *et al.*, 2015). This, combined with underrepresentation in variant databases,

results in an apparent excess of rare variants in individuals of African ancestry (Auton *et al.*, 2015; Lek *et al.*, 2016). One consequence of this is that it is more challenging to identify pathogenic variants; thus there is a clinical and ethical urgency for improvement in representing diverse populations in variant databases, particularly given the different prevalence of certain diseases in populations of different ancestries (Petrovski and Goldstein, 2016; Carss *et al.*, 2017). Some cohorts are starting to address this problem; for example, ExAC is more ancestrally diverse than prior large cohorts (Lek *et al.*, 2016), and the United States Precision Medicine All of Us Initiative is also focusing on diverse populations. There are also some smaller population-specific initiatives such as the African Genome Variation Project, the South Asian Genome project, and East London Genes and Health (Gurdasani *et al.*, 2015; Chambers *et al.*, 2014; Narasimhan *et al.*, 2016).

Other population-specific considerations in this field include the fact that on average individuals of South Asian and Middle Eastern ancestry have higher autozygosity as a result of endogamy (Nakatsuka *et al.*, 2017), resulting in a higher proportion of homozygous variants. When studying these endogamous populations, analysis should not be restricted to detecting homozygous inherited variants, as the Deciphering Developmental Disorders Study (2017) recently showed that 6% of children with a developmental disorder whose parents were first cousins or closer had a plausible pathogenic *de novo* mutation. Variant frequency can also vary considerably depending on ancestry. Although there may be exceptions, in general an allele tolerated at a high allele frequency in one population is unlikely to be pathogenic in other populations for most studies of severe clinical phenotypes (Lek *et al.*, 2016). Thus, when we are interested in understanding the effects of rare variants, it is important to consider the highest allele frequency observed across a diverse collection of ancestry stratified controls rather than just the global population minor allele frequency.

For severe and dominant Mendelian disorders the highest pathogenic variant discovery yield is achieved when focusing on ultra-rare variants, defined as variants not observed in control cohorts (Petrovski *et al.*, 2017; Allen *et al.*, 2017; Zhu *et al.*, 2017; Genovese *et al.*, 2016; Ganna *et al.*, 2018). For more common and complex disorders (e.g. familial hypercholesterolemia) the tolerated allele frequency should be less conservative. However, understanding precisely what the ideal cut-off should be is difficult, given that the prevalence of a pathogenic variant in population reference cohorts depends on many factors, including disease prevalence, genetic and allelic heterogeneity, mode of inheritance, penetrance, severity and age of onset of disease (Minikel *et al.*, 2016; Whiffin *et al.*, 2017; Bennett *et al.*, 2017). Already, for some well-established genes, more sophisticated tools to address these shortcomings have been developed. For example, CardioClassifier facilitates interpretation of variants that might be associated with inherited cardiac conditions. It assesses whether the population frequency of the variants is consistent with pathogenicity (Whiffin *et al.*, 2018).

In summary, the increasing size of available human reference cohorts has revolutionised variant interpretation through increasing the resolution and accuracy of rare-variant frequency estimates. Also, by characterising the typical patterns of genetic variation in large human population cohorts, they have facilitated the identification of genes and within-gene regions that are intolerant to certain classes of variation that in turn allows identification of exceptional genotypes in an individual patient. Moving forward, these cohorts are likely to increase still further in size. Also, as the community moves more towards WGS, our understanding of typical patterns of non-coding genetic variation will start to catch up with that of coding, improving understanding of regulatory regions. Other future improvements will include accessibility to ancestrally more diverse cohorts, and more phenotype-specific tools like CardioClassifier.

27.4 Functional Assays of Genetic Variation

Another source of evidence used for variant interpretation is experimental data generated to understand the functional consequence(s) of a gene disruption or specific variant effect. Experimental methods are typically employed in three major contexts: first, for further investigation of previously reported disease-associated variants; second, for assessing pathogenicity of a novel variant in a known disease-associated gene that is amenable to functional evaluation; and third, for assessing the probability that a novel candidate gene is associated with disease. Scientists should strongly consider which context their question falls into to employ the proper experimental design. We also outline some suggestions for careful interpretation of functional validation data and discuss the importance of distinguishing between evidence at the gene and the variant level. An extensive discussion of commonly used experimental techniques is beyond the scope of this chapter.

The first-line commonly adopted experimental approach is the use of cellular models to investigate the function of a gene or variant. Cellular models are usually either *in vitro* immortalised cell lines or cells donated by the patient. Experiments using *in vitro* cell lines might involve investigating the localisation, structure and/or function of the product of the gene of interest in its unaltered state compared to its mutated state to provide gene-level and variant-level evidence of pathogenicity. Typically, the wildtype version of a gene of interest is introduced to a plasmid. PCR-based mutagenesis then allows scientists to introduce the variant of interest. The plasmid can then be transfected into the cell lines (Paquet *et al.*, 2016). Although it is likely that this technique will be replaced by introduction of the variant into the nuclear DNA by gene editing techniques, the traditional heterologous expression system still has certain advantages such as allowing for strong expression of the gene of interest via a cytomegalovirus promoter.

Advantages of using *in vitro* immortalised cell lines include the fact that biological samples from the patient are not required, and that inaccessible cell types, such as neuronal cells, can be investigated. Cell lines may be easier to work with than patient-derived cells as they have undergone selection for stability in experimental conditions. Also, this approach allows the effect of a variant to be assessed independently of the genetic context within the patient's genome, which can allow the researcher to disentangle these two factors. Disadvantages include that they have gone through many generations of cell division and selection so they become divergent from the original tissue sample, which raises questions about the generalisability of conclusions from these studies (Landry *et al.*, 2013; Seim *et al.*, 2017; Kaur and Dufour, 2012).

Another approach is to study the localisation, structure or function of the gene or variant of interest in cells from the patient themselves (either primary cells or a stable cell line made from the primary cells). The advantage of this approach is that the model is closer to the biological reality. Disadvantages include the fact that patients may be unable or unwilling to donate the necessary biological material, which depends on the phenotype and the expression profile of the gene of interest. If the gene of interest is specifically expressed in an inaccessible tissue such as neuronal tissue, direct sampling may not be possible. Induced pluripotent stem cells offer the potential to study inaccessible cell types in patient cells, but these are currently not without their own challenges (Aoi, 2016). Here, another promising prospect is that of human organoids, which provide a better representation of disease tissue and cellular context than two-dimensional cultures (Lancaster *et al.*, 2013). Even if a patient's sample can be obtained, establishing a stable cell line from the primary cells is not always possible, and primary cells themselves have a limited life span. Another problem is that the effect of the variant of interest cannot easily be distinguished from other factors. Gene editing technology such as CRISPR

provides the prospect of a solution to this challenge by reverting a variant back to the reference allele, providing an isogenic control.

Studying animal models, particularly mouse and zebrafish models, is another common experimental approach. For example, one of the most popular approaches to experimental validation of PTVs is to generate a knockout mouse that has reduced or eliminated expression of the gene of interest and to study the phenotype that results to assess how similar it is to the patient phenotype. This has been very fruitful, particularly for interpretation of novel candidate disease-associated genes (Brown *et al.*, 2018). The main advantage of this approach is that it allows investigation of the localisation, structure and/or function of the gene or variant of interest at the organismal level rather than the cellular level. Disadvantages include concerns with certain experimental methods, such as non-specific effects of morpholino knockdown in zebrafish confounding results (Gerety and Wilkinson, 2011). Also, model organisms might not be appropriate if the aetiology of the phenotype of interest is not conserved between humans and the proposed model. Additionally, there are ethical implications, and the investment of time and money required is generally greater than that required for experiments in cellular model systems. We describe some resources available to explore already available experimental results from animal models in Section 27.6.

The limitations of typical experimental workflows include that they are often a bottleneck in terms of time, and there is a lot of inconsistency in how different pathogenic variants in the same gene are investigated, which makes it difficult to compare the effect of different variants based on read-outs from different laboratories. An alternative is to carry out multiplexed assays for variant effect (MAVEs): high-throughput screens of variants in a disease-associated gene (Starita *et al.*, 2017). This allows an assessment of the effect of all possible variants regardless of known pathogenicity, and has already been performed for a handful of human disease-associated genes, including *PPARG*, *PTEN*, and *TPMT* (Majithia *et al.*, 2016; Matreyek *et al.*, 2018). For many genes, MAVEs would be expected to complement, if not replace, existing *in silico* based predictors of variant deleteriousness.

The reliability of experimental validation can be categorised into three main groups. For the first group of genes, their reliable functional readouts permit confident establishment of a causal link between a patient's genotype and their phenotype. Some of the best examples can be found among immunodeficiency phenotypes (Casanova *et al.*, 2014). In the second group reside genes that have accessible functional assays; however, it becomes important to distinguish between a statistically significant difference and a clinically relevant difference. For example, it has been common in literature to interpret a missense variant resulting in a minor percentage change in channel current as functional validation of a variant's effect on the channel. Yet, for many of these genes, there has been no reported effort to generate functional readouts (e.g. current change of channels) for general population variants of a similar predicted effect. A clear demonstration that missense polymorphisms in the same gene do not result in major changes to functional read-out scores, compared to the changes observed in reported pathogenic variants, will enable us to better appreciate the value of functional read-outs in the interpretation of causality for that gene. However, to date, this is rarely performed. Finally, for some genes or phenotypes there are no well-established functional assays to characterise the effects of variants found in patients. In such cases, the community must depend heavily on the remaining sources of supportive evidence, including genic prioritisation scores (D.G. MacArthur *et al.*, 2014).

In summary, experimental assays of the functional effect of genetic variation are an important source of information during variant interpretation, whether the chosen method is use of *in vitro* cell lines, patient cells, animal models, or indeed another method not discussed here such as transcriptomic or proteomic analysis of the patient. Care must be taken to interpret the results correctly, bearing in mind the limitations of the selected method, the aims

of the experiment, and whether the evidence provided is at the gene level or variant level. Experimental controls accessed from the general population are critical, but rarely performed. In the next section we will discuss use of existing information about gene function. This includes resources available to explore efforts to generate and phenotype model organisms on a large scale, and how they relate to the human phenotypes.

27.5 Leveraging Existing Information about Gene Function Including Human and Model Phenotype Resources

Here we introduce another source of evidence for variant interpretation, which is leveraging prior information about known gene function from both humans and model organisms from literature. In this section we highlight some relevant resources, including databases and ontologies, and summarise these in Table 27.4.

Online databases (such as OMIM, Genetics Home Reference, ClinGen and Orphanet) provide information about the clinical characteristics of genetic disorders, as well as the reported mode of inheritance, class of pathogenic variants, and sometimes the prevalence and expected age of onset. OMIM is often the first stop for researchers seeking an overview of the role of a gene in human disease. Typical information in OMIM includes a brief description of gene function, history of published research into the gene, gene structure and function, and importantly an overview of human phenotypes caused by clinically relevant variants in the gene. It includes links to phenotype-level pages, each of which has detailed clinical information describing the phenotypic spectrum of the disorder, age of onset, and others. It is updated daily, manually curated, and is regarded as the most valuable resource of its kind.

Genetics Home Reference is a similar database but it is directed less at researchers and more at patients/consumers. ClinGen is a similar resource that defines the clinical relevance of genes and variants for use in precision medicine and research (Rehm *et al.*, 2015). ClinGen runs several useful resources including Knowledge Base, which contains information about disease-associated genes including associated phenotypes, and whether there is any evidence of haploinsufficiency or triplosensitivity. Orphanet is a European online database specifically focused on rare diseases (Rath *et al.*, 2012). It also contains an inventory of orphan drugs and a directory of expert resources including links to relevant clinical trials and patient organisations.

Such databases are invaluable for interpreting genomic variants. However, there are four critical issues to be aware of. First, the concordance between the phenotype described in the resource and the phenotype of the patient in question must be assessed by clinical experts. For highly specific, well-characterised phenotypes this is generally a straightforward proposition, and high concordance adds confidence that the identified novel variant is clinically relevant. However, for genes with non-specific heterogeneous phenotypes such as intellectual disability, this is not necessarily the case. If the patient does have a consistent phenotype, but also additional features not previously described, it is often assumed that an expansion to the known phenotypic spectrum has been discovered. This may be the case; however, alternative interpretations include that the patient may have a 'blended phenotype' with undiscovered mutations in another gene (Posey *et al.*, 2016). Therefore, the confidence ascribed to this source of evidence depends on the phenotypic specificity.

Second, the mode of inheritance described in the resource should be consistent with that of the patient in question. Extra care is recommended when interpreting genes that have been reported as both dominant and recessive for the same phenotype. While this will happen for some genes, for others additional study to rule out recessive inheritance or to independently support a dominant disease mechanism may be required.

Table 27.4 Selected databases containing gene- and variant-level data

Class	Name	Full name	Website	Reference
Human disease-specific databases of genes	OMIM Genetics Home Reference	Online Mendelian Inheritance in Man Genetics Home Reference	https://omim.org/ https://ghr.nlm.nih.gov/	Amberger <i>et al.</i> (2015) Mitchell and McCray (2003)
	ClinGen Orphanet	Clinical Genome Resource Orphanet	https://www.clinicalgenome.org/ http://www.orpha.net	Rehm <i>et al.</i> (2015) Rath <i>et al.</i> (2012)
Non-disease-specific databases of information on genes or proteins	GeneCards UniProt	GeneCards UniProt	https://www.genecards.org/ http://www.uniprot.org/	Rebhan <i>et al.</i> (1997) UniProt Consortium (2018)
	GWAS catalog GTEx	NHGRI-EBI genome-wide association study catalog Genotype-Tissue Expression project	https://www.ebi.ac.uk/gwas/ https://www.gtexportal.org/home/	MacArthur <i>et al.</i> (2017) GTEx Consortium <i>et al.</i> (2017)
Human locus-specific databases of variants	BioGRID	Biological General Repository for Interaction Datasets	https://thebiogrid.org/	Chatr-Aryamontri <i>et al.</i> (2017)
	STRING Open Targets	STRING Open Targets	https://string-db.org/ https://www.opentargets.org/	Szklarczyk <i>et al.</i> (2015) Koscielny <i>et al.</i> (2017)
	ClinVar HGMD Decipher	ClinVar Human Gene Mutation Database Decipher	https://www.ncbi.nlm.nih.gov/clinvar/ http://www.hgmd.cf.ac.uk/ac/index.php https://decipher.sanger.ac.uk/	Landrum <i>et al.</i> (2018) Stenson <i>et al.</i> (2017) Firth <i>et al.</i> (2009)
	CFTR2	The Clinical and Functional TTranslation of CFTR	http://cftr2.org	Castellani (2013)
	LOVD RettBASE	Leiden Open Variation Database 3.0 Rett Syndrome Variation Database	http://www.lovd.nl/3.0/home http://mecp2.chw.edu.au/	Fokkema <i>et al.</i> (2011) Krishnaraj <i>et al.</i> (2017)
	MGI ZFIN OMIA	Mouse Genome Informatics Zebrafish Information Network Online Mendelian Inheritance in Animals	http://www.informatics.jax.org/ https://zfin.org/ http://omia.org/home/	Smith <i>et al.</i> (2018) Howe <i>et al.</i> (2013) Lenffer <i>et al.</i> (2006)
Variant prioritisation tool	Exomiser	Exomiser	https://www.sanger.ac.uk/science/tools/exomiser	Bone <i>et al.</i> (2016)
Ontology	HPO	Human Phenotype Ontology	http://human-phenotype-ontology.github.io/	Köhler <i>et al.</i> (2017)
Data sharing framework	GA4GH	Global Alliance for Genomics & Health	http://genomicsandhealth.org	Rahimzadeh <i>et al.</i> (2016)

Third, the variant mechanism described in the resource (e.g. loss of function or gain of function) should be considered. For example, identifying a novel *de novo* missense variant in a gene where the gene–trait association is solely characterised by dominant PTVs would require further functional data to confirm that any novel missense *de novo* mutations do indeed have an effect equivalent to a null allele.

Finally, databases can be outdated and contaminated with inaccurate information. Therefore, thorough literature searches are recommended and periodic reanalyses of updated gene- and variant-associated literature will continue to improve diagnostic yields (Need *et al.*, 2017; Walsh *et al.*, 2017).

As well as databases that are designed for disease-associated genes, there are non-disease-specific sources of information about gene or protein function that can help gauge the plausibility of a putative novel disease-associated gene. These include GeneCards, which brings together gene-centric data from multiple sources including information about gene function, structure, regulation, interactors, and expression; UniProt (UniProt Consortium, 2018) which contains protein-centric data including protein function, structure, interactors and expression; the NHGRI-EBI GWAS catalog, which contains associations discovered by common-variant GWAS (MacArthur *et al.*, 2017); GTEx, which contains extensive data about gene expression, including associations with genotype (GTEx Consortium *et al.*, 2017); and finally, BioGRID and STRING, which both house databases of molecular interactions between genes and/or proteins (Chatr-Aryamontri *et al.*, 2017; Szklarczyk *et al.*, 2015).

So far in this section we have concentrated on gene-level knowledge. There are also databases containing variant-level data that are critical to variant interpretation. These resources rely on widespread data sharing. The most commonly used are ClinVar, HGMD and Decipher. Information associated with each variant typically includes phenotype, interpretation of pathogenicity, and reference in literature. ClinVar is a freely accessible database of variants submitted directly by researchers and clinicians (Landrum *et al.*, 2018). HGMD is a database of variants mined from the literature and then manually curated; it requires a fee for full access (Stenson *et al.*, 2017). Decipher is a freely accessible database of variants that originally focused on copy number variants, but has now expanded to include SNVs and indels identified in children with developmental disorders (Firth *et al.*, 2009). There are also locus-specific or disease-specific databases of pathogenic variants, which contain more detailed information and can be more extensively curated than the generic databases. These include CFTR2, Leiden Open Variation Database 3.0 and RettBASE (Fokkema *et al.*, 2011; Krishnaraj *et al.*, 2017). A more comprehensive list can be found on the EBI HGNC website (see the web resources list in Section 27.8). Variant databases are not always concordant with each other; therefore, it is best to use multiple resources. When possible, it is advisable to read the original publication to ensure the authors' findings and interpretations are accurately represented in the database (Dorschner *et al.*, 2013).

As discussed in Section 27.4, animal models can be a valuable source of evidence for variant interpretation. Large consortia have generated model organisms such as knockouts on a large scale and performed phenotyping, allowing researchers to learn the effect of knocking out their gene of interest. Existing animal model resources include Mouse Genome Informatics (MGI), the Zebrafish Information Network (ZFIN), and Online Mendelian Inheritance in Animals (OMIA) (Smith *et al.*, 2018; Howe *et al.*, 2013). Whether the animal model accurately reflects the human biology and thus resulting phenotypes when knocking out a specific gene remains an open question in most cases. However, the more reflective the animal phenotype is to the patient presentation (including phenotypes in the matching anatomical systems), and the more specific the observed animal phenotype is, the greater the confidence.

An interesting example of a tool that brings together prior knowledge of both human and animal model gene-centric data to assist variant interpretation is Exomiser, which effectively

ranks genetic variants based on the patient's phenotypic similarity to known gene-associated phenotypes, and model organism phenotypes of orthologs (Bone *et al.*, 2016). To efficiently cross-reference data sources in this way requires a consistent and standardised way of referring to phenotypes. Human Phenotype Ontology (HPO) is a standardised vocabulary for describing human phenotypes and the relationships between them (Köhler *et al.*, 2017). Describing phenotypes in this way is more restrictive than using free text, but this common metadata language and standardisation allows phenotypic descriptions of patients to be effectively analysed (Greene *et al.*, 2016).

In this section we have highlighted some databases relating to gene function that can be used to aid variant interpretation. These include disease-specific gene-centric databases such as OMIM, non-disease specific databases such as GeneCards, variant-level databases such as HGMD and ClinVar, model organism databases such as MGI, and finally Exomiser as an example of a tool that aims to pull these data together to rank genes in order of predicted pathogenicity.

27.6 Holistic Variant Interpretation

In this section we bring together the various sources of evidence for variant interpretation: understanding variant consequence, understanding context from human reference cohorts, experimental validation, and leveraging prior information about gene function and disease associations. We describe some existing guidelines, frameworks and tools that leverage multiple sources of evidence to help standardise variant interpretation.

Consistency in interpreting variant pathogenicity is of upmost importance in clinical genetics. However, there is high variability in variant interpretation between different individuals and laboratories that is exacerbated by the increasing number of variant interpretation tools and resources, because each analyst tends to assign their own weight to different sources of evidence based on differing preferences, experience and training, which can result in differing interpretations of pathogenicity (Pepin *et al.*, 2016; Bland *et al.*, 2018). Attempts have been made to standardise aspects of the variant interpretation process, most notably detailed guidelines recommended by the ACMG-AMP (Richards *et al.*, 2015). These have been widely adopted by the community, and have by one measure improved consistency of variant interpretation from 34% to 71%, which is promising, but there is clearly scope for further standardisation and automation (Amendola *et al.*, 2016; Pepin *et al.*, 2016).

Efforts to refine, improve, and expand the ACMG-AMP recommendations continue, led by the ClinGen Sequence Variant Interpretation Working Groups. For example, it has been recommended that criteria referring to 'reputable sources' be removed from the guidelines; Bayesian classification has been used to highlight logical inconsistencies; classification of edge cases has been clarified; for some phenotypes much more specific versions of the guidelines have been developed by expert panels; and phenotype-specific recommendations on appropriate allele frequency filtering cut-offs have been developed (Kelly *et al.*, 2018; Biesecker and Harrison, 2018; Tavtigian *et al.*, 2018; Nykamp *et al.*, 2017; Whiffin *et al.*, 2018). Recently, similar recommendations for somatic variants in cancer have been published (Li *et al.*, 2017).

There are important distinctions in variant interpretation between the clinical genetic context and the research study context. The purpose of a clinical genetic test is typically to identify the pathogenic variant in a known disease-associated gene and return the result to the patient for the purposes of diagnosis, management, prognosis, and reproductive decisions. Therefore, formal and highly regulated procedures, including strict adherence to established guidelines, are required. A related question is whether secondary findings from clinical genetic tests should

be returned to patients. It is vitally important that variants that constitute secondary findings are appropriately interpreted to prevent unnecessary medical interventions, yet it is a topic about which there is still much disagreement within the community (Green *et al.*, 2013; Vears *et al.*, 2018). In comparison, the purpose of a research study can be more broad-ranging and can include improving understanding of disease biology, identifying phenotypic expansions and identifying novel disease-associated genes. Additional differences between clinical and research interpretation include the degree to which each stage of the process is regulated, and the fact that results from research studies are not always returned to the patients and cannot directly influence patient management. Research-based variant interpretation is, however, of critical importance as it has the greatest contribution to expanding our knowledge base of genotype–phenotype correlations. The ACMG-AMP guidelines should be adhered to even in the research context when using the term ‘pathogenic’ to describe variants that may ultimately be used for clinical variant interpretation. In practice, some studies straddle both the research and clinical fields. For example, a research study may choose to perform variant interpretation according to the strict ACMG guidelines and under the supervision of senior clinical geneticists, so that variants deemed to be pathogenic or likely pathogenic can be subsequently confirmed in an accredited clinical laboratory, evaluated by a clinical geneticist and reported back to the patient’s physician who may choose to return the result to the patient.

In summary, standardised guidelines on variant interpretation are vital for maintaining scientific rigour. Such guidelines have been published (D.G. MacArthur *et al.*, 2014; Richards *et al.*, 2015) and are continually extended, revised and improved. These are most important in the context of ascribing pathogenicity to a variant from a clinical genetic test, but are also helpful in a research context. There are at least three remaining challenges in this area. First, there are important clinical implications of changing variant interpretation guidelines. Variants may be reclassified after they have been returned to patients, which has psychological implications for them, and logistical, ethical and legal implications for researchers regarding the prospect of reanalyses (Thorogood *et al.*, 2017; Taber *et al.*, 2018). Second, manual assessment of each criterion required by the ACMG-AMP recommendations is time-consuming and can lead to subjectivity. Therefore, refinement of automated pathogenicity classification tools that consider all the required sources of evidence such as ClinGen’s pathogenicity calculator are needed (Patel *et al.*, 2017). Third, open-source tools that aggregate and display data for variant interpretation remain an unmet need within the community.

27.7 Future Challenges and Closing Remarks

There has been substantial progress in variant interpretation in recent years, driven largely by the increase in the scale of sequencing and in data sharing. However, there are also looming challenges that we outline in this section. We must embrace emerging methods including structuring the differing data sources into connected knowledge graph representations, shifting away from the conventional sequential evaluation of evidence to more holistic evaluations, and adopting more machine learning in the derivation of expert systems. Additionally, we must invest more in sequencing larger and more ancestrally diverse cohorts, and be open to sharing data to improve variant interpretation sufficiently to fully realise the potential of genomics to transform patient care, drug development, and basic research.

There are also other classes of variants that we have not discussed, and for which the variant interpretation task remains extremely challenging. These include modifier variants, SVs and non-coding variants. Modifier variants can account for some of the phenotypic heterogeneity that is often observed between individuals who have the same causal variant, and they might

also affect the degree to which individuals respond to pharmaceutical treatment. Identifying clinically relevant modifier variants and unravelling their biological role is essential to understand the heterogeneity of disease and the variability in response to therapy. Nevertheless, this remains a challenging task, partly because modifiers will require larger case collections, and because environmental and epigenetic factors can contribute to phenotypic heterogeneity.

Structural variants include deletions, insertions, tandem duplications, dispersed duplications, inversions, translocations, and complex structural variants. Collectively, these have been reported to account for more differences between any two individuals than SNVs and indels, and are known to influence species evolution, gene expression, phenotypic traits, cancer, complex disease and Mendelian disease (Zhang *et al.*, 2009; Bailey and Eichler, 2006; Sudmant *et al.*, 2015; Conrad *et al.*, 2010; Chiang *et al.*, 2017; Carss *et al.*, 2017; Sanchis-Juan *et al.*, 2018). Yet they remain very challenging to identify and interpret from high-throughput sequencing data, so they are often either ignored or only a subset of them are interrogated in any given study. Reasons for this include variable sequencing coverage (especially for exome sequencing), low sensitivity and/or specificity of callers especially at repetitive regions, and lack of community-wide agreement on best practice for SV calling, definition, validation, and merging across samples (Sudmant *et al.*, 2015; Carss *et al.*, 2017). As the importance of this class of variants is increasingly being highlighted, there are efforts to address these challenges. For example, long-read high-throughput sequencing offers the potential to improve sensitivity of calling at repetitive regions, and many new and improved SV calling algorithms and analytical tools are being published, including some that adopt machine learning approaches (Merker *et al.*, 2018; Belyeu *et al.*, 2018; Antaki *et al.*, 2018).

Pathogenic non-coding variation is known to contribute to Mendelian disease (Spielmann and Mundlos, 2016; Short *et al.*, 2018). Interpretation of non-coding variants is still considered very challenging. Nevertheless, with non-coding variants increasingly being annotated with information that can aid variant interpretation, including functional domains from epigenomics and constrained regions from WGS in large human populations, combined with plummeting costs of WGS, non-coding variant interpretation is an active area of research (Kawaji *et al.*, 2017; Stunnenberg *et al.*, 2016; Petrovski *et al.*, 2015; di Julio *et al.*, 2018; Gussow *et al.*, 2017; ENCODE Project Consortium, 2012). One increasingly important source of information that could facilitate variant interpretation, particularly in the non-coding space, is the growing collections of large-scale omics data sets, including transcriptomics, proteomics, epigenomics, metabolomics and microbiomics. The optimum method to integrate these multidimensional data sets with genomic data in order to realise their full potential remains a challenging open question, recently reviewed by Hasin *et al.* (2017).

Understanding the genomic context of a variant requires large human reference cohorts, as discussed in Section 27.3. To improve our understanding of the distribution of variation, even larger cohorts, specifically addressing the under-representation of non-European populations, will be required. A large amount of valuable sequencing data and associated phenotypic data has already been generated but remains locked away in clinical laboratories, research institutes, and genetic testing companies. Lack of data sharing may be due to commercial sensitivity or privacy concerns, including a fear of reidentification (Takashima *et al.*, 2018; Wang *et al.*, 2017). Attempts to increase data sharing include ClinGen and the Global Alliance for Genomics & Health (GA4GH) which aim to create a centralised repository of genomic and clinical data to improve clinical interpretation and patient care.

Tools that allow sharing of candidate genes without requiring researchers to share sequence data include GeneMatcher (Sobreira *et al.*, 2015). These facilitate information sharing across various groups by connecting researchers studying the same gene. This has enabled fast-tracking novel gene–trait associations courtesy of multiple affected families being coidentified

across the globe, and there are many such instances where this has been successful (Myers *et al.*, 2017; Ito *et al.*, 2018). The single most important caveat with this opportunity is that careful consideration regarding the other lines of evidence, such as population genetics signatures and functional characterisation of genetic lesions, is perhaps more important than ever. Several other models of data sharing fall in between the two extremes of sharing nothing and publicly sharing the sequencing data themselves. These include the sharing of metadata but not individual-level data, and a proportional tiered approach involving managed data access agreements (Wright *et al.*, 2016).

With the growing number of examples where precision medicine positively influences patient management comes the increasing importance of correctly classifying genetic variants. This is especially true in established disease-associated genes where, without a proper framework, the criteria for declaring causality can often become relaxed. The current aspiration in human genomics is to systematise precision diagnostics, and an overarching priority is to develop statistical frameworks to quantitatively and objectively assess the confidence that specific genotypes found in established disease-associated genes are pathogenic. It remains an open question whether the time is right to derive a single comprehensive, non-sequential, flexible expert system that incorporates all the various sources of evidence into one probability calculation, turning disparate data and expert judgements into a single weight of evidence.

There have been earlier efforts relevant to this aspiration (Wilfert *et al.*, 2016; Hu *et al.*, 2013). Some were published before recent access to large-scale human sequencing reference cohorts or lacked integration of the orthogonal lines of evidence now considered key in routine variant interpretation. The argument to focus on this key challenge is the maturity of the data generation platforms, our increased understanding of the diverse types of evidence relevant to decision-making, and our access to entirely novel information, including access to empirical patterns of genetic variation in the human population. Even components as fundamental as the achievable genome-wide minor allele frequency resolution have become more informed in recent years, and as we generate more sequence data we will continue to improve our resolution at the low end of the allele frequency spectrum. Singleton observations remain the largest variant group in existing large reference cohorts amassing over 100,000 individuals; and our understanding of their true allele frequencies remains limited by the total population size. Nonetheless, we argue that the time is right for development of integrative statistical frameworks incorporating the various sources of evidence discussed in this chapter. Also, various novel multi-omics sources such as transcriptomics, proteomics, metabolomics, microbiomes and radiomics are now frequently discussed data types whose technical and analytic platforms are maturing, and advances are being made in interpretation of output data.

This ability to synthesise all the evidence together is expected to transform variant interpretation and would enable the community to achieve consistent conclusions about the pathogenicity of variants in a standardised and objective manner. Many of these data sources will continue to mature and edge cases will arise. For these reasons it is expected that the framework would need to be sufficiently flexible to allow for refinement of classification and to accommodate more common and complex genetic disorders, which would require additional development to understand relevance of variants contributing to risk in more genetically complex settings.

In this chapter we have outlined four sources of evidence for variant interpretation. Robust variant interpretation is a vital strand in understanding genome biology in health and disease. This understanding forms the bridge between genomics and the health-care revolution that was promised at the time of the publication of the first draft of the human genome (Collins, 1999; Collins and McKusick, 2001) At the 10-year anniversary of this landmark achievement, some disappointment was expressed that the promised health-care revolution had not yet been realised, along with doubt that it ever would be (Ball, 2010). However, the translation of genomic

science into health care took decades rather than years because of the complexity of the human genome, the high initial costs of sequencing, the size of the cohorts required to truly understand the distribution of human variation, our limited knowledge of the genetic architecture for many simple and complex traits, the patients and experiments required to translate sequence into candidate drug targets, and the time it takes to conduct clinical trials. Yet, the recent success of several health-care interventions that would not have been possible without genomics, including gene therapy for *RPE65*-mediated inherited retinal dystrophy, chimeric antigen receptor T-cell therapy for B-cell lymphoblastic leukaemia, and oligonucleotides that induce exon skipping to treat muscular dystrophy, (Russell *et al.*, 2017; Maude *et al.*, 2018; Goemans *et al.*, 2016), along with increased investment in genomics research for target identification, target validation and companion molecular diagnostics by the pharmaceutical industry, suggests that we are finally realising that initial promise of the human genome project.

27.8 Web Resources

Some commonly used variant interpretation tools:

NNSPLICE	http://www.fruitfly.org/seq_tools/splice.html
HSF	http://www.umd.be/HSF3/
MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html
TraP	http://trap-score.org/
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
SIFT	http://sift.jcvi.org/
LRT	http://www.genetics.wustl.edu/jflab/lrt_query.html
PhyloP	http://compgen.cshl.edu/phast/background.php
PhastCons	http://compgen.cshl.edu/phast/background.php
GERP++	http://mendel.stanford.edu/SidowLab/downloads/gerp/
Condel	http://bbgllab.irbbarcelona.org/fannsdb/
MutationTaster	http://www.mutationtaster.org/
eXtasy	http://extasy.esat.kuleuven.be/
CADD	http://cadd.gs.washington.edu/home
REVEL	https://sites.google.com/site/revelgenomics/
VEST	http://karchinlab.org/apps/appVest.html
DANN	https://cbcl.ics.uci.edu/public_data/DANN/
cBioPortal	http://www.cbioportal.org/
MuPIT	http://mupit.icm.jhu.edu
mCSM	http://structure.bioc.cam.ac.uk/mcsm
SDM	http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php
DUET	http://structure.bioc.cam.ac.uk/duet
LOFTEE	https://github.com/konradjk/loftee
IGV	http://software.broadinstitute.org/software/igv/
dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP

Some commonly used human reference cohorts:

1000 Genomes	http://www.internationalgenome.org/
NHLBI Exome Sequencing Project	http://evs.gs.washington.edu/EVS/

UK10K	https://www.uk10k.org/
ExAC	http://exac.broadinstitute.org/
Bravo TOPMed	https://bravo.sph.umich.edu/freeze5/hg38/
gnomAD	http://gnomad.broadinstitute.org/

Some commonly used scores of variant intolerance:

RVIS	http://genic-intolerance.org/
Constraint	http://exac.broadinstitute.org/
LoF depletion	http://genic-intolerance.org/
pLI	http://exac.broadinstitute.org/
LoFtool	https://academic.oup.com/bioinformatics/article/33/4/471/2525582
subRVIS	https://github.com/igm-team/subrvvis
MTR	http://mtr-viewer.mdhs.unimelb.edu.au
Regional missense	https://www.biorxiv.org/content/early/2017/06/12/148353
z-score	
ncRVIS	http://genic-intolerance.org/
Orion	http://www.genomic-orion.org/
CDTS	http://www.hli-opendata.com/noncoding/

Some commonly used databases of information about gene- and variant-level function:

OMIM	https://omim.org/
Genetics Home Reference	https://ghr.nlm.nih.gov/
ClinGen	https://www.clinicalgenome.org/
Orphanet	http://www.orpha.net
GeneCards	https://www.genecards.org/
UniProt	http://www.uniprot.org/
GWAS catalog	https://www.ebi.ac.uk/gwas/
GTEX	https://www.gtexportal.org/home/
BioGRID	https://thebiogrid.org/
STRING	https://string-db.org/
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php
Decipher	https://decipher.sanger.ac.uk/
CFTR2	http://cftr2.org
LOVD	http://www.lovd.nl/3.0/home
RettBASE	http://mecp2.chw.edu.au/
MGI	http://www.informatics.jax.org/
ZFIN	https://zfin.org/
Online Mendelian Inheritance in Animals	http://omia.org/home/
Exomiser	https://www.sanger.ac.uk/science/tools/exomiser
HPO	http://human-phenotype-ontology.github.io/
GA4GH	http://genomicsandhealth.org
Catalog of gene-specific mutation databases	ftp://ftp.ebi.ac.uk/pub/databases/genenames/lsdb_links.txt.gz

References

- Acuna-Hidalgo, R., Bo, T., Kwint, M.P., van de Vorst, M., Pinelli, M., Veltman, J.A., Hoischen, A., Vissers, L.E.L.M. and Gilissen, C. (2015). Post-zygotic point mutations are an underrecognized source of de novo genomic variation. *American Journal of Human Genetics* **97**(1), 67–74.
- Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* **76**(1), 7.20.
- Allen, A.S., Bellows, S.T., Berkovic, S.F., et al. (2017). Ultra-rare genetic variation in common epilepsies: A case-control sequencing study. *Lancet Neurology* **16**(2), 135–143.
- Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**(Database issue), D789–D798.
- Amendola, L.M., Jarvik, G.P., Leo, M.C., et al. (2016). Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. *American Journal of Human Genetics* **98**(6), 1067–1076.
- Antaki, D., Brandler, W.M. and Sebat, J. (2018). SV2: Accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* **34**(10), 1774–1777.
- Aoi, T. (2016). 10th anniversary of iPS cells: The challenges that lie ahead. *Journal of Biochemistry* **160**(3), 121–129.
- Auton, A., Abecasis, G.R., Altshuler, D.M., et al. (2015). A global reference for human genetic variation. *Nature* **526**(7571), 68–74.
- Bailey, J.A. and Eichler, E.E. (2006). Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nature Reviews Genetics* **7**(7), 552–564.
- Ball, P. (2010). Bursting the genomics bubble. *Nature*, 31 March. Available at: [http://www.nature.com/doifinder/10.1038/news.\(2010\).145](http://www.nature.com/doifinder/10.1038/news.(2010).145).
- Belyeu, J.R., Nicholas, T.J., Pedersen, B.S., Sasani, T.A., Havrilla, J.M., Kravitz, S.N., Conway, M.E., Lohman, B.K., Quinlan, A.R. and Layer, R.M. (2018). SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *Gigascience* **7**(7).
- Bennett, C.A., Petrovski, S., Oliver, K.L. and Berkovic, S.F. (2017). ExACTly zero or once: A clinically helpful guide to assessing genetic variants in mild epilepsies. *Neurology. Genetics* **3**(4), e163.
- Biesecker, L.G. and Harrison, S.M. (2018). The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genetics in Medicine*. doi: 10.1038/gim.2018.42.
- Bland, A., Harrington, E.A., Dunn, K., Pariani, M., Platt, J.C.K., Grove, M.E. and Caleshu, C. (2018). Clinically impactful differences in variant interpretation between clinicians and testing laboratories: A single-center experience. *Genetics in Medicine* **20**(3), 369–373.
- Bone, W.P., Washington, N.L., Buske, O.J., et al. (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine* **18**(6), 608–617.
- Brown, S.D.M., Holmes, C.C., Mallon, A.-M., Meehan, T.F., Smedley, D. and Wells, S. (2018). High-throughput mouse phenomics for characterizing mammalian gene function. *Nature Reviews. Genetics* **19**(6), 357–370.
- Cars, K.J., Arno, G., Erwood, M., et al. (2017). Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. *American Journal of Human Genetics* **100**(1).
- Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**(Suppl. 3), S3.
- Casanova, J.-L., Conley, M.E., Seligman, S.J., Abel, L. and Notarangelo, L.D. (2014). Guidelines for genetic studies in single patients: LESSONS from primary immunodeficiencies. *Journal of Experimental Medicine* **211**(11), 2137–2149.

- Castellani, C. and CFTR2 Team (2013). CFTR2: How will it help care? *Paediatric Respiratory Reviews* **14**(Suppl. 1), 2–5.
- Cerami, E., Gao, J., Dogrusoz, U., et al. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**(5), 401–404.
- Chambers, J.C., Abbott, J., Zhang, W., et al. (2014). The South Asian genome. *PLoS ONE* **9**(8), e102645.
- Chatr-Aryamontri, A., Oughtred, R., Boucher, L., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research* **45**(D1), D369–D379.
- Chiang, C., Scott, A.J., Davis, J.R., et al. (2017). The impact of structural variation on human gene expression. *Nature genetics* **49**(5), 692–699.
- Chun, S. and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research* **19**(9), 1553–1561.
- Cirulli, E.T. and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**(6), 415–425.
- Cirulli, E.T., Lasseigne, B.N., Petrovski, S., et al. (2015). Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science* **347**(6229), 1436–1441.
- Collins, F.S. (1999). Shattuck lecture – medical and societal consequences of the Human Genome Project. *New England Journal of Medicine* **341**(1), 28–37.
- Collins, F.S. and McKusick, V.A. (2001). Implications of the Human Genome Project for medical science. *Journal of the American Medical Association* **285**(5), 540–544.
- Collins, F.S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine* **372**(9), 793–795.
- Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J. and Hurles, M.E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature Genetics* **42**(5), 385–391.
- Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S. and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**(7), 901–913.
- Cummings, B.B., Marshall, J.L., Tukiainen, T., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine* **9**(386).
- Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**(12), e1001025.
- de Ligt, J., Willemse, M.H., van Bon, B.W.M., et al. (2012). Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine* **367**(20), 1921–1929.
- Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**(7642), 433–438.
- Defesche, J.C., Gidding, S.S., Harada-Shiba, M., Hegele, R.A., Santos, R.D. and Wierzbicki, A.S. (2017). Familial hypercholesterolemia. *Nature Reviews Disease Primers* **3**, 17093.
- Desmet, F.O., Hamroun, D., Lalande, M., Collod-Béroud, G., Clastres, M. and Béroud, C. (2009). Human Splicing Finder: An online bioinformatics tool to predict splicing signals. *Nucleic Acids Research* **37**(9), e67.
- di Iulio, J., Bartha, I., Wong, E.H.M., et al. (2018). The human noncoding genome defined by genetic diversity. *Nature Genetics* **50**(3), 333–337.
- Di Taranto, M.D., Benito-Vicente, A., Giacobbe, C., Uribe, K.B., Rubba, P., Etxebarria, A., Guardamagna, O., Gentile, M., Martín, C. and Fortunato, G. (2017). Identification and in vitro characterization of two new PCSK9 gain of function variants found in patients with familial hypercholesterolemia. *Scientific Reports* **7**(1), 15282.

- DiStefano, M.T., Hemphill, S.E., Cushman, B.J., *et al.* (2018). Curating clinically relevant transcripts for the interpretation of sequence variants. *The Journal of Molecular Diagnostics* **20**(6), 789–801.
- Dorschner, M.O., Amendola, L.M., Turner, E.H., *et al.* (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *American Journal of Human Genetics* **93**(4), 631–640.
- Eilbeck, K., Quinlan, A. and Yandell, M. (2017). Settling the score: Variant prioritization and Mendelian disease. *Nature Reviews Genetics* **18**(10), 599–612.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74.
- Enns, G.M., Shashi, V., Bainbridge, M., *et al.* (2014). Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genetics in Medicine* **16**(10), 751–758.
- Epi4K Consortium, Epilepsy Phenome/Genome Project, Allen, A.S., *et al.* (2013). De novo mutations in epileptic encephalopathies. *Nature* **501**(7466), 217–221.
- Fadista, J., Oskolkov, N., Hansson, O. and Groop, L. (2017). LoFtool: A gene intolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* **33**(4), 471–474.
- Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M. and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics* **84**(4), 524–533.
- Flanagan, S.E., Patch, A.-M. and Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers* **14**(4), 533–537.
- Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J. and den Dunnen, J.T. (2011). LOVD v.2.0: The next generation in gene variant databases. *Human Mutation* **32**(5), 557–563.
- Forbes, S.A., Beare, D., Boutselakis, H., *et al.* (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45**(D1), D777–D783.
- Franciolli, L.C., Polak, P.P., Koren, A., *et al.* (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics* **47**(7), 822–826.
- Futema, M., Plagno, V., Li, K.W., *et al.* (2014). Whole exome sequencing of familial hypercholesterolaemia patients negative for LDLR/APOB/PCSK9 mutations. *Journal of Medical Genetics* **51**(8), 537–544.
- Ganna, A., Satterstrom, F.K., Zekavat, S.M., *et al.* (2018). Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *American Journal of Human Genetics* **102**(6), 1204–1211.
- Gelfman, S., Wang, Q., McSweeney, K.M., *et al.* (2017). Annotating pathogenic non-coding variants in genic regions. *Nature Communications* **8**(1), 236.
- Gelsi-Boyer, V., Trouplin, V., Adélaïde, J., *et al.* (2009). Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *British Journal of Haematology* **145**(6), 788–800.
- Genovese, G., Kähler, A.K., Handsaker, R.E., *et al.* (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine*, **371**(26), 2477–2487.
- Genovese, G., Fromer, M., Stahl, E.A., *et al.* (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience* **19**(11), 1433–1441.
- Gerety, S.S. and Wilkinson, D.G. (2011). Morpholino artifacts provide pitfalls and reveal a novel role for pro-apoptotic genes in hindbrain boundary development. *Developmental Biology* **350**(2), 279–289.

- Goemans, N.M., Tulinius, M., van den Hauwe, M., Kroksmark, A.-K., Buyse, G., Wilson, R.J., van Deutekom, J.C., de Kimpe, S.J., Lourbakkos, A. and Campion, G. (2016). Long-term efficacy, safety, and pharmacokinetics of drisapersen in Duchenne muscular dystrophy: Results from an open-label extension study. *PLoS ONE* **11**(9), e0161955.
- Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S. and Sunyaev, S. (2013). Sequencing studies in human genetics: Design and interpretation. *Nature Reviews Genetics* **14**(7), 460–470.
- González-Pérez, A. and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics* **88**(4), 440–449.
- Green, R.C., Berg, J.S., Grody, W.W., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine* **15**(7), 565–574.
- Greene, D., Richardson, S. and Turro, E. (2016). Phenotype similarity regression for identifying the genetic determinants of rare diseases. *American Journal of Human Genetics* **98**(3), 490–499.
- GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods Groups—Analysis Working Group, et al. (2017). Genetic effects on gene expression across human tissues. *Nature* **550**(7675), 204–213.
- Guillermín, Y., Lopez, J., Chabane, K., Hayette, S., Bardel, C., Salles, G., Sujobert, P., Huet, S. (2018). What does this mutation mean? The tools and pitfalls of variant interpretation in lymphoid malignancies. *International Journal of Molecular Sciences* **19**(4).
- Gurdasani, D., Carstensen, T., Tekola-Ayele, F., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**(7534), 327–3232.
- Gussow, A.B., Petrovski, S., Wang, Q., Allen, A.S. and Goldstein, D.B. (2016). The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biology* **17**(1), 9.
- Gussow, A.B., Copeland, B.R., Dhindsa, R.S., Wang, Q., Petrovski, S., Majoros, W.H., Allen, A.S. and Goldstein, D.B. (2017). Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS ONE* **12**(8), e0181604.
- Hansen, N.F., Gartner, J.J., Mei, L., Samuels, Y. and Mullikin, J.C. (2013). Shimmer: Detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics* **29**(12), 1498–1503.
- Hasin, Y., Seldin, M. and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology* **18**(1), 83.
- Hoischen, A., van Bon, B.W.M., Rodríguez-Santiago, B., et al. (2011). De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nature Genetics* **43**(8), 729–731.
- Hong, D.-H., Pawlyk, B.S., Adamian, M. and Li, T. (2004). Dominant, gain-of-function mutant produced by truncation of RPGR. *Investigative Ophthalmology & Visual Science* **45**(1), 36–41.
- Hood, R.L., Lines, M.A., Nikkel, S.M., et al. (2012). Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *American Journal of Human Genetics* **90**(2), 308–313.
- Hoskinson, D.C., Dubuc, A.M. and Mason-Suarez, H. (2017). The current state of clinical interpretation of sequence variants. *Current Opinion in Genetics & Development* **42**, 33–39.
- Howe, D.G., Bradford, Y.M., Conlin, T., et al. (2013). ZFIN, the Zebrafish Model Organism Database: Increased support for mutants and transgenics. *Nucleic Acids Research* **41**(Database issue), D854–D860.
- Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G. and Yandell, M. (2013). VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology* **37**(6), 622–634.

- Huang, X.-F., Wu, J., Lv, J.-N., Zhang, X. and Jin, Z.-B. (2015). Identification of false-negative mutations missed by next-generation sequencing in retinitis pigmentosa patients: A complementary approach to clinical genetic diagnostic testing. *Genetics in Medicine* **17**(4), 307–311.
- Ioannidis, N.M., Rothstein, J.H., Pejaver, V., et al. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics* **99**(4), 877–885.
- Ito, Y., Carss, K.J.K.J., Duarte, S.T.S.T., et al. (2018). De novo truncating mutations in WASF1 cause intellectual disability with seizures. *American Journal of Human Genetics* **103**(1), 144–153.
- Kaur, G. and Dufour, J.M. (2012). Cell lines: Valuable tools or useless artifacts. *Spermatogenesis* **2**(1), 1–5.
- Karczewski, K., Francioli, L., Tiao, G., Cummings, B., Alföldi, J., Wang, Q., Collins, R., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 531210; doi: <https://doi.org/10.1101/531210>.
- Kawaji, H., Kasukawa, T., Forrest, A., Carninci, P. and Hayashizaki, Y. (2017). The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Scientific Data* **4**, 170113.
- Kelly, M.A., Caleshu, C., Morales, A., et al. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: Recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. *Genetics in Medicine* **20**(3), 351–359.
- Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**(3), 310–315.
- Köhler, S., Vasilevsky, N.A., Engelstad, M., et al. (2017). The Human Phenotype Ontology in 2017. *Nucleic Acids Research* **45**(D1), D865–D876.
- Koscielny, G., An, P., Carvalho-Silva, D., et al. (2017). Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research* **45**(D1), D985–D994.
- Krishnaraj, R., Ho, G. and Christodoulou, J. (2017). RettBASE: Rett syndrome database update. *Human Mutation* **38**(8), 922–931.
- Lambert, C.G. and Black, L.J. (2012). Learning from our GWAS mistakes: From experimental design to scientific method. *Biostatistics* **13**(2), 195–203.
- Lancaster, M.A., Renner, M., Martin, C.-A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P. and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. *Nature* **501**(7467), 373–379.
- Landrum, M.J., Lee, J.M., Benson, M., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**(D1), D1062–D1067.
- Landry, J.J.M., Pyl, P.T., Rausch, T., et al. (2013). The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda, Md.)* **3**(8), 1213–1224.
- Lek, M., Karczewski, K.J., Minikel, E.V., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**(7616), 285–91. doi: 10.1038/nature19057.
- Lenffer, J., Nicholas, F.W., Castle, K., Rao, A., Gregory, S., Poidinger, M., Mailman, M.D. and Ranganathan, S. (2006). OMIA (Online Mendelian Inheritance in Animals): An enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Research* **34**(Database issue), D599–D601.
- Li, M.M., Datto, M., Duncavage, E.J., et al. (2017). Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *Journal of Molecular Diagnostics* **19**(1), 4–23.

- Liu, X., Wu, C., Li, C. and Boerwinkle, E. (2016). dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation* **37**(3), 235–241.
- MacArthur, D.G., Balasubramanian, S., Frankish, A., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**(6070), 823–828.
- MacArthur, D.G., Manolio, T.A., Dimmock, D.P., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**(7497), 469–476.
- MacArthur, J., Bowler, E., Cerezo, M., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**(D1), D896–D901.
- MacArthur, J.A.L., Morales, J., Tully, R.E., et al. (2014). Locus Reference Genomic: Reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Research* **42**(Database issue), D873–D878.
- Majithia, A.R., Tsuda, B., Agostini, M., et al. (2016). Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics* **48**(12), 1570–1575.
- Matreyek, K.A., Starita, L.M., Stephany, J.J., et al. (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genetics* **50**(6), 874–882.
- Maude, S.L., Laetsch, T.W., Buechner, J., et al. (2018). Tisagenlecleucel in children and young adults with B-cell lymphoblastic leukemia. *New England Journal of Medicine* **378**(5), 439–448.
- McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**(5), 356–369.
- Merker, J.D., Wenger, A.M., Sneddon, T., et al. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine* **20**(1), 159–163.
- Milligan, C.J., Li, M., Gazina, E.V., et al. (2014). KCNT1 gain of function in 2 epilepsy phenotypes is reversed by quinidine. *Annals of Neurology* **75**(4), 581–90.
- Minikel, E.V., Vallabh, S.M., Lek, M., et al. (2016). Quantifying prion disease penetrance using large population control cohorts. *Science Translational Medicine* **8**(322), 322ra9.
- Mitchell, J.A. and McCray, A.T. (2003). The Genetics Home Reference: A new NLM consumer health resource. In *AMIA ... Annual Symposium Proceedings*. American Medical Informatics Association, Bethesda, MD.
- Myers, C.T., Stong, N., Mountier, E.I., et al. (2017). De Novo mutations in PPP3CA cause severe neurodevelopmental disease with seizures. *American Journal of Human Genetics* **101**(4), 516–524.
- Nakatsuka, N., Moorjani, P., Rai, N., et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. *Nature Genetics*, **49**(9), 1403–1407.
- Narasimhan, V.M., Hunt, K.A., Mason, D., et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**(6284), 474–477.
- Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., McDonald, M.T., Meisler, M.H. and Goldstein, D.B. (2012). Clinical application of exome sequencing in undiagnosed genetic conditions. *Journal of Medical Genetics* **49**(6), 353–361.
- Need, A.C., Shashi, V., Schoch, K., Petrovski, S. and Goldstein, D.B. (2017). The importance of dynamic re-analysis in diagnostic whole exome sequencing. *Journal of Medical Genetics* **54**(3), 155–156.
- Niknafs, N., Kim, D., Kim, R., Diekhans, M., Ryan, M., Stenson, P.D., Cooper, D.N. and Karchin, R. (2013). MuPIT interactive: Webserver for mapping variant positions to annotated, interactive 3D structures. *Human Genetics* **132**(11), 1235–1243.
- Niroula, A. and Vihinen, M. (2016). Variation interpretation predictors: Principles, types, performance, and choice. *Human Mutation* **37**(6), 579–597.
- Nykamp, K., Anderson, M., Powers, M., et al. (2017). Sherloc: A comprehensive refinement of the ACMG-AMP variant classification criteria. *Genetics in Medicine* **19**(10), 1105–1117.

- Pagani, L., Schiffels, S., Gurdasani, D., et al. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *American Journal of Human Genetics* **96**(6), 986–991.
- Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K.M., Gregg, A., Noggle, S. and Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**(7601), 125–129.
- Patel, R.Y., Shah, N., Jackson, A.R., et al. (2017). ClinGen Pathogenicity Calculator: A configurable system for assessing pathogenicity of genetic variants. *Genome Medicine* **9**(1), 3.
- Pe'er, I., Yelensky, R., Altshuler, D. and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology* **32**(4), 381–385.
- Pepin, M.G., Murray, M.L., Bailey, S., Leistritz-Kessler, D., Schwarze, U. and Byers, P.H. (2016). The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories. *Genetics in Medicine* **18**(1), 20–24.
- Petrovski, S. and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biology* **17**(1), 16–18.
- Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics* **9**(8).
- Petrovski, S., Gussow, A.B., Wang, Q., Halvorsen, M., Han, Y., Weir, W.H., Allen, A.S. and Goldstein, D.B. (2015). The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genetics* **11**(9).
- Petrovski, S., Küry, S., Myers, C.T., et al. (2016). Germline de novo mutations in GNB1 cause severe neurodevelopmental disability, hypotonia, and seizures. *American Journal of Human Genetics* **98**(5), 1001–1010.
- Petrovski, S., Todd, J.L., Durheim, M.T., et al. (2017). An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* **196**(1), 82–93.
- Pires, D.E. V., Ascher, D.B. and Blundell, T.L. (2014a). DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research* **42**(Web server issue), W314–W319.
- Pires, D.E. V., Ascher, D.B. and Blundell, T.L. (2014b). mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**(3), 335–342.
- Piton, A., Redin, C. and Mandel, J.-L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *American Journal of Human Genetics* **93**(2), 368–383.
- Poplin, R., Ruano-Rubio, V., DePristo, M.A., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint, bioRxiv 201178.
- Popovic, D., Sifrim, A., Davis, J., Moreau, Y. and De Moor, B. (2015). Problems with the nested granularity of feature domains in bioinformatics: The eXtasy case. *BMC Bioinformatics* **16**(Suppl. 4), S2.
- Posey, J.E., Rosenfeld, J.A., James, R.A., et al. (2016). Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genetics in medicine* **18**(7), 678–685.
- Quang, D., Chen, Y. and Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**(5), 761–763.
- Raghavan, N.S., Brickman, A.M., Andrews, H., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. *Annals of Clinical and Translational Neurology* **5**(7), 832–842.

- Rahbari, R., Wuster, A., Lindsay, S.J., et al. (2015). Timing, rates and spectra of human germline mutation. *Nature Genetics* **48**(2), 126–133.
- Rahimzadeh, V., Dyke, S.O.M. and Knoppers, B.M. (2016). An international framework for data sharing: Moving forward with the Global Alliance for Genomics and Health. *Biopreservation and Biobanking* **14**(3), 256–259.
- Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B. and Ayme, S. (2012). Representation of rare diseases in health information systems: The Orphanet approach to serve a wide range of end users. *Human Mutation* **33**(5), 803–808.
- Rauch, A., Wieczorek, D., Graf, E., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: An exome sequencing study. *Lancet* **380**(9854), 1674–1682.
- Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997). GeneCards: Integrating information about genes, proteins and diseases. *Trends in Genetics* **13**(4), 163.
- Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997). Improved splice site detection in Genie. *Journal of Computational Biology* **4**(3), 311–323.
- Rehm, H.L., Berg, J.S., Brooks, L.D., et al. (2015). ClinGen – the Clinical Genome Resource. *New England Journal of Medicine* **372**(23), 2235–2242.
- Richards, S., Aziz, N., Bale, S., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**(5), 405–423.
- Rivas, M.A., Pirinen, M., Conrad, D.F., et al. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**(6235), 666–669.
- Russell, S., Bennett, J., Wellman, J.A., et al. (2017). Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: A randomised, controlled, open-label, phase 3 trial. *Lancet* **390**(10097), 849–860.
- Sabatine, M.S., Giugliano, R.P., Wiviott, S.D., et al. (2015). Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. *New England Journal of Medicine* **372**(16), 1500–1509.
- Samocha, K.E., Robinson, E.B., Sanders, S.J., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**(9), 944–950.
- Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M. and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. Preprint, bioRxiv 148353.
- Sanchis-Juan, A., Stephens, J., French, C.E., et al. (2018). Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome medicine* **10**(1), 95.
- Schwarz, J.M., Rödelsperger, C., Schuelke, M. and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**(8), 575–576.
- Seim, I., Jeffery, P.L., Thomas, P.B., Nelson, C.C. and Chopin, L.K. (2017). Whole-genome sequence of the metastatic PC3 and LNCaP human prostate cancer cell lines. *G3 (Bethesda, Md.)* **7**(6), 1731–1741.
- Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C. and Telenti, A. (2018). Identification of misclassified ClinVar variants via disease population prevalence. *American Journal of Human Genetics* **102**(4), 609–619.
- Short, P.J., McRae, J.F., Gallone, G., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**(7698), 611–616.
- Siepel, A., Bejerano, G., Pedersen, J.S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**(8), 1034–1050.

- Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardeshtiravani, A., Sakai, R., Konings, P., Vermeesch, J.R., Aerts, J., De Moor, B. and Moreau, Y. (2013). eXtasy: Variant prioritization by genomic data fusion. *Nature Methods* **10**(11), 1083–1084.
- Smith, C.L., Blake, J.A., Kadin, J.A., Richardson, J.E., Bult, C.J. and Mouse Genome Database Group (2018). Mouse Genome Database (MGD)-2018: Knowledgebase for the laboratory mouse. *Nucleic Acids Research* **46**(D1), D836–D842.
- Sobreira, N., Schietecatte, F., Valle, D. and Hamosh, A. (2015). GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Human Mutation* **36**(10), 928–930.
- Spielmann, M. and Mundlos, S. (2016). Looking beyond the genes: The role of non-coding variants in human disease. *Human Molecular Genetics* **25**(R2), R157–R165.
- Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J. and Fowler, D.M. (2017). Variant interpretation: Functional assays to the rescue. *American Journal of Human Genetics* **101**(3), 315–325.
- Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D. and Cooper, D.N. (2017). The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics* **136**(6), 665–677.
- Stunnenberg, H.G., International Human Epigenome Consortium and Hirst, M. (2016). The International Human Epigenome Consortium: A blueprint for scientific collaboration and discovery. *Cell* **167**(5), 1145–1149.
- Sudmant, P.H., Rausch, T., Gardner, E.J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* **526**(7571), 75–81.
- Szklarczyk, D., Franceschini, A., Wyder, S., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* **43**(Database issue), D447–D452.
- Taber, J.M., Klein, W.M.P., Lewis, K.L., Johnston, J.J., Biesecker, L.G. and Biesecker, B.B. (2018). Reactions to clinical reinterpretation of a gene variant by participants in a sequencing study. *Genetics in Medicine* **20**(3), 337–345.
- Takashima, K., Maru, Y., Mori, S., Mano, H., Noda, T. and Muto, K. (2018). Ethical concerns on sharing genomic data including patients' family members. *BMC Medical Ethics* **19**(1), 61.
- Taliun, D., Harris, D., Kessler, M., Carlson, J., Szpiech, Z., Torres, R., Gagliano Taliun, S., et al. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv 563866; doi: <https://doi.org/10.1101/563866>.
- Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M. and Biesecker, L.G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine* **20**(9), 1054–1060.
- Telenti, A., Pierce, L.C.T., Biggs, W.H., et al. (2016). Deep Sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **113**(42), 11901–11906.
- Thorogood, A., Cook-Deegan, R. and Knoppers, B.M. (2017). Public variant databases: liability? *Genetics in Medicine* **19**(7), 838–841.
- Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**(2), 178–192.
- Traynelis, J., Silk, M., Wang, Q., Berkovic, S.F., Liu, L., Ascher, D.B., Balding, D.J. and Petrovski, S. (2017). Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Research* **27**(10), 1715–1729.

- Trehearne, A. (2016). Genetics, lifestyle and environment. UK Biobank is an open access resource following the lives of 500,000 participants to improve the health of future generations. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **59**(3), 361–367.
- Trudeau, M.M., Dalton, J.C., Day, J.W., Ranum, L.P.W. and Meisler, M.H. (2006). Heterozygosity for a protein truncation mutation of sodium channel SCN8A in a patient with cerebellar atrophy, ataxia, and mental retardation. *Journal of Medical Genetics* **43**(6), 527–30.
- Turnbull, C., Scott, R.H., Thomas, E., et al. (2018). The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *British Medical Journal (Clinical Research Ed.)* **361**, k1687.
- UniProt Consortium (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **46**(5), 2699.
- Van der Auwera, G.A., Carneiro, M.O., Hartl, C., et al. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.
- Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M. and Ng, P.C. (2016). SIFT missense predictions for genomes. *Nature Protocols* **11**(1), 1–9.
- Vaz-Drago, R., Custódio, N. and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. *Human Genetics* **136**(9), 1093–1111.
- Vears, D.F., Sénéchal, K., Clarke, A.J., et al. (2018). Points to consider for laboratories reporting results from diagnostic genomic sequencing. *European Journal of Human Genetics* **26**(1), 36–43.
- Veeramah, K.R., O'Brien, J.E., Meisler, M.H., et al. (2012). De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *American Journal of Human Genetics* **90**(3), 502–510.
- Vervoort, R., Lennon, A., Bird, A.C., Tulloch, B., Axton, R., Miano, M.G., Meindl, A., Meitinger, T., Ciccodicola, A. and Wright, A.F. (2000). Mutational hot spot within a new RPGR exon in X-linked retinitis pigmentosa. *Nature Genetics* **25**(4), 462–466.
- Walsh, R., Thomson, K.L., Ware, J.S., et al. (2017). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in Medicine* **19**(2), 192–203.
- Walter, K., Min, J.L.J.L., Huang, J., et al. (2015). The UK10K project identifies rare variants in health and disease. *Nature* **526**(7571), 82–90.
- Wang, S., Jiang, X., Singh, S., Marmor, R., Bonomi, L., Fox, D., Dow, M. and Ohno-Machado, L. (2017). Genome privacy: Challenges, technical approaches to mitigate risk, and ethical considerations in the United States. *Annals of the New York Academy of Sciences* **1387**(1), 73–83.
- Whiffin, N., Minikel, E., Walsh, R., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine* **19**(10), 1151–1158.
- Whiffin, N., Walsh, R., Govind, R., et al. (2018). CardioClassifier: Disease- and gene-specific computational decision support for clinical genome interpretation. *Genetics in Medicine* **20**(10), 1245–1254.
- Wilfert, A.B., Chao, K.R., Kaushal, M., Jain, S., Zöllner, S., Adams, D.R. and Conrad, D.F. (2016). Genome-wide significance testing of variation from single case exomes. *Nature Genetics* **48**(12), 1455–1461.
- Worth, C.L., Preissner, R. and Blundell, T.L. (2011). SDM – a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research* **39**(Web server issue), W215–W222.
- Wright, C.F., Hurles, M.E. and Firth, H.V. (2016). Principle of proportionality in genomic data sharing. *Nature Reviews Genetics* **17**(1), 1–2.
- Wu, J. and Jiang, R. (2013). Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *ScientificWorldJournal* **2013**, 675851.

- Yeo, G. and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* **11**(2–3), 377–394.
- Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics* **10**, 451–481.
- Zhu, X., Petrovski, S., Xie, P., et al. (2015). Whole-exome sequencing in undiagnosed genetic diseases: Interpreting 119 trios. *Genetics in Medicine* **17**(10), 774–781.
- Zhu, X., Padmanabhan, R., Copeland, B., et al. (2017). A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLoS Genetics* **13**(11), e1007104.
- Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America* **111**(4), E455–E464.

28

Prediction of Phenotype from DNA Variants

M.E. Goddard,¹ T.H.E. Meuwissen,² and H.D. Daetwyler³

¹ Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Victoria, Australia, and Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia

² Norwegian University of Life Sciences, Ås, Norway

³ Agriculture Victoria, AgriBio, Bundoora, Victoria, Australia, and School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, Australia

Abstract

Complex or quantitative traits are typically controlled by thousands of polymorphic sites in the genome and by environmental factors. It would be useful in agriculture and in medicine to be able to predict the genetic value or future phenotype of individuals from information on the alleles they carry at variable DNA sites. This can be done by using a panel of single nucleotide polymorphisms (SNPs) that cover the whole genome so that all sites which cause variation in the trait are in linkage disequilibrium with one or more of the SNPs. A prediction equation is derived from a training population in which individuals have been assayed for the SNP genotypes and recorded for the trait. The most accurate methods fit all marker or SNP effects simultaneously and treat these effects as random variables drawn from a specified distribution. This prior distribution may be a simple normal distribution or a mixture of distributions including a spike at zero so that some SNPs have no effect on the trait. The accuracy of the prediction depends on many variables, especially the sample size of the training population and the effective population size (N_e) of the population in which predictions are to be made. Recent N_e is important because it controls the range of linkage disequilibrium among variable sites in the genome, including those that cause variation in the trait. We consider the nature of genetic variation affecting phenotype, the DNA polymorphism data available for prediction, methods of prediction and their accuracy, and examples of the use of prediction of phenotype.

28.1 Introduction

There are many situations in which we would like to predict the phenotype of an individual. In humans we might wish to predict whether or not an individual will suffer from a heart attack in the next year. In livestock and plants we might wish to predict the performance of the offspring of existing individuals so that those with the most desirable offspring can be chosen as parents. In this case it is the additive genetic value or breeding value of the existing plants and animals that we wish to predict. The variation in phenotype among individuals is due to variation in their DNA sequence and to variation in other factors affecting the phenotype that are put under the broad heading of 'environmental' factors. Knowledge of the alleles that an individual carries

at DNA polymorphisms can, of course, only predict the genetic value of the individual for a trait. Therefore, the accuracy with which the phenotype is predicted by DNA polymorphisms is limited because the impact of environmental factors is ignored. However, this limitation may be overcome by combining genetic predictors with predictors of environmental influences. Here, we will limit ourselves to the genetic predictors. In the case where we are predicting the effects of natural or artificial selection, the prediction of genetic effects suffices since these are the effects that cause the genetic change or adaptation of the population.

In this chapter we consider the nature of genetic variation affecting phenotype, the DNA polymorphism data available for prediction, methods of prediction and their accuracy, and examples of the use of prediction of phenotype.

28.2 Genetic Variation Affecting Phenotype

Some phenotypes can be explained by mutations at a single gene. For instance, in humans the disease phenylketonuria can be caused by mutations in the *PAH* gene for the enzyme phenylalanine hydroxylase. If the mutations that cause the phenotype are known, they can be used to assist in diagnosis, to predict the future phenotype of the individual, and to detect carriers of a recessive mutation (i.e. to predict the phenotype of offspring of the individual). However, many traits are not like this. Complex traits are affected by many genes, where each gene only explains a relatively small proportion of variation in the trait, and by environmental factors. In some cases, these many sources of variation result in a continuously variable trait such as height in humans that varies from short to tall. In other cases, despite similar underlying complexity, we only observe the phenotype in one or a limited number of states. For instance, we may classify individuals as either unaffected by a disease or suffering from that disease. In this case the many genetic and non-genetic factors contribute to a continuously variable susceptibility to the disease even though we only observe the outcome as 'affected' or 'unaffected'. The task of predicting genetic values and hence phenotypes for a complex trait is more difficult than for traits caused by a single mutation. In theory, if we knew all polymorphisms that affect the trait, we could estimate the genetic value associated with all possible combinations of alleles at these sites and hence predict genetic values. In practice this is not possible, since we do not know all causal polymorphisms and there may be too many to estimate all these effects accurately. The remainder of this review will concentrate on more practical methods for predicting genetic value for complex traits.

Over the last decade, the availability of assays for thousands of single nucleotide polymorphisms (SNPs) has greatly increased our understanding of the genetics of complex traits. These assays, based on microarrays containing a panel of random SNPs ('SNP chips'), have been widely used to conduct genome-wide association studies (GWASs; see **Chapter 21**). In a GWAS, a sample of individuals are recorded for the trait of interest and assayed for their genotype at a panel of SNPs. The intention is that the panel of SNPs covers the whole genome so that polymorphisms affecting the trait (causal variants) will be in linkage disequilibrium (LD) with at least one SNP on the panel. This LD will generate an association between the trait and the SNP which is detected by the analysis of the results. As well as mapping many causal variants for complex traits, GWASs have led to general conclusions about the architecture of complex traits which we now briefly summarise.

Most complex traits are affected by thousands of polymorphisms, most of which have very small effects (Yang *et al.*, 2010; Wood *et al.*, 2014). For instance, Robinson *et al.* (2017) found that 75% of the variance of body mass index (BMI) explained by SNPs was due to SNPs that

individually explain less than 0.01% of the genetic variance. Consequently, large sample sizes are needed to estimate these effects with any precision, and this explains why earlier association studies with smaller samples yielded unreliable results. To protect against false positives due to multiple testing, a stringent significance threshold ($p < 5 \times 10^{-8}$) is used when testing for an association between any one SNP and the trait. Since only a limited fraction of the variance is due to variants with large effects, the stringent significance tests imply that the SNPs declared significant explain little of the known genetic variance (Manolio *et al.*, 2009) because many SNPs with real associations with the trait fail to reach significance. However, if the variance explained by all SNPs together is estimated, it is closer to the total genetic variance, but typically still less than that estimated from family studies (Yang *et al.*, 2010). There are possible biases in both the SNP-based and family-based estimates of genetic variance, but one likely explanation for the discrepancy between the two (the ‘missing heritability’) is that the SNPs are not in complete LD with all the causal variants, possibly because some of the causal variants have rare alleles (Yang *et al.*, 2010). The effects of incomplete LD between SNPs and causal variants on prediction of genetic effects are discussed below.

Theory and experimental results suggest that most of the genetic variation for complex traits in outbreeding populations is due to the additive effects of individual alleles (Hill *et al.*, 2008). That is, non-additive effects such as dominance and epistasis explain a minority of the genetic variance. This is fortunate for prediction because it means that we can use predictors that linearly combine the effects of all alleles and do not need to estimate the effect of every possible combination of alleles.

28.3 Data on DNA Polymorphisms Used for Prediction of Genetic Effects

Until recently the success of prediction of genetic effects was limited due to the small number of polymorphisms that could be assayed. Polymorphisms were chosen for experiments because an assay existed (e.g. blood groups) or because they were in a candidate gene, and tested for association with the trait of interest. Only a few polymorphisms found in this way proved useful for prediction. They were comparatively rare cases where a single polymorphism had a large effect on the trait (e.g. resistance to infectious pancreatic necrosis virus in Atlantic salmon: Houston *et al.*, 2008; Moen *et al.*, 2015). The relatively low marker density of the available assays (e.g. microsatellites) led to the use of linkage experiments instead of association studies. Within a family, a panel of 200 microsatellites can be enough to detect linkage between a microsatellite and a causal polymorphism. Linkage between a marker and a causal variant is indicated when there is an association between them within a family even though there is no consistent association across the whole population. A method of using this linkage for prediction of phenotype was proposed (Fernando and Grossman, 1989) but seldom used because, among other reasons, the phase of linkage has to be estimated for every family where a prediction is to be made. The availability of SNP chips with much larger numbers of genetic markers overcame this problem because the higher LD between the SNP and the causal polymorphism implies a more consistent association between the SNP and the trait within that population. In theory, a complete genome sequence should be superior to an SNP panel for prediction because the causal variants should be included in the sequence and hence one does not need to rely on LD. In practice, this benefit may be reduced if some classes of sequence variants (e.g. structural variants) are not included in the data. Currently, prediction is usually based on sequence data (imputed or assayed) or SNP panels and/or individual polymorphisms thought to be associated with the trait.

28.4 Prediction of Additive Genetic Values

Under a squared error loss function, the best prediction of additive genetic values or breeding values (\mathbf{g}) that can be made from a specified set of data is simply the expected value of \mathbf{g} conditional on the data ($E(\mathbf{g} | \text{data})$). Suppose that the data consist of the phenotypes (\mathbf{y}) and M SNP genotypes (\mathbf{Z}) from a sample of N individuals, and we assume a linear model

$$\mathbf{y} = \mathbf{g} + \mathbf{e} \quad \text{and} \quad \mathbf{g} = \mathbf{Z}\mathbf{b}, \quad (28.1)$$

where \mathbf{y} is an $N \times 1$ vector of phenotypic values, \mathbf{g} an $N \times 1$ vector of breeding values, $\mathbf{e} \sim N(\mathbf{0}, I\sigma_e^2)$ an $N \times 1$ vector of non-genetic or environmental effects, \mathbf{Z} an $N \times M$ matrix of SNP genotypes and \mathbf{b} an $M \times 1$ vector of effects of the SNPs on breeding value. Finding the best predictor of \mathbf{g} is equivalent to finding the best predictor of \mathbf{b} , so we want to estimate \mathbf{b} by $\hat{\mathbf{b}} = E(\mathbf{b} | \mathbf{y}, \mathbf{Z})$ and $\hat{\mathbf{g}} = \mathbf{Z}\hat{\mathbf{b}}$. We have

$$E(\mathbf{b} | \mathbf{y}, \mathbf{Z}) = \frac{\int \mathbf{b} f(\mathbf{y} | \mathbf{b}, \mathbf{Z}) f(\mathbf{b}) d\mathbf{b}}{\int f(\mathbf{y} | \mathbf{b}, \mathbf{Z}) f(\mathbf{b}) d\mathbf{b}}, \quad (28.2)$$

where $f(\mathbf{b})$ is the prior probability density of \mathbf{b} and $f(\mathbf{y} | \mathbf{b}, \mathbf{Z})$ the probability density of \mathbf{y} conditional on \mathbf{b} and \mathbf{Z} .

Thus equation (28.2) assumes that the effects of SNPs on the trait are random variables. Since there are often more predictors (i.e. SNPs) than data points, treating the SNP effects as fixed effects leads to unstable estimates of \mathbf{b} and low accuracy of prediction (Meuwissen *et al.*, 2001). This is a standard Bayesian perspective but it is also justified from a frequentist perspective – there are thousands of SNP effects, so it is reasonable to consider any one of them as a sample from a population of values. Implementation of equation (28.2) requires that we specify the distribution of \mathbf{b} (i.e. $f(\mathbf{b})$). Choice of $f(\mathbf{b})$ leads to a variety of different methods of prediction.

Meuwissen *et al.* (2001) considered three options for $f(\mathbf{b})$ and the resulting methods of prediction. If $f(\mathbf{b})$ is assumed to be a normal distribution with constant variance ($\mathbf{b} \sim N(\mathbf{0}, I\sigma_b^2)$), the method is an example of best linear unbiased prediction (BLUP; note that the predictor \mathbf{b} is a linear function of the data \mathbf{y}). This assumes that all SNP effects are small. An alternative that allows some SNPs to have bigger effects is to assume that \mathbf{b} follows a t distribution. This leads to solutions for $\hat{\mathbf{b}}$ which are no longer linear in \mathbf{y} , and Meuwissen *et al.* called the method Bayes A. Both these methods assume that no SNPs have zero effect on the trait, although the modal effect size is zero and many SNPs have effects close to zero. Since it is expected that many SNPs are not in LD with causal variants, Meuwissen *et al.* allowed for zero effects by assuming that \mathbf{b} could follow a mixture of zero and a t distribution and called this method Bayes B. Bayes A and B were implemented by Markov chain Monte Carlo methods which require more computing time than BLUP.

Subsequently several other possibilities for the distribution of \mathbf{b} have been tried (de los Campos *et al.*, 2013). For instance, Bayes C (Verbyla *et al.*, 2009; Habier *et al.*, 2011) assumes a mixture of zero and a normal distribution, while Bayes R (Erbe *et al.*, 2012) assumes a mixture of zero and three normal distributions with increasing variance to flexibly model a range of possible distributions. If the distribution of \mathbf{b} is assumed to be a Laplace distribution the analysis is a Bayesian lasso. Other prior distributions of SNP effects lead to other Bayesian methods such as those of Zhou *et al.* (2013) (BLSMM) and Speed and Balding (2014) (MultiBLUP). It is possible that the distribution of \mathbf{b} depends on the part of the genome in which an SNP is located. For instance, LDAK (Speed *et al.*, 2017) makes the effect dependent on the local LD of the genomic region and MultiBLUP allows regions with higher and lower variance to be discovered by the analysis. The SNP genotypes can also be coded in multiple ways. For instance, they

can be coded simply by the number of reference alleles, resulting in genotypes coded 0, 1 or 2. Alternatively, these values can be standardised so that all SNP genotypes have a variance of 1.0. This implies larger $|\mathbf{b}|$ values for SNPs with low minor allele frequency (MAF). An alternative is to group SNPs by MAF and LD and estimate parameters for each group (Yang *et al.*, 2015). Although these choices have large effects on some parameter estimates, they do not seem to affect the accuracy of prediction greatly.

An intuitive way to think about the differences between these methods is as follows. When an SNP effect is treated as a random variable the estimate is similar to a fixed effect estimate that has been shrunk or regressed back towards zero. In BLUP the relative amount of shrinkage is the same regardless of the estimated size of the effect. However, the nonlinear or Bayesian methods do not shrink all estimates equally. Typically, large estimates are shrunk less than small estimates, since large estimates are more likely to reflect true effects. Consequently, under BLUP all SNP effect estimates are somewhat similar in magnitude, whereas under nonlinear models some SNP effect estimates remain large and others are shrunk almost to zero. SNPs that are a long distance from a causal variant are typically in lower LD with it than SNPs that are close to the causal variant, and, consequently, these distant SNPs tend to receive smaller effect estimates under a nonlinear model than under BLUP.

Simulation studies, not surprisingly, find the accuracy of prediction is maximised when the assumed distribution of \mathbf{b} matches the distribution used to generate the data (Daetwyler *et al.*, 2010; Hayes *et al.*, 2010). In real data we do not know the distribution of \mathbf{b} . Tests on real data on livestock find that the nonlinear or Bayesian methods give accuracy as high as or higher than BLUP (Erbe *et al.*, 2012; Kemper *et al.*, 2015; Moser *et al.*, 2015; MacLeod *et al.*, 2016) and that there is little consistent difference between the various nonlinear methods. The nonlinear methods tend to outperform BLUP for traits with polymorphisms of large effect segregating or when multiple breeds are combined or when the prediction is tested in a breed not included in the training population. These results can be explained as follows. When there are some polymorphisms of large effect segregating, the distribution of SNP effects is closer to that assumed by the nonlinear methods and so they give higher accuracy. Within a breed of livestock, LD extends over millions of base pairs of DNA and consequently the effect of a causal variant can be predicted by a linear combination of SNPs over 1 Mbp or more of DNA. However, the phase of long-range LD varies between breeds (De Roos *et al.*, 2008; Vander Jagt *et al.*, 2018) so, in a mixture of breeds, a prediction that gives more weight to SNPs near the causal variants is better than one that gives all SNPs approximately equal weight. Nevertheless, the good performance of BLUP emphasises the large number of causal variants that contribute to variation in most complex traits.

Although it seems logical to start from equation (28.2) since it gives the best prediction, some heuristic methods are also commonly used. In human genetics a common method is called the ‘polygenic risk score’ (PRS). This method selects SNPs that are significant at some threshold and then estimates the effect of each SNP by fitting the SNP as a fixed effect in a model with no other SNPs fitted or with all other SNPs fitted as random effects using BLUP. Consequently, the effect of each SNP tends to capture the effects of all causal variants with which it is in LD. Each causal variant may be captured more than once by different SNPs so that the score exaggerates the differences between individuals. On the other hand, if only a few SNPs are chosen much of the total genetic variance may be missed and the selected SNPs are usually chosen because they had the biggest effects out of a set of many thousands of SNPs, which implies that their estimates suffer from the winner’s curse and hence are overestimated (Beavis, 1998). So the final predictor will exaggerate the differences between individuals. This can be cured by regressing back the final predictor towards the mean so its variance is reduced. This PRS prediction index would be similar to BLUP if the LD surrounding all SNPs were similar. This is not the case.

However, Moser *et al.* (2015) found the prediction index to give accuracies in between BLUP and the nonlinear Bayesian methods. The PRS has the advantage that it can be calculated from summary data (i.e. estimated SNP effects) without access to individual-level data. However, other methods can also be applied to summary-level data provided that a small reference data set is available from which to estimate LD among the SNPs.

Another general-purpose prediction method that can be applied to SNP data is partial least squares (PLS). PLS is a linear predictor and neglects linear components of the genotypes that are of lesser importance to the prediction of the phenotype. BLUP does this also, but instead of neglecting such components completely it down-weights components that are less important based on the variance that they explain. PLS and BLUP result therefore in similar predictions (Hastie *et al.*, 2001).

Machine learning methods have also been applied to prediction but do not seem to have any advantage over the linear model methods described above. These methods could include non-additive interactions between alleles and loci in the prediction of genetic value. Attempts have also been made to include these non-additive effects in conventional models, but they do not generally increase prediction accuracy (Aliloo *et al.*, 2016). An exception is that inclusion of an effect of overall heterozygosity to account for inbreeding depression may be useful (Aliloo *et al.*, 2016).

Two methods of coding the SNP genotypes are in common use (Strandén and Christensen, 2011). One is by the number of reference alleles at each SNP so that the three genotypes are coded 0, 1 and 2. Alternatively, these numbers can be standardised so that all SNP variables have the same variance (by dividing by the square root of the heterozygosity of the SNP). When used in combination with BLUP, the first coding implies that SNP effects follow the same distribution regardless of allele frequency and so SNPs with high MAF explain more variance. If we standardise, then we assume that SNPs with low MAF explain the same variance as SNPs with high MAF because they have larger effects per allele. Tests of the relationship between MAF and allele effects suggest that the truth lies in between the two alternatives (Speed *et al.*, 2017; Zeng *et al.*, 2018). In practice, the accuracy of prediction does not seem to be much affected by the coding chosen, perhaps because rare SNPs contribute little to the prediction.

28.4.1 An Equivalent Model

The BLUP model (28.1) has an explicit term for every SNP and could be called a random regression model. An equivalent model uses the SNP genotypes to describe the relationships among the individuals; that is,

$$\mathbf{y} = \mathbf{g} + \mathbf{e}, \quad V(\mathbf{g}) = \mathbf{Z}\mathbf{Z}'\sigma_b^2. \quad (28.3)$$

This form may be computationally less demanding when the number of individuals (N) is less than the number of SNPs (M) because $\mathbf{Z}\mathbf{Z}'$ is an $N \times N$ matrix whereas the mixed model equations, based on equation (28.1) and containing the SNP effects explicitly, yield an $M \times M$ matrix. The form of the equations based on (28.3) is referred to as a genomic BLUP (GBLUP; VanRaden, 2008). Estimation of the additive genetic values (\mathbf{g}) in this model requires inversion of the $N \times N$ genomic relationship matrix (GRM, $\mathbf{Z}\mathbf{Z}'$) which is computationally demanding if N is large. However, Misztal and Legarra (2017) point out that the rank of $\mathbf{Z}\mathbf{Z}'$ is likely to be much less than N and this can be used to efficiently invert the matrix.

28.4.2 Single-Step BLUP

It is often the case that not all individuals with phenotypic records also have SNP genotypes. When the ungenotyped individuals have known relatedness with genotyped individuals, the GBLUP model can be extended to include them (Aguilar *et al.*, 2010; Christensen and Lund,

2010). This model is called single-step GBLUP. It is equivalent to using a linear mixed model to predict the missing genotypes from the genotyped relatives of an ungenotyped individual. A problem with the single-step methods is that the relationship matrix derived from pedigrees (A) and the relationship matrix derived from SNP genotypes ($Z'Z$) typically use different bases (i.e. the A matrix base is the animals at the top of the pedigree and the GRM base is the population whose allele frequencies are used to standardise the genotypes in the Z matrix). Often all genotyped individuals are used to calculate these allele frequencies so the base is implicitly the population of genotyped individuals. However, a more logical base would be that used in the A matrix, that is the individuals that are ancestors of the other genotyped individuals. Then the base individuals appear unrelated and the more recent individuals show relationships due to pedigree connections. However, estimation of allele frequencies in the base population is difficult if not all base animals are genotyped. Legarra *et al.* (2015) propose a solution. They assume that the allele frequencies of the markers in a hypothetical base population are all 0.5. They calculate the GRM with these allele frequencies and then adjust the A matrix so that it uses the same base. The method deals with genetic groups such as breeds by incorporating founder populations (meta-founders) that are derived from the hypothetical base but can be related to each other as estimated from the GRM. The use of base allele frequencies of 0.5 means that the inbreeding estimated from the GRM agrees with that estimated from the average homozygosity of individuals.

Imputation of missing genotypes by using a linear mixed model is not the best way to impute missing genotypes, especially for ungenotyped ancestors with many genotyped descendants. Meuwissen *et al.* (2015) use segregation analysis to impute genotypes for these ungenotyped ancestors and find that this leads to higher accuracy of Estimated Breeding Values (EBVs) than single-step GBLUP. However, the computational efficiency of single-step GBLUP makes it attractive for data sets with many ungenotyped individuals. In some cases the gain in accuracy from single step (i.e. from including ungenotyped individuals) is not great but it can account for selection and avoid double counting and make all EBVs of genotyped and ungenotyped individuals comparable (Misztal and Legarra, 2017).

Fernando *et al.* (2014, 2016) and Liu *et al.* (2014) have described single-step methods that directly estimate the SNP effects rather than the breeding values and which may be computationally advantageous in some circumstances – for instance, if the number of SNPs (M) is less than the number of individuals (N) and the $M \times M$ matrix $Z'Z$ is small enough to be held in core memory.

28.4.3 Multiple Traits

Another advantage of the single-step GBLUP approach is that it can easily be extended to multi-trait analyses, which tend to be beneficial when traits are genetically correlated. The gain in prediction accuracy is largest when highly correlated traits are recorded on different individuals so that the multi-trait analysis effectively increases the size of the training population (Calus and Veerkamp, 2011). In the Bayesian context this can also be implemented but requires assumptions about the distribution of the number of traits affected by an SNP and whether some trait pairs are more likely to be affected by the same SNPs than others. It may be assumed here that if the traits are analysed together in a single analysis, they are highly related traits, and if an SNP affects one of the traits, it is also expected to affect the others. In effect the assumption is that if it is clear that an SNP affects one of the traits in the Bayesian analysis, it is assumed that its effect on the other traits is non-zero, although the data may still result in an estimate that is close to zero (Kemper *et al.*, 2018). Whether the latter assumption or alternative assumptions about the SNPs coaffecting traits yields best predictions will depend on how close these assumptions are to the true underlying genetic model.

In order to implement multi-trait methods estimates of genetic correlations are needed. When individual-level data are available, restricted maximum likelihood is frequently used. When only summary-level data are available, recently developed methods can be used (Vilhjálmsson *et al.*, 2015).

In some cases a trait that is not of importance in its own right may still improve predictions of a correlated and important trait. For instance, intermediate phenotypes that connect a causal polymorphism to the important phenotype may lead to greater accuracy of prediction, especially if the causal variant has a large effect on the intermediate phenotype (Kemper *et al.*, 2016).

28.4.4 Gene Expression

One particular group of traits is the measurement of mRNA concentration of each gene in a given cell type or tissue. Here the trait (gene expression) is not of interest in itself but may help identify causal variants for other traits or at least the gene through which the causal variants acts. This concept relies on the assumption that some causal variants for conventional phenotypes are actually expression quantitative trait loci (Zhu *et al.*, 2016).

28.4.5 Using External Information

In estimating the effect of SNPs, one might wish to use external information. Two kinds of external data are information on which genes affect a given trait and information about sites in the genome that have evidence for function. The latter is most useful when using sequence data so that the direct effect of the polymorphism should be the effect in the model. For instance, one might assume that polymorphisms that affect the amino acid sequence of a protein are likely to have a greater effect than synonymous coding polymorphisms. The model can account for this by splitting the SNPs into two or more groups and assuming a different distribution of effects for each group (e.g. Speed and Balding, 2014). That is,

$$\mathbf{y} = \mathbf{g} + \mathbf{e}, \quad \mathbf{g} = \mathbf{Z}_1 \mathbf{b}_1 + \mathbf{Z}_2 \mathbf{b}_2.$$

In BLUP $\mathbf{b}_1 \sim N(\mathbf{0}, I\sigma_{b_1}^2)$ and $\mathbf{b}_2 \sim N(\mathbf{0}, I\sigma_{b_2}^2)$, with the variances estimated from the data. In a method called Bayes RC, \mathbf{b}_1 and \mathbf{b}_2 are both distributed as a mixture of four normal distributions but with different mixing proportions which are estimated from the data (MacLeod *et al.*, 2016). Simulation studies show that taking advantage of this external information can increase prediction accuracy, but there is limited evidence of this from real data.

28.5 Factors Affecting Accuracy of Prediction

Once phenotypic and genetic data have been collected, the best way to assess the accuracy of a prediction algorithm is to apply it to a new, independent set of individuals (not used in training the algorithm) and to compare the predicted and observed phenotypes. However, we would like to understand the factors controlling accuracy before all data are collected, so that the best decisions on the data to be collected and the method of analysis can be made. If we have collected genotypes but not phenotypes and we propose to use BLUP for the analysis, we can predict the accuracy from the theory of linear mixed models. If the SNPs collectively explain all the genetic variance and if all SNPs are in linkage equilibrium and genotypes have been standardised so that all genotypes have variance 1, then the correlation squared between the true effect of an SNP (\mathbf{b}) and the estimated effect ($\hat{\mathbf{b}}$) is

$$R^2 = \frac{N}{N + \lambda}, \tag{28.4}$$

where N is the number of individuals with phenotype and genotype in the training set and $\lambda = \sigma^2/\sigma_b^2$ (Daetwyler *et al.*, 2008; Goddard, 2009). In this artificial case the accuracy of estimating the breeding value ($\mathbf{g} = \mathbf{Z}\mathbf{b}$) is the same as the accuracy of estimating the SNP effect (\mathbf{b}). In reality the SNPs are not in linkage equilibrium. Fortunately, this simple approach can be made more realistic by considering the effects of independent chromosome segments each containing many SNPs but modelled as one ‘effective’ marker. In a random mating population of constant effective population size (N_e), the number of effective chromosome segments (M_e) can be predicted from the LD among markers:

$$M_e = \frac{1}{\sum r^2},$$

where r^2 is the squared correlation between a pair of markers (a measure of LD) and the sum is over all pairs of markers in the data. Alternatively, M_e can be estimated from N_e and the number of chromosomes in the genome (n) and their average length in morgans (L):

$$M_e = \frac{2nLN_e}{\log(2N_eL)},$$

assuming $r^2 = 1/(4N_e c + 1)$ for a population with inbreeding due to a reduction in N_e , or

$$M_e = \frac{2nLN_e}{\log(N_e L)},$$

assuming $r^2 = 1/(4N_e c + 2)$ for a population at constant N_e , where c is the distance between markers in morgans.

These analytical predictions of accuracy are based on standardised genotypes so that loci explain the same variance regardless of MAF. If effect sizes are assumed to be independent of MAF, then loci with high MAF explain more variance than loci with low MAF. Under these circumstances, accuracy is predicted to be higher because the effects of SNPs with high MAF are estimated more accurately and explain more variance (Goddard, 2009).

However, in real populations with changing N_e and non-random mating, these predictions of M_e and hence accuracy are only approximate. However, they are still useful. For instance, they predict correctly that the accuracy of genomic prediction is higher in breeds of livestock that have small recent N_e than in humans with large recent N_e . This occurs despite similar levels of polymorphism and historical N_e and emphasises the importance of recent N_e in creating long-range LD equivalent to low M_e .

In equation (28.4), σ^2 is the phenotypic variance, ignoring any reduction in error variance achieved by fitting all markers simultaneously. Daetwyler *et al.* (2008) correct this by using $\sigma^2(1 - r^2)$ instead of σ^2 . Elsen (2017) gives another derivation of the same result.

If the markers do not explain all of the genetic variance for a trait, then the accuracy is further reduced. If the markers explain a fraction ρ^2 of the genetic variance, then the maximum accuracy with which the genetic value of an individual can be predicted from the markers alone is ρ (Erbe *et al.*, 2013). If the markers have similar properties, such as MAF, to the causal variants then $\rho^2 = M_e/(M_e + M)$ (Yang *et al.* 2010). However, if the LD between causal variants and markers is less than the LD between markers themselves, then ρ^2 is less than this. This difference in LD is expected if causal variants have lower MAF than the markers, as appears to occur to some extent (Speed *et al.*, 2017; Zeng *et al.*, 2018).

Putting together some of these points, the prediction of the squared accuracy (R^2) can be written as (MacLeod *et al.* 2014)

$$R^2 = \frac{\rho^2 \theta}{\theta + 1 - h^2 R^2}, \quad (28.5)$$

where $\theta = \rho^2 Nh^2/M_e$. Although not an explicit formula for R^2 , this equation makes clear how the residual variance decreases as R^2 increases.

The above predictions of accuracy apply to using BLUP to estimate genetic value of individuals. If nonlinear or Bayesian methods are used the effects of N and h^2 are similar to those described above for BLUP.

The benefit of using Bayesian methods instead of BLUP depends on the distribution of the effects of 'effective chromosome segments'. If the number of causal variants exceeds the number of effective chromosome segments (M_e) then there may be no advantage of Bayesian methods over BLUP because the distribution of the effects of chromosome segments may be close to normal, which is the prior assumed by BLUP (Daetwyler *et al.*, 2010). In cattle this happens within a breed where the number of causal variants appears to be similar to M_e . However, even in this situation, the Bayesian methods are superior to BLUP if there are some causal variants of large effect, because then the distribution of chromosome segment effects is no longer normal. This occurs, for instance, in the trait of cattle milk fat concentration, where a polymorphism in the *DGAT* gene explains ~40% of the genetic variance (Grisart *et al.*, 2002, 2004), and in human diseases such as rheumatoid arthritis where the MHC region has a large effect (Moser *et al.*, 2015). When data from multiple breeds are combined, the effective number of chromosome segments is much larger than in a single breed (e.g. $M_e = 60,000$). Consequently, if the number of individuals and the number of markers remain constant, the accuracy of prediction is reduced for two reasons. Firstly, the proportion of genetic variance explained by the SNPs ($M/(M_e + M)$) is reduced. This is another way of saying that a panel of 50,000 SNPs is not dense enough to guarantee that all genetic variance is explained by the SNP. Secondly, the parameter $\theta = \rho^2 Nh^2/M_e$ in equation (28.5) is reduced. Therefore, in a multi-breed population a denser SNP panel or sequence information will increase accuracy. Another consequence of the high M_e is that the number of causal variants is less than M_e and hence Bayesian methods have an advantage over BLUP.

An implication the limited effective number of segments is that increasing marker density does not always increase accuracy of prediction. Brøndum *et al.* (2015) and VanRaden *et al.* (2017) found that the use of full genome sequence data increased accuracy by approximately 2–3% even in Holstein cattle, where the effective number of segments is small and the number of genotyped animals is large. Others have found no improvement for Holsteins (van Binsbergen *et al.*, 2014). This limited effect of using sequence data may be due in part to the use of sequence imputed from SNP genotypes. Inevitably the imputation is not 100% accurate, and this inaccuracy is reflected in a loss of accuracy in the EBVs.

Over generations, chromosome segments whose effects have been estimated will be broken up by recombination. Consequently, the accuracy of a prediction equation is expected to decline over generations. Simulation studies show that the accuracy can be maintained by continually updating the training population or by using nonlinear prediction methods that give more weight to markers close to the causal variants (MacLeod *et al.*, 2016).

28.6 Other Uses of the Bayesian Genomic Selection Models

Usually GWASs are analysed by fitting one SNP at a time and consequently an estimated effect represents not only that of the SNP itself but also of all SNPs in LD with it. Therefore, the models used for genomic selection, which fit all SNPs simultaneously, should give more accurate estimates of SNP effects and mapping of causal variants (Fernando and Garrick, 2013). MacLeod *et al.* (2016) demonstrate the advantage of Bayes RC over one-SNP-at-a-time GWASs when mapping variants for milk production. If there are many SNPs in high LD, the posterior

probability that each one is the causal variant may be low. However, by adding the posterior probabilities together, it is possible to show that the group of SNPs has high probability of including a causal variant. A similar method is to calculate the variance of local EBV for a genomic window where the local EBV is the prediction of genetic value based on the SNPs in this window. Large variances indicate a causal variant somewhere in the window (Kemper *et al.*, 2015). Posterior probabilities can also be combined across traits to implement a multi-trait meta-analysis approach that seeks to identify pleiotropic loci affecting multiple traits (Bolormaa *et al.*, 2017).

These genomic selection models describe a model of the effects of polymorphisms throughout the genome. Therefore, the estimates of the parameters of these models are a description of the genetic architecture of the trait (MacLeod *et al.*, 2016). For instance, by including a mixture of four normal distributions Bayes R can provide an estimate of the distribution of the effects of causal variants. However, this description is dependent on the restrictions of the model and almost certainly differs from the true distribution. With this caveat, almost all quantitative traits examined appear to be due to a large number (thousands) of causal variants, and this conclusion is supported by other approaches to the same question.

28.7 Examples of Genomic Prediction

28.7.1 Cattle

In cattle, genomic prediction within a breed based on ~50,000 SNPs using BLUP works well, provided the training population is greater than 10,000, and it has been widely adopted. In such situations, rates of genetic progress have doubled since the introduction of genomic prediction (García-Ruiz *et al.*, 2016). However, there are many situations where a training population of more than 10,000 cattle is not available within each breed and for all important traits. In this case, it is tempting to combine the training population from multiple breeds. Unfortunately, BLUPs of genetic value from a prediction model trained in one breed give low accuracy when applied to another breed. Within a breed, LD extends over megabase-pair distances so that the effective number of chromosome segments whose effect must be estimated is small (e.g. $M_e = 4000$). However, the phase of LD varies between breeds (de Roos *et al.*, 2008) and so the M_e across breeds is about 60,000. Therefore it makes sense to use denser SNPs or sequence data and to apply Bayesian methods of prediction. These changes lead to small gains in accuracy, especially when applied to multiple breed populations.

Even within the Holstein breed, VanRaden *et al.* (2017) found a 2.7 percentage unit increase in accuracy for dairy traits from using sequence data instead of 50k SNP chips.

A further gain in accuracy should be available from using external information provided this information identifies categories of polymorphism that are enriched for causal variants. Macleod *et al.* (2016) classified sequence variants as non-synonymous coding variants in genes connected with lactation, other variants near these genes and all other variants. The first two classes were enriched for variants affecting milk yield but the gain in accuracy of prediction was small. This could be due to the fact that, despite the enrichment, the majority of genetic variance (81%) was still due to the many polymorphisms in the third class. Thus, to increase accuracy, we need external information that identifies variants explaining a large proportion of the genetic variance.

28.7.2 Humans

In GWASs of humans for traits and diseases, power has been increased by meta-analysis of many individual data sets without combining the individual-level data. That is, only the SNP

effects, estimated one SNP at a time, and their standard errors are combined. The final estimates of SNP effects obtained in this way can be used to construct a PRS. However, it would be better to estimate all SNP effects simultaneously, as done in BLUP and the Bayesian methods. Yang *et al.* (2012) moved in this direction by using summary GWAS statistics together with a small reference population of people with individual-level genotype data available from which LD could be estimated and assumed to be the same in the samples used for the GWAS. This approach can also be used in a BLUP; Vilhjálmsson *et al.* (2015) show that this yields higher accuracy than a PRS. Maier *et al.* (2018) have extended this to a multi-trait BLUP and found that the accuracy of prediction is increased considerably for some traits (e.g. type 2 diabetes) by the multi-trait approach.

28.8 Conclusions

The genetic value for complex traits, and hence phenotype, of individuals can be predicted from their genotype at genetic markers such as SNPs and from whole-genome sequence data. The variation in most complex traits is caused by thousands of polymorphisms, mostly of very small effect, scattered throughout the genome. Consequently, the panel of genetic markers must be dense enough so that all causal variants are in linkage disequilibrium with one or more markers. Understanding the accuracy of prediction is aided by viewing the problem as estimating the effect on the trait of chromosomal segments which may contain no, one or more causal variants. The length of these effective chromosomal segments depends on the LD in the population. If LD extends over only short distances, the number of effective segments is high and their effects are small and hence the accuracy of estimating their effect is low. In this case, the accuracy can be increased by increasing the size of the training population with genotypes and phenotypes, by using denser markers or full genome sequence, by nonlinear or Bayesian statistical methods, by multi-trait analysis, and by using functional information about the sites in the genome.

References

- Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. and Lawlor, T.J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* **93**(2), 743–752.
- Aliloo, H., Pryce, J.E., González-Recio, O., Cocks, B.G. and Hayes, B.J. (2016). Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. *Genetics Selection Evolution* **48**(1), 8.
- Beavis, W.D. (1998). QTL analyses: Power, precision, and accuracy. In A.H. Paterson (ed.), *Molecular Dissection of Complex Traits*. CRC Press, Boca Raton, FL, pp. 145–162.
- Bolormaa, S., Swan, A.A., Brown, D.J., Hatcher, S., Moghaddar, N., van der Werf, J.H., Goddard, M.E. and Daetwyler, H.D. (2017). Multiple-trait QTL mapping and genomic prediction for wool traits in sheep. *Genetics Selection Evolution* **49**(1), 62.
- Brøndum, R.F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D. and Lund, M.S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science* **98**(6), 4107–4116.
- Calus, M. and Veerkamp, R. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* **43**(1), 26.
- Christensen, O.F. and Lund, M.S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**(1), 2.

- Daetwyler, H.D., Villanueva, B. and Woolliams, J.A. (2008). Accuracy of Predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**(10), e3395.
- Daetwyler, H.D., Pong-Wong, R., Villanueva, B. and Woolliams, J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics* **185**, 1021–1031.
- de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. and Calus, M.P.L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**(2), 327–345.
- De Roos, A.P.W., Hayes, B.J., Spelman, R.J. and Goddard, M.E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**(3), 1503–1512.
- Elsen, J.-M. (2017). An analytical framework to derive the expected precision of genomic selection. *Genetics Selection Evolution* **49**(1), 95.
- Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A. and Goddard, M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science* **95**(7), 4114–4129.
- Erbe, M., Gredler, B., Seefried, F.R., Bapst, B. and Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE* **8**(12), e81046.
- Fernando, R.L. and Garrick, D. (2013). Bayesian methods applied to GWAS. In C. Gondro, J. van der Werf and B. Hayes (eds), *Genome-Wide Association Studies and Genomic Prediction*. Humana Press, New York, pp. 237–274.
- Fernando, R.L. and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21**(4), 467.
- Fernando, R.L., Dekkers, J.C. and Garrick, D.J. (2014). A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution* **46**(1), 1–13.
- Fernando, R.L., Cheng, H., Golden, B.L. and Garrick, D.J. (2016). Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. *Genetics Selection Evolution* **48**(1), 96.
- García-Ruiz, A., Cole, J.B., VanRaden, P.M., Wiggans, G.R., Ruiz-López, F.J. and Van Tassell, C.P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences of the United States of America* **113**(28), E3995–4004.
- Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* **136**(2), 245–257.
- Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M. and Snell, R. (2002). Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**(2), 222–231.
- Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.J., Kvasz, A., Mni, M., Simon, P., Frere, J.M., Coppieters, W. and Georges, M. (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* **101**(8), 2398–2403.
- Habier, D., Fernando, R., Kizilkaya, K. and Garrick, D. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**(1), 186.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

- Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J. and Goddard, M.E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics*, **6**(9), e1001139.
- Hill, W.G., Goddard, M.E. and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics* **4**(2), e1000008.
- Houston, R.D., Haley, C.S., Hamilton, A., Guy, D.R., Tinch, A.E., Taggart, J.B., McAndrew, B.J. and Bishop, S.C. (2008). Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*). *Genetics* **178**(2), 1109–1115.
- Kemper, K.E., Reich, C.M., Bowman, P.J., vander Jagt, C.J., Chamberlain, A.J., Mason, B.A., Hayes, B.J. and Goddard, M.E. (2015). Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution* **47**(1), 29.
- Kemper, K.E., Littlejohn, M.D., Lopdell, T., Hayes, B.J., Bennett, L.E., Williams, R.P., Xu, X.Q., Visscher, P.M., Carrick, M.J. and Goddard, M.E. (2016). Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC Genomics* **17**(1), 858.
- Kemper, K.E., Bowman, P.J., Hayes, B.J., Visscher, P.M. and Goddard, M.E. (2018). A multi-trait Bayesian method for mapping QTL and genomic prediction. *Genetics Selection Evolution* **50**(1), 10.
- Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. and Misztal, I. (2015). Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. *Genetics* **200**(2), 455–468.
- Liu, Z., Goddard, M.E., Reinhardt, F. and Reents, R. (2014). A single-step genomic model with direct estimation of marker effects. *Journal of Dairy Science* **97**(9), 5833–5850.
- MacLeod, I.M., Bowman, P.J., Vander Jagt, C.J., Haile-Mariam, M., Kemper, K.E., Chamberlain, A.J., Schrooten, C., Hayes, B.J. and Goddard, M.E. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* **17**(1), 144.
- Maier, R.M., Zhu, Z., Lee, S.H., Trzaskowsk, M., Ruderfer, D.M., Stahl, E.A., Ripke, S., Wray, N.R., Yang, J., Visscher, P.M. and Robinson, M.R. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature Communications* **9**(1), 989.
- Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753.
- Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4), 1819–1829.
- Meuwissen, T.H.E., Svendsen, M., Solberg, T. and Ødegård, J. (2015). Genomic predictions based on animal models using genotype imputation on a national scale in Norwegian red cattle. *Genetics Selection Evolution* **47**(1), 79.
- Misztal, I. and Legarra, A. (2017). Invited review: Efficient computation strategies in genomic selection. *Animal* **11**(5), 731–736.
- Moen, T., Torgersen, J., Santi, N., Davidson, W.S., Baranski, M., Ødegård, J., Kjøglum, S., Velle, B., Kent, M., Lubieniecki, K.P., et al. (2015). Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in Atlantic salmon. *Genetics* **200**(4), 1313–1326.
- Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R. and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLOS Genetics* **11**(4), e1004969.

- Robinson, M., English, G., Moser, G., Lloyd-Jones, L.R., Triplett, M.A., Zhu, Z., Nolte, I., Van Vliet-Ostaptchouk, J., Snieder, H., Esko, T. *et al.* (2017). Genotype–covariate interaction effects and the heritability of adult body mass index. *Nature Genetics* **49**, 1174–1181.
- Speed, D. and Balding, D.J. (2014). MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Research* **24**(9), 1550–1557.
- Speed, D., Cai, N., UCLEB Consortium, Johnson, M.R., Nejentsev, S. and Balding, D.J. (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics* **49**, 986–992.
- Strandén, I. and Christensen, O.F. (2011). Allele coding in genomic evaluation. *Genetics Selection Evolution* **43**(1), 25.
- van Binsbergen, R., Bink, M.C., Calus, M.P., van Eeuwijk, F.A., Hayes, B.J., Hulsegege, I. and Veerkamp, R.F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* **46**(1), 41.
- Vander Jagt, C.J., Chamberlain, A.J., Schnabel, R.D., Hayes, B.J., Daetwyler, H.D. (2018). Which is the best variant caller for large whole-genome sequencing datasets? In *Proceedings of the World Congress in Genetics Applied to Livestock Production*, Auckland, NZ.
- VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**(11), 4414–4423.
- VanRaden, P.M., Tooker, M.E., O'Connell, J.R., Cole, J.B. and Bickhart, D.M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution* **49**(1), 32.
- Verbyla, K.L., Hayes, B.J., Bowman, P.J. and Goddard, M.E. (2009). Technical note: Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research* **91**, 307–311.
- Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., *et al.* (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics* **97**(4), 576–592.
- Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., *et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186.
- Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., *et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565.
- Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., *et al.* (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369.
- Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., *et al.* (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47**, 1114.
- Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., *et al.* (2018). Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics* **50**(5), 746–753.
- Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9**(2), e1003264.
- Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., *et al.* (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481.

Disease Risk Models

Allison Meisner¹ and Nilanjan Chatterjee²

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Abstract

The popularity of disease risk models has risen in recent years due in part to an increase in researchers' ability to measure many disease risk factors and a concomitant growing interest in 'precision medicine' approaches. These models seek to combine different sources of information to provide an estimate of an individual's disease risk. In this chapter we provide an overview of absolute disease risk models, including the steps involved in their development and evaluation. We discuss the use of genetic risk factors in these models, focusing in particular on polygenic risk scores. We review efforts to develop and evaluate breast cancer risk models, including those that incorporate genetic information. We close with a discussion of future directions and remaining challenges in this area.

29.1 Introduction and Background

29.1.1 Disease Risk Models and Their Applications

Our understanding of the role of various risk factors, including demographic, lifestyle, environmental, clinical, and genetic factors, in the development of complex diseases continues to grow. There is considerable interest in translating this knowledge into improvements in human health through the development of models that can be used to estimate an individual's risk of a disease D . In practice, an individual's probability of developing D is difficult to interpret and is influenced by unknown characteristics of the individual (i.e., stochastic factors); thus, it is not a particularly meaningful quantity. Instead, the 'risk' of D for an individual with a given risk factor profile can be defined as the proportion of individuals in the population with the same risk factor profile who develop D . A disease risk model is used to define similarity among individuals as it becomes difficult to match individuals exactly when there are a variety of risk factors under consideration.

Disease risk models hold great promise, as estimates of disease risk can be used to counsel individuals, leading to improvements in decision-making and spurring lifestyle and behavior changes, and to develop stratified approaches to disease prevention at the population level by optimally weighing risks, costs, and benefits of interventions (Antoniou and Easton, 2006; Domchek *et al.*, 2003; Freedman *et al.*, 2005; Lloyd-Jones, 2010; Usher-Smith *et al.*, 2015).

In the future, it is expected that disease risk models will play a central role in the development of ‘precision medicine’ approaches to disease prevention, an overarching goal of several recent large initiatives such as the All of Us Research Program undertaken in the United States by the National Institutes of Health (2018).

Disease risk models can be applied to both primary prevention, that is, preventing or delaying the development of D , and secondary prevention, where the goal is to detect D early and prevent progression (Chatterjee *et al.*, 2016). Specific applications of disease risk models include identifying high- (or low-)risk individuals who may be recommended for (or against) enhanced surveillance, medication, or other interventions, each with their own risks and benefits; aiding in decision-making in terms of both whether to intervene and which intervention to use; planning and designing prevention trials; informing estimates of risk and benefit for a particular intervention; and quantifying population-level burden, cost, and impact of interventions (Freedman *et al.*, 2005; Gail, 2011; Temple, 2010).

A model’s clinical utility relates to the degree of risk stratification it provides; in other words, measures of clinical utility evaluate the degree to which a model can assign individuals to sufficiently different risks so as to provide actionable information at the individual or population level (Figure 29.1) (Chatterjee *et al.*, 2016). In developing population-level policies, it is often desirable to identify risk thresholds that can be used to stratify the population by category of risk. These thresholds will depend upon the risk–benefit implications of the resulting decisions, and should be considered in the evaluation of the model (Chatterjee *et al.*, 2016). As the extent of risk stratification provided by a model is related to the amount of variability in estimated risks, the inclusion of additional risk factors in the model can offer improvements in its ability to identify individuals at different levels of risk (Chatterjee *et al.*, 2016). Consequently, and in light of the fact that knowledge of risk factors continues to evolve, disease risk models should be frequently updated and evaluated in order to incorporate new information (Chatterjee *et al.*, 2016).

Throughout this chapter, we will use the terms ‘case’ and ‘control’, where cases are individuals who develop or experience the disease D in a specific time interval ($D = 1$) and controls are those who do not ($D = 0$).

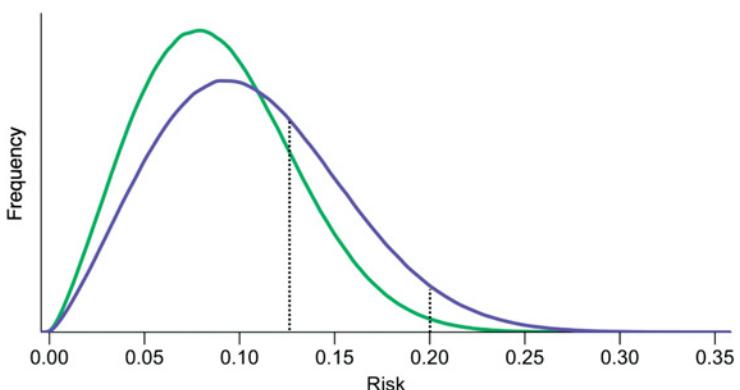


Figure 29.1 Stratification of lifetime risk for a relatively common disease such as breast cancer. Suppose moderate action (e.g., encouraging lifestyle changes) is taken for risks between 12.5% and 20%, and more serious action (such as a therapeutic intervention) is taken for risks above 20%. The two curves represent risk distributions with different degrees of risk stratification. In particular, the distribution represented by the blue curve involves a larger proportion of the population taking some action than the distribution represented by the green curve. The variance of the risk distribution represented by the blue curve is 44% larger than that represented by the green curve, leading to differences in the extent of risk stratification. As additional risk factors are identified and included in the risk model, more spread in the distribution of risk leads to increases in the clinical utility of the model. Adapted with permission from Chatterjee *et al.* (2016), Springer Nature.

29.1.2 Examples of Available Disease Risk Models

Numerous disease risk models have been developed; here, we discuss two established risk models, one in breast cancer and one in cardiovascular disease.

One of the earliest disease risk models was the Gail model, which can be used to estimate an individual's age-specific risk of breast cancer using information on several epidemiologic risk factors, such as age at menarche, number of previous breast biopsies, age at first birth, and number of first-degree relatives with breast cancer (Gail *et al.*, 1989). This model and subsequent updates have since been widely used for counseling individuals in clinical practice, designing prevention trials, and weighing the risks and benefits of preventive therapy such as tamoxifen (Freedman *et al.*, 2003; Gail, 2011; Gail *et al.*, 1999).

Another well-known disease risk model is the Framingham Risk Score, which was developed using data from a large cohort study in Framingham, Massachusetts and provides estimates of 10-year risk of cardiovascular disease (CVD) (D'Agostino *et al.*, 2008). The model on which the Score is based includes epidemiologic risk factors such as age, cholesterol, blood pressure, treatment for hypertension, smoking status, and the presence of diabetes (Figure 29.2). The Framingham Risk Score is generally used by primary care physicians to provide preventive care to patients. The developers of the Score also proposed using repeat assessments of CVD risk to monitor changes in risk due to treatment or lifestyle modifications.

29.1.3 Incorporating Genetic Factors

Most established disease risk models include traditional epidemiologic risk factors, such as environmental, demographic, and clinical variables. However, many complex diseases are due to a combination of genetic factors, environmental variables, and other medical conditions (Chatterjee *et al.*, 2016). For most diseases, models with known epidemiologic risk factors have only modest risk stratification ability (Chatterjee *et al.*, 2016). Furthermore, as many epidemiologic risk factors and/or their effects on disease risk can change over time, using models with epidemiologic risk factors alone to provide long-term estimates of risk may be challenging (Chatterjee *et al.*, 2016). Finally, recent genome-wide association studies (GWAS) have led to the discovery of scores of genetic variants associated with disease risk. As the costs of genotyping and/or sequencing continue to drop, incorporation of genetic markers into disease risk models will become increasingly relevant for clinical application. Thus, there is interest in combining genetic and epidemiologic risk factors to develop disease risk models (Chatterjee *et al.*, 2016; Freedman *et al.*, 2005). One approach for incorporating genetic risk factors into a disease risk model is to combine genetic variants into a polygenic risk score (PRS), which is then included in the disease risk model as a continuous risk factor. In this chapter we focus on a particular type of genetic variant: single nucleotide polymorphisms (SNPs), which have been shown in recent studies to contribute substantially to disease heritability. To date, GWAS have focused mostly on common variants, typically defined as markers that have a minor allele frequency of at least 5% in one or more major ethnic populations. SNPs can also be classified as high- or low-penetrance, reflecting the strength of their association with a particular disease.

Some genetic models are already being used in high-risk families, for example, to estimate risk of breast and ovarian cancer in individuals with rare, high-penetrant *BRCA1* or *BRCA2* mutations (Chatterjee *et al.*, 2016). A number of recent studies have evaluated the cost–benefit implications of testing for mutations in *BRCA1/BRCA2* and other high-risk genes at the population level (Gabai-Kapara *et al.*, 2014; King *et al.*, 2014; Manchanda *et al.*, 2014, 2018; Palomaki, 2015). Interest is now moving beyond the use of a narrow set of rare, high-penetrant variants; in particular, it may be possible to have a broad public health impact if factors contributing to

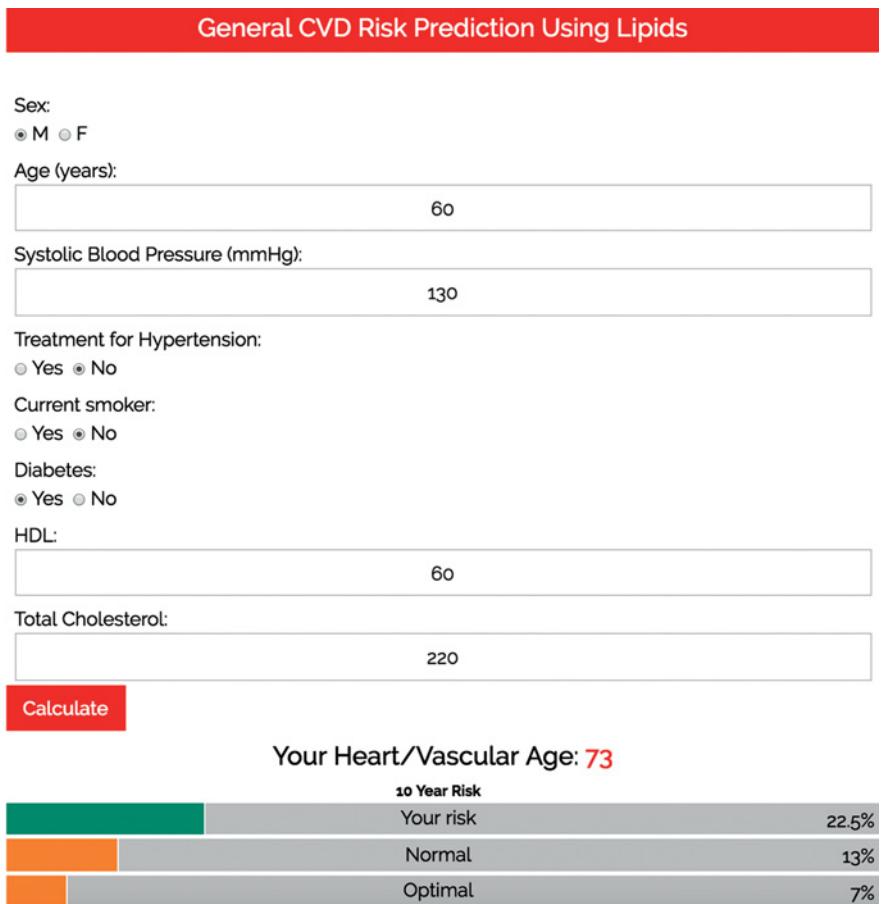


Figure 29.2 An example application of the Framingham Risk Score online calculator (<https://www.framinghamheartstudy.org/fhs-risk-functions/cardiovascular-disease-10-year-risk/>). For a 60-year-old, non-smoking male with a systolic blood pressure of 130 mmHg, no treatment for hypertension, diabetes, a high-density lipoprotein (HDL) cholesterol level of 60 mg/dL, and a total cholesterol of 220 mg/dL, the estimated 10-year risk of CVD (i.e. the estimated risk of CVD by age 70) is 22.5%.

risk in the general population can be identified (Chatterjee *et al.*, 2016). There is evidence that for many diseases, common SNPs have the potential to explain a large proportion of heritability in the general population and so may be useful for risk stratification (Chen *et al.*, 2014; Gusev *et al.*, 2014; Lee *et al.*, 2011, 2012, 2013; Lu *et al.*, 2014; Sampson *et al.*, 2015). However, there is also evidence that in many cases, a large number of SNPs, each with a relatively small effect, are associated with the disease (Chatterjee *et al.*, 2016); such diffuse genetic architecture will require very large GWAS in order to develop effective genetic disease risk models. While large GWAS have been or are being conducted for many common diseases, such investigations of rarer diseases may not be completed for some time.

29.2 Absolute Risk Model

Models intended for clinical use should provide estimates of absolute risk, accounting for the underlying incidence rate of D , to be clinically useful. Absolute risk is defined as the proportion

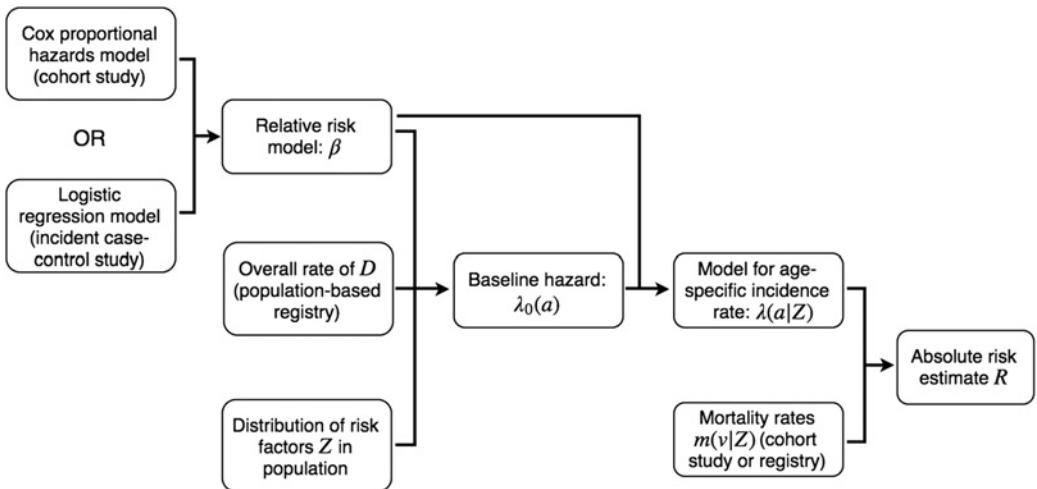


Figure 29.3 A flow chart illustrating the steps to build an absolute risk model.

of asymptomatic individuals developing or experiencing D over a specified time interval, given a set of risk factors and the presence of potential competing risks (Gail *et al.*, 1989). Constructing absolute risk models involves identifying risk factors, using a model to combine these factors to characterize relative risk, and then estimating absolute risk (Figure 29.3) (Chatterjee *et al.*, 2016; Gail *et al.*, 1989). Specifically, estimates of absolute risk are based on the relative risk model, the baseline incidence rate of D in the population, and the mortality rate in the population (Chatterjee *et al.*, 2016).

For many outcomes, particularly late-onset chronic diseases, it will be important to account for age as it is typically the strongest risk factor (Chatterjee *et al.*, 2016). The proportional hazards model can be used to estimate age-specific incidence rates, which in turn provide estimates of absolute risk over a certain time interval (Chatterjee *et al.*, 2016). For a given set of risk factors Z , denote the conditional age-specific incidence rate at age a , that is, the probability of D at age a given that D has not occurred prior to that age, as $\lambda(a|Z)$. We can consider the following model for $\lambda(a|Z)$:

$$\lambda(a|Z) = \lambda_0(a) \exp(\beta^\top Z), \quad (29.1)$$

which stipulates that $\lambda(a|Z)$ is given by the product of the relative risk model $\exp(\beta^\top Z)$ and the baseline hazard $\lambda_0(a)$ (Cox, 1972). In particular, this model allows each risk factor to act multiplicatively on $\lambda(a|Z)$. While this model does not allow β , the coefficients for Z , to vary with a , more flexible models could include an interaction term between age and Z , effectively allowing the coefficients for Z to vary with a (Gail *et al.*, 1989; Maas *et al.*, 2016a). Following the development of the proportional hazards model, the probability of D over the age interval $[a, a+s]$ given that D has not occurred by age a , i.e., the absolute risk, is

$$R_{a,a+s} = \int_a^{a+s} \lambda(u|Z) \exp\left(-\int_a^u \{\lambda(v|Z) + m(v|Z)\} dv\right) du, \quad (29.2)$$

where $m(v|Z)$ is the age-specific mortality from other causes (Gail *et al.*, 1989). This formula is based on competing risks methods and essentially involves summing (over ages between a and $a+s$) the probability of D at age u conditional on being alive and not having experienced D up to that point (Chatterjee *et al.*, 2016; Gail *et al.*, 1989). It is important to account for age-specific competing hazards as the observable risk of D could be substantially reduced in the presence of

competing risks, especially at older ages (Chatterjee *et al.*, 2016; Maas *et al.*, 2016a). For most diseases, using overall rates of mortality from competing risks $m(v)$ is reasonable provided the risk factors Z do not exert large effects on mortality (Chatterjee *et al.*, 2016). Likewise, if it is reasonable to assume D has a modest effect on mortality, overall mortality rates can be used; in situations where this assumption is not tenable, more sophisticated modeling approaches have been pursued (Ganna and Ingelsson, 2015). Note that a similar approach was used to construct the Gail breast cancer model described in Section 29.1.2 (Gail *et al.*, 1989).

Different sources of information may be utilized to construct a model for absolute risk. For the relative risk model ($\exp(\beta^\top Z)$), two possibilities exist: (1) use data from a prospective cohort study to estimate hazard ratios using a Cox proportional hazards model; and (2) use data from incident case-control studies to approximate hazard ratios via logistic regression with adjustment for age using fine categories (Cox, 1972; Prentice *et al.*, 1978). The latter approach is particularly well suited to risk factors such as SNPs, which tend to have weak effects and are generally not related to survival or likelihood of study participation, making the odds ratio a good approximation to the hazard ratio (Chatterjee *et al.*, 2016). For modeling risk associated with non-genetic factors, however, use of case-control studies requires more caution as selection bias, differential recall, and reverse causality can lead to substantial bias. Case-control studies that include prevalent cases or do not follow a population-based design are particularly problematic (Hill *et al.*, 2003; Wacholder *et al.*, 1992a). A representative cohort study or population-based registry can be used to estimate the baseline hazard, that is, the incidence rate of D for a baseline risk profile, $\lambda_0(a)$, based on the overall rate (Chatterjee *et al.*, 2016; Cox, 1972; Gail *et al.*, 1989; Maas *et al.*, 2016b). Rates of mortality from competing risks ($m(v)$) can also be obtained from cohort studies or population-based registries (Chatterjee *et al.*, 2016). Using registries to estimate disease incidence and mortality rates improves generalizability: if the relative risk model can be applied broadly, the absolute risk model can simply be updated for a new population by calibrating the baseline risk to the new population (using information on the overall disease incidence rate and the risk factor distribution in the new population) and incorporating the mortality rate for the new population (Chatterjee *et al.*, 2016).

29.2.1 General Software for Building Absolute Risk Models

The iCARE software program, available as an R package, was developed to synthesize information from several sources to construct an absolute risk model, streamlining the steps described in Figure 29.3 and allowing for easy updating of models and tailoring to different populations (Maas *et al.*, 2016b). iCARE combines a relative risk model, age-specific rates of D and rates of mortality from competing risks, and the distribution of risk factors in the population of interest (Maas *et al.*, 2016b). In particular, the software uses overall age-specific rates of D (typically obtained from registry data), the relative risk model, and the risk factor distribution (from, for example, population-based health surveys) to estimate the baseline hazard, thereby calibrating the model to a given population (Maas *et al.*, 2016b). This allows the model to be easily updated for new populations (Maas *et al.*, 2016b). In addition, iCARE readily handles missing data by using the provided risk factor distribution to perform imputation and can incorporate independent SNPs via summary-level data, namely odds ratio estimates and allele frequencies, removing the need for genotype-level information (Maas *et al.*, 2016b).

Other R packages exist for building disease risk models, including riskRegression (Gerds *et al.*, 2017) and PredictABEL (Kundu *et al.*, 2015). The riskRegression package uses the proportional hazards framework described earlier to estimate cumulative incidence, but its capacity to update a given model for a new population is limited relative to the iCARE package. In addition, the riskRegression package cannot easily accommodate missing data, nor can it incorporate

SNPs based on summary-level data alone. PredictABEL uses logistic regression to estimate the risk of D given Z , which is not equivalent to the absolute risk as the latter concerns risk over a specific time period. In addition, the PredictABEL package cannot incorporate missing data as easily as the iCARE package.

29.3 Building a Polygenic Risk Score

It is useful to conceptualize the underlying true PRS for a given genetic model. For example, under an additive model, the true PRS could be defined as the total genetic burden of the disease, given by a weighted linear combination of the variants contributing to the genetic burden of the disease, where the weights represent the true effects of the variants in the population under the given model. As the true PRS is not observable, an estimated PRS, S_E , is typically formulated as a weighted sum of risk alleles for a collection of SNPs: $S_E = \sum_i w_i X_i$, where w_i is some (estimated) weight for the i th SNP and $X_i \in \{0, 1, 2\}$ indicates the number of copies of the risk allele for the i th SNP (Chatterjee *et al.*, 2016). Thus, to build a PRS, both a set of SNPs and a weight for each SNP are needed (Chatterjee *et al.*, 2016).

This form for the PRS is motivated by the log-linear risk model, $\log[P(D = 1|X)] = \sum_i w_i X_i$, and corresponds to each SNP having a multiplicative effect on the risk of D . There is evidence that for many diseases, the additive effects of common variants explain a substantial proportion of heritability (Chen *et al.*, 2014; Gusev *et al.*, 2014; Lee *et al.*, 2011, 2012, 2013; Lu *et al.*, 2014; Sampson *et al.*, 2015), indicating that PRS constructed by linearly combining common SNPs may offer meaningful risk stratification. Furthermore, multiplicative effects on risk across many SNPs, even when the individual effects are not large, can provide meaningful risk stratification (Maas *et al.*, 2016a). However, as we will discuss in Section 29.3.1, the performance of an estimated PRS may be affected by errors in the selection of SNPs for inclusion in the PRS and/or errors in the weights used to combine the SNPs (w_i). After it has been estimated, the PRS is included in the relative risk model described in Section 29.2; the idea is that the risk conferred by the available SNPs is (ideally) captured by the PRS, which is then combined with other risk factors (Chatterjee *et al.*, 2016).

The weights w_i used in the PRS are typically an estimate of the association between the i th SNP and D based on, for example, a regression model. Thus, it is important to consider the impact of population stratification, which could influence coefficient estimates (Mak *et al.*, 2017). In addition, if data from different populations are combined for estimation, the resulting estimates could mask heterogeneity in effect sizes (Mak *et al.*, 2017).

29.3.1 Expected Performance

The utility of a PRS depends upon the heritability of D , as this is related to the variability of the true PRS, which determines the degree of genetic risk stratification that is possible (Chatterjee *et al.*, 2016). Utility also depends on the residual effects of family history and other risk factors and available strategies for prevention and early detection (Chatterjee *et al.*, 2016). Thus, heritability is useful for understanding the limits of a PRS, but utility is influenced by other factors as well (Chatterjee *et al.*, 2016).

Importantly, even if heritability estimates indicate that SNPs could ultimately be useful for estimating risk, the PRS must include the precise effect of all SNPs in order to fully capture their utility (Chatterjee *et al.*, 2016). In practice, however, the estimated PRS is not equal to the true PRS due to imprecision of estimation and selection procedures and imperfect tagging of causal SNPs (Chatterjee *et al.*, 2016). This issue is particularly acute for diseases with extreme

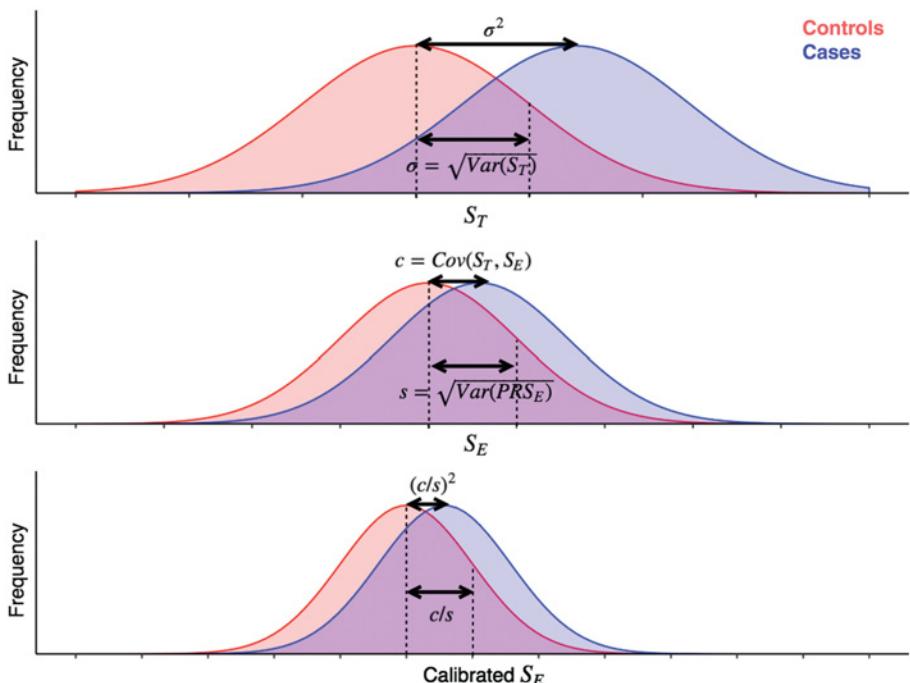


Figure 29.4 Distribution of S_T , S_E , and the calibrated estimated PRS ($\frac{c}{s^2} S_E$) in cases and controls. Adapted with permission from Chatterjee *et al.* (2016), Springer Nature.

polygenic architecture where heritability is spread across many SNPs (Chatterjee *et al.*, 2013; Dudbridge, 2013; Lee *et al.*, 2013; Stahl *et al.*, 2012; Talmud *et al.*, 2015). In particular, for the same narrow-sense GWAS heritability, that is, the heritability explained by the additive effects of common tagging variants (i.e. those from GWAS), the existence of more susceptibility SNPs means the individual effect sizes are smaller, so power is lower and the utility of the PRS is diminished (Chatterjee *et al.*, 2016; Manolio *et al.*, 2009; Park and Dunson, 2010; Park *et al.*, 2010, 2011; Yang *et al.*, 2010, 2011; Zuk *et al.*, 2012). As a result, for most diseases, the variants discovered thus far explain only a small proportion of heritability (Franke *et al.*, 2010; Jostins and Barrett, 2011; Kraft and Hunter, 2009; Lango Allen *et al.*, 2010; Pharoah *et al.*, 2008; Speliotis *et al.*, 2010; Teslovich *et al.*, 2010; Van Hoek *et al.*, 2008; Wacholder *et al.*, 2010); this shortfall is referred to as the issue of ‘missing heritability’ (Manolio *et al.*, 2009; Zuk *et al.*, 2012). With infinite sample sizes and all common tagging variants genotyped or accurately imputed, PRS will be able to explain narrow-sense GWAS heritability (Chatterjee *et al.*, 2013, 2016). The utility of a risk model with a PRS can also depend on the algorithm used for SNP selection and the inclusion of other risk factors such as family history (Chatterjee *et al.*, 2013). Although family history generally has modest utility on its own, models with family history and a PRS can offer better performance than the PRS alone, particularly for rare, highly familial diseases (Chatterjee *et al.*, 2013).

A good deal of theory has been developed to understand the performance of a PRS (Figure 29.4). Consider the setting where D is rare. Suppose $P(D = 1|X) = \exp(S_T)$, where S_T is the true PRS and X is the set of susceptibility SNPs (Chatterjee *et al.*, 2016; Pharoah *et al.*, 2002). One can view $\text{Var}(S_T) = \sigma^2$ as a measure of the heritability of D on the log-risk scale since this quantity is related to the relative risk of D in monozygotic twins (Chatterjee *et al.*, 2016;

Pharoah *et al.*, 2002). This means that an individual SNP's contribution to the variance of the PRS and thus the heritability of D is a function of the SNP's allele frequency and its weight in the PRS (Mak *et al.*, 2017). By the central limit theorem, it is generally reasonable to assume $S_T \sim N(0, \sigma^2)$ and, if D is rare, $(S_T|D=0) \sim N(0, \sigma^2)$ (Chatterjee *et al.*, 2016; Pharoah *et al.*, 2002). Such a log-normal distribution for genetic risk has been found to fit well to breast cancer data and comports with the intuition that as the number of variants increases, their additive effects will yield a continuous distribution of log risk (Pharoah *et al.*, 2002). It can be shown that this log-normal distribution of risk yields $(S_T|D=1) \sim N(\sigma^2, \sigma^2)$; that is, the distribution of the true PRS in cases is normal with the same variance as in controls but with the mean shifted by an amount equal to the variance (Chatterjee *et al.*, 2016; Pharoah *et al.*, 2002). This means that the degree of separation in S_T for cases and controls is determined by σ^2 , the variation in log risk in the population and a measure of heritability (Chatterjee *et al.*, 2016; Pharoah *et al.*, 2002). Thus, the ability of S_T to discriminate between cases and controls, which can be evaluated with measures like the area under the receiver operating characteristic (ROC) curve (AUC) and is one measure of a model's ability to stratify risk, is related to heritability (Chatterjee *et al.*, 2016; Janssens *et al.*, 2006; Pharoah *et al.*, 2002; So *et al.*, 2011; Wray *et al.*, 2010).

In practice, the estimated PRS, S_E , is available. It has been shown that the risk stratification ability of S_E depends on $r = c/s$, where $s^2 = \text{Var}(S_E)$ and $c = \text{Cov}(S_E, S_T)$ (Chatterjee *et al.*, 2013; Dudbridge, 2013). In particular, as with S_T , it is reasonable to assume $S_E \sim N(0, s^2)$, which yields $(S_E|D=0) \sim N(0, s^2)$ and $(S_E|D=1) \sim N(c, s^2)$ (Chatterjee *et al.*, 2013; Dudbridge, 2013). As more SNPs that contribute to S_T are included in S_E , c will increase even if some SNPs that do not contribute to S_T are included, though the inclusion of such SNPs will increase s^2 (Chatterjee *et al.*, 2016). These conditional distributions for S_E give $\text{logit}[P(D=1|S_E)] = \alpha_0 + \frac{c}{s^2}S_E$; thus, r^2 determines the discriminatory ability of the calibrated estimated PRS, $\frac{c}{s^2}S_E$, since $\left(\frac{c}{s^2}S_E|D=0\right) \sim N(0, r^2)$ and $\left(\frac{c}{s^2}S_E|D=1\right) \sim N(r^2, r^2)$ (Chatterjee *et al.*, 2016). Normal theory can also be used to determine the discriminatory ability of a model with both S_E and family history (Chatterjee *et al.*, 2013).

As noted above, the degree of risk stratification provided by PRS is limited by the heritability of the disease; the same is true of the risk stratification ability of individual genetic factors (Pharoah *et al.*, 2002; Wray *et al.*, 2010). In other words, greater risk stratification is possible for more heritable traits (Chatterjee *et al.*, 2016). Data from family studies are commonly used to estimate heritability (Chatterjee *et al.*, 2013). Heritability estimates may be problematic due to bias from confounding in family studies, differential recall or knowledge of family history in case-control studies, inaccurate self-reported family history, and issues in applying certain modeling approaches to estimate heritability (Chang *et al.*, 2006; Golan *et al.*, 2014; Kerber and Slattery, 1997; Mitchell *et al.*, 2013; Speed *et al.*, 2012). Note also that regardless of disease heritability, the ability to stratify risk will generally be less for rare diseases than for more common diseases.

29.3.2 Standard Approach to Constructing a PRS: LD Clumping and p -Value Thresholding

A common approach to constructing a PRS uses GWAS results to select independent SNPs that reach genome-wide significance (typically a p -value of less than 5×10^{-8}) and weights the selected SNPs via their estimated regression coefficients, $w_i = \hat{\beta}_i$ (Figure 29.5) (Chatterjee *et al.*, 2016). Since this approach uses estimated coefficients from GWAS data, publicly available GWAS summary statistics can be used to construct the PRS. This approach typically yields PRS with small to modest discriminatory ability (Bao *et al.*, 2013; Chatterjee *et al.*, 2016; Krarup *et al.*, 2015; Scott *et al.*, 2013). For instance, one study of type 2 diabetes found that a PRS based

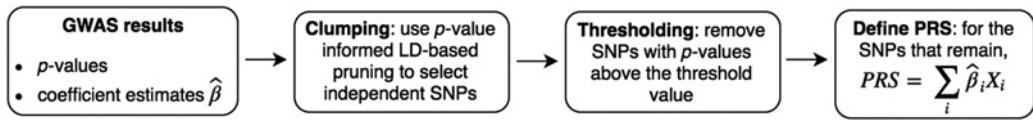


Figure 29.5 An illustration of the standard approach to constructing a PRS.

on 65 SNPs yielded an AUC of 0.60 (Talmud *et al.*, 2015) while a study of a 77-SNP breast cancer PRS reported an AUC of 0.62 (Mavaddat *et al.*, 2015). Likewise, a study of a 27-SNP colorectal cancer PRS found AUCs of 0.55–0.60 (Hsu *et al.*, 2015).

The genome-wide significance threshold often used for including SNPs in a PRS is quite stringent, and it has been shown that using a more liberal threshold can improve the utility of PRS for highly polygenic diseases such as schizophrenia (International Schizophrenia Consortium, 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), bipolar disorder (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011), and multiple sclerosis (International Multiple Sclerosis Genetics Consortium, 2010). The optimal threshold, in terms of the performance of the PRS in independent data, will depend upon the sample size and the genetic architecture of the trait (Chatterjee *et al.*, 2013; Dudbridge, 2013). A validation sample or cross-validation techniques could be used to find a suitable *p*-value threshold (Chatterjee *et al.*, 2016). Gains in accuracy resulting from searching for an optimal *p*-value threshold are expected to be modest with current sample sizes, though bigger gains may be possible for highly heritable and extremely polygenic traits (Chatterjee *et al.*, 2013; International Multiple Sclerosis Genetics Consortium, 2010; International Schizophrenia Consortium, 2009; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). When selecting a *p*-value threshold, the trade-off between sensitivity and specificity of SNP selection and its impact on performance is an important consideration. When PRS are constructed using estimated coefficients from GWAS, a more liberal threshold will lead to the inclusion of more true positive SNPs, but performance will ultimately diminish as more false positive SNPs are included (Chatterjee *et al.*, 2013). Broadly, the optimal *p*-value threshold is likely to be more liberal than the genome-wide significance threshold, though this will depend on the GWAS sample size and the underlying genetic architecture (Chatterjee *et al.*, 2013).

The approach described in Figure 29.5 involves removing correlated SNPs from consideration due to the potentially adverse effect of correlation on the performance of the PRS (Wu *et al.*, 2013). Essentially, if several SNPs on a locus are in high linkage disequilibrium (LD) with one another and are all included in the PRS without properly accounting for the joint effects of the SNPs, the contribution of that locus to the PRS will be exaggerated (Mak *et al.*, 2017). One common approach for identifying SNPs for inclusion is association-informed LD-based pruning, or clumping (Purcell *et al.*, 2007; Wray *et al.*, 2014). This involves sorting SNPs based on their marginal *p*-values and, based on this ranked list, iteratively removing SNPs in LD with a more highly ranked SNP (i.e. a more significant SNP). Typically, a fairly stringent LD threshold (e.g. flagging SNPs with $r^2 > 0.05$ as being in LD) is needed as inclusion of a large number of weakly correlated SNPs could lead to a PRS with poor performance. LD clumping is implemented in the program PLINK (Purcell *et al.*, 2007). When *p*-values are used to select SNPs for inclusion in the PRS ('thresholding'), performance is generally improved by first clumping the SNPs (Mak *et al.*, 2017). However, LD-based pruning in general could lead to the removal of SNPs in high LD but independently associated with *D*, thereby limiting the heritability explained (Chatterjee *et al.*, 2016; Vilhjálmsson *et al.*, 2015).

In the approach described in Figure 29.5, results from GWAS data, namely, *p*-values and coefficient estimates, are used to select and weight SNPs, respectively, in creating a PRS. However,

since only SNPs reaching a particular significance threshold are included, the corresponding coefficient estimates are biased away from the null due to the winner's curse (Shi *et al.*, 2016). This may in turn affect the performance of the PRS. Recently, a method has been proposed to construct a PRS while correcting for the winner's curse by using a shrinkage estimator (Shi *et al.*, 2016). This correction has been shown to lead to PRS with improved performance.

29.3.3 Advanced Approaches to Constructing a PRS

The utility of PRS built using LD clumping and *p*-value thresholding is generally less than is theoretically achievable given the heritability explained by SNPs (International Schizophrenia Consortium, 2009; Ripke *et al.*, 2013; Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Stahl *et al.*, 2012), motivating the consideration of alternative methods.

One such approach is LDpred, which estimates the posterior mean effect size for individual SNPs given a point-normal mixture prior on the effect sizes under a joint model that incorporates all SNPs simultaneously (Vilhjálmsson *et al.*, 2015). In particular, LDpred combines the prior on the genetic architecture and information on LD from a reference panel to calculate the posterior mean effect sizes using GWAS summary data (Vilhjálmsson *et al.*, 2015). Under certain modeling assumptions, this yields an optimal predictor of the disease in terms of bias and variance (Vilhjálmsson *et al.*, 2015). The point-normal mixture prior accommodates non-infinitesimal genetic architecture, i.e. that where only a fraction of the SNPs are associated with the disease (Vilhjálmsson *et al.*, 2015). Intuitively, the external LD information is used to approximately infer the underlying joint effects of the SNPs from the summary-level data (Yang *et al.*, 2012) and the point-normal mixture model performs regularization, which is often needed to fit high-dimensional models.

Improvements in performance have been seen for LDpred compared with clumping and *p*-value thresholding; in particular, increases in the Nagelkerke R^2 from 20.1% to 25.3% for schizophrenia and from 9.8% to 12% for multiple sclerosis have been observed (Vilhjálmsson *et al.*, 2015). Furthermore, it has been shown that as sample size increases, the prediction accuracy (R^2) of a PRS built with LDpred converges to the heritability explained by the SNPs (Vilhjálmsson *et al.*, 2015). In general, the advantage of LDpred over clumping and thresholding will grow with sample size, particularly for highly polygenic traits (Vilhjálmsson *et al.*, 2015). On the other hand, if the model for LD is misspecified, the performance of the PRS will be diminished (Vilhjálmsson *et al.*, 2015). Additionally, the LD reference panels on which the method relies may not be adequate for rare variants (Vilhjálmsson *et al.*, 2015).

Lassosum is another approach that builds a PRS by fitting a regularized model for the joint effects of SNPs using summary-level data and external LD information (Mak *et al.*, 2017). The key difference between lassosum and LDpred is the use of the lasso penalty function in lassosum (Mak *et al.*, 2017); thus, lassosum is expected to perform well if the effect sizes follow a double exponential prior. Improvements in accuracy for lassosum relative to clumping and *p*-value thresholding as well as LDpred have been found in some instances (Mak *et al.*, 2017). For example, for Crohn's disease, the AUC for lassosum was 0.71, while it was 0.68 and 0.65 for LDpred and clumping and thresholding, respectively; likewise, for rheumatoid arthritis, the AUC for lassosum was 0.70, compared to 0.68 for LDpred and 0.67 for clumping and thresholding (Mak *et al.*, 2017). In general, algorithms based on lasso-type (Tibshirani, 1996) approaches that do not require pre-selection of SNPs may do better than the standard approach described in Section 29.3.2, but improvements are typically small (Chatterjee *et al.*, 2013). Lassosum, similar to LDpred, is expected to offer gains over clumping and thresholding due to the incorporation of independent effects through explicit modeling of LD. The relative performance of

LDpred and lassosum, on the other hand, will depend on whether the true effect size distribution is closer to a point-normal mixture distribution or a double exponential distribution. Essentially, LDpred and lassosum are expected to perform well under certain effect size distributions, but may perform poorly otherwise. More flexible approaches have been proposed, including Dirichlet process regression, which uses the Dirichlet process to generate a prior for the effect size distribution, meaning that an effect size distribution can be inferred from the available data (Zeng and Zhou, 2017).

Pleiotropic, functional, and annotation information can be used to inform selection of SNPs for inclusion in a PRS via differential priors for different SNPs (Chatterjee *et al.*, 2016). Various mixed models, Bayesian methods, and penalized methods that allow external information to be incorporated through priors for the effect size distribution have been proposed (Golan and Rosset, 2014; Speed and Balding, 2014; Zhou *et al.*, 2013). These methods typically assume that the effect sizes have a symmetric distribution centered at zero, where the spread of the distribution is defined by one or more variance parameter(s), depending upon the degree of complexity of the model (Chatterjee *et al.*, 2016). This distribution serves to shrink the estimated effect sizes towards the null to improve the bias–variance trade-off, as these both lead to imprecision in the PRS (Chatterjee *et al.*, 2016). The form of the prior dictates the degree of shrinkage (Chatterjee *et al.*, 2016). These methods have been shown to improve estimation of risk of coronary artery disease, Crohn's disease, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes, increasing the correlation between observed outcomes and estimated risks from 0.21 to 0.30, on average (Speed and Balding, 2014).

In general, more work needs to be done on how to accommodate rare variants in these methods, particularly for approaches that rely on, for example, LD reference panels (Vilhjálmsson *et al.*, 2015). Finally, several methods have been proposed that use genotype data to construct a PRS without assuming independent SNPs, though they cannot be used with GWAS summary statistics (Abraham *et al.*, 2013, 2014; Erbe *et al.*, 2012; Habier *et al.*, 2011; Meuwissen *et al.*, 2001; Ogutu *et al.*, 2012; Pirinen *et al.*, 2013; Szymczak *et al.*, 2009; Zhou *et al.*, 2013).

29.4 Combining PRS and Epidemiologic Factors

Once a PRS has been constructed, it needs to be combined with epidemiologic risk factors (Chatterjee *et al.*, 2016). In particular, a model for the joint effects of the PRS and the other risk factors must be formulated, which involves characterizing risk for individual factors and considering interactions (Chatterjee *et al.*, 2016). Ideally, in order to avoid selection and recall biases (Wacholder *et al.*, 1992a,b,c), data from a prospective cohort study should be used to estimate risk associated with epidemiologic risk factors (Chatterjee *et al.*, 2016). However, estimation of multiplicative interactions, which tends to be less sensitive to selection bias, could be done in a wider range of studies (Chatterjee *et al.*, 2016; Wacholder *et al.*, 2002). Additionally, case-only analyses can be used to investigate interactions with greater power, provided certain assumptions, such as independence of the genetic and environmental risk factors, are met (Chatterjee and Carroll, 2005; Piegorsch *et al.*, 1994; Umbach and Weinberg, 1997). Although the omission of interaction terms may not have a large impact on the ability of the model to discriminate between cases and controls (Aschard *et al.*, 2012), it can lead to discrepancies between observed proportions of cases and estimated risks, particularly at the extremes of the risk distribution. In general, as tests of individual interaction terms may have low power, it is important to assess the overall fit of the model.

Family history should be considered for inclusion as an epidemiologic factor as the PRS will generally only explain a portion of risk associated with family history (Chatterjee *et al.*, 2013,

2016). Typically, in population-based studies of unrelated individuals, a binary variable is used to model family history, representing presence in a first-degree relative (Chatterjee *et al.*, 2013, 2016). This approach is useful in the general population, though more granular and extensive family history data could improve the ability of the model to stratify risk, particularly in highly affected families (Chatterjee *et al.*, 2013; Lee *et al.*, 2014; Mazzola *et al.*, 2015). When family history and a PRS are modeled together, the effect of family history will be reduced to the degree that heritability is explained by the PRS (Maas *et al.*, 2016b).

Recall that PRS typically include common variants with weak or modest individual effects. For diseases for which family studies have identified rare variants with strong effects, including both the PRS, which is useful in the general population, and these rare variants, which are important in highly affected families, in the risk model may be a useful strategy. For instance, combining the PRS and the results of gene panel testing of rare variants in a model yields a single risk score that is useful in both the general population and in families. Furthermore, in individuals with rare variants, the PRS may provide additional information, leading to meaningful risk stratification within this group, as has been demonstrated for male breast and prostate cancer (Lecarpentier *et al.*, 2017).

29.5 Model Validation

Broadly, model validation compares model estimates to what is observed (Taylor *et al.*, 2008). The clinical utility of a model (discussed in Section 29.6), regardless of how it is measured, will depend upon the statistical validity of the model. Thus, evaluating the validity of a disease risk model is essential. For disease risk models, model validity is typically assessed by considering calibration, the ability of a model to provide unbiased estimates of risk for individuals with different risk factor profiles, and discrimination, the degree to which the model can differentiate between cases and controls (Freedman *et al.*, 2005). In general, a valid model will be well calibrated and have at least modest discriminatory ability. The validity of a model can be influenced by bias resulting from model misspecification, bias in parameter estimates, uncertainty due to small sample size, and/or issues of generalizability (Gail *et al.*, 1989). It is essential that model validity be evaluated in a large, representative, and independent cohort of healthy individuals. Data from nested case-control studies (i.e. case-control studies sampled from prospective cohorts) can also be used, but doing so requires using sampling weights to determine the underlying disease rate (Chatterjee *et al.*, 2016).

Typically, calibration, the degree to which estimated risks match the observed proportion of individuals who develop D , is evaluated by creating strata based on estimated risk and, within each stratum, comparing the average of the estimated risks to the proportion of individuals who develop or experience D in a given time interval (Chatterjee *et al.*, 2016). Graphical displays are useful as they illustrate patterns of miscalibration (Steyerberg, 2009). Various statistical tests can be used to measure calibration either overall or at the extremes of risk (Song *et al.*, 2015; Steyerberg, 2009). Good calibration is particularly important at the extremes of estimated risk, where estimates have the potential to have the greatest clinical relevance (Song *et al.*, 2015). It is possible to have over- or underestimation of risk in some or all strata owing to different causes of miscalibration (Figure 29.6) (Chatterjee *et al.*, 2016). Poor calibration in certain strata may be found if, for example, the relative risk model is misspecified due to the omission of nonlinear or non-additive effects (Chatterjee *et al.*, 2016). On the other hand, if the disease incidence rate in the population used to develop the model does not reflect that in the population to which the model is applied, the result may be consistently over- or underestimated risks across strata (Chatterjee *et al.*, 2016). Since the degree to which the observed proportions and estimated risks agree can differ by various factors, such as ethnicity and geography, it is important to

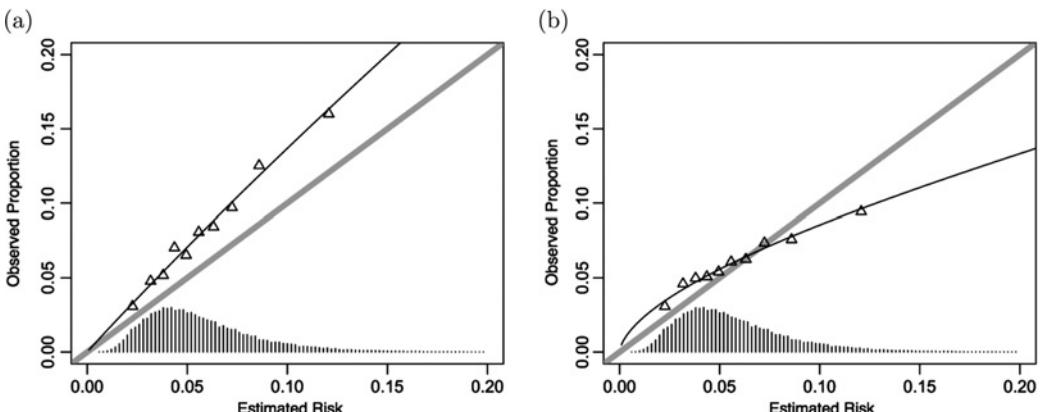


Figure 29.6 Example calibration plots based on simulated data. These plots present the observed proportion of cases versus the estimated risk for ten evenly-sized strata defined by estimated risk (denoted by triangles). Perfect calibration is indicated by the gray diagonal line. The solid line is logistic calibration, which is based on the regression model $\text{logit}[P(D = 1|\hat{\pi})] = \beta_0 + \beta_1\hat{\pi}$, where $\hat{\pi}$ is the estimated risk. The histogram on the x -axis is the distribution of estimated risk. In (a), the model is applied to a population with a different incidence rate (6.1% in the population used to fit the model versus 8.3% in the population to which the model is applied), leading to underestimation of risk in all strata. In (b), the model is applied to a population where the underlying relative risk model is misspecified; in particular, the incidence rate is nearly identical in the two populations (6.1% and 6.0%), but $\text{logit}[P(D = 1|\hat{\pi})] = \beta_0 + 0.6\hat{\pi}$ in the new population. This results in miscalibrated estimates of risk in certain strata. These plots were made using the `val.prob` function in the R package `rms`.

assess calibration in different populations at different times (Chatterjee *et al.*, 2016). If a model is found to have poor calibration, recalibration, which typically involves scaling and/or shifting estimates of risk, is possible (Freedman *et al.*, 2005).

Measures like the ROC curve and the AUC can be used to evaluate the ability of the model to discriminate between cases and controls. The AUC depends upon the number of risk factors included in the model and their association with D , as well as issues such as imprecision in parameter estimates and model misspecification. For a well-calibrated PRS and a rare disease, the discriminatory ability of a model depends on the variability of the PRS in the population (see Section 29.3.1). For long-term outcomes, for which risks are generally harder to estimate accurately, a high AUC (e.g., above 0.80) may not be expected. While the assessment of calibration requires data from a cohort study (or data from a nested case-control study with sampling weights), it has recently been shown that data from two-phase studies can be used to evaluate the AUC (Choudhury *et al.*, 2017).

29.6 Evaluating Clinical Utility

Following validation, a model should be evaluated for its potential utility in clinical or public health applications (Chatterjee *et al.*, 2016). It is important to evaluate clinical utility in order to determine whether estimated risks can be effectively used to inform prevention or treatment strategies (Chatterjee *et al.*, 2016). Broadly speaking, clinical utility is a function of the variability in risk estimates generated by the model for the population to which it will be applied, provided the model is well calibrated (Chatterjee *et al.*, 2016). In general, the most informative criteria for judging clinical utility will depend upon the application (Chatterjee *et al.*, 2016).

A high AUC is often considered necessary for a model to be useful. However, the AUC has several limitations and reliance on this measure alone can lead to false conclusions about the clinical utility of a model. First, the measure itself, which is equivalent to the probability that

a randomly selected case has an estimated risk greater than that of a randomly selected control, does not have a direct clinical interpretation. Other measures, such as the proportion of the population and the proportion of cases whose estimated risks exceed clinically meaningful risk thresholds (Pfeiffer and Gail, 2011; Pharoah *et al.*, 2002; So *et al.*, 2011), can be used to obtain more direct insight into the clinical utility of risk models. Furthermore, the AUC does not depend on the incidence rate of the disease in the population. The ability of models to stratify risk, however, depends heavily on the rate of the disease in the population. For instance, it has been shown that a breast cancer disease risk model involving a PRS and epidemiologic risk factors with only modest discriminatory ability (i.e. an AUC of 0.65) can provide sufficient stratification of risk so as to be useful for identifying high-risk individuals (Maas *et al.*, 2016a). On the other hand, for a less common disease, such as ovarian cancer, a model with similar discriminatory power is expected to produce less risk stratification and thus have less clinical utility.

Ideally, in order to maximize benefit and minimize harm, a disease risk model should identify a small portion of the total population that includes a majority of cases (Chatterjee *et al.*, 2016). Only models with very high AUCs (e.g. above 0.9) can achieve such optimal targeting of individuals. In practice, risk models with more modest AUCs may be able to identify a substantial fraction of individuals at high risk that could be targeted for more invasive intervention. However, a large majority of cases in the population are still expected to fall outside this high-risk group, requiring broader prevention efforts to reduce the total burden of the disease in the population (Park *et al.*, 2012).

For diseases with recommended primary or secondary prevention strategies (e.g. statins for heart disease and mammographic screening for breast cancer, respectively) based on a specific risk threshold, it is useful to evaluate the proportion of the population and the proportion of future cases that would be eligible for such prevention strategies based on a new disease risk model (Naylor and Vasan, 2016; US Preventive Services Task Force, 2016). Similarly, there has been some recent work on approaches to assess the value of adding new risk factors to a model. One set of methods considers reclassification, which relies on risk thresholds to cross-classify subjects (Chatterjee *et al.*, 2016). Different types of reclassification indices quantify the extent of useful reclassification provided by a new model (Pencina *et al.*, 2008, 2011). However, these measures can be abstract and may not have a direct clinical interpretation as they do not relate to any measure of the net benefit of using the new model (Chatterjee *et al.*, 2016). An index without thresholds has been proposed, but it has been found to be susceptible to false positive results, incorrectly suggesting the new model offers improved performance (Kerr *et al.*, 2014; Pepe *et al.*, 2014).

Other ways to assess the clinical utility of a model include a plot of the proportion of the population with estimated risks above a threshold and the proportion of cases with estimated risks above this threshold vs. the threshold value (Figure 29.7) or a plot of the proportion of cases above a risk threshold against the proportion of the population above this threshold (Pharoah *et al.*, 2002). As many disease risk models are intended to be deployed clinically to aid in decision-making, it may also be important to evaluate whether use of the model influences behavior on the part of clinicians and/or patients and, importantly, patient outcomes (Usher-Smith *et al.*, 2015). Other possible measures of clinical utility will be discussed in the context of the breast cancer example in Section 29.7.

29.7 Example: Breast Cancer

Several models for breast cancer risk have been proposed over the last thirty years, including models with and without genetic risk factors. One of the first such models is the Gail model, introduced in Section 29.1.2 (Gail *et al.*, 1989). This model uses epidemiologic risk factors

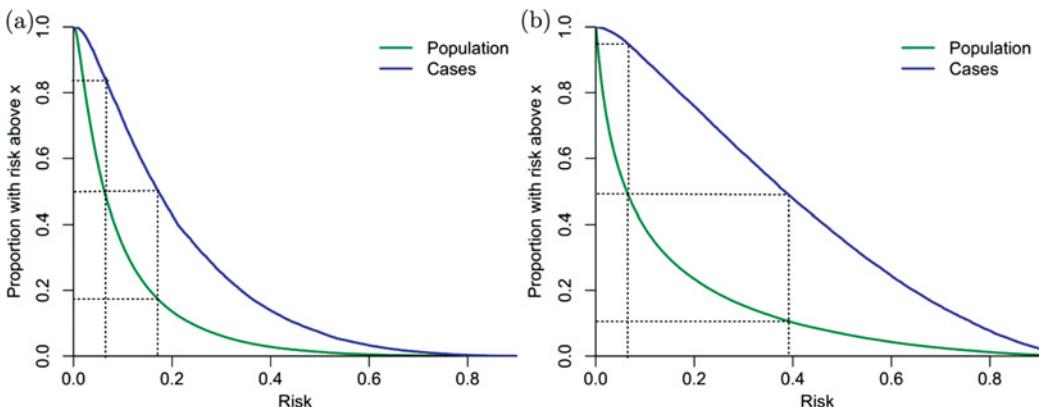


Figure 29.7 These plots present the proportion of individuals with estimated risks above a threshold versus the threshold for a given disease risk model. The results are presented separately for the population and for cases. The disease risk model in (a) has an AUC of 0.78. In order to identify 50% of cases, a risk threshold of 17% should be used; such a threshold would in turn identify 18% of the population overall. Furthermore, 50% of the population has a risk above 6.3%, including 85% of cases. The plot in (b) presents the results for a disease risk model with an AUC of 0.90. Here, a risk threshold of 39% identifies 50% of cases and 11% of the population. In addition, 50% of the population has a risk above 6.3%, including 95% of cases. Adapted with permission from Pharoah *et al.* (2002), Springer Nature.

to estimate absolute breast cancer risk. The first version of this model, BCRAT1, uses age at menarche, number of previous breast biopsies, age at first live birth, and number of first-degree relatives with breast cancer to estimate a woman's absolute risk of breast cancer within a particular time period given her current age (Gail *et al.*, 1989). This model was later updated, giving the BCRAT2 model, which includes mammographic density (Chen *et al.*, 2006). Although the inclusion of this strong risk factor yielded improvements in the discriminatory ability of the model (an AUC of 0.64 for BCRAT2 compared to an AUC of 0.60 for BCRAT1) (Chen *et al.*, 2006), BCRAT2 requires imaging data; for this reason, BCRAT1 is still widely used in clinical settings.

Linkage studies of breast cancer in highly affected families and subsequent positional cloning have identified the major genes *BRCA1* and *BRCA2*, which are known to contain a spectrum of rare high-penetrant mutations (Hall *et al.*, 1990; Wooster *et al.*, 1994). While these mutations are common in highly affected families, they explain only a small fraction of the cases that arise in the general population. Thus, Pharoah *et al.* (2002) investigated the potential of a polygenic model to provide risk stratification in the general population. Based on previous epidemiologic studies, Pharoah *et al.* assumed the familial relative risk among siblings was 2, and used this to estimate the variability of the underlying PRS on the log-risk scale, giving 1.2 (i.e. $\sigma^2 = 1.2$). This estimate of variability is also supported by estimates from a polygenic model fit to data from affected families. Pharoah *et al.* concluded that this variability in genetic risk was of a large enough degree to be useful. They estimated that for a PRS capturing all of this variation in genetic risk, the 50% of individuals at highest risk would include 88% of breast cancer cases. Furthermore, Pharoah *et al.* noted that even if a PRS captured only 50% of the genetic risk, the polygenic approach would be preferred over a model with standard epidemiologic risk factors alone.

The International Breast Cancer Intervention Study (IBIS) model (Tyrer *et al.*, 2004) uses information on epidemiologic risk factors to provide an estimate of breast cancer risk, accounting for the likelihood of having a particular *BRCA1/2* genotype. This model uses extensive family history information to estimate the risk of a particular *BRCA1/2* genotype and uses these

estimates in conjunction with information on clinical factors (e.g., age at menarche, height, weight, menopause status, number of births, and age at first birth) to estimate a woman's risk of breast cancer over a certain time interval, given her current age.

The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model estimates breast cancer risk based on major gene effects and the effect of residual familial aggregation due to an unobserved polygenic component (Antoniou *et al.*, 2004, 2008a; Lee *et al.*, 2016). The BOADICEA model has a similar form to the absolute risk models discussed in Section 29.2: $\lambda(a|G, P) = \lambda_0(a) \exp(G(a) + P(a))$, where $G(a)$ is the major gene (*BRCA1/BRCA2*, *PALB2*, *CHEK2*, and *ATM*) effect at age a and $P(a) \sim N(0, \sigma^2)$ is the polygenic effect at age a (Antoniou *et al.*, 2001, 2008a; Mavaddat *et al.*, 2010). Parameter estimates were obtained via complex segregation analysis of data from highly affected families, accounting for ascertainment and using population-based breast cancer incidence rates to calibrate the baseline hazard (Antoniou *et al.*, 2001, 2004, 2008a). Such complex segregation analysis relies on disease information from family members within a pedigree. The BOADICEA model has been validated in families and is used primarily in the management of familial breast cancer (Antoniou *et al.*, 2008b; National Collaborating Centre for Cancer, 2013; Riley *et al.*, 2012; Smith *et al.*, 2012). Recent updates to the model allow cancer incidence to vary by region and ethnicity and include breast cancer subtypes as different endpoints (Lee *et al.*, 2014; Mavaddat *et al.*, 2010). When applying the model to new populations, several issues must be considered, including whether the effects of the major genes on risk are the same in different populations and whether the distribution of the polygenic component is similar across populations (Lee *et al.*, 2014).

Recent studies have investigated the utility of PRS based on SNPs discovered in GWAS in the general population (Darabi *et al.*, 2012; Gail, 2008; Hüsing *et al.*, 2012; Maas *et al.*, 2016a; Mavaddat *et al.*, 2015; Wacholder *et al.*, 2010). Most recently, Maas *et al.* (2016a) investigated whether low-penetrant common SNPs, in combination with epidemiologic risk factors, could be useful in guiding strategies for breast cancer prevention. The approach involved using data from nested case-control studies sampled from a consortium of prospective cohort studies to fit a multivariate logistic regression model for a 92-SNP PRS and a number of epidemiologic risk factors. The PRS was constructed in two steps. First, genotype data on 24 SNPs was used to construct a PRS based on a logistic regression model. Then published odds ratios for an additional 68 SNPs for which genotype data were not available were used to simulate a 68-SNP PRS, which was then added to the 24-SNP PRS. Models for absolute risk were then constructed by combining odds ratio estimates, the distribution of epidemiologic risk factors from national health surveys, age-specific breast cancer rates from registry data, and information on mortality from registry data.

To demonstrate the utility of the model, Maas *et al.* showed the distribution of absolute risk in the population of white women in the USA according to three models: a model with epidemiologic risk factors alone, the PRS alone, and a model with epidemiologic risk factors and the PRS. They showed that including the PRS provided substantially stronger stratification of absolute risk than was obtained using the model with epidemiologic risk factors alone. As a specific application, they showed that a combined model can identify 16% of the population who are age 40, and hence may not be recommended for screening under current guidelines, but have an estimated risk that is higher than the risk of an average 50-year-old woman, who would generally be recommended for screening. Similarly, they showed that a combined model could identify 32% of women who are 50 years old, but have an estimated risk that is below that of an average 40-year-old woman.

Furthermore, to demonstrate the potential utility of the combined absolute risk model for counseling women on lifestyle interventions, Maas *et al.* assessed the distribution of estimated

risk due to modifiable factors in categories defined by the estimated risk from non-modifiable factors. They found that women at high risk from non-modifiable factors could have risks comparable to an average woman in the population if they made healthier lifestyle choices, that is, did not drink or smoke, maintained a low body mass index, and did not use menopausal hormone replacement therapy. In general, Maas *et al.* found greater spread in risk from modifiable risk factors at higher levels of risk due to non-modifiable factors. This translates into a higher proportion of cancers that could be prevented by modifying risk factors among those at higher levels of risk due to non-modifiable factors. Specifically, they found that nearly 30% of cancers could be prevented by risk factor modification and nearly 20% of these cancers were in women in the highest decile of non-modifiable risk.

29.8 Discussion

29.8.1 Future Directions

It is anticipated that GWAS sample sizes will continue to grow, allowing the discovery of more risk variants, including both common and rare variants, and increasing the precision of estimates of risk associated with individual variants. Additionally, developments in the use of electronic medical records, the creation of tools for big data management and analysis, and advances in methods for collecting and analyzing biological specimens will lead to larger studies (Chatterjee *et al.*, 2016). To reach their full potential, PRS must include common, low-frequency, and rare variants (Chatterjee *et al.*, 2016). Our understanding of rare and low-frequency variants, which may explain a substantial fraction of heritability and, consequently, variability in risk (Mancuso *et al.*, 2016), will continue to increase through large-scale sequencing and imputation-based studies (Chatterjee *et al.*, 2016). Rare, high-penetrant variants are particularly important in highly affected families (Chatterjee *et al.*, 2016). The development of more sophisticated methods to account for LD, incorporate functional annotation of SNPs, and/or perform pleiotropic analysis across related traits is also likely to lead to improvements in the utility of PRS (Andreassen *et al.*, 2013; Chatterjee *et al.*, 2016; Chung *et al.*, 2014; Korte *et al.*, 2012; Li *et al.*, 2014; Mak *et al.*, 2017; Vilhjálmsson *et al.*, 2015).

Further research is needed on methods to evaluate clinical utility and issues related to implementation of risk models in general (Freedman *et al.*, 2005). In terms of implementation, it will be important to consider how clinicians convey risk to their patients, how decisions are made by patients, how those decisions are influenced by disease risk models, and the effect of such decisions on behavior (Freedman *et al.*, 2005). In addition, it will be important in the future to incorporate patient preferences into model development, validation, and evaluation (Freedman *et al.*, 2005).

For diseases with common and strong risk factors, incorporation of a PRS into a disease risk model is not expected to yield large improvements in model performance in terms of the AUC (Clayton, 2009). However, the inclusion of a PRS could lead to the identification of a substantial number of individuals at the extremes of disease risk who could be recommended for or against interventions, which each have risks and benefits. For diseases with no known risk factors or only weak risk factors, developing disease risk models based on a PRS may offer new opportunities for risk-based interventions.

29.8.2 Challenges

The development, validation, evaluation, and deployment of disease risk models with genetic factors face several challenges. Issues related to the implications of widespread genetic testing,

the regulation of genetic testing, and standards for clinical utility must be addressed (Easton *et al.*, 2015; Evans *et al.*, 2015; Gabai-Kapara *et al.*, 2014; Grosse and Khoury, 2006; King *et al.*, 2014; Thomas *et al.*, 2015). Acceptance of widespread genetic testing may hinge on whether the results lead to 'useful action' (Pharoah *et al.*, 2002). In addition, the use of disease risk models in practice requires integrating these tools into the existing healthcare system (Rogowski *et al.*, 2009) and evaluating the impact of risk-based interventions and procedures (Chatterjee *et al.*, 2016). This includes the development of user-friendly implementations of disease risk models; in some regions (e.g. the European Union) such implementations are classified as medical devices, requiring regulatory approval prior to clinical use. The consequences of using disease risk models must be considered at both the individual level and the population level, including the effects on patient outcomes and the economic repercussions of widespread adoption (Usher-Smith *et al.*, 2015). Additional challenges to clinical implementation include choosing a risk model, choosing when and where risk should be estimated, understanding and overcoming barriers to use, and choosing thresholds for intervention (Usher-Smith *et al.*, 2015).

Furthermore, many disease risk models are developed, validated, and/or evaluated using predominantly Caucasian data sets, making application to other ethnicities potentially problematic (Freedman *et al.*, 2005). In particular, methods for developing absolute risk models, including constructing a PRS, for different ethnic populations will require increasing attention. As data from large GWAS are available in Caucasian samples and genetic risks for many complex traits appear to be similar across populations, PRS for non-Caucasian populations could be built efficiently by borrowing information from earlier studies. However, for some traits there is evidence that a PRS developed in one population may behave differently when applied to a new population, indicating that caution is required (Márquez-Luna *et al.*, 2017; Martin *et al.*, 2017). Another challenge is a lack of population-based registry data for many non-cancer outcomes, leading to challenges in developing absolute risk models for these traits (Freedman *et al.*, 2005).

As indicated in Section 29.8.1, there is evidence that very large sample sizes will be needed to substantially improve the utility of PRS, even for diseases where much of the heritability is due to common SNPs (Chatterjee *et al.*, 2013). This could be particularly challenging for relatively rare diseases, where accruing a sufficient number of cases will require sustained effort over many years. Finally, as mentioned in Section 29.1.3, models with epidemiologic risk factors in addition to PRS face additional potential challenges as these factors can change over time. Repeated measurements may therefore be required to accurately assess risk, necessitating expensive longitudinal cohort studies and methods for dynamic risk estimation (Chatterjee *et al.*, 2016).

References

- Abraham, G., Kowalczyk, A., Zobel, J. and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology* **37**(2), 184–195.
- Abraham, G., Tye-Din, J., Bhalala, O., Kowalczyk, A., *et al.* (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genetics* **10**(2), e1004137.
- Andreassen, O., Thompson, W., Schork, A., Ripke, S., *et al.* (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics* **9**(4), e1003455.
- Antoniou, A. and Easton, D. (2006). Risk prediction models for familial breast cancer. *Future Medicine* **2**(2), 257–274.

- Antoniou, A., Pharoah, P., McMullan, G., Day, N., et al. (2001). Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. *Genetic Epidemiology* **21**(1), 1–18.
- Antoniou, A., Pharoah, P., Smith, P. and Easton, D. (2004). The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British Journal of Cancer* **91**(8), 1580–1590.
- Antoniou, A., Cunningham, A., Peto, J., Evans, D., et al. (2008a). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. *British Journal of Cancer* **98**(8), 1457–1466.
- Antoniou, A., Hardy, R., Walker, L., Evans, D., et al. (2008b). Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: Validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *Journal of Medical Genetics* **45**(7), 425–431.
- Aschard, H., Chen, J., Cornelis, M., Chibnik, L., et al. (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *American Journal of Human Genetics* **90**(6), 962–972.
- Bao, W., Hu, F., Rong, S., Rong, Y., et al. (2013). Predicting risk of type 2 diabetes mellitus with genetic risk models on the basis of established genome-wide association markers: A systematic review. *American Journal of Epidemiology* **178**(8), 1197–1207.
- Chang, E., Smedby, K., Hjalgrim, H., Glimelius, B. and Adami, H.-O. (2006). Reliability of self-reported family history of cancer in a large case-control study of lymphoma. *Journal of the National Cancer Institute* **98**(1), 61–68.
- Chatterjee, N. and Carroll, R. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**(2), 399–418.
- Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., et al. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genetics* **45**(4), 400–405.
- Chatterjee, N., Shi, J. and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**(7), 392–406.
- Chen, G.-B., Lee, S., Brion, M., Montgomery, G., et al. (2014). Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Human Molecular Genetics* **23**(17), 4710–4720.
- Chen, J., Pee, D., Ayyagari, R., Graubard, B., et al. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *Journal of the National Cancer Institute* **98**(17), 1215–1226.
- Choudhury, P., Chaturvedi, A. and Chatterjee, N. (2017). Evaluating discriminatory accuracy of models using partial risk-scores in two-phase studies. Preprint, arXiv:1710.04379.
- Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2014). GPA: A statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics* **10**(11), e1004787.
- Clayton, D. (2009). Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genetics* **5**(7), e1000540.
- Cox, R.D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**(2), 187–220.
- D'Agostino, Sr, R., Vasan, R., Pencina, M., Wolf, P., et al. (2008). General cardiovascular risk profile for use in primary care. *Circulation* **117**(6), 743–753.
- Darabi, H., Czene, K., Zhao, W., Liu, J., et al. (2012). Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. *Breast Cancer Research* **14**(1), R25.

- Domchek, S., Eisen, A., Calzone, K., Stopfer, J., et al. (2003). Application of breast cancer risk prediction models in clinical practice. *Journal of Clinical Oncology* **21**(4), 593–601.
- Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* **9**(3), e1003348.
- Easton, D., Pharoah, P., Antoniou, A., Tischkowitz, M., et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *New England Journal of Medicine* **372**(23), 2243–2257.
- Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science* **95**(7), 4114–4129.
- Evans, B., Burke, W. and Jarvik, G. (2015). The FDA and genomic tests – getting regulation right. *New England Journal of Medicine* **372**(23), 2258–2264.
- Franke, A., McGovern, D., Barrett, J., Wang, K., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42**(12), 1118–1125.
- Freedman, A., Graubard, B., Rao, S., McCaskill-Stevens, W., et al. (2003). Estimates of the number of US women who could benefit from tamoxifen for breast cancer chemoprevention. *Journal of the National Cancer Institute* **95**(7), 526–532.
- Freedman, A., Seminara, D., Gail, M., Hartge, P., et al. (2005). Cancer risk prediction models: A workshop on development, evaluation, and application. *Journal of the National Cancer Institute* **97**(10), 715–723.
- Gabai-Kapara, E., Lahad, A., Kaufman, B., Friedman, E., et al. (2014). Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. *Proceedings of the National Academy of Sciences of the United States of America* **111**(39), 14205–14210.
- Gail, M. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute* **100**(14), 1037–1041.
- Gail, M. (2011). Personalized estimates of breast cancer risk in clinical practice and public health. *Statistics in Medicine* **30**(10), 1090–1104.
- Gail, M., Brinton, L., Byar, D., Corle, D., et al. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**(24), 1879–1886.
- Gail, M., Costantino, J., Bryant, J., Croyle, R., et al. (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute* **91**(21), 1829–1846.
- Ganna, A. and Ingelsson, E. (2015). 5 year mortality predictors in 498,103 UK Biobank participants: A prospective population-based study. *Lancet* **386**(9993), 533–540.
- Gerds, T.A., Scheike, T.H., Blanche, P. and Ozenne, B. (2017). riskRegression: Risk regression models and prediction scores for survival analysis with competing risks. R package version 1.4.3.
- Golan, D. and Rosset, S. (2014). Effective genetic-risk prediction using mixed models. *American Journal of Human Genetics* **95**(4), 383–393.
- Golan, D., Lander, E. and Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America* **111**(49), E5272–E5281.
- Grosse, S. and Khoury, M. (2006). What is the clinical utility of genetic testing? *Genetics in Medicine* **8**(7), 448–450.
- Gusev, A., Lee, S., Trynka, G., Finucane, H., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics* **95**(5), 535–552.

- Habier, D., Fernando, R., Kizilkaya, K. and Garrick, D. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**(1), 186.
- Hall, J., Lee, M., Newman, B., Morrow, J., et al. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**(4988), 1684–1689.
- Hill, G., Connelly, J., Hébert, R., Lindsay, J. and Millar, W. (2003). Neyman's bias re-visited. *Journal of Clinical Epidemiology* **56**(4), 293–296.
- Hsu, L., Jeon, J., Brenner, H., Gruber, S., et al. (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* **148**(7), 1330–1339.
- Hüsing, A., Canzian, F., Beckmann, L., Garcia-Closas, M., et al. (2012). Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *Journal of Medical Genetics* **49**(9), 601–608.
- International Multiple Sclerosis Genetics Consortium. (2010). Evidence for polygenic susceptibility to multiple sclerosis – the shape of things to come. *American Journal of Human Genetics* **86**(4), 621–625.
- International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**(7256), 748–752.
- Janssens, A., Aulchenko, Y., Elefante, S., Borsboom, G., et al. (2006). Predictive testing for complex diseases using multiple genes: Fact or fiction? *Genetics in Medicine* **8**(7), 395.
- Jostins, L. and Barrett, J. (2011). Genetic risk prediction in complex disease. *Human Molecular Genetics* **20**(R2), R182–R188.
- Kerber, R. and Slattery, M. (1997). Comparison of self-reported and database-linked family history of cancer data in a case-control study. *American Journal of Epidemiology* **146**(3), 244–248.
- Kerr, K., Wang, Z., Janes, H., McClelland, R., et al. (2014). Net reclassification indices for evaluating risk prediction instruments: A critical review. *Epidemiology* **25**(1), 114–121.
- King, M.-C., Levy-Lahad, E. and Lahad, A. (2014). Population-based screening for BRCA1 and BRCA2: 2014 Lasker Award. *Journal of the American Medical Association* **312**(11), 1091–1092.
- Korte, A., Vilhjálmsson, B., Segura, V., Platt, A., et al. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44**(9), 1066–1071.
- Kraft, P. and Hunter, D. (2009). Genetic risk prediction – are we there yet? *New England Journal of Medicine* **360**(17), 1701–1703.
- Krarup, N., Borglykke, A., Allin, K., Sandholt, C., et al. (2015). A genetic risk score of 45 coronary artery disease risk variants associates with increased risk of myocardial infarction in 6041 Danish individuals. *Atherosclerosis* **240**(2), 305–310.
- Kundu, S., Aulchenko, Y.S. and Janssens, A. (2015). PredictABEL: Assessment of risk prediction models. R package version 1.2.2.
- Lango, H., Allen, Estrada, K., Lettre, G., Berndt, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**(7317), 832–838.
- Lecarpentier, J., Silvestri, V., Kuchenbaecker, K., Barrowdale, D., et al. (2017). Prediction of breast and prostate cancer risks in male BRCA1 and BRCA2 mutation carriers using polygenic risk scores. *Journal of Clinical Oncology* **35**(20), 2240–2250.
- Lee, A., Cunningham, A., Kuchenbaecker, K., Mavaddat, N., et al. (2014). BOADICEA breast cancer risk prediction model: Updates to cancer incidences, tumour pathology and web interface. *British Journal of Cancer* **110**(2), 535–545.
- Lee, A., Cunningham, A., Tischkowitz, M., Simard, J., et al. (2016). Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. *Genetics in Medicine* **18**(12), 1190.

- Lee, S., Wray, N., Goddard, M. and Visscher, P. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**(3), 294–305.
- Lee, S., DeCandia, T., Ripke, S., Yang, J., et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44**(3), 247–250.
- Lee, S., Harold, D., Nyholt, D., ANZGene Consortium, et al. (2013). Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Human Molecular Genetics* **22**(4), 832–841.
- Li, C., Yang, C., Gelernter, J. and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human Genetics* **133**(5), 639–650.
- Lloyd-Jones, D. (2010). Cardiovascular risk prediction: Basic concepts, current status, and future directions. *Circulation* **121**(15), 1768–1777.
- Lu, Y., Ek, W., Whiteman, D., Vaughan, T., et al. (2014). Most common 'sporadic' cancers have a significant germline genetic component. *Human Molecular Genetics* **23**(22), 6112–6118.
- Maas, P., Barrdahl, M., Joshi, A., Auer, P., et al. (2016a). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncology* **2**(10), 1295–1302.
- Maas, P., Wheeler, W., Brook, M., Check, D., et al. (2016b). iCARE: An R package to build and apply absolute risk models. Preprint, bioRxiv 079954.
- Mak, T., Porsch, R., Choi, S., Zhou, X. and Sham, P. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41**(6), 469–480.
- Manchanda, R., Legood, R., Burnell, M., McGuire, A., et al. (2014). Cost-effectiveness of population screening for BRCA mutations in Ashkenazi Jewish women compared with family history-based testing. *Journal of the National Cancer Institute* **107**(1), 380.
- Manchanda, R., Patel, S., Gordeev, V., Antoniou, A., et al. (2018). Cost-effectiveness of population-based BRCA1, BRCA2, RAD51C, RAD51D, BRIP1, PALB2 mutation testing in unselected general population women. *Journal of the National Cancer Institute* **110**(7), 714–725.
- Mancuso, N., Rohland, N., Rand, K., Tandon, A., et al. (2016). The contribution of rare variation to prostate cancer heritability. *Nature Genetics* **48**(1), 30–35.
- Manolio, T., Collins, F., Cox, N., Goldstein, D., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753.
- Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Price, A. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology* **41**(8), 811–823.
- Martin, A., Gignoux, C., Walters, R., Wojcik, G., et al. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics* **100**(4), 635–649.
- Mavaddat, N., Rebbeck, T., Lakhani, S., Easton, D. and Antoniou, A. (2010). Incorporating tumour pathology information into breast cancer risk prediction algorithms. *Breast Cancer Research* **12**(3), R28.
- Mavaddat, N., Pharoah, P., Michailidou, K., Tyrer, J., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *Journal of the National Cancer Institute* **107**(5).
- Mazzola, E., Blackford, A., Parmigiani, G. and Biswas, S. (2015). Recent enhancements to the genetic risk prediction model BRCAPO. *Cancer Informatics* **14**(S2), 147–157.
- Meuwissen, T., Hayes, B. and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4), 1819–1829.
- Mitchell, R., Brewster, D., Campbell, H., Porteous, M., et al. (2004). (2013). Accuracy of reporting of family history of colorectal cancer. *Gut* **53**(2), 291–295.

- National Collaborating Centre for Cancer. *Familial Breast Cancer: Classification and Care of People at Risk of Familial Breast Cancer and Management of Breast Cancer and Related Risks in People with a Family History of Breast Cancer*. National Collaborating Centre for Cancer, Cardiff (2013).
- National Institutes of Health. All of Us Research Program (2018). URL <https://allofus.nih.gov/>.
- Naylor, M. and Vasan, R. (2016). Recent update to the US cholesterol treatment guidelines: A comparison with international guidelines. *Circulation* **133**(18), 1795–1806.
- Ogutu, J., Schulz-Streeck, T. and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings* **6**(S2), S10.
- Palomaki, G. (2015). Is it time for BRCA1/2 mutation screening in the general adult population? Impact of population characteristics. *Genetics in Medicine* **17**(1), 24.
- Park, J.-H. and Dunson, D. (2010). Bayesian generalized product partition model. *Statistica Sinica* **20**, 1203–1226.
- Park, J.-H., Wacholder, S., Gail, M., Peters, U., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics* **42**(7), 570–575.
- Park, J.-H., Gail, M., Weinberg, C., Carroll, R., et al. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences of the United States of America* **108**(44), 18026–18031.
- Park, J.-H., Gail, M., Greene, M. and Chatterjee, N. (2012). Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: An evaluation of seven common cancers. *Journal of Clinical Oncology* **30**(17), 2157–2162.
- Pencina, M., D'Agostino, Sr, R., D'Agostino, Jr, R. and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**(2), 157–172.
- Pencina, M., D'Agostino, Sr, R. and E. Steyerberg. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* **30**(1), 11–21.
- Pepe, M., Janes, H. and Li, C. (2014). Net risk reclassification p values: Valid or misleading? *Journal of the National Cancer Institute* **106**(4).
- Pfeiffer, R. and Gail, M. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67**(3), 1057–1065.
- Pharoah, P., Antoniou, A., Bobrow, M., Zimmern, R., et al. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* **31**(1), 33–36.
- Pharoah, P., Antoniou, A., Easton, D. and Ponder, B. (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. *New England Journal of Medicine* **358**(26), 2796–2803.
- Piegorsch, W., Weinberg, C. and Taylor, J. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**(2), 153–162.
- Pirinen, M., Donnelly, P. and Spencer, C. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annals of Applied Statistics* **7**(1), 369–390.
- Prentice, R., Kalbfleisch, J., Peterson, A., Jr, Flournoy, N., et al. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**(4), 541–554.
- Psychiatric GWAS Consortium Bipolar Disorder Working Group. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* **43**(10), 977–983.

- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3), 559–575.
- Riley, B., Culver, J., Skrzynia, C., Senter, L., et al. (2012). Essential elements of genetic cancer risk assessment, counseling, and testing: Updated recommendations of the National Society of Genetic Counselors. *Journal of Genetic Counseling* **21**(2), 151–161.
- Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45**(10), 1150–1159.
- Rogowski, W., Grosse, S. and Khoury, M. (2009). Challenges of translating genetic tests into clinical and public health practice. *Nature Reviews Genetics* **10**(7), 489–495.
- Sampson, J., Wheeler, W., Yeager, M., Panagiotou, O., et al. (2015). Analysis of heritability and shared heritability based on genome-wide association studies for 13 cancer types. *Journal of the National Cancer Institute* **107**(12), djv279.
- Schizophrenia Psychiatric Genome-Wide Association Study Consortium. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**(10), 969–976.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**(7510), 421–427.
- Scott, I., Seegobin, S., Steer, S., Tan, R., et al. (2013). Predicting the risk of rheumatoid arthritis and its age of onset through modelling genetic risk variants with smoking. *PLoS Genetics* **9**(9), e1003808.
- Shi, J., Park, J.-H., Duan, J., Berndt, S., et al. (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genetics* **12**(12).
- Smith, R., Cokkinides, V. and Brawley, O. (2012). Cancer screening in the United States, 2012. *CA: A Cancer Journal For Clinicians* **62**(2), 129–142.
- So, H.-C., Kwan, J., Cherny, S. and Sham, P. (2011). Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *American Journal of Human Genetics* **88**(5), 548–565.
- Song, M., Kraft, P., Joshi, A., Barrdahl, M. and Chatterjee, N. (2015). Testing calibration of risk models at extremes of disease risk. *Biostatistics* **16**(1), 143–154.
- Speed, D. and Balding, D.J. (2014). MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Research* **24**(9), 1550–1557.
- Speed, D., Hemani, G., Johnson, M.R. and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**(6), 1011–1021.
- Speliotis, E., Willer, C., Berndt, S., Monda, K., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**(11), 937–948.
- Stahl, E., Wegmann, D., Trynka, G., Gutierrez-Achury, J., et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* **44**(5), 483–489.
- Steyerberg, E. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer-Verlag, New York.
- Szymczak, S., Biernacka, J., Cordell, H., González-Recio, O., et al. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology* **33**(S1), S51–S57.
- Talmud, P., Cooper, J., Morris, R., Dudbridge, F., et al. (2015). Sixty-five common genetic variants and prediction of type 2 diabetes. *Diabetes* **64**(5), 1830–1840.
- Taylor, J., Ankerst, D. and Andridge, R. (2008). Validation of biomarker-based risk prediction models. *Clinical Cancer Research* **14**(19), 5977–5983.
- Temple, R. (2010). Enrichment of clinical study populations. *Clinical Pharmacology & Therapeutics* **88**(6), 774–778.

- Teslovich, T., Musunuru, K., Smith, A., Edmondson, A., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**(7307), 707–713.
- Thomas, D., James, P. and Ballinger, M. (2015). Clinical implications of genomics for cancer risk genetics. *Lancet Oncology* **16**(6), e303–e308.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**(1), 267–288.
- Tyrer, J., Duffy, S. and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine* **23**(7), 1111–1130.
- Umbach, D. and Weinberg, C. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine* **16**(15), 1731–1743.
- US Preventive Services Task Force. Breast Cancer: Screening Recommendation Summary (2016). URL <https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening1/>.
- Usher-Smith, J., Emery, J., Hamilton, W., Griffin, S. and Walter, F. (2015). Risk prediction tools for cancer in primary care. *British Journal of Cancer* **113**(12), 1645–1650.
- Van Hoek, M., Dehghan, A., Witteman, J., van Duijn, C. et al. (2008). Predicting type 2 diabetes based on polymorphisms from genome-wide association studies. *Diabetes* **57**(11), 3122–3128.
- Vilhjálmsson, B., Yang, J., Finucane, H., Gusev, A., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American Journal of Human Genetics* **97**(4), 576–592.
- Wacholder, S., McLaughlin, J., Silverman, D. and Mandel, J. (1992a). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology* **135**(9), 1019–1028.
- Wacholder, S., Silverman, D., McLaughlin, J., and Mandel, J. (1992b). Selection of controls in case-control studies. II. Types of controls. *American Journal of Epidemiology* **135**(9), 1029–1041.
- Wacholder, S., Silverman, D., McLaughlin, J., and Mandel, J. (1992c). Selection of controls in case-control studies. III. Design options. *American Journal of Epidemiology* **135**(9), 1042–1050.
- Wacholder, S., Chatterjee, N. and Hartge, P. (2002). Joint effect of genes and environment distorted by selection biases: Implications for hospital-based case-control studies. *Cancer Epidemiology, Biomarkers & Prevention* **11**(9), 885–889.
- Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., et al. (2010). Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine* **362**(11), 986–993.
- Wooster, R., Neuhausen, S., Mangion, J., Quirk, Y., et al. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**(5181), 2088–2090.
- Wray, N., Yang, J., Goddard, M. and Visscher, P. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genetics* **6**(2), e1000864.
- Wray, N., Lee, S., Mehta, D., Vinkhuyzen, A., et al. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry* **55**(10), 1068–1087.
- Wu, J., Pfeiffer, R. and Gail, M. (2013). Strategies for developing prediction models from genome-wide association studies. *Genetic Epidemiology* **37**(8), 768–777.
- Yang, J., Benyamin, B., McEvoy, B., Gordon, S., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**(7), 565–569.
- Yang, J., Manolio, T., Pasquale, L., Boerwinkle, E., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**(6), 519–525.
- Yang, J., Ferreira, T., Morris, A., Medland, S., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**(4), 369–375.

- Zeng, P. and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications* **8**(1), 456.
- Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9**(2), e1003264.
- Zuk, O., Hechter, E., Sunyaev, S. and Lander, E. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* **109**(4), 1193–1198.

30

Bayesian Methods for Gene Expression Analysis

Alex Lewin,¹ Leonardo Bottolo,^{2,3,4} and Sylvia Richardson^{3,4}

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

²Department of Medical Genetics, University of Cambridge, Cambridge, UK

³The Alan Turing Institute, London, UK

⁴MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Abstract

We review the use of Bayesian methods for analysing gene expression data, from microarrays and bulk RNA sequencing, focusing on methods which select groups of genes on the basis of their expression in RNA samples derived under different experimental conditions. We first describe Bayesian methods for estimating gene expression levels from the intensity measurements obtained from analysis of microarray images and from RNA-sequencing reads. We next discuss the issues involved in assessing differential gene expression between experimental conditions, including models for classifying the genes as differentially expressed or not. Finally, we review the extensive Bayesian literature on univariate and multivariate gene-selection models, used for finding genetic signatures of a given disease, and discuss their application for detecting quantitative trait loci (QTLs) and expression QTLs.

30.1 Introduction

High-throughput technologies such as DNA microarrays and next-generation sequencing have emerged over the last twenty years as major sources of information for functional genomics. These technologies enable researchers to capture one of the fundamental processes in molecular biology, the transcription of genes into messenger RNA (mRNA) that will be subsequently translated to form proteins. This process is called gene expression. By quantifying the amount of transcription, it is possible to identify the genes that are expressed in different types of cells, different tissues, and to understand the cellular processes in which they intervene. However, transforming the huge quantity of data currently produced in high-throughput experiments into useful knowledge is not trivial, and research into ways of interpreting this rich body of data has become an active area, involving statistics, machine learning and computer science.

In this chapter we focus on data from microarrays and bulk RNA-sequencing. These high-throughput experiments represent data comparable to that obtained by performing tens of thousands of ‘experiments’ of a similar type in parallel. The ‘experiments’ on a given array or put through the same sequencer will share certain characteristics related to the manufacturing process of the particular array or sequencer and the extraction and handling of the biological

sample. The primary interest from a biological perspective is in comparing expression levels between samples from different biological conditions of interest (e.g. cancerous against non-cancerous cells), and the challenge is identifying differences that are related to the biology of the samples rather than to technical experimental variation.

Many of the characteristic features of experiments involving microarrays and high-throughput sequencing render them particularly well suited to a flexible modelling strategy within the Bayesian framework. The aim of this chapter is to focus on the contribution that Bayesian methods offer and to highlight this by discussing in detail the steps taken for modelling the variability in gene expression data at several levels. In particular, we focus on the framework of Bayesian hierarchical modelling as a generic model building strategy in which unobserved quantities are organized into a small number of discrete levels with logically distinct and scientifically interpretable functions and probabilistic relationships between them that capture inherent features of the data. The hierarchy of levels makes it particularly suitable for modelling gene expression data, which arises from a number of processes and is affected by many sources of variability.

One of the most important aspects of Bayesian hierarchical modelling is the sharing of information across parallel units. For example, gene expression experiments used by biologists to study fundamental processes of activation/suppression frequently involve genetically modified animals or specific cell lines, and such experiments are typically carried out only with a small number of biological samples. It is clear that this small amount of replication makes standard estimates of gene variability unstable. By assuming exchangeability across the genes, inference is strengthened by borrowing information from comparable units.

Another strength of the Bayesian framework is the propagation of uncertainty through the model. Due to the many sources of systematic variation between arrays or sequence libraries, gene expression data is often processed through a series of steps, each time estimating and subtracting effects in order to make samples comparable. The final result of this process can be over-confident inference, as the uncertainty associated with each step is ignored. In a Bayesian model it is straightforward to include each of these effects simultaneously, thus retaining the correct level of uncertainty on the final estimates. Further, when including in the model structured priors that are associated with classification (e.g. mixture priors), estimates of uncertainty of the classification are obtained along with the fit of the model.

We start in Sections 30.2 and 30.3 with an overview of Bayesian methods for estimating gene or transcript expression levels from arrays and sequencing experiments. These sections include work on normalization across arrays and sequence libraries, and choices for the likelihood function for array and sequencing data. Section 30.4 discusses the issues involved in assessing differential gene expression between two conditions. We include a brief explanation of mixture models, and some discussion of different decision rules used to choose the lists of genes considered to be differentially expressed. Finally, Section 30.5 reviews the current work on multivariate methods for finding subsets of genes that can predict and classify univariate and multivariate phenotypes. This work includes models for detecting quantitative trait loci (QTLs), genetic variants associated with univariate phenotypes, and eQTLs (expression QTLs), which are variants associated with multivariate expression phenotypes.

Throughout the chapter, we discuss Bayesian variable selection methods as well as Bayesian shrinkage methods. In both cases, the emphasis is on parsimony of the multivariate model in order to enhance interpretation. Inference in Bayesian models is made in either an empirical or fully Bayesian framework. In the case of fully Bayesian models, Markov chain Monte Carlo (MCMC) is usually used to estimate the posterior distribution of the model, though approximate methods such as variational Bayes are starting to make an appearance. We do not go into details of these procedures, except in the case of non-standard algorithms.

30.2 Modelling Microarray Data

30.2.1 Modelling Intensities

Microarrays use hybridization reactions to select out specific RNA sequences from a biological sample. Each known transcript sequence is represented at a particular position on the array. The concentration of a particular sequence in the original sample is represented by the intensity of fluorescently labelled molecules. Intensity measurements from image analysis are used to construct a summary measure of the expression of each gene or transcript of interest.

Many models have used a linear framework for modelling log-transformed expression levels x_{gcr} for gene g , experimental condition c and replicate array r :

$$\log(x_{gcr}) = \mu_{gc} + \gamma_{cr} + \epsilon_{gcr}, \quad (30.1)$$

where μ_{gc} represents the mean level of expression of gene g for condition c , γ_{cr} is a normalization term for the array containing the replicate r sample of condition c , and ϵ_{gcr} is the residual. Expression measures have been observed to have increasing variability with increasing value (Schadt *et al.*, 2000), so they are often modelled on the log scale. Sometimes a shifted log transform is used, as for example in Gottardo *et al.* (2006b).

The main parameters of interest are the μ_{gc} , which represent the expression across different experimental conditions. Priors on differential expression parameters for both microarray and sequencing data are discussed in Section 30.4. The advantage of the Bayesian framework is that these parameters are estimated simultaneously with the sources of noise and biases involved in the data. The priors for bias and variation from microarrays are discussed in the following sections.

30.2.2 Gene Variability

There are many sources of variation in microarray data. It is possible to put replicate RNA samples from the same individual on different microarrays, but this is usually considered unnecessary as it has been observed that these so-called technical replicates show very high correlation. Instead different arrays are generally hybridized with samples taken from different individuals. Thus the variability incorporated in the error term ϵ_{gcr} in equation (30.1) represents the biological variability. It is generally accepted that different genes show different levels of biological variability, thus parameters in the distributions for the errors will typically depend on the gene index.

Several Bayesian models in the literature assume normal errors (Lönnstedt & Speed, 2003; Baldi & Long, 2001; Bhattacharjee *et al.*, 2004; Lewin *et al.*, 2006). Gottardo *et al.* (2006a) use a t distribution (bivariate for cDNA data) to accommodate more outlying data points. Newton *et al.* (2001, 2004) give the data a gamma likelihood rather than the lognormal implied by equation (30.1). Simple model-checking techniques suggest the gamma and lognormal families are equally suitable for gene expression data.

Since the numbers of individuals for each experimental condition are often small, independent estimates of gene variance parameters would be unstable. Therefore gene variances σ_{gc}^2 are usually shrunk, by assuming exchangeability across genes (and sometimes conditions). Both empirical Bayes (Lönnstedt & Speed, 2003) and fully Bayesian methods (Lewin *et al.*, 2006; Gottardo *et al.*, 2006a) which rely on MCMC algorithms for inference have been used. Rather than allowing a separate variance for each gene, Bhattacharjee *et al.* (2004) allow gene variances to take one of three values, estimated as part of the model, as an alternative way of sharing information across genes. By contrast, Baldi & Long (2001) allow gene variances to depend on

expression level, by making the variances exchangeable among genes with similar expression levels (defined by a window on the expression level) and estimating these using empirical Bayes methods.

30.2.3 Normalization

Microarray data show systematic differences between expression levels found on different arrays (e.g. Schadt *et al.*, 2000). This may be taken into account during preprocessing, but generally empirical differences in gene expression are still found between arrays. Often the systematic effect is such that there is a nonlinear relationship between the expression levels on different arrays.

Much work has been done in the classical statistical literature on different methods of accounting for these systematic nonlinear differences (normalizing). These usually involve a transformation of the data before it is analysed with another method. Most work on Bayesian models for gene expression has also assumed that this process has been done beforehand.

Some exceptions are those proposed by Parmigiani *et al.* (2002) and Gottardo *et al.* (2006b). Both of these include a constant term in a linear model, estimated in a fully Bayesian manner. Bhattacharjee *et al.* (2004) and Lewin *et al.* (2006) extend this idea to model a normalization term γ_{gcr} as a nonlinear function of expression level. Bhattacharjee *et al.* (2004) use a piecewise linear function of gene expression level. Due to marginalization over the joint posterior, posterior estimates of the normalization term will be reasonably smooth functions of expression level. Lewin *et al.* (2006) use a quadratic spline for normalization as a function of expression level. They show that transforming the data first rather than modelling the normalization simultaneously with the other unknown quantities can introduce bias, as the gene expression levels have to be estimated and thus have variability, as in measurement error problems (Carroll *et al.*, 1998).

30.3 Modelling RNA-Sequencing Reads

30.3.1 Alignments for RNA-Sequencing Data

In contrast to gene expression arrays, RNA-sequencing (RNA-seq) data sets are not restricted to studying predefined sets of genes. Instead, following reverse transcription and amplification of the resulting cDNA fragments, they are randomly fragmented and sequenced using short read sequencing technology. The methods for RNA-seq data described in this chapter require a reference sequence to align reads to. This may be a genome, transcriptome or set of genomes or transcriptomes. A reference genome is a set of chromosome sequences of DNA. A reference transcriptome is a set of predefined transcripts: this is typically the set of known isoforms of all genes. The simplest RNA-seq models can be used only with genome alignments, and thus infer expression at the gene level (aggregated over all possible isoforms). Many models below are designed to distinguish different isoforms and thus require transcriptome alignments. And there are a small number of models that aim to quantify different alleles of a particular gene or isoform (known as allele specific expression). These methods typically need two reference genomes or two reference transcriptomes for alignment.

A genome or transcriptome may be constructed *de novo* from a new set of generated DNA or RNA sequence reads, and can then be used in the models below as an alternative reference. For the transcriptome, there are algorithms which produce an alternative transcriptome by augmenting an established reference with newly assembled isoforms and transcripts

(Trapnell *et al.*, 2012). Once this has been done the new transcriptome can be used for alignment in exactly the same way as the original reference would have been used. This is advocated in Maretty *et al.* (2014) and Aguiar *et al.* (2017), but several other models described below could be used in the same way for isoform discovery.

All of the models below therefore assume that each read has been aligned to the reference using an existing alignment algorithm (e.g. Kim *et al.*, 2013). Most alignment algorithms provide a quality score for the mapping which indicates how confidently a read can be mapped. This information is used to calculate probabilities of given reads having arisen from several different possible transcripts, which may be from different parts of the genome or represent multiple transcripts of the same gene.

If we are only interested in gene-level inference then restricting data to reads mapped to a unique position in the genome with high quality is adequate, and simple methods can be used to model the data (see Section 30.3.4). The methods in Sections 30.3.2 and 30.3.3 deal with the case where a read is mapped back to multiple locations and/or its allocation to a given transcript is unclear. This is particularly pertinent when attempting to allocate a read to a gene with multiple transcripts: the short sequence length makes mapping back to a unique transcript extremely challenging and has motivated the methodological developments described here.

In this section we survey a range of likelihood functions that have been proposed in Bayesian models for data consisting of individual reads and for count data. We show how the different choices of likelihood are related to and consistent with each other.

30.3.2 Likelihood for Read-Level Data

A number of models have been developed to model the alignments of individual reads $r_i, i = 1, \dots, n$, to a transcriptome (Katz *et al.*, 2010; Glaus *et al.*, 2012; Maretty *et al.*, 2014; Nariai *et al.*, 2016; Aguiar *et al.*, 2017). Note that r_i represents the full read sequence of L bases, as these models distinguish individual reads rather than count data. These models incorporate the probability of a given read sequence coming from a particular transcript, hence can cope with reads aligned to multiple transcripts. These probabilities are calculated from the small set of alignments for each read, each with information about base-calling quality, and matches/mismatches between the observed read and the reference transcript at that alignment.

To understand and unify these models, we introduce parameters θ_t representing relative expression of transcript t , and latent variables $T_i = t$ meaning that read i originated from transcript t . The set of transcripts may be all possible transcripts in the transcriptome (e.g. Glaus *et al.*, 2012) or only isoforms of a particular gene (e.g. Katz *et al.*, 2010).

The sequencing experiment is a two-stage process: firstly, molecules are fragmented; and secondly, the fragments are sequenced. The probability of a particular read sequence r_i originating from a fragment of transcript t is given by $\mathbb{P}(T_i = t|\theta) = \theta_t$, with $\sum_{t'} \theta_{t'} = 1$. Conditional on T_i , the data-generating process for an individual read i is modelled as

$$\mathbb{P}(r_i|T_i = t) = \sum_{q=1}^{l_t-L+1} \mathbb{P}(r_i|T_i = t, \text{start at } q)\mathbb{P}(\text{start at } q|T_i = t),$$

where l_t is the length of the transcript, L is the read length and q indexes position along the transcript. Note that this model implicitly assumes only one read is generated from a given sequence of transcript in the sample. In reality, long transcripts may produce many fragments and hence many reads. The read-level models presented here can therefore be understood to model read generation from fragments, hence the expression parameter θ_t here represents fragment concentration of a given transcript t in the sample rather than molecular concentration.

The probability of a fragment starting at position q is not necessarily equal for all positions, as sequences containing more GC nucleotides are less easily sequenced. This non-uniform probability may be modelled using bias factors depending on GC content (Li *et al.*, 2010). These factors $\mathbb{P}(\text{start at } q|T_i = t)$ are calculated from separate information and fixed in these models. If this sequencing bias is not included in the model, this probability reduces to uniform probability over possible start positions: $\mathbb{P}(\text{start at } q|T_i = t) = 1/(l_t - L + 1)$, where L is the read length. The probability of a fragment being read as the observed sequence depends on the probabilities of base-calling error. First, define a binary indicator to record whether each read base matches the corresponding base in the transcriptome: $m_{kq}^{i,t} = 1$ if base k of read i matches position $q + k - 1$ of transcript t . Then the probability of the read given the start position is

$$\mathbb{P}(r_i|T_i = t, \text{start at } q) = \prod_{k=1}^L (1 - p_{ik})^{m_{kq}^{i,t}} (p_{ik})^{1 - m_{kq}^{i,t}},$$

where p_{ik} is the probability of a base-calling error given by the alignment algorithm for base k of read i . In practice, since base-calling errors are very small, $\mathbb{P}(r_i|T_i = t, \text{start at } q)$ will be negligible for any transcripts the read does not align to, and for any q except the actual read alignment start position. Thus we see that the probability of a given read sequence having arisen from a given transcript fragment can be written as

$$\mathbb{P}(r_i|T_i = t) = \mathbb{P}(r_i|T_i = t, \text{start at } q_{it}^*) \mathbb{P}(\text{start at } q_{it}^*|T_i = t) I[t \in A_i],$$

where A_i is the set of transcript alignments for read i , and q_{it}^* is the actual start position of the alignment of read i to transcript t . These probabilities are calculated from alignment data, so the $\mathbb{P}(r_i|T_i = t)$ are treated as fixed and known in these models.

The alignments, bias parameters and base-calling errors are fixed in the analysis, so we end up with the likelihood for a Bayesian model,

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^n \mathbb{P}(r_i|\theta) \\ &= \prod_{i=1}^n \sum_{t \in A_i} \mathbb{P}(r_i|T_i = t) \mathbb{P}(T_i = t|\theta_t), \end{aligned}$$

and only the priors for the expression parameters $\theta_t, t = 1, \dots, m$ need to be specified in order to complete a Bayesian model for expression analysis. This model can be readily extended to multiple samples and experimental conditions by defining different expression parameters for different conditions. Different choices of prior for the expression parameters are discussed in Section 30.4.2.

As mentioned earlier, these read-level models estimate fragment concentrations. Length corrections are made to the model parameters to produce estimates of relative concentrations of molecules. The most common adjustment is to use reads per kilobase million (RPKM) length adjustment, with molecular expression values being given by $\theta_t^* = 10^9 \times \theta_t / l_t$, where l_t is the length of transcript t (Mortazavi *et al.*, 2008). In reality there is a distribution of numbers of fragments per molecule; however, this is not usually modelled and the above scaling is used as an approximation to give reasonable expression estimates per molecule.

30.3.3 Likelihood for Transcript-Level Read Counts

If sequencing bias and base calling errors are not included in the read-level model, the term in the likelihood reduces to $\mathbb{P}(r_i|\theta) = \sum_{t \in A_i} \theta_t / (l_t - L + 1)$, where r_i is a specific sequence of

L bases. Since the probability does not depend on the specific sequence r_i , only the alignment set A_i , we can sum over all start positions to express the probability any single read/fragment originating from transcript set A as $\mathbb{P}(\text{any single read from set } A | \theta) = \sum_{t \in A} \theta_t$. Note that this still effectively assumes a single fragment/read per molecule of transcript, as in the read-level model. From now on we will switch to parameterizing the models in terms of relative molecular expression values $\theta_t^* \propto \theta_t / l_t$ as in the RPKM normalization. So we end up with $\mathbb{P}(\text{any single read from set } A | \theta^*) \propto \sum_{t \in A} \theta_t^* l_t$.

Now we can aggregate reads aligning to the same set of transcripts. Here we form a partition of the reads via their alignment sets. Define the set of unique alignment sets as $\{A'_1, \dots, A'_M\}$. These sets of transcripts do not need to be disjoint. Hence we can write the above model as

$$\mathbb{P}(\text{any read from set } A'_j | \theta^*) \propto \sum_{t \in A'_j} \theta_t^* l_t = \sum_t M_{jt} \theta_t^* l_t,$$

where we have defined a mapping function (Turro *et al.*, 2011)

$$M_{jt} = \begin{cases} 1, & \text{if transcript } t \text{ is in set } A'_j \\ 0, & \text{else.} \end{cases}$$

Assuming independence between reads, this becomes a multinomial model for read counts: $\{y_1, \dots, y_M\} \sim \text{Mult}(n, \vec{\gamma})$ where y_j is the number of reads aligning to transcript set A'_j and $\gamma_j = \frac{\sum_t M_{jt} \theta_t^* l_t}{\sum_t M_{jt} \theta_t^* l_t}$. If the total number of reads is large and γ_j is small, the multinomial can be approximated by a Poisson distribution for each transcript set, $y_j \sim \text{Pois}(n\gamma_j)$. A further approximation results in the following model (Turro *et al.*, 2011):

$$y_j \sim \text{Pois}(nl_j\gamma_j^*),$$

where l_j is the length sequence shared by the transcripts in set j , $\gamma_j^* = \sum_t M_{jt} \theta_t^*$ and θ_t^* now represents the relative molecular concentration of transcript t . The shared sequences represented by l_j are in general sub-sequences of transcript, often corresponding to one or more exons shared by multiple transcripts. Wang *et al.* (2015) have developed a similar model (using a generalized Poisson distribution) inferring over shared sections of transcripts, which they call mathematical exons. Zhao *et al.* (2018) also use this structure, extending it to a negative binomial model for read counts, in order to allow for over-dispersion between biological replicates.

These models can be fitted straightforwardly in the Bayesian framework by including latent variables X_{jt} representing unobserved read counts from subsection j of transcript t , with $\sum_t X_{jt} = y_j$ (Turro *et al.*, 2011; Wang *et al.*, 2015; Zhao *et al.*, 2018). The resulting inference is on the expression parameters θ_t^* which are the relative expressions of all possible transcripts and alternative isoforms of all genes.

Turro *et al.* (2011) fitted this model to a set of RNA-seq reads from F1 crosses of two strains of mice (CAST and C57). Two sets of F1 hybrids were compared: F1-i (initial cross with CAST mother and C57 father) and F1-r (reciprocal cross with CAST father and C57 mother). Reference transcriptomes were created for both CAST and C57 strains, and combined into a single reference for read alignment. For each sample, log fold changes, $\log(\theta_{t,\text{CAST}}^* / \theta_{t,\text{C57}}^*)$, between haplo-isoforms were estimated. Figure 30.1 shows these log fold changes for the initial crosses versus the reciprocal crosses. There are three clear groups of transcripts in the plot: *cis*-regulated transcripts, imprinted transcripts and transcripts from the X chromosome.

As with the read-level models above, these count models can be extended to model differential expression and allele-specific expression by appropriately parameterizing the expression parameters. Section 30.4.2 gives examples of priors for differential expression suitable for read

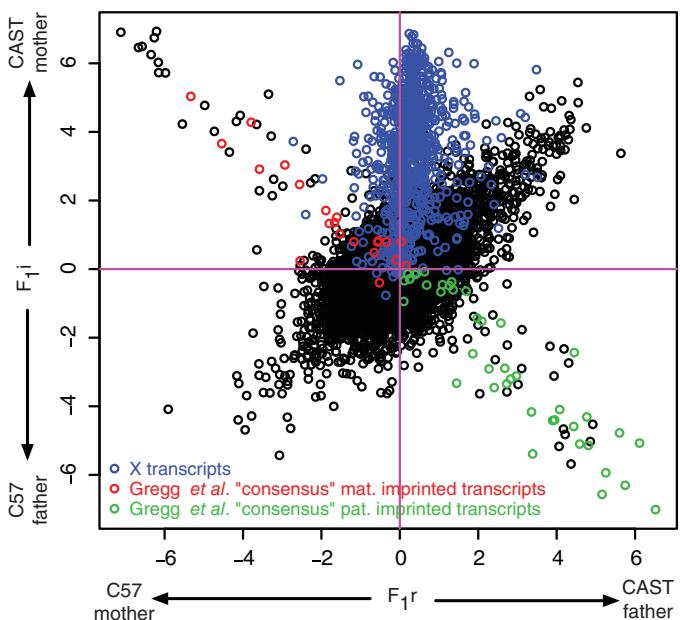


Figure 30.1 Scatter plot of log fold changes between haplo-isoforms in the reciprocal ($F1-r$) and the initial ($F1-i$) crosses, highlighting X transcripts in blue, isoforms termed ‘consensus’ maternally imprinted in red and ‘consensus’ paternally imprinted in green. Consensus imprinted isoforms were those previously identified as imprinted.

count models. Count models must include normalization since different samples may result in different numbers of total reads. The above derivation of the count model naturally includes the total number of reads in each sample, which can be seen as a simple form of normalization. More robust normalization can be achieved by replacing this by a smoothed estimate of library size such as the trimmed mean of M -values method (Robinson & Oshlack, 2010).

30.3.4 Likelihood for Gene-Level Read Counts

If the transcript sets $\{A'_1, \dots, A'_M\}$ defined above are disjoint, then the model for read counts simplifies to

$$y_g \sim Pois(nl_g \theta_g^*),$$

where the expression parameter θ_g^* may still represent a sum over different transcripts. The predominant use for this simpler model is to infer expression per gene, where each read is aligned to a single gene, and the expression θ_g^* is called gene expression, and is the aggregate over all possible isoforms of the gene.

Many Bayesian models have been proposed using this Poisson distribution as a starting point. However, read counts from biological replicates are generally observed to show over-dispersion (Marioni *et al.*, 2008). A simple approach to the modelling is to add parameters to the Poisson model to allow for the over-dispersion. This can be done by adding an extra level to the Poisson model: $\theta_g^* | \mu_g, \phi_g \sim Gam(\mu_g, \phi_g)$, which is equivalent to the negative binomial model

$$y_g | \mu_g, \phi_g \sim NegBin\left(\mu_g, \frac{nl_g}{(nl_g + \phi_g)}\right),$$

where the expression parameters of interest are now the μ_g . The parameter ϕ_g allows for counts to be more variable than expected by the Poisson distribution. Negative binomial models using various different parameterizations have been used by many people (e.g. Hardcastle & Kelly, 2010; Wu *et al.*, 2013; Chung *et al.*, 2013; Bi & Davuluri, 2013; Liu *et al.*, 2015; Vavoulis *et al.*, 2015). Other variations include using a Poisson with a lognormal prior on the θ_g^* (Gu *et al.*, 2014; Katzfuss *et al.*, 2014). Lee *et al.* (2015) model read counts at each base of a given gene, using a Poisson model with a hierarchical prior to combine information across positions at the gene level. Van De Wiel *et al.* (2013) allow explicitly for transcripts not to be expressed, by using a zero-inflated negative binomial likelihood. Ritchie *et al.* (2015) depart from the Poisson model, applying their linear regression framework (limma) to log-transformed read counts.

Many Bayesian models add a hierarchical prior on the over-dispersion parameters, providing a shrinkage or borrowing of strength across genes/isoforms. This is beneficial in gene expression experiments which tend to be performed on small numbers of replicates. For example, Wu *et al.* (2013) put a lognormal prior on the gamma scale parameter ϕ_g . Leng *et al.* (2013) categorize genes into groups based on complexity of isoform structure (e.g. the number of isoforms a gene possesses). Separate hierarchical priors are used for the different groups, to give genes with similar isoform complexity similar variances. Gu *et al.* (2014) infers expression at the exon level, and combines estimates through a hierarchical prior into an average gene expression level.

30.4 Priors for Differential Expression Analysis

One of the most widely studied problems that arises when studying gene expression data is that of differential gene expression between two experimental conditions, for example between knock-out and wildtype animals, or between cases and controls. Each microarray or RNA-seq library corresponds to one sample, which is generated under one of the experimental conditions. Hence we can think of each condition $c = 1, 2$ having a number of replicate samples $r = 1, \dots, n_c$. For read count models the data is the counts aligned to gene g in replicate r of condition c (aggregated over all isoforms). For microarrays the data is usually a log-transformed estimate of expression for gene g in replicate r of condition c . The models for array intensities discussed in Section 30.2 and for sequencing data in Section 30.3.4 are then extended to include expression parameters for each condition.

30.4.1 Differential Expression from Microarray Data

It is useful to write the expression levels in two experimental conditions as

$$\begin{aligned}\mu_{g1} &= \alpha_g - \delta_g/2, \\ \mu_{g2} &= \alpha_g + \delta_g/2,\end{aligned}\tag{30.2}$$

where α_g represents the overall expression level for gene g and δ_g represents the log differential expression. For two-colour arrays the data can be given as log fold changes between the conditions (the data is paired) and in that case there is no α_g parameter. When the data is given separately for the two conditions, α_g must be modelled. It is usually treated as a fixed effect, so no information is shared between genes for this parameter.

The fold change parameter δ_g may be given an unstructured prior (e.g. Baldi & Long, 2001; Bhattacharjee *et al.*, 2004; Lewin *et al.*, 2006). Baldi & Long (2001) and Smyth (2004) propose so-called regularized or moderated t statistics. These consist of the Bayesian posterior mean estimate of the log fold change parameter, divided by a shrunken estimate of standard deviation.

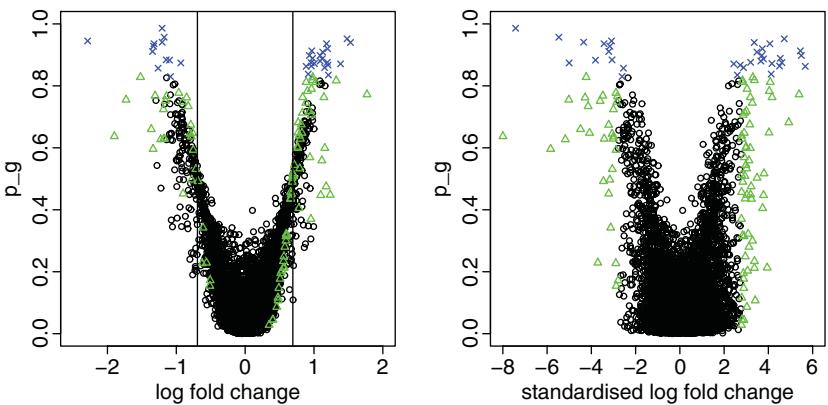


Figure 30.2 (Left) Posterior probabilities p_g against log fold difference. (Right) Posterior probabilities p_g against regularized t statistics. Genes with $p_g > 0.83$ (corresponding to an estimated false discovery rate of 10%) are shown as crosses. Triangles indicate remaining genes with regularized t statistics above 2.78.

This shrunken estimate is the square root of the posterior mean of the variance parameter, shrinkage being provided by the exchangeable prior on the variances estimated in an empirical Bayes framework.

With a non-informative prior on the δ_g , Lewin *et al.* (2006) proposed a decision rule based on a threshold δ_{cut} set according to a biologically interesting level of differential expression. Differential expression is defined as δ_g being greater than δ_{cut} , corresponding to an interval null hypothesis with the interval fixed *a priori*. The decision rule is that genes are declared to be differentially expressed if the posterior probability $p_g = \mathbb{P}(|\delta_g| > \delta_{\text{cut}} | \mathbf{x})$ is greater than some threshold probability. This rule combines statistical significance through the choice of threshold probability, and biological significance through the choice of threshold δ_{cut} . The decision rule is illustrated in Figure 30.2.

Using an unstructured prior on the differential expression parameter requires the analyst to set a pre-specified threshold of biological interest. As an alternative, many people choose to use mixture models to classify genes as differentially expressed or not. Usually this means putting a mixture prior on some measure of the difference between expression levels in the two experimental conditions. The mixture models can be classified into two groups: those which put a mixture prior on the model parameter δ_g , and those which model the data directly as a mixture. These are not intrinsically different, as the parameters δ_g could be integrated out to give a mixture model on the data, but it is convenient to describe the models separately.

In general, a finite mixture distribution for a quantity Δ_g is a weighted sum of probability distributions,

$$\Delta_g \sim \sum_{k=0}^{K-1} w_k f_k(\phi_k), \quad (30.3)$$

where the weights sum to one ($\sum_{k=0}^{K-1} w_k = 1$). Each mixture component has a certain distribution f_k , with parameters ϕ_k . The weight w_k represents the probability of Δ_g being assigned to mixture component k . In the context of differential expression, most mixture models used consist of two components ($K = 2$), one of which (f_0) can be thought of as representing the ‘null hypothesis’ of there being no differential expression. The second component corresponds to the alternative hypothesis that there is differential expression. Of course it is not necessary to see the model in terms of hypothesis testing; in the Bayesian framework it is a straightforward

procedure to classify each gene into one or other of the mixture components. This is usually done using the posterior probability of component membership (see later in this section for further discussion).

One of the earliest mixture models used in gene expression analysis was that of Efron *et al.* (2001). In this model Δ_g in equation (30.3) is a regularized t statistic t_g , one for each gene:

$$t_g \sim w_0 f_0 + w_1 f_1 \quad (30.4)$$

The densities of the mixture components are estimated nonparametrically using standard kernel density procedures. Regularized t statistics are calculated using expression data from the same experimental condition, to provide an estimate of the null component f_0 . An estimate of w_0 (which represents the proportion of genes in the null, or not differentially expressed) is obtained using empirical Bayes methods. The whole mixture distribution ($w_0 f_0 + w_1 f_1$) can be estimated using all the t_g . Thus the second component f_1 can be inferred.

A fully Bayesian version of this model has been discussed by Do *et al.* (2005). In this work, the framework of Dirichlet process mixtures (DPMs) is used to formulate a prior probability model for the distributions f_0 and f_1 . A DPM model, characterized by a base measure G^* , a scalar parameter α , and a mixing kernel to be specified, is one of the most popular nonparametric Bayesian models due to the simplicity of its representation and MCMC implementation (Escobar & West, 1995; Walker *et al.*, 1999). Do *et al.* (2005) choose base measures $G_0^* \sim N(0, \tau^2)$ and $G_1^* \sim \frac{1}{2}N(-b, \tau^2) + \frac{1}{2}N(b, \tau^2)$ for f_0 and f_1 respectively and Gaussian mixing kernels with common variance parameter σ^2 . The specification of G_1^* reflects the prior belief that differential expression in either direction is equally likely, in the absence of more specific prior information. Using the stick-breaking construction of Dirichlet priors (Sethuraman, 1994) leads to a useful representation of f_k , $k = 0, 1$, as an infinite mixture of normals:

$$f_k = \sum_{h=1}^{\infty} p_{hk} N(\mu_{hk}, \sigma^2),$$

with $\mu_{hk} \stackrel{i.i.d.}{\sim} G_k^*$, $k = 0, 1$, and the weights following the stick-breaking structure: $p_{hk} = U_h \prod_{j < h} (1 - U_j)$ with $U_h \stackrel{i.i.d.}{\sim} Beta(1, \alpha)$. In Do *et al.* (2005) all the model parameters are given hyperprior distributions, conjugate inverse gamma for τ^2 and σ^2 and conjugate normal for b , α is fixed at 1 and w_0 is given either a beta prior or a uniform prior away from 0. As in Efron *et al.* (2001), within-condition data differences are used to estimate f_0 , while between-condition differences are modelled as arising from the mixture defined in expression (30.4). The performance of this model will depend on the number of within-replicate differences that are used to calibrate f_0 and on the information introduced in the hyperprior specification. When there are only a few replicates, the mixture might be close to non-identifiability.

Broët *et al.* (2002) suggest another model using a mixture at the data level to classify genes. Here the data is first transformed with a linear model to produce normalized log fold changes d_g . The d_g are modelled using a fully Bayesian mixture of normals which includes estimation of the proportion of differentially expressed genes (the weights in the mixture). The number of components in the mixture K is not restricted to 2. There is still just one component representing the null, but several representing differentially expressed genes. This allows grouping of genes into different levels of differential expression. In fact K is not fixed in this model, but estimated, in a fully Bayesian way, using the split and merge algorithm for mixtures with an unknown number of components introduced in Richardson & Green (1997).

One practical challenge when assigning mixture distributions to model parameters (the priors), rather than on the data (likelihood), is that care must be taken to ensure identifiability of

the parameters of the mixture components. A common choice is to make the null component a point mass. This corresponds to testing the null hypothesis $\delta_g = 0$ against the two-sided alternative. Lönnstedt & Speed (2003), Lin *et al.* (2003) and Smyth (2004) use mixture priors on the parameter δ_g representing difference between conditions. Lönnstedt & Speed (2003) use a mixture of a point mass at zero and a conjugate normal prior on the δ_g . Smyth (2004) uses the same mixture model, but on data which has first been transformed using a linear model similar to that in equation (30.1) but using a robust estimation method, to obtain log fold changes. These two models are estimated using empirical Bayes methods. The proportion of true nulls is not estimated, thus these methods produce a ranking of the genes rather than an actual estimate of how many genes are differentially expressed.

Rather than putting the mixture directly on the δ_g , Newton *et al.* (2004) propose a mixture prior on the pair of parameters μ_{g1}, μ_{g2} . Their likelihood is gamma, but the μ_{gc} still represent mean expression in the two conditions. One component of the mixture has $\mu_{g1} = \mu_{g2}$ drawn from one distribution, the other has μ_{g1}, μ_{g2} drawn from two separate distributions. These distributions are estimated nonparametrically, using an EM algorithm. Gottardo *et al.* (2006a) has a similar mixture on the pair μ_{g1}, μ_{g2} , this time using normal priors, and a fully Bayesian estimation method, including estimating the proportion of differentially expressed genes. Reilly *et al.* (2003) has a model with a similar structure, but in addition incorporates prior information about certain genes being controls (and therefore not differentially expressed).

An early model for differential expression which does not employ a mixture model on the difference between the two conditions is that of Ibrahim *et al.* (2002). They model the data *in each condition* as coming from a mixture of a point mass and a lognormal distribution, the point mass representing the threshold for genes to be unexpressed. A measure for differential expression is formed from the ratio of expectation of expression in the two conditions.

In the fully Bayesian mixture models described above, decisions are usually made using the posterior probabilities of a gene being allocated to the different mixture components. The mixture example given in equation (30.3) can also be written as

$$\begin{aligned}\Delta_g | z_g &\sim w_{z_g} f_{z_g}(\phi_{z_g}). \\ \mathbb{P}(z_g = k) &= w_k.\end{aligned}\tag{30.5}$$

where the z_g are allocation parameters which label the mixture component to which gene g is assigned. The posterior probability of gene g being in component k is $\mathbb{P}(z_g = k | \mathbf{x})$.

Defining a loss function enables one to form the decision rule. First, denote the set of genes declared to be differentially expressed by S_1 and the set of genes called non-differentially expressed by S_0 . In the two-component mixture models, since there are two possible classifications for each gene, there are two possible penalties for misclassification: one for false positives, one for false negatives. If the ratio of these two penalties is λ , the same for all genes, the loss function l is proportional to

$$l \propto \sum_{g \in S_0} \mathbb{P}(z_g \neq 0 | \mathbf{x}) + \lambda \sum_{g \in S_1} \mathbb{P}(z_g = 0 | \mathbf{x}).\tag{30.6}$$

This is minimized by defining S_0 as the set of genes for which $\mathbb{P}(z_g = 0 | \mathbf{x}) \geq 1/(1 + \lambda)$, that is, genes are classified using a threshold on the posterior probabilities of classification in the mixture. Müller *et al.* (2007) discuss different possible loss functions and the decision rules they lead to.

The posterior probabilities can also be used to obtain an estimate of the false discovery rate, which is the ratio of false positives to total declared positives:

$$\widehat{FDR} = \frac{1}{|S_1|} \sum_{g \in S_1} \mathbb{P}(z_g = 0 | \mathbf{x}) \quad (30.7)$$

(see Newton *et al.*, 2004; Broët *et al.*, 2004; Müller *et al.*, 2007). This is sometimes referred to as the expected false discovery rate. An estimate of the false non-discovery rate (ratio of false negatives to total negatives) can be defined similarly. The false discovery rate is widely used in classical statistical analysis of gene expression data. It is useful to be able to give this estimate when comparing with different analysis methods and it has generally found in simulation studies that equation (30.7) gives accurate estimates of the true false discovery rate.

For mixtures of more than two components, one may consider different rules. The most obvious would be to assign genes to the component with highest probability; that is, gene g is assigned to component $k = \max_{k'} \mathbb{P}(z_g = k' | \mathbf{x})$. However, when there are more than two components, this can lead to genes being declared differentially expressed (in a particular component) when their posterior probability of being classified into that component is low. For example, with four components, a gene which has almost equal probability of being classified in all components can be declared differentially expressed (into the best component for that gene) with posterior probability of 0.26 of being in the that component. An alternative, more conservative, suggestion would be to classify into one of the components representing differential expression only those genes for which the corresponding posterior probability of belonging to that component is above a set threshold (e.g. 50% or higher), and otherwise the genes are classified into the null. Again, evaluating the associated false discovery rate of such rules will guide the choice of appropriate thresholds. Such a rule was used in a related context, that of modelling DNA copy number changes (gains or losses) in comparative genomic hybridization experiments by a spatially structured mixture model with three components (gain, loss, normal), in Broët & Richardson (2006). For typical noise to signal ratio, the authors found that classifying into the gain or loss components DNA sequences with a posterior probability above 0.8 gave good operational characteristics in this context, whereas the Bayes rule exhibited poorer performance.

When mixture models are estimated using Empirical Bayes methods, without an estimate of the number of genes in the null, the posterior probability of being allocated to the null can only be estimated up to a constant. In this situation the posterior odds ratio can be used to rank genes:

$$Odds_g = \frac{\mathbb{P}(z_g = 0 | \mathbf{x})}{\mathbb{P}(z_g \neq 0 | \mathbf{x})} \quad (30.8)$$

(Lönnstedt & Speed, 2003; Smyth, 2004).

30.4.1.1 Multi-class Data

A number of models have been proposed which extend the methods used for differential expression in two conditions to compare expression in several conditions or classes. These might be used, for example, to compare the actions of several drugs and a control sample simultaneously, or to compare different tumour samples. As with the mixture models described previously, it can be useful to describe the classification of genes in terms of null and alternative hypotheses. There are a number of different choices of alternative hypothesis for multi-class data. Here we discuss models which use hypotheses of the type ‘the gene is differentially expressed (or not) in at least one condition’ without distinguishing which condition it is. An alternative approach

is taken by Ishwaran & Rao (2005a) who, similarly to their work on differential expression, formulate the multi-class analysis as a multivariate regression problem, use variable selection and shrinkage to output lists of significant genes between any two conditions and then use these lists to highlight patterns of interest between the conditions.

A common formulation is the classical ANOVA model, which tests the null hypothesis ‘the gene has the same expression in all conditions’ against the alternative ‘the gene has differential expression in at least one condition’. This is used in a Bayesian framework by Broët *et al.* (2004), who start with an F statistic for each gene. These F statistics are transformed to the normal scale and modelled with a two-component mixture to classify genes as differentially expressed or not. The null component is a standard normal, while the alternative component is modelled semi-parametrically with a mixture of normals. This type of formulation is also suggested by Smyth (2004), who uses moderated F statistics, using shrunken estimates of variances as with the moderated t statistics.

A slightly different formulation is considered by Bochkina & Richardson (2007), who use alternative hypotheses such as ‘the gene is differentially expressed in a set of pairwise comparisons of interest’ against a compound null which is the opposite of this, that is, genes are only selected to be of interest if they show changes in a predefined set of comparisons of interest.

30.4.2 Differential Expression from RNA-Sequencing Data

30.4.2.1 Priors for Differential Gene Expression

A range of priors for gene-level differential expression have been proposed, very similar to those used in differential expression from microarray data. The most common in Bayesian models is to use a mixture prior, as this conveniently classifies genes as differentially expressed or not. Chung *et al.* (2013) developed a model for paired data and use a two-component mixture of normals for a differential expression parameter. Lee *et al.* (2015) used a three-component mixture, also for paired data. Several people use a spike-and-slab type mixture of a point mass at zero for non-differentially expressed genes and other distributions for differentially expressed genes (Van De Wiel *et al.*, 2013; Gu *et al.*, 2014; Katzfuss *et al.*, 2014). Leng *et al.* (2013) fit a mixture model at the data level, directly to read counts. An alternative way to classify differential expression is to use an unstructured prior on a fold change parameter δ_g and calculate the posterior probability of δ_g being larger than some given threshold (e.g. Wu *et al.*, 2013; Bi & Davuluri, 2013; Ritchie *et al.*, 2015). This is a useful way to incorporate the requirement that fold changes should themselves be large as well as statistically significant.

30.4.2.2 Priors for Differential Transcript Expression

The Bayesian models for individual reads typically involve a vector of transcript expression parameters $\vec{\theta}$ whose components sum to one. The model in Katz *et al.* (2010) is designed for analysing one locus (gene) at a time, hence the probabilities of transcripts being expressed are relative to that locus only. Glaus *et al.* (2012) model relative expression across the transcriptome. Both models use Dirichlet priors for the expression parameters. In these papers differential expression between conditions is assessed in a *post hoc* procedure, fitting separate models on the posterior distributions of the expression parameters. Papastamoulis & Rattray (2017) develops a model with two Dirichlet priors $\vec{\theta}_1$ and $\vec{\theta}_2$ for expression parameters in two conditions, with the constraint that $\theta_{1t} = \theta_{2t}$ for non-differentially expressed genes.

Marety *et al.* (2014) use a two-stage prior to explicitly model zero expression levels: each transcript t has a latent binary variable z_t indicating whether the transcript is expressed or not, and the non-zero expression levels have a Dirichlet prior. This model was developed for transcript discovery, by using read alignments to newly assembled transcriptomes.

30.4.2.3 Priors for Allele-Specific Expression

Nariai *et al.* (2016) extend the read-level model given in Section 30.3.2 to accommodate two reference transcriptomes from alternative alleles. For each read i , in addition to the label T_i indicating the originating transcript, the model includes a binary indicator H_i for allele 0 or 1. The model becomes

$$\mathbb{P}(r_i|\vec{\theta}, \vec{\phi}) = \sum_{t,h} \mathbb{P}(r_i|T_i = t, H_i = h)\mathbb{P}(T_i = t, H_i = h|\vec{\theta}, \vec{\phi}),$$

where the probability of the read sequence coming from a fragment of transcript t and allele h is similar to before, and the probability of the fragment is decomposed into transcript expression and allele choice:

$$\mathbb{P}(T_i = t, H_i = h|\vec{\theta}, \vec{\phi}) = \begin{cases} \phi_t \theta_t, & h = 0, \\ (1 - \phi_t) \theta_t, & h = 1, \end{cases}$$

so that the overall transcript expression is still represented by the $\vec{\theta}$ vector, and ϕ_t is the probability of one particular allele of transcript t being expressed (represents allelic preference).

León-Novelo *et al.* (2014) use a read count model for allele-specific expression, with a similar parameter structure combining overall mean expression and allelic preference.

30.5 Multivariate Gene Selection Models

In the previous sections we have discussed gene expression association studies where the aim is to find gene expression changes that relate to biological outcomes by comparing, for each gene, their differential expression under different conditions. In this section we are concerned with a different but related problem, that of using gene expression for phenotype prediction. Our aim here is to build multivariate molecular profiles based on combination of the expression of a subset of genes which can characterize different phenotypes (e.g. clinical outcomes). We are thus in the framework of multivariate regression and classification models. The specific difficulty of genomic applications is that there are typically many more covariates than samples, the so-called ‘large p (thousands of genes), small n (50–100 samples) regression paradigm’, and consequently standard regression/discrimination techniques do not apply. Further, the interest is in finding parsimonious regression models that include only small subsets of genes so that biological interpretation and validation can be attempted.

Bayesian approaches to multivariate gene selection have broadly followed two related lines of development: regression methods with covariate selection; and regression models with shrinkage priors that favour sparsity. We shall review these in turn. Mostly we shall discuss so-called supervised classification situations where the characteristics of the samples that one wants to predict are known. Variable selection can also be performed simultaneously with the task of uncovering clustering patterns of the samples, in an unsupervised manner. These methods aim to answer the following questions. First, are there genes that are irrelevant to discrimination of subgroups/clusters? Second, among the discriminatory genes, are there small subsets of them that characterize a subgroup/cluster alone?

30.5.1 Variable Selection Approach

Suppose that we have potentially p predictor variables (e.g. genes), each measured on a set of n samples: x_{gc} with $g = 1, \dots, p$ and $c = 1, \dots, n$. Thus for each predictor variable g , we have a vector of n measurements \vec{x}_g . The outcome variable, y_c , $c = 1, \dots, n$, can be continuous (e.g.

measuring a biomarker) or categorical (e.g. encoding a cancer subtype), and we denote by $\vec{\beta}$ the vector of regression parameters linking \mathbf{X} and \mathbf{Y} (these being the matrices for predictors and outcomes, respectively). Thus β_g is the regression parameter corresponding to the covariate \vec{x}_g .

Bayesian variable selection (BVS) is usually implemented through a hierarchical model, where all possible 2^p models are represented by a p -dimensional indicator variable $\vec{\gamma}$:

$$\gamma_g = \begin{cases} 1, & \text{variable (gene) } g \text{ is included,} \\ 0, & \text{variable (gene) } g \text{ is excluded.} \end{cases}$$

A prior on the model space can be specified via a prior $p(\vec{\gamma})$. A common choice is $p(\vec{\gamma}) = \prod_{g=1}^p \pi^{\gamma_g} (1 - \pi)^{1-\gamma_g}$, and by choosing π small, the number of variables selected can be controlled. Alternatively, a beta prior distribution can be assumed for π and the sparsity of the regression is controlled by the parameters of the beta prior distribution whose coefficients can be fixed by matching the *a priori* expected number (and standard deviation) of potential associations (Kohn *et al.*, 2001). A general introduction about variable-selection priors can be found in Chipman *et al.* (2001).

The most generic approach to variable selection, often referred to as stochastic search variable selection (SSVS), was taken in Mitchell & Beauchamp (1988), George & McCulloch (1993), George & McCulloch (1997), and in many subsequent papers (Clyde, 1999; Brown *et al.*, 1998, 2002) including applications on gene mapping (Yi *et al.*, 2003). SSVS differs from Kuo and Mallick's (1998) method and its applications in genetics (Sillanpää & Bhattacharjee, 2004, 2006), and from Gibbs variable selection (GVS; Dellaportas *et al.*, 2002) in the way $p(\beta_g, \gamma_g)$, the joint prior distribution of the regression parameter β_g and the indicator variable γ_g for the predictor g , is modelled (O'Hara & Sillanpää, 2009).

Much of the work on BVS was developed for linear models where y_c is continuous. In this context, authors differ in the choice of the prior distribution for $\vec{\beta}$, in particular whether the components of $\vec{\beta}$ are treated as independent or not, and whether a conjugate formulation is chosen so that the prior on $\vec{\beta}$ includes the noise parameter of the linear model. The resulting hierarchical mixture prior is referred by George & McCulloch (1997) as the 'conjugate prior'. In SSVS and GVS, the prior for the regression coefficients is formulated via a mixture. Most models define a point mass at zero for β_g when $\gamma_g = 0$, while when $\gamma_g = 1$, large variances are favored with a distribution to be specified. Either the variance can be fixed, in which case the model corresponds to a classical 'fixed-effects' model, or a prior distribution is specified on the variance τ of the regression coefficients, in which case the model coincides with a 'random-effects' model. When the model is extended to include a prior on τ , two strategies can be applied: the variance is either common to all predictors $p(\beta_g | \gamma_g = 1, \tau)$ (global shrinkage) or specific for each predictor separately $p(\beta_g | \gamma_g = 1, \tau_g)$ (local shrinkage). Application of local shrinkage can be found in Ishwaran & Rao (2003, 2005a,b) where a modified 'slab and spike' model that assumes a continuous bimodal prior for each β_g , a scale mixture of two centred normals, one having a small variance, is presented. They show that this prior is useful in finding genes that are differentially expressed in two or more conditions.

In the common case of a prior for β_g with a point mass at zero, for ease of notation, we shall denote by $\vec{\beta}_{\gamma}$ all non-zero elements of $\vec{\beta}$ and, correspondingly, we denote by \mathbf{X}_{γ} the columns of \mathbf{X} corresponding to those elements of $\vec{\gamma}$ equal to 1. A standard choice of prior for $\vec{\beta}_{\gamma}$ is the so-called g -prior (Zellner, 1986), where $\vec{\beta}_{\gamma} \sim N(0, g(\mathbf{X}_{\gamma}^T \mathbf{X}_{\gamma})^{-1})$, where g is a positive scale factor to be chosen. Several extensions have been proposed. George & Foster (2000) suggest an empirical Bayes approach for g when p is not too large, Liang *et al.* (2008) discuss the benefits of a prior on g that overcomes many of problems when g is fixed, while Cui & George (2008) and

Scott & Berger (2010) compare the two approaches. Maruyama & George (2011) generalize Zellner–Siow Cauchy priors to allow for $p_\gamma > n$, with p_γ the number of ‘active’ variables, Baragatti & Pommeret (2012) adapt the g -prior with the ridge parameter proposed by Gupta & Ibrahim (2007) to cope with situations where the covariates are highly collinear, and Krishna *et al.* (2009) extend Zellner’s g -prior for out-of-sample prediction. Finally, Ley & Steel (2012) examine the performance of the various priors on g in the context of simulated and real data. Similarly, if independent normals are specified for the components of $\vec{\beta}$ (Kuo & Mallick, 1998; Brown *et al.*, 2002; Hans *et al.*, 2007; Guan & Stephens, 2011), again a scalar has to be chosen. These choices influence the sparsity of the final regression model, and a full understanding of this aspect is the object of current research (Gelman, 2006; Polson & Scott, 2010).

Application in genetics of the above prior set-ups for finding genetic signatures of a given disease or performing quantitative trait mapping will be presented in Section 30.6.1. However, in real case studies, the independent prior associated with a global shrinkage seems to be preferred from both a statistical (Brown *et al.*, 2002; Hans *et al.*, 2007; Krishna *et al.*, 2009) and a biological (Guan & Stephens, 2011) viewpoint.

In the context of gene expression data, because we are in a ‘large p , small n ’ situation, the posterior distribution over the model space of variable dimensions is multi-modal. Moreover, full posterior inference for the entire model space of size 2^p is not feasible if p is larger than about 20. Hence, MCMC methods are rather used as stochastic search algorithms with the aim of quickly finding many regions of high posterior probability. The Markov chain needs to move quickly around the support of the posterior distribution and, as usual, it is useful to integrate out as many parameters as possible. For this reason, the ‘conjugate prior’ settings have been favoured in the linear model (George & McCulloch, 1997). In the non-conjugate case, the regression coefficients cannot be integrated out analytically. When proposing changes to $\vec{\gamma}$, a key question is how to propose sensible changes to the regression vector $\vec{\beta}$, and various strategies have been adopted. Dellaportas *et al.* (2002) assume that $\beta_g | \gamma_g = 0$ is sampled from a ‘pseudo-prior’ concentrated around the region of high posterior mass allowing the Markov chain to move to $\beta_g | \gamma_g = 1$ and explore the model space effectively. Holmes & Held (2006) propose to sample simultaneously γ_g and β_g using a Metropolis–Hastings step. In their formulation the Metropolis–Hastings acceptance probability does not depend on $\vec{\beta}$ since the conditional distributions (evaluated at the current and proposed new value of $\vec{\beta}$) cancel out with the proposal densities and this implicit marginalization leads to an efficient sampling of $\vec{\gamma}$.

Much of the application of variable selection in studies of gene expression has concerned binary or categorical variables rather than the linear model. Typically, samples are classified as good or poor prognosis or linked to different clinical sub-entities, like subtypes of cancer. There is no immediate conjugate formulation of Bayesian categorical regression, but following the approach of Albert & Chib (1993), probit regression can be efficiently implemented through the use of latent auxiliary variables z which allows integration of the regression coefficients in the full conditional distribution of the indicator variables $\vec{\gamma}$. This approach was taken by Lee *et al.* (2003) and Sha *et al.* (2004). These authors have implemented different MCMC schemes for updating $\vec{\gamma}$: Gibbs sampling for Lee *et al.* (2003), which will tend to be slow mixing, and Metropolis with add/delete/swap moves for Sha *et al.* (2004). Holmes & Held (2006) have proposed an auxiliary variable formulation of the logistic model which allows, in a similar way to the probit model, integration of the regression coefficients when updating the indicator variables $\vec{\gamma}$, thus improving mixing – although Holmes and Held’s sampler is likely to mix slowly because the auxiliary variable z is correlated with $(\beta_\gamma, \vec{\gamma})$ and a Gibbs sampler is used to update z (Lamnisos *et al.*, 2009). For both probit and logistic regression models, the calibration of the prior distribution of the regression coefficients again influences the outcome of the variable selection process. In this respect, the logistic regression, which is more commonly used for

binary regression, is easier to calibrate than the probit model as it has heavier tails and so exhibits less sensitivity. Following these considerations, Polson *et al.* (2013) and Polson & Scott (2013) have introduced a novel efficient data-augmentation algorithm for Bayesian logistic regression based on the Pólya–gamma latent variables, but valid for any binomial likelihood. In contrast to Albert & Chib (1993), their method is a scale mixture (the binomial likelihood parameterized by log odds can be represented as mixtures of Gaussians with respect to a Pólya–gamma distribution) rather than a location mixture, and Albert and Chib's truncated normals are replaced by Pólya–gamma latent variables. Similar to Albert & Chib (1993), but differently from Holmes & Held (2006) and Frühwirth-Schnatter & Frühwirth (2010), their method requires only a single layer of latent variables, which leads to the construction of a more efficient Gibbs sampler.

In general, MCMC variable selection algorithms in high dimension are difficult to implement due to slow convergence. Recent developments in stochastic simulation algorithms, such as population-based reversible jump MCMC (Jasra *et al.*, 2007) and the use of parallel tempered chains (Liang & Wong, 2000; Bottolo & Richardson, 2010), have been successfully applied (Petretto *et al.*, 2010; Bottolo *et al.*, 2011b). An alternative search algorithm, the shotgun stochastic search method, which is close in spirit to MCMC but aims to search rapidly for the most probable models, has been proposed by Hans *et al.* (2007). Titsias & Yau (2017) extended the shotgun stochastic search algorithm, allowing a larger set of models to be explored in parallel by defining the *radius* of the Hamming ball, centred on the current model and with a metric (distance) defined in the model space. Note that it is a discussion point whether to report models with high posterior probabilities and their associated variables, or to extract marginal information about the selected variables by looking at their marginal posterior probabilities of inclusion (Sha *et al.*, 2004).

Another approach is based on adaptive MCMC (Andrieu & Thoms, 2008) where the proposal distribution is tuned automatically and adaptively as the sampler runs such that a new proposal depends on the history of the visited models. Nott & Kohn (2005) notice that since γ_g is binary, the full conditional $p(\gamma_g | y, \vec{\gamma}_{\setminus g}) = E(\gamma_g | y, \vec{\gamma}_{\setminus g})$ with $\vec{\gamma} = (\gamma_g, \vec{\gamma}_{\setminus g})$. In their adaptive MCMC, the conditional expectation is approximated by the best linear predictor using the posterior mean and covariance matrix of the visited $\vec{\gamma}$ s. Recently, Ji & Schmidler (2013) have developed an algorithm based on a Metropolized independent sampler that automatically tunes the parameters of a family of mixture proposal distributions during simulation. Asymptotic convergence is guaranteed by enforcing the conditions presented in Roberts & Rosenthal (2009). Adding a point-mass density to the mixture proposal distribution, their adaptive proposal mimics the 'slab and spike' shape of the sparsity prior on the regression coefficients used in SSVS. For the linear model an alternative to MCMC is sequential Monte Carlo (Del Moral *et al.*, 2006), which has been used to search efficiently the large model space with better results than standard and adaptive MCMC methods in simulated and real examples (Schäfer & Chopin, 2013).

In the approaches presented above, *a priori* the predictors are treated in the same fashion, that is to say, potentially they all contribute to explain the outcome's variability. However, some information about the importance of the predictors is readily available at a very low cost – for instance, a simple univariate regression can highlight the most important predictors and this information can be used to construct more efficient proposal distributions. This approach was taken by Guan & Stephens (2011). When adding, their proposal is constructed by ranking the predictors with respect to some measure of association with the response, such as Bayes factors. To ensure that the move is reversible, deletion is performed uniformly at random among the variables currently in the model. This strategy enables the algorithm to move quickly to regions of high posterior mass, although multicollinearities can decrease the efficiency of the algorithm. In a highly collinear setting, Kwon *et al.* (2011) propose to tackle the problem of

the correlation among the predictors. In their algorithm, after selecting uniformly at random a predictor currently in the model, they propose to remove the most correlated predictor with the selected one. Similarly, when adding, they propose to add the least correlated predictor (within a specified set) with the random selected one. The set is defined by the covariates (not currently in the model) that are correlated with the selected predictor given a user-defined correlation threshold. To ensure irreducibility of the Markov chain, a requirement for convergence, they couple their strategy with SSVS. In this way the resulting ‘hybrid’ algorithm, while searching for data-supported models, checks whether unimportant correlated variables have been included, reducing at the same time over-fitting. In similar problems where there are multicollinearities among the predictors, a more elaborated sampling scheme which is based on the Swendsen–Wang algorithm for the Ising model has been proposed by Nott & Green (2004) for the linear model and extended for generalized linear models by Nott & Leonte (2004).

‘Slab and spike’ priors can also be used in the context of latent factor models (West, 2003; Lucas *et al.*, 2006; Carvalho *et al.*, 2008; Knowles & Ghahramani, 2011; Bhattacharya & Dunson, 2011). Modelling high-dimensional data via latent factor models is a powerful dimension reduction technique that allows the identification of patterns of covariation among genes. In their application of factor models to the analysis of gene expression data, West (2003), Lucas *et al.* (2006), Carvalho *et al.* (2008), Knowles & Ghahramani (2011), and Bhattacharya & Dunson (2011) further structure the factor loading matrix to encourage sparsity via a mixture prior with point mass at zero. A biological interpretation of the factors as potentially representing biological pathways is then derived by examining the list of genes most weighted on each factor.

We end this section by referring to the work of Tadesse *et al.* (2005), Kim *et al.* (2006) and Stingo & Vannucci (2010) where variable selection and clustering of the samples are performed simultaneously. This joint modelling is motivated by the remark that using a high-dimensional vector of gene expression to uncover clusters among the samples might not be effective, and on the contrary can tend to mask existing structure, while a more parsimonious model that selects only a small subset of covariates to inform the clustering is more easily interpretable. Such joint modelling was discussed in a Bayesian context in two related papers, which differ in their model for the clustering structure. Tadesse *et al.* (2005) formulate their clustering structure using a finite mixture of multivariate normals with a variable number of components and use reversible jump techniques to explore different structures, while Kim *et al.* (2006) exploit the computational benefits of DPMs. Stingo & Vannucci (2010) extend the method presented in Tadesse *et al.* (2005) for discriminatory analysis, incorporating into the model prior information on the relationships among the genes by using a Markov random field prior.

30.5.2 Bayesian Shrinkage with Sparsity Priors

An alternative approach to BVS for selecting a small number of regressors is to use a hierarchical formulation of the regression problem with a prior on the regression coefficients that favours sparsity. It is usually referred to as ‘adaptive shrinkage’ since the shrinkage is estimated from the data (O’Hara & Sillanpää, 2009). In contrast to BVS, where the sparsity is induced in the model by using indicators, here the same sparsity effect is obtained by specifying a prior directly on the regression coefficient β_g that approximates the ‘slab and spike’ shape. For this reason it is also known as the ‘absolutely continuous shrinkage’ approach (Griffin & Brown, 2010, 2012). It is worth noticing that different choices of priors and hierarchical structures on the regression coefficients can be interpreted as a different choice of penalties if one adopts the point of view of the penalized likelihood framework into which ridge regression and other shrinkage methods can be cast.

A general formulation that encompasses many of the models which have been proposed is that of scale mixture of normals. In this formulation, the regression coefficients β_g are given independent normal priors, $\beta_g \sim N(0, \tau_g)$, and the variances τ_g are themselves given a hyperprior distribution, $\tau_g \sim p(\tau_g)$. The specification of this prior distribution leads to different kinds of sparsity for the β s, but all priors used have in common the desirable feature that the integrated prior for β_g is a heavy-tailed distribution with a peak around zero, thus favouring only a small number of covariates to be substantially different from zero with no shrinkage for data-supported covariates with $\beta_g \neq 0$. The double exponential (Laplace) prior of the Bayesian lasso (Park & Casella, 2008; Hans, 2009) is an example of this class of prior since it can also be written as a scale mixture of normals with $p(\tau_g)$ being a one-parameter exponential distribution. Griffin & Brown (2010) show that, since in the Bayesian lasso the gamma mixing distribution τ_g has shape parameter equal to one, this choice is too restrictive and, in contrast to the 'slab and spike' formulation, it does not allow an infinite spike at zero. Instead, they introduce a more flexible normal-gamma prior distribution and extend it to a multivariate scenario with correlated regression coefficients (Griffin & Brown, 2012) with applications in categorical or ordinal regressions.

Recently, the idea of local shrinkage has been relaxed to allow for a global-local shrinkage $p(\beta_g | \tau, \lambda_g)$ (Polson & Scott, 2010, 2012; Bhadra *et al.*, 2016), meaning that there is a global hyperparameter τ that shrinks all the regression coefficients towards zero, while the heavy-tailed prior for the local hyperparameter λ_g allows some β_g to escape the shrinkage. The horseshoe prior (Carvalho *et al.*, 2009, 2010) and its extension for ultra-sparse signals (Bhadra *et al.*, 2017) belong to this class of prior distributions. Consistent with the properties of other 'absolutely continuous shrinkage' priors, the horseshoe has been developed as a 'one-group' prior. Polson & Scott (2010) coined this term to distinguish the horseshoe prior from the 'two-group' model used in the 'slab and spike' formulation. Other examples of 'one-group' global-local models include the three-parameter beta prior (Armagan *et al.*, 2011), the generalized double Pareto prior (Armagan *et al.*, 2013) and the Dirichlet–Laplace prior (Bhattacharya *et al.*, 2015).

A possible limitation of the Bayesian shrinkage approach is that there is no indicator variable to show when a variable is included in the model. Selection of a relevant predictor can be done by setting a standardized threshold, such as $\beta_g > c$, or deriving a fully Bayesian decision rule for the threshold (Mutshinda & Sillanpää, 2012). While the lack of an indicator variable could be a benefit rather than a limitation since no model search is needed, it does not allow estimation of model uncertainty, which is important within the Bayesian framework. To overcome this problem, for the Bayesian lasso, Hans (2010) computes exactly the marginal posterior probabilities of inclusion for small models. For larger models, he derives a Gibbs sampler to calculate the posterior inclusion probabilities, bridging Bayesian shrinkage to SSVS. Finally, Lykou & Ntzoufras (2013) derive a Gibbs sampler for the Bayesian lasso with the specification of the coefficients of τ_g so as to control the level of shrinkage and the variable selection procedure at the same time (see also Mutshinda & Sillanpää, 2010).

An early empirical attempt to assess the impact of the choice of the prior density for τ_g in the context of gene expression studies can be traced back to Bae & Mallick (2004). They implemented three different choices of prior for τ_g : an inverse gamma with two hyperparameters that are chosen in order to favour large variances; a Laplace prior with one parameter; and a Jeffreys prior, $p(\tau_g) \propto \tau_g^{-1}$, with no tuning of the prior parameters. They found that the Jeffreys prior induces more sparseness than the Laplace prior and yields good performance. Since no unifying theoretical results were available at that time, different procedures have been proposed. For instance, the Jeffreys prior has been used by Xu (2003) and Wang *et al.* (2005) to perform gene mapping and in Yang & Xu (2007) to perform QTL mapping for dynamic traits, whereas

Yi & Xu (2008) applied Bayesian lasso in a QTL mapping problem. Shrinkage methods have also been used for mapping QTLs in multiple continuous and discrete traits (Xu *et al.*, 2009).

General theoretical results are now available since the normal-gamma prior for the regression coefficient β_g , also investigated by Caron & Doucet (2008), is seen to be the finite-dimensional marginal distribution of a particular type of Lévy process, the variance-gamma process (Polson & Scott, 2010, 2012). This representation permits the generalization of the problem and a better understanding of the level of sparsity induced by the specific choice of the prior. From a practical point of view, it allows a more efficient and effective implementation of Bayesian shrinkage procedures in a high-dimensional set-up.

30.6 Quantitative Trait Loci

Besides the application of variable selection in microarrays or RNA-seq experiments, where the response is a binary or categorical variable measured on n individuals and the predictors are a group of genes measured on the same set of individuals, BVS has been applied also in QTL problems. The main idea of QTL mapping is to find the association between a quantitative trait, usually a continuous variable, and a categorical variable (genotype values). In this framework, Bayesian methods have been particularly successful since they are able to analyse all genetic markers together and detect small effects that would be missed by standard ‘single-SNP’ methods.

Particularly important is the determination of genetic causes of gene expression, known as expression QTL. We will review statistical methods that have been proposed to detect eQTLs. While for a single gene they coincide with continuous trait mapping models, methods for integrated eQTL analysis differ substantially from multiple QTL tools (Xu *et al.*, 2005, 2009; Banerjee *et al.*, 2008; Chen *et al.*, 2009; Sillanpää *et al.*, 2012) when a set of genes are jointly considered. We will conclude this section reviewing Bayesian partition models and their connection with multiple response models since they combine mapping tasks with clustering of gene expression into subgroups or ‘partitions’ that share similar regulatory mechanisms.

30.6.1 Single-Response Models

Several strategies have been proposed based on different combinations of (i) priors on the regression coefficients and the residual variance, (ii) the sparsity prior on the number of polygenic effects, and (iii) the search algorithm used to explore the model space. Yi *et al.* (2003) were the first to use SSVS to identify QTL; Yi (2004) employed a trans-dimensional sampling algorithm to estimate the number of QTLs, their genomic positions and the genetic effects; Yi *et al.* (2005) extended Yi (2004) to include epistatic (interaction) effects for continuous traits, while Yi & Banerjee (2009) modified it for any discrete or continuous trait; since in QTL analysis the signal is very sparse, Yi *et al.* (2007) employed MCMC for sparse regression models presented in Kohn *et al.* (2001) to improve the mixing of the MCMC sampler. Bottolo & Richardson (2010) use a combination of Zellner–Siow Cauchy priors (Liang *et al.*, 2008) for the regression coefficients and evolutionary Monte Carlo (Liang & Wong, 2000) for models exploration. In particular, they consider a parallel chain MCMC implementation that overcomes the known problems suffered by many SSVS algorithms of being trapped in a local mode of the model space’s posterior distribution, whereas the Zellner–Siow Cauchy prior allows a data-dependent level of shrinkage. Guan & Stephens (2011) propose a multi-SNP method where the parameterization of the hierarchical mixture prior induces a uniform prior density on the proportion of variance explained (PVE). This prior specification is based on the idea that a larger model size does not

necessarily lead to a larger PVE since *a priori* it may be plausible that there could be either a large number of relevant predictors in the model with small PVE, or a small number of predictors with large PVE. Moreover, they introduce a new proposal density for SSVS based on the strength of association (Bayes factor) between the response and each covariate that greatly improves the ability of the search algorithm to move quickly to regions of high posterior mass. Variational Bayes (VB) (Blei *et al.*, 2017), a class of machine learning techniques that rely on the optimization of the variational function as a proxy for the posterior distribution and scale well as data sizes increase, has been used by Carbonetto *et al.* (2012) as an alternative to MCMC. VB is particularly appealing since it does not require sampling from the posterior distribution of $\vec{\gamma}$. Instead, variable selection is obtained through the optimization of the posterior ‘slab and spike’ density.

Further extensions of Bayesian QTL models include the specification of prior information through a random Markov field prior on $\vec{\gamma}$ (Stingo *et al.*, 2011) and the inclusion of random effects (Zhou *et al.*, 2013) whose prior parameters are estimated based on the proportion of genetic variance explained. Finally, an alternative parameterization for the linear mixed model which has greater power than a linear model is proposed by Moser *et al.* (2015), where the regression coefficients are modelled as a hierarchical mixture model.

30.6.2 Multiple-Response Models

In recent years research has mainly focused on the analysis of the genetic determinants of gene expression (eQTL). Only a few methods, specifically developed for the analysis of protein or metabolite QTL, mQTL and pQTL respectively, exist. This is mainly due to the fact that methods developed for eQTL can be extended with few modifications to other omics data. However, eQTL analysis poses two main challenges. First, the genetic architecture of the regulation in gene expression is more complicated than other quantitative traits. Second, relevant information is common across genes and the analysis of one gene ‘at a time’ may lead to a loss of power.

Regarding the first problem, two different approaches have been developed to detect either *cis* effects or *trans* effects. In the *cis* analysis only SNPs that are close to the gene under investigation (usually within 1–5 Mbp) are tested for association. When detected, the effect of the *cis* marker on gene expression is large so statistical tests have sufficient power even when the sample size is relatively small. More complicated is the detection of *trans* effects since the number of ‘distal’ genetic control points is unknown and can vary with the gene under investigation. The effect size for the associated genetic markers is generally small and difficult to detect even when the sample size is large. With respect to the second problem, the joint analysis of gene expression is particularly important for the detection of *trans*-eQTL ‘hotspots’, genetic markers that exert their influence over many genes. From a biological point of view, their discovery highlights the systemic nature of these genomic locations and their pleiotropic effect. There has been some debate over the existence of *trans*-eQTL ‘hotspots’ since the effect size of the associations is generally small and can be due to confounding effects (e.g. artefacts arising from technical or environmental factors) rather than from a real genetic mechanism. Bayesian integrated methods that perform sparse association in order to discover *cis* and *trans* eQTLs while correcting for confounding effects have been proposed to solve this problem (Stegle *et al.*, 2010; Fusi *et al.*, 2012).

Recently the availability of data recorded on multiple conditions (i.e. tissues or cell types) or on multiple time points has further increased the dimension of the problem. In this framework eQTL analysis is performed between a matrix of genotype values measured on a set of individuals and a multidimensional array, or tensor, describing the expression levels across conditions

or time points for the same set of individuals. Such data is extremely informative regarding the nature and architecture of the genetic regulation; for instance, it is able to reveal whether the association between a genetic marker and a gene is conserved across tissues or cell types or, conversely, is tissue-specific or, in a time-course analysis, how the regulation evolves over different time points. From a statistical perspective, the joint analysis of multi-condition gene expression facilitates the discovery of small *trans* effects that would have gone unnoticed otherwise, especially when they are conserved across conditions or cell types. The advantage of the Bayesian paradigm and, in particular, of Bayesian hierarchical models is especially prominent in this scenario since it allows information sharing not only between genes but also between different conditions or cell types.

In the following we will review two multiple response models that have been considered for eQTL detection. They can be broadly summarized into two categories: (i) single genetic marker and multiple traits; and (ii) multiple genetic markers and multiple traits. The latter is the primary statistical tool used for ‘hotspots’ detection.

30.6.2.1 Single-Marker Multi-trait Analysis

Besides a single trait, SNPTEST (Marchini *et al.*, 2007) can be used to analyse multiple quantitative traits. The model exploits the conjugacy of the matrix variate normal distribution with the matrix variate normal-inverse-Wishart prior for which the marginal likelihood can be obtained in closed form. Similar to the single-trait analysis, both Bayes factor and marginal posterior probability of inclusion are available, indicating the strength of association of the target SNP with *all* traits under investigation. In an eQTL analysis across tissues, Flutre *et al.* (2013) relax the hypothesis of pleiotropy across *all* traits. For a target gene, their goal is to model the potential genetic association between the expression level measured in *S* tissues and a target SNP by means of *S* linear regressions linked by the residual correlation. However, since in their model there is only one predictor, their goal is to find which ‘configuration’ of tissues, if any, is associated with the target SNP. Each of the *S* regression coefficients is modelled by a ‘slab and spike’ distribution with a random effect on the slab part (Wen & Stephens, 2014). The hierarchical model is completed by the specification of a sparsity prior that controls the number of tissues associated. Making the simplifying assumption that each gene has at most one eQTL which may be active in multiple tissues, and after averaging over all possible ‘configurations’, their method provides posterior evidence of an eQTL in a *cis* region. Generalizations of this model have been considered also by Wen (2014).

30.6.2.2 Multi-Marker Multi-trait Analysis

The first solution to the problem of the joint analysis of expression traits and markers was proposed by Kendziorski *et al.* (2006). Their method combines the advantages of differential expression analysis (multiple genes, one marker) with eQTL single response models (multiple SNPs, one gene). In particular, for each gene *g*, the expression level is modelled by a mixture model over *M* markers,

$$L_g = \pi_0 f_0(\vec{y}_g) + \sum_{m=1}^M \pi_m f_m(\vec{y}_g),$$

where \vec{y}_g is the vector of observations, π_0 is the prior probability of no association with any marker, π_m is the prior probability of association with marker *m*, $f_0(\vec{y}_g)$ is the marginal likelihood for equivalently expressed genes and $f_m(\vec{y}_g)$ is the marginal likelihood for differentially expressed genes. Note that the prior probability π_m of association between gene *g* and marker *m* does not depend on *g*. Conditionally on π_0 and π_m , $m = 1, \dots, M$, the joint likelihood is

$L = \prod_{g=1}^G L_g$, where G is the number of genes considered. After introducing the latent variable $\gamma_{mg} = \{0, 1\}$, the posterior probability that a gene maps nowhere or to any of the genetic markers is calculated via Bayes' rule by using an EM algorithm. In the mixture over marker (MOM) approach a restrictive assumption is that each gene is forced to be associated at most with one marker, or not associated with any of them. Given this assumption, each transcript can be regulated by either a *cis* or a *trans* locus but not by both, which limits the applicability of the method. Gelfond *et al.* (2007) extended MOM by allowing a gene g to be associated with more than one genetic marker. They also modelled the prior probability of a transcript–marker association favouring loci that are close to a transcript to reflect the important role that genomic proximity can play in the regulation of expression.

A similar idea that the prior probability of association π_m is modulated across all genes is used also by Jia & Xu (2007). Rather than a mixture model, they use SSVS to detect marker–gene associations with a 'slab and spike' prior on the regression coefficients,

$$\beta_{mg} | \gamma_{mg}, \sigma_g^2 \sim (1 - \gamma_{mg})\delta_0(\beta_{mg}) + \gamma_{mg}N(\beta_{mg}; 0, \sigma_g^2). \quad (30.9)$$

In a second stage, they model γ_{mg} as a Bernoulli distribution with probability π_m that does not depend on g . A final level of hierarchy assigns a beta prior density on π_m . Model parameters are estimated by a straightforward Gibbs sampler. In their model, the posterior probability $p(\pi_m | \mathbf{Y})$ can be used as a measure of evidence of being a 'hotspot' since it can be interpreted as the posterior frequency of the number of genes associated with marker m .

Richardson *et al.* (2010) and Bottolo *et al.* (2011b) noticed that while the above models are able to capture relevant information shared across genes, they are not able to control the overall level of sparsity. In Jia & Xu (2007) the sparsity prior on π_m tends to favour hotspots associated with the expression of many genes, whereas in Kendziorski *et al.* (2006) sparsity is achieved by imposing the simplifying assumption that at most one marker is associated with a gene. In order to better control the level of sparsity, Richardson *et al.* (2010) and Bottolo *et al.* (2011b) propose to decompose the prior probability of association between marker m and gene g into $\pi_{mg} = \rho_m \times \pi_g$, where π_g is the prior probability (constant across markers) of association in gene g and ρ_m is the *a priori* 'propensity' for marker m to be a 'hotspot'. Thus, the role of ρ_m is to boost the probability of marker–gene associations in genomic locations where there is evidence that many genes are regulated by the same marker, without compromising the level of sparsity that is controlled by π_g . In particular, for each gene g , they assign a beta prior on π_g whose coefficients match the *a priori* expected number (and standard deviation) of potential associations, whereas on ρ_m they specify a gamma prior with $E(\rho_m) = 1$ so that $E(\gamma_{mg}) \times p$ represents the 'expert opinion about the expected number of potential associations for gene g '. Another key advantage of their method is the possibility of integrating out both the regression coefficients and the residual variances, so that the marginal likelihood is available in closed form. They call their method HESS since it links hierarchically the ESS model presented in Bottolo & Richardson (2010) and Bottolo *et al.* (2011a) for single-response QTL mapping across genes. An extension of the model for tensor representation of gene expression is presented in Lewin *et al.* (2015) where they combine eQTL detection in multiple tissues (Petretto *et al.*, 2010) with HESS. The schematic representation of their multi-tissue HESS (MT-HESS) model is presented in Figure 30.3, where q genes in r conditions are jointly analysed with p SNPs.

A similar model for integrated eQTL analysis is presented in Scott-Boyer *et al.* (2012). In their integrated hierarchical Bayesian model (iBMQ), a two-level hierarchical model is specified for the regression coefficients. In the first level, a 'point–mass mixture' (Lucas *et al.*, 2006) is introduced,

$$\beta_{mg} | \pi_{mg}, \sigma_g^2 \sim (1 - \pi_{mg})\delta_0(\beta_{mg}) + \pi_{mg}N(\beta_{mg}; 0, \sigma_g^2).$$

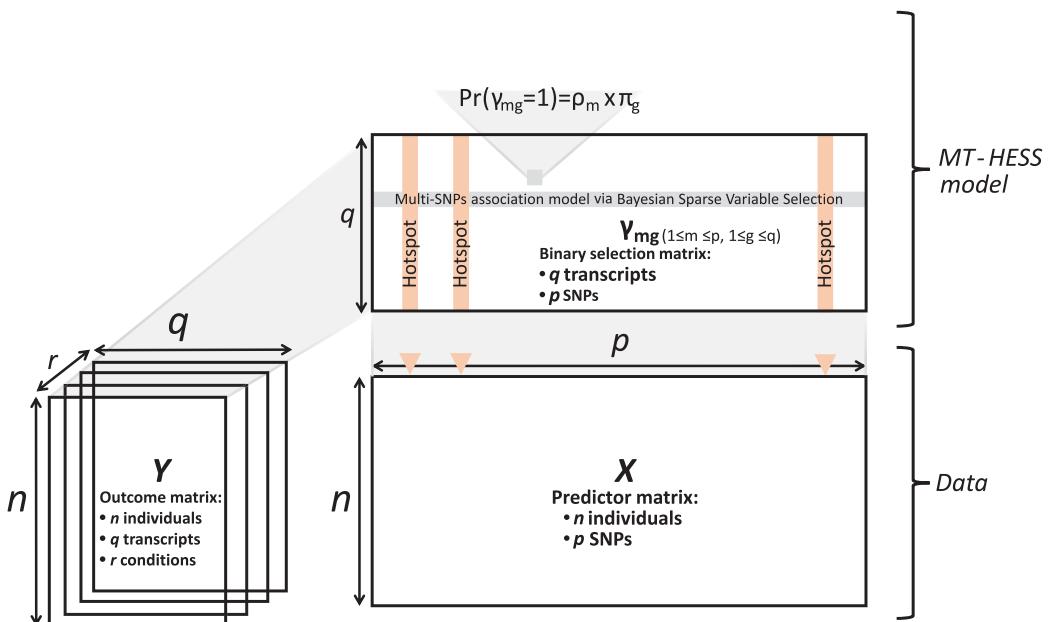


Figure 30.3 Schematic representation of the MT-HESS model. Data integration of input matrices is performed through the binary selection matrix. For each SNP–transcript pair in the binary selection matrix, the prior probability of association is described by a multiplicative model.

In the second level, they use a ‘two-group’ prior on the mixing weight π_{mg} ,

$$\pi_{mg} | \rho_m, \alpha, \beta \sim (1 - \rho_m) \delta_0(\pi_{mg}) + \rho_m \text{Beta}(\pi_{mg}; \alpha, \beta).$$

The prior specification is completed by assigning a beta prior on ρ_m , with $E(\rho_m) = a_0/(a_0 + b_0)$, where α and β are independent and exponentially distributed. Integrating out π_{mg} , the marginal model for the regression coefficients becomes

$$\beta_{mg} | \rho_m, \alpha, \beta, \sigma_g^2 \sim \left(1 - \rho_m \frac{\alpha}{\alpha + \beta}\right) \delta_0(\beta_{mg}) + \rho_m \frac{\alpha}{\alpha + \beta} N(\beta_{mg}; 0, \sigma_g^2),$$

which highlights the analogies with HESS decomposition of the prior probability of association. However, in iBHQ the boosting effect is with respect to the global ‘base-rate’ probability $\alpha/(\alpha + \beta)$, which may be difficult to elicit. Moreover, since in iBHQ $E(\gamma_{mg} | \alpha, \beta) = \{a_0/(a_0 + b_0)\} \{\alpha/(\alpha + \beta)\}$, *a priori* their model introduces dependence among the genes since $a_0/(a_0 + b_0) \neq 1$.

Recently, a VB approach for eQTL detection has been proposed by Ruffieux *et al.* (2017), where the prior distribution on the regression coefficients is a ‘slab and spike’ distribution similar to (30.9), $\mathbb{P}(y_{mg} = 1 | \pi_{mg}) = \pi_{mg}$ and $\pi_{mg} = \pi_m \sim \text{Beta}(\pi_m; a_0, b_0)$. This choice is mainly motivated by computational considerations since it ensures a closed form for their variational density, although it is not able to control sparseness as the number of markers m increases. Similar to Jia & Xu (2007), the implied marginal distribution of the regression coefficients is

$$\beta_{mg} | \sigma_g^2 \sim \left(1 - \frac{a_0}{a_0 + b_0}\right) \delta_0(\beta_{mg}) + \frac{a_0}{a_0 + b_0} N(\beta_{mg}; 0, \sigma_g^2).$$

In the methods presented above the dependence among genes is controlled by a hierarchical model on the marker–gene probability of association. Instead, Bhadra & Mallick (2013) propose to model the dependence among genes by a model inspired by the seemingly unrelated regression (SUR) model (Zellner, 1962). The basic idea is that, besides shared/common regulations, confounding effects or unexplained covariation due to unmeasured predictors make the genes correlated. Similar to the SUR, they model the residual correlation not explained by the genetic markers. However, in contrast to SUR, they use a conjugate model for the regression coefficients that allows a closed-form calculation of the marginal likelihood. With this assumption, their model coincides with the approach proposed by Brown *et al.* (1998, 2002), although in Bhadra & Mallick (2013) the residual covariance matrix is modelled by a hyper-inverse Wishart distribution (Carvalho *et al.*, 2007). Model selection is performed on the underlying graphical model encoded in the hyper-inverse Wishart distribution and among the markers analysed. In contrast to Petretto *et al.* (2010) and Marchini *et al.* (2007) where the pleiotropic effect is tested on few continuous traits or tissues, Bhadra & Mallick’s model is equivalent to a pleiotropic test for the whole set of genes considered, an assumption that does not hold in practice since ‘hotspots’ can only regulate up to a fraction of the genes. The distinction between methods that assume the same genetic model for all traits and methods that allow different genetic models for different traits is discussed also in Banerjee *et al.* (2008).

We conclude this section by describing a final method that combines partition models and multiple-response models in which the main assumption is that genes with similar expression profiles share similar regulatory mechanisms. Specifically, these models assume that coexpressed genes that are identified in the same partition are also co-regulated (i.e. the same genetic mechanism controls their variation), whereas genes in different partitions would have different regulatory mechanisms. Monni & Tadesse (2009) consider a multivariate Gaussian mixture model with an unknown number of components in order to partition the set of genes considered. The mean and the scale of each component are determined by a multiple-response regression model on a subset of predictors. Like Tadesse *et al.* (2005), this approach is equivalent to finding the subsets of predictors that discriminate between components. A null partition is permitted and corresponds to the set of genes not predicted by any markers and, conversely, markers that do not predict any genes. A more general model has been proposed by Zhang *et al.* (2010) who consider a Bayesian partition model for eQTL where correlated expression traits \mathbf{Y} and their associated set of markers \mathbf{X} are treated as a ‘module’. They introduce three sets of latent indicator variables for genes, markers and individuals’ genotypes to decouple the problem $\mathbb{P}(\mathbf{Y}|\mathbf{X})$ into $\mathbb{P}(\mathbf{Y}|\mathbf{T})$ and $\mathbb{P}(\mathbf{X}|\mathbf{T})$, where \mathbf{T} are the latent variables. Specifically, the expression level of each gene in a module is predicted by an analysis of variance model with the individuals’ genotypes as a factor. The joint distribution of the epistasis effects (haplotypes) in the same module follows a multinomial distribution whose frequency vector depends on the number of markers in the module and the individuals’ genotypes. A null partition is also permitted. Finally, a generalization of this model with a more sophisticated probabilistic description of the haplotypes’ frequency and the ability to account for the linkage disequilibrium among markers is presented in Jiang & Liu (2015).

Acknowledgements

The authors would like to thank their colleagues Angelos Alexopoulos, Marta Blangiardo, Natalia Bochkina, Peter Green, Anne-Mette Hein, Hélène Ruffieux and Ernest Turro for many insightful discussions on Bayesian modelling in gene expression analysis. The work on this chapter was supported by the ‘Stochastic Computation in the Biological Sciences’ programme at

the Isaac Newton Institute for Mathematical Sciences, the BBSRC ‘Exploiting Genomics’ grant 28EGM16093, the MRC grant MR/M013138/1 ‘Methods and tools for structural models integrating multiple high-throughput omics data sets in genetic epidemiology’ and the Alan Turing Institute under the Engineering and Physical Sciences Research Council grant EP/N510129/1.

References

- Aguiar, D., Cheng, L.-F., Dumitrascu, B., Mordelet, F., Pai, A.A. and Engelhardt, B.E. (2017). BIISQ: Bayesian nonparametric discovery of Isoforms and Individual Specific Quantification. Preprint, arXiv:1703.08260.
- Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.
- Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18**, 343–373.
- Armagan, A., Clyde, M. and Dunson, D.B. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems*. La Jolla, CA: Neural Information Processing Systems, pp. 523–531.
- Armagan, A., Dunson, D.B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23**, 119–143.
- Bae, K. and Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–3430.
- Baldi, P. and Long, A.D. (2001). A Bayesian framework for the analysis of microarray data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.
- Banerjee, S., Yandell, B.S. and Yi, N. (2008). Bayesian quantitative trait loci mapping for multiple traits. *Genetics* **179**, 2275–2289.
- Baragatti, M. and Pommeret, D. (2012). A study of variable selection using g -prior distribution with ridge parameter. *Computational Statistics & Data Analysis* **56**, 1920–1934.
- Bhadra, A., Datta, J., Polson, N.G. and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* **103**, 955–969.
- Bhadra, A., Datta, J., Polson, N.G. and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12**, 1105–1131.
- Bhadra, A. and Mallick, B.K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69**, 447–457.
- Bhattacharjee, M., Pritchard, C.C., Nelson, P.S. and Arjas, E. (2004). Bayesian integrated functional analysis of microarray data. *Bioinformatics* **20**, 2943–2953.
- Bhattacharya, A. and Dunson, D.B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306.
- Bhattacharya, A., Pati, D., Pillai, N.S. and Dunson, D.B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110**, 1479–1490.
- Bi, Y. and Davuluri, R.V. (2013). NPEBseq: Nonparametric empirical Bayesian-based procedure for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 262.
- Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.
- Bochkina, N. and Richardson, S. (2007). Tail posterior probability for inference in pairwise and multiclass gene expression data. *Biometrics* **63**, 1117–1125.
- Bottolo, L., Chadeau-Hyam, M., Hastie, D.I., Langley, S.R., Petretto, E., Tiret, L., Tregouet, D. and Richardson, S. (2011a). ESS++: A C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* **27**, 587–588.

- Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S.A., Tiret, L. and Richardson, S. (2011b). Bayesian detection of expression quantitative trait loci hot spots. *Genetics* **189**, 1449–1459.
- Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis* **5**, 583–618.
- Broët, P., Lewin, A., Richardson, S., Dalmasso, C. and Magdelenat, H. (2004). A mixture model based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20**, 2562–2571.
- Broët, P. and Richardson, S. (2006). Bayesian hierarchical model for identifying change in gene expression from microarray experiments. *Bioinformatics* **9**, 671–683.
- Broët, P., Richardson, S. and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.
- Brown, P., Vannucci, M. and Fearn, T. (1998). Multivariate Bayes variable selection and prediction. *Journal of the Royal Statistical Society, Series B* **60**, 627–641.
- Brown, P., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B* **64**, 519–536.
- Carbonetto, P., Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**, 73–108.
- Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning* (New York: Association for Computing Machinery Press), 88–95.
- Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1998). Measurement Error in Nonlinear Models. Boca Raton, FL: Chapman and Hall/CRC.
- Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.
- Carvalho, C.M., Massam, H. and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.
- Carvalho, C.M., Polson, N.G. and Scott, J.G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 73–80.
- Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.
- Chen, W., Ghosh, D., Raghunathan, T.E. and Sargent, D.J. (2009). Bayesian variable selection with joint modeling of categorical and survival outcomes: An application to individualizing chemotherapy treatment in advanced colorectal cancer. *Biometrics* **65**, 1030–1040.
- Chipman, H., George, E. and McCulloch, R. (2001). The practical implementation of Bayesian model selection (with discussion). In P. Lahiri (ed.), *Model Selection*. Beachwood, OH: Institute of Mathematical Statistics, pp. 67–134.
- Chung, L.M., Ferguson, J.P., Zheng, W., Qian, F., Bruno, V., Montgomery, R.R. and Zhao, H. (2013). Differential expression analysis for paired RNA-seq data. *BMC Bioinformatics* **14**, 110.
- Clyde, M. (1999). Bayesian model averaging and model search strategies. In J., Bernardo, J., Berger, A., Dawid, & A., Smith (eds.), *Bayesian Statistics 6*. Proceedings of the Sixth Valencia International Meeting. Oxford: Oxford University Press, pp. 157–185.
- Cui, W. and George, E.I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138**, 888–900.
- Del Moral, P., Doucet, A. and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B* **68**, 411–436.

- Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* **12**, 27–36.
- Do, K., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Applied Statistics* **54**, 627–644.
- Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.
- Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.
- Flutre, T., Wen, X., Pritchard, J. and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics* **9**, e1003486.
- Frühwirth-Schnatter, S. and Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In T. Kneib & G. Tutz (eds.), *Statistical Modelling and Regression Structures*. Heidelberg: Physica-Verlag, pp. 111–132.
- Fusi, N., Stegle, O. and Lawrence, N.D. (2012). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology* **8**, e1002330.
- Gelfond, J.A., Ibrahim, J.G. and Zou, F. (2007). Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics* **63**, 1108–1116.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515–534.
- George, E. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.
- George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.
- George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica sinica* 339–373.
- Glaus, P., Honkela, A. and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721–1728.
- Gottardo, R., Raftery, A.E., Yeung, K.Y. and Bumgarner, R.E. (2006a). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–18.
- Gottardo, R., Raftery, A.E., Yeung, K.Y. and Bumgarner, R.E. (2006b). Quality control and robust estimation for cDNA microarrays with replicates. *Journal of the American Statistical Association* **101**, 30–40.
- Griffin, J.E. and Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.
- Griffin, J.E. and Brown, P.J. (2012). Structuring shrinkage: Some correlated priors for regression. *Biometrika* **99**, 481–487.
- Gu, J., Wang, X., Halakivi-Clarke, L., Clarke, R. and Xuan, J. (2014). BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-seq data. *BMC Bioinformatics* **15**, S6.
- Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics* **5**, 1780–1815.
- Gupta, M. and Ibrahim, J.G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102**, 867–880.
- Hans, C. (2009). Bayesian LASSO regression. *Biometrika* **96**, 835–845.
- Hans, C. (2010). Model uncertainty and variable selection in Bayesian LASSO regression. *Statistics and Computing* **20**, 221–229.
- Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search for ‘large p’ regression. *Journal of the American Statistical Association* **102**, 507–516.

- Hardcastle, T.J. and Kelly, K.A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422.
- Holmes, C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.
- Ibrahim, J.G., Chen, M.-H. and Gray, R.J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97**, 88–99.
- Ishwaran, H. and Rao, J. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.
- Ishwaran, H. and Rao, J. (2005a). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association* **100**, 438–455.
- Ishwaran, H. and Rao, J. (2005b). Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730–773.
- Jasra, A., Stephens, D.A. and Holmes, C.C. (2007). Population-based reversible jump Markov chain Monte Carlo. *Biometrika* **94**, 787–807.
- Ji, C. and Schmidler, S.C. (2013). Adaptive markov chain Monte Carlo for Bayesian variable selection. *Journal of Computational and Graphical Statistics* **22**, 708–728.
- Jia, Z. and Xu, S. (2007). Mapping quantitative trait loci for expression abundance. *Genetics* **176**, 611–623.
- Jiang, B. and Liu, J.S. (2015). Bayesian partition models for identifying expression quantitative trait loci. *Journal of the American Statistical Association* **110**, 1350–1361.
- Katz, Y., Wang, E.T., Airoldi, E.M. and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009.
- Katzfuss, M., Neudecker, A., Anders, S. and Gagneur, J. (2014). Preprint, arXiv:1410.4827.
- Kendziorski, C., Chen, M., Yuan, M., Lan, H. and Attie, A.D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**, 19–27.
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36.
- Kim, S., Tadesse, M.G. and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**, 877–893.
- Knowles, D. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Annals of Applied Statistics* **5**, 1534–1552.
- Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* **11**, 313–322.
- Krishna, A., Bondell, H.D. and Ghosh, S.K. (2009). Bayesian variable selection using an adaptive powered correlation prior. *Journal of Statistical Planning and Inference* **139**, 2665–2674.
- Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā, Series B* 65–81.
- Kwon, D., Landi, M.T., Vannucci, M., Issaq, H.J., Prieto, D. and Pfeiffer, R.M. (2011). An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics & Data Analysis* **55**, 2807–2818.
- Lamnisos, D., Griffin, J.E. and Steel, M.F. (2009). Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics* **18**, 592–612.
- Lee, J., Ji, Y., Liang, S., Cai, G. and Müller, P. (2015). Bayesian hierarchical model for differential gene expression using RNA-seq data. *Statistics in Biosciences* **7**, 48–67.
- Lee, K., Sha, N., Dougherty, E., Vannucci, M. and Mallick, B. (2003). Gene selection: A Bayesian variable selection approach. *Bioinformatics* **19**, 90–97.

- Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M.G., Haag, J.D., Gould, M.N., Stewart, R.M. and Kendziorski, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043.
- León-Novelo, L.G., McIntyre, L.M., Fear, J.M. and Graze, R.M. (2014). A flexible Bayesian method for detecting allelic imbalance in RNA-seq data. *BMC Genomics* **15**, 920.
- Lewin, A., Richardson, S., Marshall, C., Glazier, A. and Aitman, T. (2006). Bayesian modelling of differential gene expression. *Biometrics* **62**, 1–9.
- Lewin, A., Saadi, H., Peters, J.E., Moreno-Moral, A., Lee, J.C., Smith, K.G., Petretto, E., Bottolo, L. and Richardson, S. (2015). MT-HESS: An efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics* **32**, 523–532.
- Ley, E. and Steel, M.F. (2012). Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* **171**, 251–266.
- Li, J., Jiang, H. and Wong, W.H. (2010). Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biology* **11**, R50.
- Liang, F., Paulo, R., Molina, G., Clyde, M.A. and Berger, J.O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423.
- Liang, F. and Wong, W.H. (2000). Evolutionary Monte Carlo: Applications to C_p model sampling and change point problem. *Statistica Sinica* **10**, 317–342.
- Lin, Y., Reynolds, P. and Feingold, E. (2003). An empirical Bayesian method for differential expression studies using one-channel microarray data. *Statistical Applications in Genetics and Molecular Biology* **2**, 8.
- Liu, F., Wang, C. and Liu, P. (2015). A semi-parametric Bayesian approach for differential expression analysis of RNA-seq data. *Journal of Agricultural, Biological, and Environmental Statistics* **20**, 555–576.
- Lönnstedt, I. and Speed, T. (2003). Replicated microarray data. *Statistica Sinica* **12**, 31–46.
- Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. and West, M. (2006). Sparse statistical modelling in gene expression genomics. In K.-A., Do, P., Müller, & M., Vannucci (eds.), *Bayesian Inference for Gene Expression and Proteomics*. Cambridge: Cambridge University Press, pp. 155–176.
- Lykou, A. and Ntzoufras, I. (2013). On Bayesian LASSO variable selection and the specification of the shrinkage parameter. *Statistics and Computing* **23**, 361–390.
- Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.
- Marett, L., Sibbesen, J.A. and Krogh, A. (2014). Bayesian transcriptome assembly. *Genome Biology* **15**, 501.
- Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**, 1509–1517.
- Maruyama, Y. and George, E.I. (2011). Fully Bayes factors with a generalized g-prior. *Annals of Statistics* **39**, 2740–2765.
- Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.
- Monni, S. and Tadesse, M.G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis* **4**, 413–436.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621.

- Moser, G., Lee, S.H., Hayes, B.J., Goddard, M.E., Wray, N.R. and Visscher, P.M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genetics* **11**, e1004969.
- Müller, P., Parmigiani, G. and Rice, K. (2007). FDR and Bayesian multiple comparison rules. In J.M., Bernardo, M.J., Bayarri, J.O., Berger, A.P., Dawid, D., Heckerman, A.F.M., Smith & M., West (eds.), *Bayesian Statistics 8*. Oxford: Oxford University Press, pp. 349–370.
- Mutshinda, C.M. and Sillanpää, M.J. (2010). Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* **186**, 1067–1075.
- Mutshinda, C.M. and Sillanpää, M.J. (2012). A decision rule for quantitative trait locus detection under the extended Bayesian LASSO model. *Genetics* **192**, 1483–1491.
- Nariai, N., Kojima, K., Mimori, T., Kawai, Y. and Nagasaki, M. (2016). A Bayesian approach for estimating allele-specific expression from RNA-Seq data with diploid genomes. *BMC Genomics* **17**, 2.
- Newton, M., Kendziorski, C., Richmond, C., Blattner, F. and Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.
- Newton, M., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5**, 155–176.
- Nott, D.J. and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics* **13**, 141–157.
- Nott, D.J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92**, 747–763.
- Nott, D.J. and Leonte, D. (2004). Sampling schemes for Bayesian variable selection in generalized linear models. *Journal of Computational and Graphical Statistics* **13**, 362–382.
- O'Hara, R.B. and Sillanpää, M.J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis* **4**, 85–117.
- Papastamoulis, P. and Rattray, M. (2017). Bayesian estimation of differential transcript usage from RNA-seq data. *Statistical Applications in Genetics and Molecular Biology* **16**, 387.
- Park, T. and Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* **103**, 681–686.
- Parmigiani, G., Garrett, E., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society, Series B* **64**, 717–736.
- Petretto, E., Bottolo, L., Langley, S.R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T.J., Cook, S.A., et al. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology* **6**, e1000737.
- Polson, N.G. and Scott, J.G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* **9**, 501–538.
- Polson, N.G. and Scott, J.G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society, Series B* **74**, 287–311.
- Polson, N.G. and Scott, J.G. (2013). Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika* **100**, 459–471.
- Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108**, 1339–1349.
- Reilly, C., C., W. and Rutherford, M. (2003). A method for normalizing microarrays using genes that are not differentially expressed. *Journal of the American Statistical Association* **98**, 868–878.
- Richardson, S., Bottolo, L. and Rosenthal, J.S. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics* **9**, 539–569.

- Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B* **59**, 731–792.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47.
- Roberts, G.O. and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367.
- Robinson, M.D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25.
- Ruffieux, H., Davison, A.C., Hager, J. and Irincheeva, I. (2017). Efficient inference for genetic association studies with multiple outcomes. *Biostatistics* **18**, 618–636.
- Schadt, E., Li, C., Su, C. and Wong W. (2000). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry* **80**, 192–202.
- Schäfer, C. and Chopin, N. (2013). Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing* **23**, 163–184.
- Scott, J.G. and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* **38**, 2587–2619.
- Scott-Boyer, M.P., Imholte, G.C., Tayeb, A., Labbe, A., Deschepper, C.F. and Gottardo, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology* **11**.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.
- Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C. and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60**, 812–819.
- Sillanpää, M., Pikkukookana, P., Abrahamsson, S., Knürr, T., Fries, A., Lerceteau, E., Waldmann, P. and García-Gil, M. (2012). Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling. *Heredity* **108**, 134.
- Sillanpää, M.J. and Bhattacharjee, M. (2004). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**, 427–439.
- Sillanpää, M.J. and Bhattacharjee, M. (2006). Association mapping of complex trait loci with context-dependent effects and unknown context variable. *Genetics* **174**, 1597–1611.
- Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.
- Stegle, O., Parts, L., Durbin, R. and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology* **6**, e1000770.
- Stingo, F.C., Chen, Y.A., Tadesse, M.G. and Vannucci, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Annals of Applied Statistics* **5**, 1978–2002.
- Stingo, F.C. and Vannucci, M. (2010). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27**, 495–501.
- Tadesse, M., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.
- Titsias, M.K. and Yau, C. (2017). The Hamming ball sampler. *Journal of the American Statistical Association* **112**, 1–14.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562.

- Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L.J., Richardson, S. and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology* **12**, R13.
- Van De Wiel, M.A., Leday, G.G., Pardo, L., Rue, H., Van Der Vaart, A.W. and Van Wieringen, W.N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14**, 113–128.
- Vavoulis, D.V., Francescatto, M., Heutink, P. and Gough, J. (2015). DGEclust: Differential expression analysis of clustered count data. *Genome Biology* **16**, 39.
- Walker, S., Damine, P., Laud, P. and Smith, A. (1999). Bayesian nonparametric inference for distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–527.
- Wang, H., Zhang, Y.-M., Li, X., Masinde, G.L., Mohan, S., Baylink, D.J. and Xu, S. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**, 465–480.
- Wang, Z., Wang, J., Wu, C. and Deng, M. (2015). Estimation of isoform expression in RNA-seq data using a hierarchical Bayesian model. *Journal of Bioinformatics and Computational Biology* **13**, 1542001.
- Wen, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70**, 73–83.
- Wen, X. and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene-environment interactions. *Annals of Applied Statistics* **8**, 176.
- West, M. (2003). Bayesian factor regression models in the ‘large p, small n’ paradigm. In J., Bernardo, M., Bayarri, J., Berger, A., Dawid, D., Heckerman, A., Smith, & M., West (eds.), *Bayesian Statistics 7*. Proceedings of the Seventh Valencia International Meeting. Oxford: Oxford University Press, pp. 733–742.
- Wu, H., Wang, C. and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243.
- Xu, C., Li, Z. and Xu, S. (2005). Joint mapping of quantitative trait loci for multiple binary characters. *Genetics* **169**, 1045–1059.
- Xu, C., Wang, X., Li, Z. and Xu, S. (2009). Mapping QTL for multiple traits using Bayesian statistics. *Genetics Research* **91**, 23–37.
- Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.
- Yang, R. and Xu, S. (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **176**, 1169–1185.
- Yi, N. (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**, 967–975.
- Yi, N. and Banerjee, S. (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* **181**, 1101–1113.
- Yi, N., George, V. and Allison, D.B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**, 1129–1138.
- Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D. and Yandell, B.S. (2007). An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**, 1865–1877.
- Yi, N. and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**, 1045–1055.
- Yi, N., Yandell, B.S., Churchill, G.A., Allison, D.B., Eisen, E.J. and Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**, 1333–1344.

- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*. Amsterdam: North-Holland, pp. 233–243.
- Zhang, W., Zhu, J., Schadt, E.E. and Liu, J.S. (2010). A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computational Biology* **6**, e1000642.
- Zhao, L., Wu, W., Feng, D., Jiang, H. and Nguyen, X. (2018). Bayesian analysis of RNA-seq data using a family of negative binomial models. *Bayesian Analysis* **13**, 411–436.
- Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9**, e1003264.

31

Modelling Gene Expression Dynamics with Gaussian Process Inference

Magnus Rattray,¹ Jing Yang,¹ Sumon Ahmed,¹ and Alexis Boukouvalas²

¹Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine & Health, University of Manchester, UK

²PROWLER.io, Cambridge, UK

Abstract

Gaussian process (GP) inference provides a flexible nonparametric probabilistic modelling framework. We present examples of GP inference applied to time series gene expression data and for single-cell high-dimensional ‘snapshot’ expression data. We provide a brief overview of GP inference and show how GPs can be used to identify dynamic genes, infer degradation rates, model replicated and clustered time series, model stochastic single-cell dynamics, and model perturbations or branching in time series data. In the case of single-cell expression data we present a scalable implementation of the Gaussian process latent variable model, which can be used for dimensionality reduction and pseudo-time inference from single-cell RNA-sequencing data. We also present a recent approach to inference of branching dynamics in single-cell data. To scale up inference in these applications we use sparse variational Bayesian inference algorithms to deal with large matrix inversions and intractable likelihood functions.

31.1 Introduction

Gaussian process (GP) regression was initially introduced as a nonparametric model of spatial data, but GPs are now widely applied for multivariate regression, classification, reinforcement learning and dimensionality reduction (Rasmussen and Williams, 2006). Their popularity stems from a useful combination of modelling flexibility with inferential tractability. Modelling flexibility is enabled by the covariance function, which can capture rich statistical relationships between data points. Inferential tractability comes from the ability to integrate over Gaussian distributions in high dimensions. GPs provide useful latent variables in non-Gaussian models, with tractable inference for large-scale problems achieved through recent developments in approximate inference. In this chapter we focus on the application of GP methods to model gene expression dynamics both from time series experiments and single-cell snapshot assays. We consider a range of modelling scenarios, from the application of standard GP regression approaches for identifying dynamic or periodic genes, to more problem-specific GP models that incorporate mRNA degradation, clustering or branching dynamics.

31.1.1 Covariance Function

A GP describes a distribution over functions. Functions evaluated at any finite set of points will follow a multivariate Gaussian distribution. Initially we will consider a one-dimensional function $f(x)$ where x represents time,

$$f \sim \mathcal{GP}(\mu, k),$$

in which $\mu = \mu(x)$ is the mean function and $k = k(x, x')$ is the covariance function, often referred to as the kernel function. The mean function is simply the mean of function values at any particular time x ,

$$\mu(x) = E[f(x)],$$

while the covariance function is the covariance of function values at any two times x and x' ,

$$k(x, x') = E[f(x)f(x')] - E[f(x)]E[f(x')].$$

The covariance function plays a more fundamental role in GP modelling than the mean function, and in many cases the mean function can be set to zero.

The covariance function comes from some parametric family which determines typical properties of the samples $f(x)$. For example, a popular choice for regression is the squared exponential covariance function

$$k(x, x') = \alpha \exp\left(\frac{-(x-x')^2}{2l}\right). \quad (31.1)$$

Figure 31.1(a) shows a function sampled from a GP with this covariance function, which is infinitely differentiable and smooth. This choice is popular in regression over data that is thought to come from a smooth underlying model; for example, bulk gene expression time course data is averaged over millions of cells and may therefore be expected to change smoothly in time. The covariance function has two parameters. The amplitude α determines the scale of the functions (i.e. the marginal variance of the function at a specific value of x). The length-scale (or in our case time-scale) l determines how frequently the function crosses the zero line on average. As $l \rightarrow \infty$ samples approach straight lines while as $l \rightarrow 0$ samples approach white noise, which is a completely uncorrelated Gaussian process.

Alternatively, the Ornstein–Uhlenbeck (OU) process covariance function is given by

$$k(x, x') = \alpha \exp\left(\frac{-|x-x'|}{l}\right), \quad (31.2)$$

with an L1 norm replacing the L2 norm in the exponent. Figure 31.1(b) shows a function sampled from a GP with this covariance function. Samples are continuous but they are now rough and non-differentiable. Dynamically this can be thought of as a process with finite velocities but infinite acceleration. We will see later that the OU process can model single-cell gene expression data, where intrinsic fluctuations are not averaged away as they are in bulk gene expression data. The two covariance functions above are limiting cases of a more general Matérn covariance function which can be used to vary the roughness of the samples (Rasmussen and Williams, 2006).

There is much interest in periodic oscillations in biological systems, with circadian rhythms, the cell cycle and various ultradian rhythmic processes the subject of intensive research. GPs provide very natural models for periodic functions. Figure 31.1(c) shows samples from a covariance function that generates smooth periodic functions (MacKay, 1998), while Figure 31.1(d) shows samples from a quasi-periodic OU process (Westerman *et al.*, 2009; Phillips *et al.*, 2017).

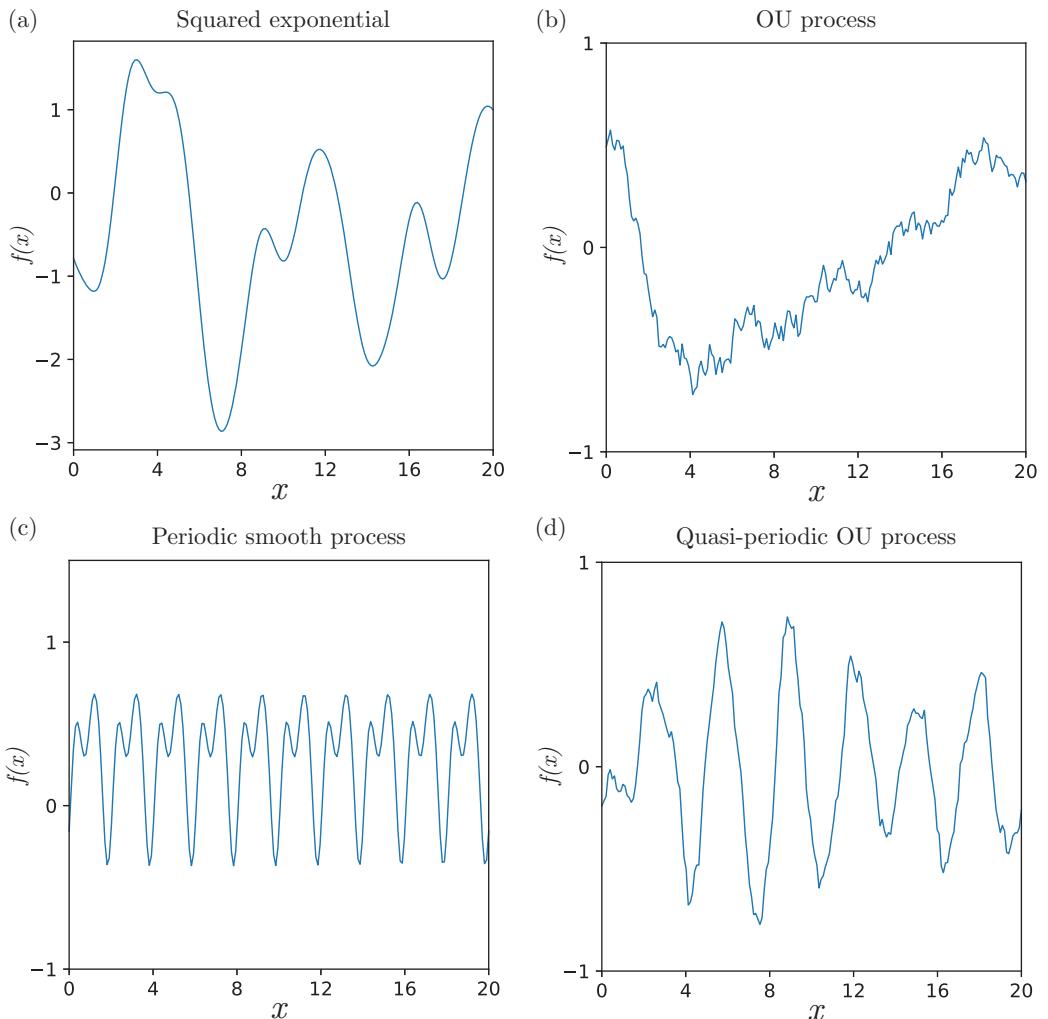


Figure 31.1 Samples from four classes of covariance function: (a) squared exponential; (b) Ornstein–Uhlenbeck process; (c) periodic smooth process; (d) quasi-periodic OU process.

The functions in Figure 31.1 are all stationary, with covariance functions that depend only on the distance between time points $|x - x'|$. This stationarity assumption can break down in certain applications. For example, gene expression time course data may be collected after a perturbation leading to a rapid initial transient phase before settling down to a constant value asymptotically. Non-stationary alternatives have therefore been developed which can better model changes in amplitude or length-scale of gene expression data over time (see, for example, Heinonen *et al.*, 2014).

An appropriate candidate set of covariance functions can be chosen using application domain knowledge, while statistical model selection can be used to select the best one from this candidate set. For example, the roughness or periodicity properties of a system may be suggested from first-principles modelling (as shown in Section 31.3.1) or the experimental design may suggest a hierarchical data structure (as shown in Section 31.2.3). Statistical model selection can then be addressed using standard likelihood-based or Bayesian model selection strategies,

with recent methods to estimate out-of-sample prediction accuracy showing great promise (Vehtari *et al.*, 2017). A nice feature of GP models is that the sample paths can be analytically integrated out to obtain a marginal likelihood that depends on relatively few parameters (see Section 31.1.2). This is an attractive feature of GPs for both maximum likelihood and Bayesian integration approaches, since there are relatively few parameters to optimise or integrate over using numerical methods.

31.1.2 Inference

Given a finite set of noise-corrupted measurements at different times, we are interested in which underlying functions are most likely to have generated the observed data. If we assume that the covariance function is known then this is very easy to do with a GP, because we can condition and marginalise exactly with Gaussian distributions.

In the regression setting, we have a data set D with regressors $\mathbf{X} = \{x_n\}_{n=1}^N$ and corresponding real-valued targets $\mathbf{Y} = \{y_n\}_{n=1}^N$. In the case of time course data the regressors are an ordered vector such that $x_n \geq x_{n-1}$, but there is no restriction on the spacing since GPs operate over a continuous domain. We allow the case $x_n = x_{n-1}$ since that provides a simple way to incorporate replicates. We assume that measurement noise in \mathbf{Y} , denoted by ϵ , is independently Gaussian distributed $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ and the underlying model for \mathbf{Y} as a function of \mathbf{X} is $f(\cdot)$, so that

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon, \quad (31.3)$$

where $f(\mathbf{X})$ represents a sample from a GP evaluated at all the times in the vector \mathbf{X} . Our prior modelling assumption is that the function f is drawn from a GP prior with zero mean and covariance function $k(x, x')$. The probability of the data \mathbf{Y} under the model is obtained by integrating out the function $f(\mathbf{X})$,

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}) &= \int \mathcal{N}(\mathbf{Y}|\mathbf{f}, \sigma^2 \mathbf{I}) \mathcal{N}(\mathbf{f}|\mathbf{0}, K(\mathbf{X}, \mathbf{X})) d\mathbf{f} \\ &= \mathcal{N}(\mathbf{Y}|\mathbf{0}, K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}), \end{aligned} \quad (31.4)$$

where we have written $\mathbf{f} = f(\mathbf{X})$ and $K(\mathbf{X}, \mathbf{X})$ is the $N \times N$ covariance matrix with elements $k(x_n, x_m)$ determined by the covariance function.

A typical regression analysis will be focused on a new input x_* and its prediction f_* . Based upon Gaussian properties (Rasmussen and Williams, 2006) the posterior distribution of f_* given data \mathbf{Y} is $f_* | \mathbf{Y} \sim \mathcal{N}(\mu_*, C_*)$, with

$$\begin{aligned} \mu_* &= K(\mathbf{X}, x_*)^T (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{Y}, \\ C_* &= K(x_*, x_*) - K(\mathbf{X}, x_*)^T (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} K(\mathbf{X}, x_*). \end{aligned}$$

This is the posterior prediction of the function f at a specific time point x_* but is easily generalised to the full functional posterior distribution, showing that the posterior function is another GP (Rasmussen and Williams, 2006). We see above that the mean prediction is a weighted sum over data with weights larger for nearby points in a manner determined by the covariance function. The posterior covariance captures our uncertainty in the inference of f_* which will typically be reduced as we incorporate more data. Figure 31.2 shows an example of regression with a squared exponential covariance function. In Figure 31.2(a) we show some samples from the prior, and in Figure 31.2(b) the posterior distribution is fitted to four observations. In this case, the data are observed without noise ($\sigma^2 = 0$) but we still have uncertainty because many functions are consistent with the data. The posterior shows which functions are likely given the data and our prior belief in the underlying function. The prior expects functions to be smooth and

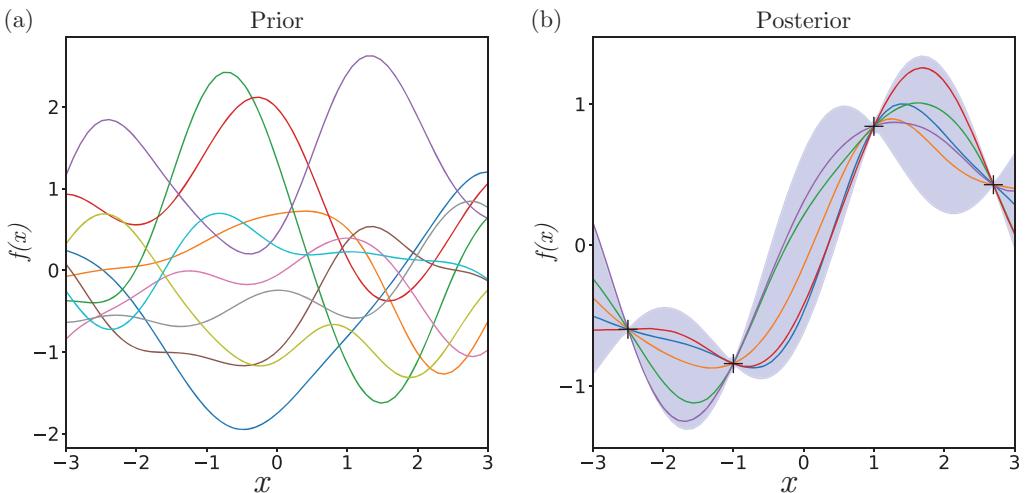


Figure 31.2 (a) Ten samples drawn from a GP with a squared exponential covariance function with hyperparameters $\alpha = 1$ and $l = 1$. (b) Ten samples from the posterior distribution after observing four data points without any observation noise ($\sigma^2 = 0$). The functions are constrained to pass through the data but the posterior distribution captures our uncertainty away from the data. The shading shows two standard deviations of posterior distribution at each time.

not to change very rapidly, and therefore our uncertainty increases gradually as we move away from the data.

We often refer to the parameters of the covariance function (including the noise variance) as hyperparameters, since the function $f(x)$ itself can be considered a functional parameter of the model. The log likelihood of the hyperparameters $L(\theta)$ is the logged probability of the data in equation (31.4),

$$\begin{aligned} L(\theta) &= \log \mathcal{N}(\mathbf{Y}|\mathbf{0}, K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}) \\ &= -\frac{N}{2} \log(2\pi) - \frac{1}{2} \log \det(\sigma^2 \mathbf{I} + \mathbf{K}) - \frac{1}{2} \mathbf{Y}^T (\sigma^2 \mathbf{I} + \mathbf{K})^{-1} \mathbf{Y}, \end{aligned} \quad (31.5)$$

where we have written $\mathbf{K} = K(\mathbf{X}, \mathbf{X})$. This likelihood function has a complex form and may be multimodal so that hyperparameter inference by either maximum likelihood or Bayesian inference requires numerical optimisation or integration methods. Gradient-based methods for optimisation (e.g. quasi-Newton or conjugate gradient) or Bayesian inference (e.g. Hamiltonian Monte Carlo (HMC)) are implemented in a number of popular GP inference software packages. In this chapter we have used the python packages GPy (<https://sheffieldml.github.io/GPy/>) and GPflow (<https://github.com/GPflow>), as well as the DEtime R package (<https://github.com/ManchesterBioinference/DEtime>).

31.2 Applications to Bulk Time Series Expression Data

Most gene expression experiments involve measuring expression in bulk samples that typically contain many millions of cells, for example microarray or RNA-sequencing data derived from tissue, cell culture or whole organisms. In this case, averaging over many cells will remove the intrinsic stochasticity of gene expression within individual cells and we would expect time course data to follow a smooth trajectory, although measurements will include experimental

sources of variation that have to be taken into account. We typically model these sources as independent and identically distributed (i.i.d.) noise added to the GP function, although in some cases it is advantageous to use more structured models of variation across time series replicates (discussed in Section 31.2.3).

31.2.1 Identifying Differential Expression in Time

Kalaitzis and Lawrence (2011) introduce a simple approach to identify whether genes measured in a time course experiment show significant evidence of changes in time. A GP model is used to determine whether there is evidence for smooth temporal changes. By fitting the covariance function's length-scale parameter, one can determine the likelihood under a GP model and compare with the likelihood under a constant model, in which case all variation is modelled as white noise. Figure 31.3 shows this approach applied to one gene from a time series gene expression data set from Lewis *et al.* (2015). In Figure 31.3(a) we fit a GP model and in Figure 31.3(b) we show the fit of a constant model, with associated credible regions for the fitted model in each case. In this example the likelihood of the GP model is much higher, providing significant evidence for dynamics in this gene. Note that the likelihood here is for the model with the function $f(x)$ marginalised out (sometimes referred to as the marginal likelihood, equation (31.5)) and therefore it can reasonably be used for model selection as complexity of the function $f(x)$ is controlled for by Bayesian model averaging.

The likelihood ratio between the GP and white noise model provides a simple means to rank genes in terms of differential expression (DE) across time. To select a significance threshold one could therefore use a penalised likelihood approach (e.g. the BIC score) to penalise the more complex model or use a likelihood ratio test with an associated false discovery rate (FDR) threshold (see Phillips *et al.*, 2017, for an example of this FDR approach to discovering periodic genes). A fully Bayesian treatment would require a numerical approach to integrate over the hyperparameters.

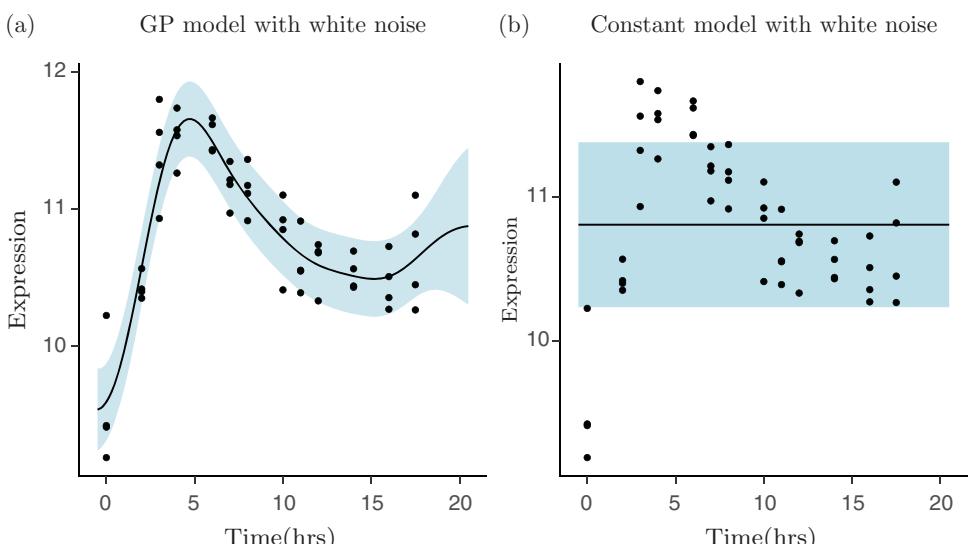


Figure 31.3 In (a) we apply GP regression to expression data from arabidopsis after infection by a plant pathogen (Lewis *et al.*, 2015) for gene *CATMA1a00010* and we compare it to the fit to a constant model shown in (b). For this gene there is strong evidence of differential expression across time.

As well as looking at differential expression, GPs have also been used to investigate changes in splicing over time (Topa and Honkela, 2016) and Huang and Sanguinetti (2016) combine a GP model of temporal change in transcript abundance with a transcript inference algorithm modelling RNA-sequencing read data.

31.2.2 Identifying Changes between Two Time Course Experiments

In some cases a time course experiment is done over two different conditions, for example to compare between a wildtype and mutant strain of an organism. In such a ‘two-sample’ experiment we may be interested in determining whether two time series are different. The above approach can then easily be extended to compare models in which the two time series data sets come from the same or different underlying GP functions. However, it is often more informative to also identify the time periods where the two samples differ. Stegle *et al.* (2010) introduced a GP-based method for identifying regions of DE between two samples based on a mixture of two GP regression models in which data at each time point can be assigned to the same function or two different functions. A simpler strategy, based on fitting two different GP functions to each time series and developing test statistics to identify regions of DE, was introduced by Heinonen *et al.* (2014) who also improved performance through use of a non-stationary covariance function.

As an alternative approach, Yang *et al.* (2016) develop a method to detect the first point where two time series begin to differ. This is done through a covariance function for a branching process. First we write down the joint covariance function for two GP functions $f(x) \sim \mathcal{GP}(0, k)$ and $g(x) \sim \mathcal{GP}(0, k)$ which are constrained to cross at a specific time $x = x_p$ so that $f(x_p) = g(x_p)$,

$$\begin{Bmatrix} K_{ff} & K_{fg} \\ K_{gf} & K_{gg} \end{Bmatrix} = \begin{Bmatrix} K(\mathbf{X}, \mathbf{X}) & \frac{K(\mathbf{X}, x_p)K(\mathbf{X}, x_p)^T}{k(x_p, x_p)} \\ \frac{K(\mathbf{X}, x_p)K(\mathbf{X}, x_p)^T}{k(x_p, x_p)} & K(\mathbf{X}, \mathbf{X}) \end{Bmatrix}. \quad (31.6)$$

Consider a control time course data set \mathbf{Y}^c and a perturbed time course data set \mathbf{Y}^p . Before x_p we model these two data sets as noise-corrupted versions of the same underlying mean function $f(x) \sim \mathcal{GP}(0, k)$,

$$\begin{aligned} y^c(x_n) &\sim \mathcal{N}(f(x_n), \sigma^2), \\ y^p(x_n) &\sim \mathcal{N}(f(x_n), \sigma^2), \quad \text{for } x_n \leq x_p. \end{aligned}$$

After x_p the mean function for y^c stays intact while the mean function for y^p changes to follow $g(x)$,

$$\begin{aligned} y^c(x_n) &\sim \mathcal{N}(f(x_n), \sigma^2), \\ y^p(x_n) &\sim \mathcal{N}(g(x_n), \sigma^2), \quad \text{for } x_n > x_p, \end{aligned}$$

where f and g are constrained to cross at x_p and therefore follow the GP given by the covariance in equation (31.6).

The model is fitted using the standard regression approach described in Section 31.1.2. The perturbation time x_p is a hyperparameter of the covariance function for this model along with the length-scale and amplitude of the functions f and g . The length-scale and amplitude can be

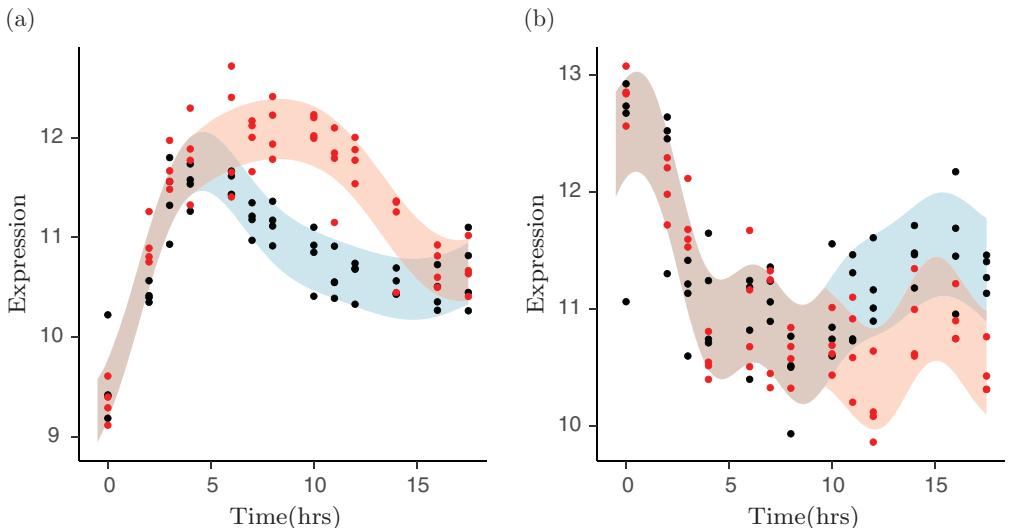


Figure 31.4 Perturbation time inference using the gene expression time series data from Lewis *et al.* (2015) which was analysed using a branching GP model by Yang *et al.* (2016). Arabidopsis gene expression dynamics was studied after infection by a wildtype pathogen (black points) and compared with infection by a mutated pathogen (red points). In (a) we show an example of an early DE gene *CATMA1a00010* from Figure 31.3, while (b) shows a late DE gene *CATMA1a00060*. The GP model fit is shown with x_p set at the mean posterior estimate in each case.

reasonably estimated by fitting independent regression models to the data from the two conditions with shared length-scale and amplitude parameters estimated by maximum likelihood. This then leaves us with the inference problem for x_p . As this is a one-dimensional problem we can estimate the posterior by a simple histogram approach,

$$p(x_p | \mathbf{Y}^c, \mathbf{Y}^p) \simeq \frac{p(\mathbf{Y}^c, \mathbf{Y}^p | x_p)}{\sum_{x_p=x_{\min}}^{x_p=x_{\max}} p(\mathbf{Y}^c, \mathbf{Y}^p | x_p)}, \quad (31.7)$$

which avoids the need to use complex optimisation or integration schemes.

Figure 31.4 shows an example application of this model to a two-sample gene expression data set from an experiment with arabidopsis (Lewis *et al.*, 2015). The experiment involves infecting the plant with a pathogen to produce the control time series and infecting the plant with a mutated strain of the pathogen to produce the perturbation time series. The model identifies the gene in Figure 31.4(a) as early DE (posterior median $x_p = 3.9$ hours) while the gene in Figure 31.4(b) diverges later in the time series (posterior median $x_p = 9.6$ hours). Yang *et al.* (2016) use the model to rank genes in terms of their perturbation time, to help understand the sequence of events underlying the immune response to infection.

As well as returning the posterior over the perturbation time, the above model can also provide evidence for whether the two time series differ (after any time) or are statistically indistinguishable. Figure 31.5 shows examples of data with a perturbation in the middle of the time course (a) and with no perturbation (b). The posterior distribution of the perturbation time is shown on the top in each case. If the perturbation time is inferred closer to the start then it is more likely that the two time course profiles are truly distinct, whereas a perturbation time inferred at the end of the time course indicates the two time profiles are very similar to each other and less likely to differ. We can make a decision over whether or not there is a bifurcation

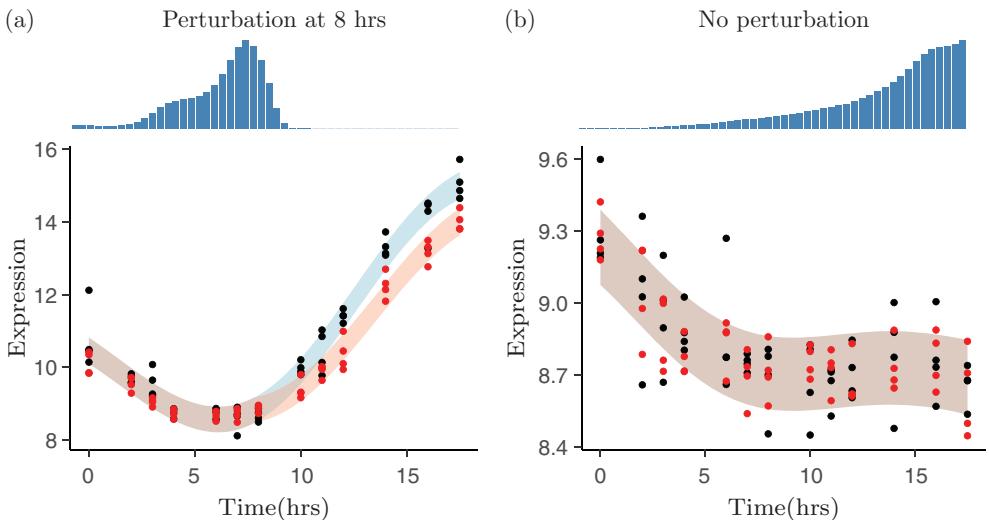


Figure 31.5 Two examples of the posterior distribution of the perturbation time (upper) and GP regression model fit based upon the maximum *a posteriori* estimate of the perturbation time (lower). In (a) we show data (gene CATMA1a00045 from Lewis *et al.* (2015)) with a perturbation introduced halfway along the time range, while in (b) we show data without any perturbation (gene CATMA1a00180). When there is no perturbation then the posterior tends to increase towards the end of the time range. A Bayes factor (equation 31.8) can be used to determine support for whether the data exhibits any bifurcations.

by considering the Bayes factor between a model with or without a perturbation (Boukouvalas *et al.*, 2018). The logged Bayes factor between a model with or without branching is given by

$$r_g = \log \frac{p(0 < x_p < x_{\max} | \mathbf{Y}^c, \mathbf{Y}^p)}{p(x_p = x_{\max} | \mathbf{Y}^c, \mathbf{Y}^p)} \\ = \log \left[\frac{1}{N_b} \sum_{x_p=x_{\min}}^{x_p=x_{\max}} p(\mathbf{Y}^c, \mathbf{Y}^p | x_p) \right] - \log [p(\mathbf{Y}^c, \mathbf{Y}^p | x_{\max})], \quad (31.8)$$

where N_b is the number of bins in the histogram approximation to the posterior and setting $x_p = x_{\max}$ is equivalent to having no perturbation at all. Here, we are assuming equal prior probability for having a perturbation (at any time before x_{\max} with equal probability) or having no perturbation. We can see from this expression that if the height of the posterior at the final time is greater than the average of the posterior over all other times, as in Figure 31.5(b), then the probability of a bifurcation event under the model is less than 0.5. In this example there is very strong evidence for a perturbation in Figure 31.5(a) ($r_g = 31.73$) and quite strong evidence for no perturbation in Figure 31.5(b) ($r_g = -1.33$).

31.2.3 Hierarchical Models of Replicates and Clusters

In some cases the assumption of biological or technical variation as i.i.d. white noise is not justified and can lead to sub-optimal modelling. For example, biological time course replicates may be collected at different times, leading to large between-replicate variation and an associated batch effect. Similarly, different genes within a cluster should not be treated as white noise around the cluster mean as each gene will have its own different underlying profile within the cluster. Hensman *et al.* (2013) introduced a hierarchical Gaussian process to capture both of

these effects. In the case of time course replicates we model the mean underlying profile of gene n shared by all replicates using a GP. We then model the replicates r as samples from another GP distributed around the shared profile,

$$\begin{aligned} g_n &\sim \mathcal{GP}(0, k_g), \\ f_{nr} &\sim \mathcal{GP}(g_n, k_f). \end{aligned}$$

This simple model allows for variation in the profile across replicates and is also a powerful approach to transfer information between replicates collected at different times. For example, Hensman *et al.* (2013) showed how eight replicated developmental time course data sets could be jointly modelled, with different replicates covering different stages of development (Kalinka *et al.*, 2010). By adding an additional layer in the hierarchy one can also model the shared profile of a cluster c ,

$$\begin{aligned} h_c &\sim \mathcal{GP}(0, k_h), \\ g_{nc} &\sim \mathcal{GP}(h_c, k_g), \\ f_{ncr} &\sim \mathcal{GP}(g_{nc}, k_f), \end{aligned}$$

where there is now a GP for each cluster c , each gene n within the cluster and each replicate r . This hierarchical approach can then be combined with a model-based clustering method to infer the cluster assignments of genes. An efficient Dirichlet process mixture model implementation was developed by Hensman *et al.* (2015) using variational inference techniques and was applied to cluster circadian expression data in Gossan *et al.* (2013) using periodic covariance functions.

31.2.4 Differential Equation Models of Production and Degradation

Any linear transformation of a GP is another GP. This holds true for integral solutions of linear ordinary differential equations (ODEs) which contain GP functions and allows GPs to be used within simple ODE models. For example, consider a model of mRNA $m(x)$ being produced with transcription rate $f(x)$ at time x . We can write

$$\frac{dm}{dx} = f(x) - \delta m,$$

where δ is the mRNA degradation rate. The solution to this linear ODE is

$$m(x) = m(0)e^{-\delta x} + \int_0^x e^{\delta(u-x)}f(u)du.$$

We see that the mRNA concentration is a weighted integral of the production rate. Figure 31.6 shows two scenarios where, for the same transcription rate function, the mRNA time profiles can be very different. In Figure 31.6(a) the degradation rate is relatively high and $m(x)$ and $f(x)$ therefore have similar shapes. In Figure 31.6(b) we see that for low degradation rate, $m(x)$ will integrate $f(x)$ and has a qualitatively different profile with a much later peak in expression.

We can place a GP prior on $f \sim \mathcal{GP}(0, k_f)$, and since $m(x)$ is a linear functional of $f(x)$ they form a two-dimensional GP. If we choose the squared exponential covariance for k_f (recall equation (31.1)) then the covariance for $[f, m]$ can be worked out in closed form (Lawrence *et al.*, 2007). Figure 31.6 shows two examples of samples drawn from this GP, with a different degradation rate hyperparameter used in each case.

This model can be adapted to a number of different scenarios. Barenco *et al.* (2006) showed how the above ODE could be used to model multiple targets of a transcription factor protein, in order to infer its activity and discover other target genes. Lawrence *et al.* (2007) then applied GP inference to the same task which avoided Markov chain Monte Carlo (MCMC) sampling of the

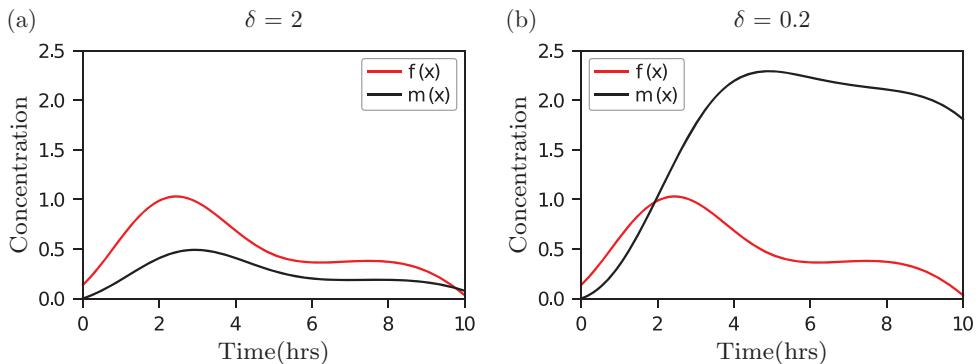


Figure 31.6 Comparison of mRNA profile $m(x)$ generated with the same transcription rate function $f(x)$ but using two different degradation rates. With high degradation ((a), $\delta = 2$) the mRNA and transcription rates have similar shape, but with low degradation ((b), $\delta = 0.2$) the mRNA level peaks much later. Here, the transcription rate function is in fact a sample drawn from a GP and $m(x)$ and $f(x)$ are therefore joint samples from a bivariate GP.

latent function and therefore more computationally efficient inference. Honkela *et al.* (2010) used a similar approach to identify transcription factor targets using data from an embryonic development time course experiment, but they extended the model to an additional layer to account for protein production and degradation. That work was further extended to ODE models with a nonlinear dependence of target gene expression to multiple regulatory transcription factors, requiring the development of MCMC techniques for GP inference and hyperparameter estimation (Titsias *et al.*, 2012). By including a delay parameter in the model, Honkela *et al.* (2015) used a similar ODE model to identify genes with a significant delay between nascent mRNA production and mature mRNA accumulation through joint analysis of RNA-sequencing and Pol-II ChIP-sequencing time course data.

31.3 Modelling Single-Cell Data

Single cells present several new challenges for inference. Gene expression is intrinsically stochastic at the single-cell level and therefore trajectory data at single-cell resolution, available from live cell imaging experiments, has to be modelled with care. One useful approach to modelling this data is to apply a linear noise approximation, in which stochastic oscillations are modelled as a GP. Live cell imaging can only be done for one or two genes simultaneously and is not scalable to genome-wide studies. Single-cell genomics enables genome-wide expression profiling, but because the experiments are destructive it is impossible to profile the same cell at different times. However, inference can be used to try and uncover gene expression dynamics by inferring where each cell lies in some *pseudotemporal* ordering. GPs provide one approach to inferring such a pseudotemporal ordering, as well as being useful in more general dimensionality reduction of single-cell data.

31.3.1 Modelling Single-Cell Trajectory Data

The chemical master equation (CME) provides a description of stochastic biochemical reactions which can be used to model gene expression dynamics in single cells. The CME can be simulated using the Gillespie algorithm (Gillespie, 1977) and provides samples of the stochastic dynamics, but it is not usually possible to compute the likelihood of data sampled

from a CME and therefore inference with such models is difficult. The linear noise approximation (LNA) has been used as an approximation to the CME which is valid for sufficiently large numbers of molecules (Komorowski *et al.*, 2009; Fearnhead *et al.*, 2014). In the LNA the dynamics is approximated by a GP with a mean function determined by an ODE and the covariance described as a Gauss–Markov stochastic process with variance also determined by an ODE. If noise-corrupted data are collected at discrete times then inference can be carried out using a Kalman filter (Kalman, 1960; Jazwinski, 1970). The GP nature of the process also allows inference for data collected through applying any linear function to the process. For example, if microscopy data is collected over long periods in studies using a luciferase reporter then the data can be modelled as the integral of the underlying process and inference with the LNA remains tractable (Folia and Rattray, 2018).

Once the deterministic part of the dynamics has reached a fixed point then the GP follows a multivariate OU process (recall Figure 31.1) which is a stationary Gauss–Markov process. In systems with negative feedback this multivariate Gauss–Markov process can exhibit oscillations, even when the deterministic part has converged to a stable fixed point, and therefore oscillations can be induced by the stochasticity present in a single-cell system (Galla, 2009). For example, consider a linear damped oscillator which can be written as a Langevin equation (Westermark *et al.*, 2009),

$$\begin{aligned}\frac{dx}{dt} &= -\lambda x - \omega y + \xi_x, \\ \frac{dy}{dt} &= \omega x - \lambda y + \xi_y,\end{aligned}$$

where ξ_x and ξ_y are zero-mean white noise with covariance $k_\xi(t, t') = 2D\delta(t - t')$. In this case the variables x and y have the same covariance function,

$$k(t, t') = \frac{D}{\lambda} \exp(-\lambda|t - t'|) \cos(\omega|t - t'|).$$

Figure 31.1(d) shows a sample from a GP with this covariance function. The samples are rough and approximately periodic, but oscillations gradually shift in phase over time so that they are not precisely periodic. This could represent a regulatory network where gene x is repressed by gene y while gene y is activated by gene x (after normalising expression to have mean zero). Without stochasticity ($D = 0$) the system converges to a fixed point and the covariance is zero – the oscillations are only caused by finite molecule numbers in such a system and are not seen in a large system size limit.

Phillips *et al.* (2017) used GP inference to identify stochastic oscillations in single-cell microscopy data by fitting a GP model with the above covariance function and comparing it with a non-oscillatory OU process with covariance function given by equation (31.2). The Hes1 transcription factor is known to exhibit negative autoregulation with delay which can lead to oscillations (Monk, 2003) and under some conditions these oscillations only persist due to the presence of stochastic fluctuations in single cells (Galla, 2009). Using GP-based inference, Phillips *et al.* (2017) found that single cells were likely to exhibit stochastic oscillations of Hes1 expression while expression data from a constitutive promoter was never classified as an oscillator, showing that the GP-based approach has good specificity in this application.

31.3.2 Dimensionality Reduction and Pseudotime Inference

Single-cell RNA-sequencing experiments have genome-wide coverage and therefore produce high-dimensional data sets with substantial biological and technical variation. GPs can be used

for dimensionality reduction of multivariate data by treating the regressors \mathbf{X} as parameters (or latent variables) to be inferred along with the functions $f(\mathbf{X})$. Recall the GP model in equation (31.3). Consider a multivariate GP regression model for many data dimensions y_i , with $i = 1 \dots d$, each with their own GP function f_i ,

$$y_{in} = f_i(x_n) + \epsilon_{in}.$$

In the case of pseudotime inference $\mathbf{X} = [x_n]$ is a vector, but more generally it will live in some low-dimensional space into which we would like to project our data. We treat \mathbf{X} as a latent variable that has to be inferred along with the functions f_i and associated covariance hyperparameters. This is the Gaussian process latent variable model (GPLVM) which is a popular probabilistic approach for nonlinear dimensionality reduction (Lawrence, 2005; Titsias and Lawrence, 2010).

The log likelihood can be worked out similarly to standard GP regression in equation (31.5), except that \mathbf{X} is now a parameter of the model,

$$L(\theta, \mathbf{X}) = -\frac{ND}{2} \log(2\pi) - \frac{D}{2} \log \det(\sigma^2 \mathbf{I} + \mathbf{K}) - \frac{1}{2} \text{tr} [(\sigma^2 \mathbf{I} + \mathbf{K})^{-1} \mathbf{Y} \mathbf{Y}^T],$$

where \mathbf{K} has elements $k(x_n, x_m)$ that depend on \mathbf{X} through the covariance function. In the original formulation of the GPLVM the latent points \mathbf{X} were optimised by maximum likelihood (Lawrence, 2005), but later the Bayesian GPLVM (BGPLVM) was introduced which placed a prior on \mathbf{X} and a variational Bayesian inference algorithm was used to approximate the posterior distribution over the latent space (Titsias and Lawrence, 2010).

The GPLVM has been used to visualise single-cell gene expression in a number of studies (Buettner and Theis, 2012; Buettner *et al.*, 2015; Zwiessele and Lawrence, 2016). It has also been used to sample pseudotime trajectories in order to quantify uncertainty in pseudotime (Campbell and Yau, 2016). In the case of single-cell time series experiments, where cells are collected at multiple capture times, then the prior on the latent variable \mathbf{X} can incorporate this capture time information to improve pseudotime inference (Reid and Wernisch, 2016). In Figure 31.7 we show how including capture time information in the prior of the BGPLVM can be used to align one of the latent dimensions with time in a developmental time course data set (Ahmed *et al.*, 2018). Cells were captured from single embryos at different times in mouse embryonic development (Guo *et al.*, 2010). At the 32-cell stage they begin to differentiate into different cell types. We see that aligning one axis with time makes the latent space more

Figure 31.7 We project single-cell gene expression data from Guo *et al.* (2010) onto two latent dimensions using PCA and the BGPLVM with two different choices of prior: (a) PCA; (b) BGPLVM with zero-mean prior for all latent points; (c) BGPLVM with prior mean in one latent dimension based on capture times. For further details, see Ahmed *et al.* (2018).

Downloaded from https://onlinelibrary.wiley.com/doi/ by Columbia University Libraries, Wiley Online Library on [27/05/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interpretable, with the other dimension capturing differences in cell type. We can also see that the nonlinear dimensionality reduction of the GPLVM leads to a cleaner separation of cell stages and types than a linear approach such as principal components analysis (PCA). Ahmed *et al.* (2018) show that using a two-dimensional latent space with pseudotime on one dimension leads to a higher correlation between pseudotime and capture time than using a one-dimensional pseudotime latent space with the same capture time informed prior on that dimension.

The BGPLVM approach to pseudotime estimation implemented by Campbell and Yau (2016) and Reid and Wernisch (2016) made use of MCMC or HMC sampling over \mathbf{X} . These sampling-based approaches do not scale up to inference over large droplet-based single-cell RNA-sequencing experiments which can profile tens of thousands of cells. The GrandPrix package used in Figure 31.7 (Ahmed *et al.*, 2018) uses a more computationally efficient variational inference scheme (Titsias and Lawrence, 2010) and is implemented using the GPflow package (Matthews *et al.*, 2017) which adapts the TensorFlow package to GP inference, allowing scalability to multiple cores is sufficient cores and optionally also graphics processing units.

31.3.3 Modelling Branching Dynamics with Single-Cell RNA-Sequencing Data

Figure 31.7 shows an example of cells differentiating into different cell types during development. Recent developments in single-cell RNA-sequencing allow gene expression to be profiled in thousands of cells. In some cases, the cells being profiled are undergoing differentiation but there are no time labels in the data. For example, a sample may contain a continuum of stem cells, differentiated cells and intermediates. In that case pseudotime methods can be used to investigate differentiation by fitting models of branching dynamics. A number of algorithms have been proposed to discover branching dynamics in cellular trajectories from single-cell expression data (Haghverdi *et al.*, 2016; Street *et al.*, 2017; Qiu *et al.*, 2017a). Lönnberg *et al.* (2017) used the overlapping mixture of Gaussian processes (OMGP) model (Lázaro-Gredilla *et al.*, 2012) to identify cellular branching dynamics after pseudotime inference, by identifying where in pseudotime the cells are better described as coming from two profiles rather than one.

The methods in Section 31.2.2 can also be extended to model branching dynamics if cell-to-branch assignments are learned in a similar way to the OMGP model. Consider a set of GP functions $F = \{f_1, f_2, \dots, f_M\}$ which are branches in a tree following a covariance of the type defined in Section 31.2.2. Then define $Z \in \{0, 1\}^{N \times M}$ to be binary variables which determine the assignment of N cells to M branches and which have to be inferred along with F . In the simplest case of a single branching event $M = 3$ and x_b is the pseudotime of branching, which we consider to be specific to a particular gene. By applying a global branching and pseudotime algorithm (e.g. Monocle 2; Qiu *et al.*, 2017a) we can gain some prior information about each gene's branching dynamics. Consider that x_g is the global branching point in pseudotime but that genes may branch before or after that point, or possibly not exhibit any branching. If $x_b < x_g$ then the global branching provides no information about which branch the cell belongs to for $x_b < x < x_g$. If $x_b > x_g$ then we can use the inferred global branching to increase the probability of a cell being assigned to a particular branch. This approach was recently used to identify whether individual gene expression shows branching dynamics and whether the branching is early or late in pseudotime (Boukouvalas *et al.*, 2018). Inference is not exactly tractable in this class of models but Boukouvalas *et al.* (2018) developed an efficient sparse variational inference algorithm which generalises the OMGP inference algorithm (Lázaro-Gredilla *et al.*, 2012) to the case where the functions in F are not independent.

The scalability of the model to large data sets is achieved in Boukouvalas *et al.* (2018) by using an inducing point sparse approximation. The key bottleneck of applying GP models to large data sets is that the full covariance inversion, required at each iteration of hyperparameter

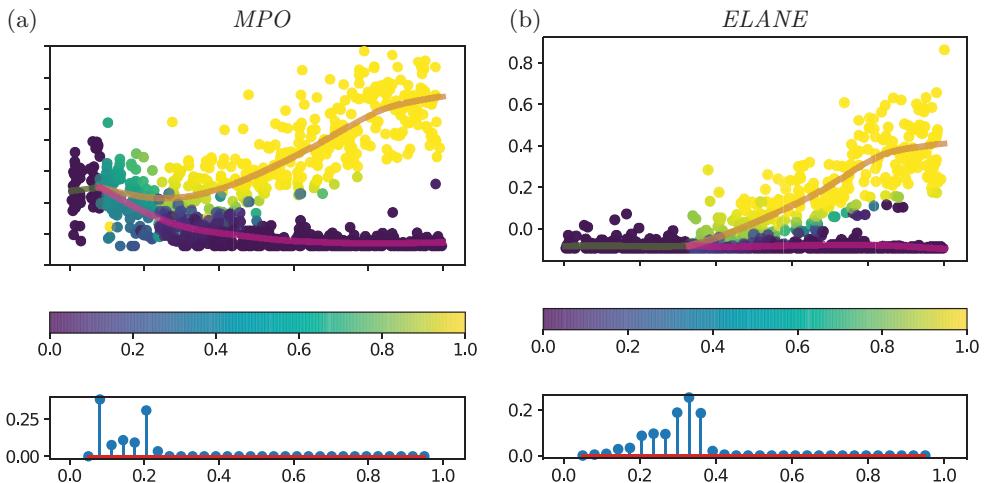


Figure 31.8 Haematopoiesis single-cell gene expression data from Paul *et al.* (2015) is shown for genes (a) *MPO* and (b) *ELANE*. Monocle 2 was used to infer pseudotime for each cell and the BGP model was used to identify the branching time for each gene. For each gene the posterior cell assignment is shown in top subpanel. In the bottom subpanel the posterior branching time is shown. For further details about the BGP model see Boukouvalas *et al.* (2018).

optimisation, scales cubically, $O(N^3)$, where N is the number of data points, or alternatively cells in the case of single-cell data. In contrast, the inducing point approximation defines a number of auxiliary variables, termed inducing points, that trade off fidelity to the full GP and computational speed. Specifically, for k inducing points the inference scales as $O(k^2N)$ rather than $O(N^3)$, where typically $k \ll N$. Where the input dimension is low (e.g. time series data) it is often sufficient to have a relatively small number of inducing points; in the present application we found $k = 30$ to be sufficient and little improvement was obtained when increasing k further in our synthetic data studies (Boukouvalas *et al.*, 2018). Bauer *et al.* (2016) provide more general insights into the performance of alternative inducing point approaches.

In Figure 31.8 we apply the branching GP (BGP; Boukouvalas *et al.*, 2018) model to single-cell RNA-sequencing haematopoietic stem cell data from Paul *et al.* (2015). The data consists of 4423 cells, and Monocle 2 was used to infer the global cellular branching and pseudotime for each cell (Qiu *et al.*, 2017a). We show the inferred gene-specific branching dynamics for two genes, one of which is inferred to branch earlier than the global branching (Figure 31.8(a)) and another which is inferred to branch later (Figure 31.8(b)). Below the genes we show the posterior over the inferred branching time which is computed using the histogram approach in equation (31.7). Boukouvalas *et al.* (2018) show that the GP approach can better deal with cases where branching differs from the global branching than the spline-based approach implemented in the BEAM package (Qiu *et al.*, 2017b) which does not model cell-to-branch assignment probabilistically.

31.4 Conclusion

We have presented a number of ways in which GP inference methods can be used to make inferences about gene expression dynamics. There are many ongoing challenges. In most cases described here we have modelled data as coming from a Gaussian distribution, but this is a poor assumption in many interesting applications. For example, single-cell RNA-sequencing

data can contain large numbers of zero measurements and count-based likelihoods are more suitable for modelling this class of data. Count-based likelihoods are available for standard GP regression and can easily be implemented for other GP models if using MCMC inference – for example, using the popular probabilistic programming language Stan (Carpenter *et al.*, 2016). However, to scale up to large numbers of single cells we have adopted more computationally efficient sparse variational inference algorithms, and count-based likelihoods are not yet available for variational inference in the BGP or GPLVM models described here.

The BGP method is used to infer gene-specific branching after applying another algorithm for inferring pseudotime and the global cellular branching. This approach does not fully take into account errors and uncertainty in this initial cell labelling stage prior to gene-specific modelling. The model does allow branch labels to change through inference since the global labels are treated as a prior. However, we do assume that pseudotimes are known without error, when in reality pseudotime inference is associated with high levels of uncertainty (Campbell and Yau, 2016). An interesting extension would therefore be to combine the BGP and GPLVM models to jointly model branching and infer latent manifolds from single-cell expression data, taking all sources of error into account through use of a unified model.

Acknowledgements

The ideas and results described here are the product of numerous interactions with our close collaborators, including Neil Lawrence, James Hensman and Antti Honkela. MR and JY are supported by MRC award MR/N00017X/1; MR and AB are supported by MRC award MR/M008908/1; MR is supported by a Wellcome Trust Investigator Award; SA was supported by a Commonwealth PhD Scholarship.

References

- Ahmed, S., Rattray, M. and Boukouvalas, A. (2018). GrandPrix: Scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* **35**(1), 47–54.
- Barencro, M., Tomescu, D., Brewer, D., Callard, R., Stark, J. and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology* **7**(3), R25.
- Bauer, M., van der Wilk, M. and Rasmussen, C.E. (2016). Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA, pp. 1533–1541.
- Boukouvalas, A., Hensman, J. and Rattray, M. (2018). BGP: Identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biology* **19**:65.
- Buettner, F. and Theis, F.J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* **28**(18), i626–i632.
- Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C. and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**(2), 155–160.
- Campbell, K.R. and Yau, C. (2016). Order under uncertainty: Robust differential expression analysis using probabilistic models for pseudotime inference. *PLoS Computational Biology* **12**(11), e1005212.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A., *et al.* (2016). Stan: A probabilistic programming language. *Journal of Statistical Software* **20**(2), 1–37.

- Fearnhead, P., Giagos, V. and Sherlock, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics* **70**(2), 457–466.
- Folia, M.M. and Rattray, M. (2018). Trajectory inference and parameter estimation in stochastic models with temporally aggregated data. *Statistics and Computing* **28**(5), 1053–1072.
- Galla, T. Intrinsic fluctuations in stochastic delay systems: Theoretical description and application to a simple model of gene regulation. (2009). *Physical Review E* **80**(2), 021909.
- Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry* **81**(25), 2340–2361.
- Gossan, N., Zeef, L., Hensman, J., Hughes, A., Bateman, J.F., Rowley, L., Little, C.B., Piggins, H.D., Rattray, M., Boot-Handford, R.P., et al. (2013). The circadian clock in murine chondrocytes regulates genes controlling key aspects of cartilage homeostasis. *Arthritis & Rheumatology* **65**(9), 2334–2345.
- Guo, G., Huss, M., Tong, G.Q., Wang, C., Sun, L.L., Clarke, N.D. and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell* **18**(4), 675–685.
- Haghverdi, L., Buettner, M., Alexander Wolf, F., Buettner, F. and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13**(10), 845.
- Heinonen, M., Guipaud, O., Milliat, F., Buard, Valérie, Micheau, Béatrice, Tarlet, G., Benderitter, M., Zehraoui, F. and d'Alché Buc, F. (2014). Detecting time periods of differential gene expression using Gaussian processes: An application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics* **31**(5), 728–735.
- Hensman, J., Lawrence, N.D. and Rattray, M. (2013). Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics* **14**(1), 252.
- Hensman, J., Rattray, M. and Lawrence, N.D. (2015). Fast nonparametric clustering of structured time-series. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(2), 383–393.
- Honkela, A., Girardot, C., Hilary Gustafson, E., Liu, Ya-H., Furlong, E.E., Lawrence, N.D. and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences of the United States of America* **107**(17), 7793–7798.
- Honkela, A., Peltonen, J., Topa, H., Charapitsa, I., Matarese, F., Grote, K., Stunnenberg, H.G., Reid, G., Lawrence, N.D. and Rattray, M. (2015). Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proceedings of the National Academy of Sciences of the United States of America* **112**(42), 13115–13120.
- Huang, Y. and Sanguinetti, G. (2016). Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32**(19), 2965–2972.
- Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New York.
- Kalaitzis, A.A. and Lawrence, N.D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics* **12**(1), 180.
- Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M. and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.
- Komorowski, M., Finkenstdt, B., Harper, C.V. and Rand, D.A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**, 1–10.

- Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research* **6**, 1783–1816.
- Lawrence, N.D., Sanguinetti, G. and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In B. Schölkopf, J.C. Platt, and T. Hoffman (eds), *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, pp. 785–792.
- Lázaro-Gredilla, M., Vaerenbergh, S.V. and Lawrence, N.D. (2012). Overlapping mixtures of Gaussian processes for the data association problem. *Pattern Recognition* **45**(4), 1386–1395.
- Lewis, L.A., Polanski, K., de Torres-Zabala, M., Jayaraman, S., Bowden, L., Moore, J., Penfold, C.A., Jenkins, D.J., Hill, C., Baxter, L., Kulasekaran, S., Truman, W., Littlejohn, G., Prusinska, J., Mead, A., Steinbrenner, J., Hickman, R., Rand, D., Wild, D.L., Ott, S., Buchanan-Wollaston, V., Smirnoff, N., Beynon, J., Denby, K. and Grant, M. (2015). Transcriptional dynamics driving mamp-triggered immunity and pathogen effector-mediated immunosuppression in arabidopsis leaves following infection with *Pseudomonas syringae* pv tomato dc3000. *Plant Cell* **27**(11), 3038–3064.
- Lönnberg, T., Svensson, V., James, K.R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M.S., Fogg, L.G., Nair, A.S., Liligeto, U., et al. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Science Immunology* **2**(9).
- MacKay, D.J. (1998). Introduction to gaussian processes. In C.M. Bishop (ed.), *Neural Networks and Machine Learning*. Springer. pp. 133–166.
- Matthews, A., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z. and Hensman, J. (2017). GPflow: A Gaussian process library using Tensorflow. *Journal of Machine Learning Research* **18**(40), 1–6.
- Monk, N.A.M. (2003). Oscillatory expression of Hes1, p53, and NF- κ B driven by transcriptional time delays. *Current Biology* **13**(16), 1409–1413.
- Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**(7), 1663–1677.
- Phillips, N.E., Manning, C., Papalopulu, N. and Rattray, M. (2017). Identifying stochastic oscillations in single-cell live imaging time series using Gaussian processes. *PLoS Computational Biology* **13**(5), e1005479.
- Qiu, X., Hill, A., Ma, Yi-A. and Trapnell, C. (2017a). Single-cell mRNA quantification and differential analysis with Census. *Nature Methods* **14**, 309–315.
- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A. and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979–982.
- Rasmussen, C.E. and Williams, C.K. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- Reid, J.E. and Wernisch, L. (2016). Pseudotime estimation: Deconfounding single cell time series. *Bioinformatics* **32**(19), 2973–2980.
- Stegle, O., Denby, K.J., Cooke, E.J., Wild, D.L., Ghahramani, Z. and Borgwardt, K.M. (2010). A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology* **17**(3), 355–367.
- Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E. and Dudoit, S. (2017). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. Preprint, bioRxiv 128843.
- Titsias, M. and Lawrence, N.D. (2010). Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851.

- Titsias, M.K., Honkela, A., Lawrence, N.D. and Rattray, M. (2012). Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Systems Biology* **6**(1), 53.
- Topa, H. and Honkela, A. (2016). Analysis of differential splicing suggests different modes of short-term splicing regulation. *Bioinformatics* **32**(12), i147–i155.
- Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**(5), 1413–1432.
- Westermark, Pål O., Welsh, D.K., Okamura, H. and Herzl, H. (2009). Quantification of circadian rhythms in single cells. *PLoS Computational Biology* **5**(11), e1000580.
- Yang, J., Penfold, C.A., Grant, M.R. and Rattray, M. (2016). Inferring the perturbation time from biological time course data. *Bioinformatics* **32**(19), 2956–2964.
- Zwijsen, M. and Lawrence, N.D. (2016). Topslam: Waddington landscape recovery for single cell experiments. Preprint, bioRxiv 057778.

32

Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models

Marco Grzegorczyk¹ and Dirk Husmeier²

¹Bernoulli Institute (BI), Faculty of Science and Engineering, Rijksuniversiteit Groningen, Groningen, Netherlands

²School of Mathematics & Statistics, University of Glasgow, Glasgow, United Kingdom

Abstract

In statistical genomics and systems biology non-homogeneous dynamic Bayesian networks (NH-DBNs) have become an important tool for learning regulatory networks and signalling pathways from post-genomic data, such as gene expression time series. This chapter gives an overview of various state-of-the-art NH-DBN models with a variety of features. All NH-DBNs presented here have in common that they are Bayesian models that combine linear regression with multiple change point processes. The NH-DBN models can be used for learning the network structures of time-varying regulatory processes from data, where the regulatory interactions are subject to temporal change. We conclude this chapter with an illustration of the methodology on two applications, related to morphogenesis in *Drosophila* and synthetic biology in yeast.

32.1 Introduction

Molecular pathways consisting of interacting proteins underlie the major functions of living cells. A central goal of molecular biology is therefore to understand the regulatory mechanisms of gene transcription and protein synthesis, and the invention of DNA microarrays and deep sequencing technologies, by which the transcription levels of thousands of genes can be measured simultaneously, mark an important breakthrough in this endeavour. Several approaches to the reverse engineering of genetic regulatory networks from gene expression data have been explored. At the most refined level of detail is a mathematical description of the biophysical processes in terms of a system of coupled differential equations that describe, for example, the processes of transcription factor binding, diffusion, and RNA degradation; see, for instance, Chen *et al.* (1999) and Wilkinson (2006). While such low-level dynamics are critical to a complete understanding of regulatory networks, they require detailed specifications of both the relationship between the interacting entities and the parameters of the biochemical reaction, such as reaction rates and diffusion constants. Zak *et al.* (2002) found that a system of ordinary differential equations describing a regulatory network of three genes with their respective mRNA and protein products is not identifiable when only gene expression data are observed, and that rich data, including detailed information on protein–DNA interactions, are needed to ensure

identifiability of the parameters that determine the interaction structure. Vyshevimirsky and Girolami (2008) successfully demonstrated the computation of the marginal likelihood for Bayesian ranking of biochemical system models described by systems of coupled differential equations. However, the computation costs are substantial. Detailed prior knowledge about plausible candidate pathways is therefore indispensable, and the approach is infeasible for *ab initio* reconstruction of regulatory networks.

At the other extreme of the spectrum are coarse-scale approaches based on pairwise association scores, such as the correlation or mutual information of the time-varying expression levels (Butte and Kohane, 2000, 2003). The underlying conjecture is that co-expression is indicative of co-regulation; thus, such associations may identify genes that have similar functions or are involved in related biological processes. The disadvantage, however, is that the identification of such pairwise associations does not provide insight into the regulation processes in the holistic context of the biological system and does not indicate, for example, whether an interaction between two genes is direct or mediated by other genes, or whether a gene is a regulator or regulatee.

A promising compromise between these two extremes is the approach of Bayesian networks (BNs) and dynamical Bayesian networks (DBNs). Following up on the pioneering work by Friedman *et al.* (2000), Hartemink *et al.* (2001) and Husmeier (2003), these models have received substantial attention from the computational biology community as statistically inferable abstract representations of gene regulatory networks.

In particular, DBNs have become a popular tool for learning gene regulatory networks from gene expression time series, and the general DBN model is described in Section 32.2.2. A drawback of DBNs is that they do not allow the network parameters to change in time. For cellular networks the strengths of the regulatory interactions often depend on unobserved factors, such as cellular, environmental and experimental conditions. This renders the traditional homogeneous DBNs too restrictive for most of the applications in the field of statistical genomics. As a consequence, various non-homogeneous DBNs (NH-DBNs) have been proposed in the computational biology literature. The class of NH-DBNs can be divided into two conceptual groups: NH-DBNs which allow only the network parameters to vary in time (see, for example, Grzegorczyk and Husmeier, 2011) and NH-DBNs which also allow the network structure to be time-dependent (see, for example, Robinson and Hartemink, 2010; Lèbre *et al.*, 2010; Dondelinger *et al.*, 2013). The latter class of NH-DBNs yields substantial flexibility, which is important for studying morphogenesis, or for analysing expression time series that cover different stages of an organism's life cycle. However, in most applications within the same stage of the life cycle, the network structure is unlikely to change, and the situation is more comparable to a traffic flow network, where it is not the road system (the network) that changes with time (changes between off-peak and rush hours), but the intensities of the traffic flows (the network parameters). NH-DBN models with changing network structures are therefore only briefly reviewed in Section 32.2.6, and the focus of this chapter is on NH-DBNs that only allow the network parameters to change.

Most of the proposed NH-DBN models use a changepoint process to divide the data into disjoint segments. Alternative data segmentation methods, such as free mixture models (MIX) and hidden Markov models (HMMs), are briefly reviewed in Section 32.2.5. The NH-DBNs can infer the data segmentation, the network structure and the segment-specific network parameters from the data. One standard ('uncoupled') NH-DBN model instantiation is described in Section 32.2.3. The drawback of 'uncoupled' NH-DBNs is that the available time series are often rather short and that the network parameters have to be learned for each segment separately. This can lead to inflated inference uncertainties or overfitting. To address this bottleneck, sequential (Grzegorczyk and Husmeier, 2012) and global (Grzegorczyk and Husmeier, 2013)

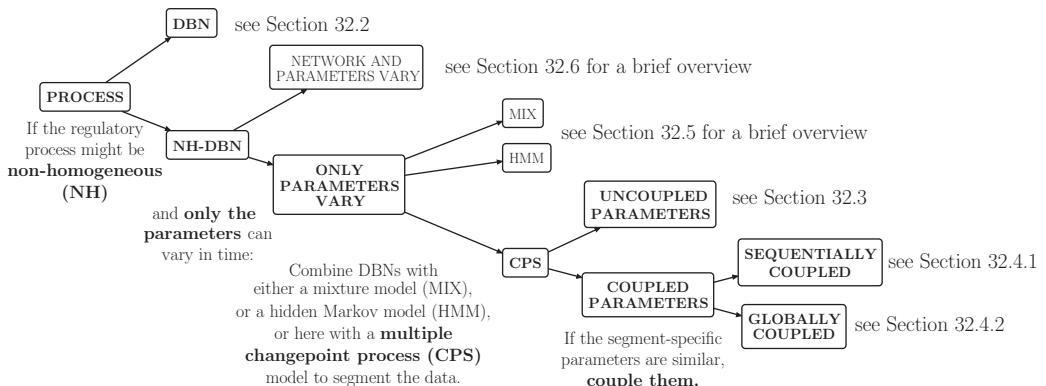


Figure 32.1 Overview and chapter outline, showing how the various non-homogeneous DBN models from the recent literature are related, and where they are reviewed in this chapter.

coupling mechanisms for the segment-specific parameters have been proposed. NH-DBN models with coupled network parameters are described in Section 32.2.4.

Figure 32.1 provides an overview of various state-of-the-art NH-DBN models with references to those subsections (all part of Section 32.2) in which the respective models are described or reviewed. Practical application examples, related to embryogenesis in *Drosophila* and synthetic biology in yeast, can be found in Section 32.3. The chapter concludes in Section 32.4 with a summary.

32.2 Methodology

32.2.1 Dynamic Bayesian Networks (DBN)

Dynamic Bayesian networks are a popular class of models for learning the dependencies between random variables from time series data. In DBNs a dependency between two random variables X and Y is typically interpreted in terms of a regulatory interaction with a time delay. A directed edge from X to Y , symbolically $X \rightarrow Y$, indicates that the value of variable Y at any timepoint t depends on the realization of X at the previous timepoint $t - 1$. Typically, various variables X_1, \dots, X_k have a regulatory effect on Y , and the relationship between X_1, \dots, X_k and Y can be described in terms of a linear regression model,

$$y_{t+1} = \beta_0 + \beta_1 x_{t,1} + \dots + \beta_k x_{t,k} + u_{t+1} \quad (t = 1, \dots, T),$$

where $T + 1$ is the number of observed timepoints, T is the effective number of datapoints that can be used for the regression model, y_{t+1} is the value of Y at timepoint $t + 1$, $x_{t,j}$ is the value of covariate X_j at timepoint t , β_0, \dots, β_k are regression coefficients with β_0 being the intercept, and u_2, \dots, u_{T+1} are independent realizations of a Gaussian noise variable with mean 0. In the regression model terminology Y is called the response and the variables X_1, \dots, X_k are called the covariates for Y . In DBNs X_1, \dots, X_k are called the regulators (parent nodes) of the regulatee (child node) Y , and there is a directed edge from each X_j to Y , symbolically $X_j \rightarrow Y$ ($j = 1, \dots, k$).

For DBN applications the typical situation is that N variables Z_1, \dots, Z_N have been measured at $T + 1$ equidistant timepoints $t = 1, \dots, T + 1$, and the goal is to infer the interactions among them. The inference results are represented compactly in the form of a network. In the network each variable Z_i is represented as a node, and directed edges between the nodes indicate

all interactions among them. When ‘self-loop’ edges, such as $Z_i \rightarrow Z_i$, are ruled out, there are $N(N - 1)$ possible interactions (directed edges). Because of the time lag, there is no need to keep the acyclicity constraint, which has to be kept in (static) Bayesian networks. Ruling out self-loops, there are $2^{N(N-1)}$ valid networks for N nodes.

Because of the time lag, inference for DBNs is conceptually easier than for BNs. Unlike in BNs, the covariates (parent nodes) can be determined separately for each node Z_i . To determine the parent nodes of Z_i , variable Z_i takes the role of the response in a regression model where the other $N - 1$ domain variables $\{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N\}$ are the potential covariates and all interactions are subject to a time lag. The goal is to infer a covariate set $\pi_i \subset \{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N\}$ for Z_i . Each Z_j in π_i is a parent of Z_i ; that is, the network has the edge from $Z_j \rightarrow Z_i$ if and only if $Z_j \in \pi_i$.

DBN inference thus corresponds to the inference of N separate regression models. The system of inferred covariate sets π_1, \dots, π_N can be interpreted as a network. There is the edge $Z_j \rightarrow Z_i$ in the network if and only if $Z_j \in \pi_i$ ($i, j \in \{1, \dots, N\} : i \neq j$).

For the following subsections we now introduce some generic notation. \mathbf{D} denotes an N -by- $(T + 1)$ data matrix whose rows correspond to the N variables and whose columns correspond to the timepoints $t = 1, \dots, T + 1$. $\mathbf{D}_{i,t}$ is the value of Z_i at timepoint t . In the i th regression model $Y := Z_i$ is the response, and there are $n := N - 1$ potential covariates: $X_1 := Z_1, \dots, X_{i-1} := Z_{i-1}, X_i := Z_{i+1}, \dots, X_n := Z_N$. The regression model has to be inferred from T datapoints \mathcal{D}_t ($t = 1, \dots, T$), where each \mathcal{D}_t contains the values of the potential covariates: $x_{t,1} := \mathbf{D}_{1,t}, \dots, x_{t,i-1} := \mathbf{D}_{i-1,t}, x_{t,i} := \mathbf{D}_{i+1,t}, \dots, x_{t,n} := \mathbf{D}_{N,t}$ and a shifted response value $y_{t+1} = \mathbf{D}_{i,t+1}$.

32.2.2 Bayesian Linear Regression

Consider a Bayesian linear regression model with response Y and $\pi = \{X_1, \dots, X_k\}$ a set of k covariates. There are T datapoints $\mathcal{D}_1, \dots, \mathcal{D}_T$, and \mathcal{D}_t contains a response value y_{t+1} and values of the covariates $x_{t,j}$ ($j = 1, \dots, k$). We build the vector \mathbf{y} of response values, a design matrix \mathbf{X} with a first column of 1s for the intercept, and a regression coefficient vector β with the $k + 1$ regression coefficients:

$$\mathbf{y} = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_{T+1} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ 1 & x_{2,1} & \dots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{T,1} & \dots & x_{T,k} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

We have not made explicit that \mathbf{X} and β both depend on π . \mathbf{X} is a T -by- $(1 + |\pi|)$ matrix, which has a first column of 1s for the intercept and then a column for each covariate $X_j \in \pi$ ($j = 1, \dots, k$), filled with the values $x_{1,j}, \dots, x_{T,j}$ of that particular covariate. The $(j + 1)$ th element β_j of β is the regression coefficient for the j th covariate $X_j \in \pi$.

We assume a multivariate Gaussian likelihood of the form

$$\mathbf{y} | (\beta, \sigma^2, \pi) \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}), \tag{32.1}$$

where \mathbf{I} denotes the identity matrix, and σ^2 is a noise variance parameter which we assume to have an inverse gamma distribution,

$$\sigma^{-2} \sim GAM(a_\sigma, b_\sigma),$$

with hyperparameters a_σ and b_σ . On β we impose a multivariate Gaussian prior,

$$\beta | (\sigma^2, \lambda^2, \pi) \sim \mathcal{N}(\mu, \sigma^2 \lambda^2 \mathbf{I}), \tag{32.2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{k+1}$ is the prior expectation vector, whose $(j+1)$ th element is the prior expectation of β_j ($j = 0, \dots, k$), and λ^2 is a ‘signal-to-noise ratio’ parameter on which we also impose an inverse gamma distribution, $\lambda^{-2} \sim GAM(a_\lambda, b_\lambda)$. Re-employing the parameter σ^2 in equation (32.2) yields a fully conjugate prior in both $\boldsymbol{\beta}$ and σ^2 ; this allows both parameter groups to be integrated out in the likelihood, that is, the marginal likelihood $p(\mathbf{y}|\lambda^2, \boldsymbol{\pi})$ to be computed (see, for example, Gelman *et al.*, 2004, Sections 3.3 and 3.4)). For the posterior distribution of the parameters we have

$$p(\boldsymbol{\beta}, \sigma^2, \lambda^2 | \mathbf{y}, \boldsymbol{\pi}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\pi}) \cdot p(\boldsymbol{\beta}|\sigma^2, \lambda^2, \boldsymbol{\pi}) \cdot p(\sigma^2) \cdot p(\lambda^2).$$

A sample $(\boldsymbol{\pi}_{(r)}, \boldsymbol{\beta}_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2)_{r=1, \dots, R}$ from the posterior distribution can be obtained by Markov chain Monte Carlo (MCMC) simulations. For the model, the full conditional distributions can be computed analytically, so that Gibbs sampling can be applied. Given a fixed covariate set $\boldsymbol{\pi}$, we initialize, for example, $\boldsymbol{\pi}_{(0)} = \boldsymbol{\pi}$, $\boldsymbol{\beta}_{(0)} = \mathbf{0}$, $\sigma_{(0)}^2 = 1$, and $\lambda_{(0)}^2 = 1$, and then we iteratively resample the parameters from their full conditionals with the pseudocode provided in Table 32.1.

For the marginal likelihood, with $\boldsymbol{\beta}$ and σ^2 integrated out, the marginalization rule from (Bishop, 2006, Section 2.3.7) yields

$$p(\mathbf{y}|\lambda^2, \boldsymbol{\pi}) = \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-\frac{T}{2}} (2b_\sigma)^{a_\sigma}}{\det(\mathbf{C})^{1/2}} (2b_\sigma + (\mathbf{y} - \mathbf{X}\boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}))^{-\left(\frac{T}{2} + a_\sigma\right)}, \quad (32.3)$$

where $\mathbf{C} := \mathbf{I} + \lambda^2 \mathbf{X} \mathbf{X}^T$ depends on λ^2 , and \mathbf{X} and $\boldsymbol{\mu}$ depend on $\boldsymbol{\pi}$.

In realistic applications the covariates have to be inferred from the data. In DBNs there are $n := N - 1$ potential covariates and the relevant subset $\boldsymbol{\pi} \subset \{X_1, \dots, X_n\}$ has to be found. A common prior assumption is that all covariate sets $\boldsymbol{\pi}$ with up to \mathcal{F} covariates are equally likely, while all other parent sets have zero prior probability:

$$p(\boldsymbol{\pi}) = \begin{cases} \frac{1}{c}, & \text{if } |\boldsymbol{\pi}| \leq \mathcal{F}, \\ 0, & \text{if } |\boldsymbol{\pi}| > \mathcal{F}, \end{cases} \quad \text{where } c = \sum_{i=0}^{\mathcal{F}} \binom{n}{i}. \quad (32.4)$$

Imposing a ‘fan-in’ restriction \mathcal{F} has the practical advantage that it restricts the system of possible covariate sets. For gene regulatory networks a ‘fan-in’ restriction of $\mathcal{F} := 3$ can also be biologically motivated, as few genes are regulated by more than three genes, and no restriction

Table 32.1 Pseudocode. Bayesian linear regression inference (for DBNs), given a fixed covariate set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}$. Iteratively the parameters are resampled from their full conditional distribution. See text of Section 32.2.2 for further details

Initialization: For example, $\boldsymbol{\pi}_{(0)} = \boldsymbol{\pi}$, $\boldsymbol{\beta}_{(0)} = \mathbf{0}$, and $\sigma_{(0)}^2 = \lambda_{(0)}^2 = 1$.

Iterations: For $r = 1, \dots, R$:

(1a) For σ^2 use a collapsed Gibbs sampling step with $\boldsymbol{\beta}$ being integrated out; i.e. sample $\sigma_{(r)}^{-2}$ from

$$GAM\left(\alpha_\sigma + \frac{T}{2}, \beta_\sigma + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})^T \left(\mathbf{I} + \lambda_{(r-1)}^2 \mathbf{X} \mathbf{X}^T\right)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\mu})\right). \text{Invert } \sigma_{(r)}^{-2} \text{ to obtain } \sigma_{(r)}^2.$$

(2a) Sample $\boldsymbol{\beta}_{(r)}$ from $\mathcal{N}\left([\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}^T \mathbf{X}]^{-1} (\lambda_{(r-1)}^{-2} \boldsymbol{\mu} + \mathbf{X}^T \mathbf{y}), \sigma_{(r)}^2 [\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}^T \mathbf{X}]^{-1}\right)$

(3a) Sample $\lambda_{(r)}^{-2}$ from $GAM\left(\alpha_\lambda + \frac{(|\boldsymbol{\pi}_{(r)}|+1)}{2}, \beta_\lambda + \frac{1}{2} \sigma_{(r)}^{-2} (\boldsymbol{\beta}_{(r)} - \boldsymbol{\mu})^T (\boldsymbol{\beta}_{(r)} - \boldsymbol{\mu})\right)$. Invert $\lambda_{(r)}^{-2}$ to obtain $\lambda_{(r)}^2$.

(4a) Keep the covariate set $\boldsymbol{\pi}$ fixed, i.e. set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}$, so that \mathbf{X} and $\boldsymbol{\mu}$ stay unchanged.

is imposed on the outdegree (i.e. we have no ‘fan-out’ restriction), thus allowing for hubs and central regulators.

Using the marginal likelihood from equation (32.3), we obtain the posterior

$$p(\boldsymbol{\pi}, \lambda^2 | \mathbf{y}) \propto p(\mathbf{y} | \lambda^2, \boldsymbol{\pi}) \cdot p(\boldsymbol{\pi}) \cdot p(\lambda^2). \quad (32.5)$$

Given λ^2 , Metropolis–Hastings MCMC steps can be used to sample covariate sets $\boldsymbol{\pi}$ from the posterior distribution. Typically three types of move are implemented. In the deletion move (D) we randomly select one $X_j \in \boldsymbol{\pi}$ and propose to remove it from $\boldsymbol{\pi}$. In the addition move (A) we randomly select one $X_j \notin \boldsymbol{\pi}$ and we propose to add it to $\boldsymbol{\pi}$. In the exchange move (E) we randomly select one $X_j \in \boldsymbol{\pi}$ and we propose to replace it by a randomly selected $X_w \notin \boldsymbol{\pi}$. Each move yields a new candidate covariate set $\boldsymbol{\pi}_*$, and the move proposes to replace the current $\boldsymbol{\pi}$ by $\boldsymbol{\pi}_*$. When randomly selecting the type of move, the acceptance probability for the move is

$$A(\boldsymbol{\pi}, \boldsymbol{\pi}_*) = \min \left\{ 1, \frac{p(\mathbf{y} | \lambda^2, \boldsymbol{\pi}_*)}{p(\mathbf{y} | \lambda^2, \boldsymbol{\pi})} \cdot \frac{p(\boldsymbol{\pi}_*)}{p(\boldsymbol{\pi})} \cdot HR \right\}, \quad \text{where } HR = \begin{cases} \frac{|\boldsymbol{\pi}|}{n - |\boldsymbol{\pi}_*|}, & \text{for D,} \\ \frac{n - |\boldsymbol{\pi}|}{|\boldsymbol{\pi}_*|}, & \text{for A,} \\ 1, & \text{for E,} \end{cases} \quad (32.6)$$

n is the number of potential covariates, and $|\cdot|$ denotes the cardinality.

If the covariate set has to be inferred from the data, step (4a) of the MCMC algorithm in Table 32.1 has to be replaced by a Metropolis–Hastings step on the covariate set; see Table 32.2.

The marginal posterior probability that X_j is a covariate for Y is

$$p_j := p(X_j \rightarrow Y | \mathbf{y}) = \sum_{\boldsymbol{\pi}} I_{\{X_j \in \boldsymbol{\pi}\}}(\boldsymbol{\pi}) \cdot \int p(\boldsymbol{\pi}, \lambda^2 | \mathbf{y}) d\lambda^2,$$

where the sum is over all possible covariate sets, and the indicator function $I(\cdot)$ is 1 if $X_j \in \boldsymbol{\pi}$ and 0 otherwise. As p_j is not computationally feasible, the sample $(\boldsymbol{\pi}_{(r)}, \beta_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2)_{r=1, \dots, R}$ is used to approximate p_j . The estimator is the fraction of covariate sets that contain X_j :

$$\hat{p}_j := \hat{p}(X_j \rightarrow Y | \mathbf{y}) = \frac{1}{R} \sum_{r=1}^R I_{\{X_j \in \boldsymbol{\pi}_{(r)}\}}(\boldsymbol{\pi}_{(r)}).$$

By imposing a threshold $\phi \in [0, 1]$ one can obtain a concrete covariate set $\hat{\boldsymbol{\pi}}_\phi$ for Y ,

$$\hat{\boldsymbol{\pi}}_\phi = \{X_j : \hat{p}_j > \phi\}, \quad (32.7)$$

containing only those covariates whose marginal posterior probability exceeds ϕ .

In a DBN domain the regression model is applied to each domain variable Z_i ($i = 1, \dots, N$) separately. In the i th regression model Z_i takes the role of the response Y and

Table 32.2 Pseudocode. New step (4a) for the pseudocode in Table 32.1 to infer the covariate set $\boldsymbol{\pi}$

-
- (4a) New step (4a), which proposes to change the covariate set:
- Choose the type of move (D, A or E) and a new candidate set $\boldsymbol{\pi}_*$, as described in the main text.
 - Propose to move from $\boldsymbol{\pi}_{(r-1)}$ to $\boldsymbol{\pi}_*$. Accept $\boldsymbol{\pi}_*$ with the acceptance probability given in equation (32.6), using $\lambda^2 = \lambda_{(r)}^2$ and $\boldsymbol{\pi} = \boldsymbol{\pi}_{(r-1)}$.
 - Draw a random number $u \in [0, 1]$. If $u < A(\boldsymbol{\pi}_{(r-1)}, \boldsymbol{\pi}_*)$, set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_*$. Otherwise leave the covariate set unchanged, i.e. set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_{(r-1)}$.
 - If the covariate set has changed, update \mathbf{X} and $\boldsymbol{\mu}$ correspondingly.
-

$Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N$ are the covariates X_1, \dots, X_n with $n = N - 1$. The marginal posterior probability $p_{j,i}$ that there is an edge from Z_j to Z_i is estimated by

$$\hat{p}_{j,i} := \hat{p}(Z_j \rightarrow Z_i | \mathcal{D}) = \frac{1}{R} \sum_{r=1}^R I_{\{Z_j \in \boldsymbol{\pi}_{i(r)}\}}(\boldsymbol{\pi}_{i(r)}),$$

where $\boldsymbol{\pi}_{i(r)}$ is the r th covariate set sampled for response Z_i . Given a threshold ϕ , a network prediction is obtained by extracting all edges whose probability exceeds ϕ . There is an edge from Z_j to Z_i if and only if $\hat{p}_{j,i} > \phi$.

32.2.3 Bayesian Piecewise Linear Regression (NH-DBN)

The Bayesian regression model in Section 32.2.2 corresponds to a *homogeneous* DBN, as it is based on the assumption that the regression coefficients β_0, \dots, β_k stay constant across time. Especially for applications in statistical genomics and systems biology, where one important objective is to learn gene regulatory networks from gene expression data, this assumption is often not appropriate. Many gene regulatory processes are subject to temporal changes, and the application of a homogeneous DBN can then lead to biased results and erroneous conclusions. It was therefore proposed (e.g. Lèbre *et al.*, 2010) to replace the linear regression model by a piecewise linear regression model. The key idea is to divide the T datapoints into H changepoint-separated disjoint segments, and to model the data within each segment h ($h = 1, \dots, H$) by a linear model with segment-specific regression coefficients $\beta_{h,0}, \dots, \beta_{h,k}$. Replacing the linear model by a piecewise linear model yields a *non-homogeneous* dynamic Bayesian network (NH-DBN).

As in Section 32.2.2, we first consider a generic regression model with response Y and $\boldsymbol{\pi} = \{X_1, \dots, X_k\}$ being the covariate set. Each datapoint D_t ($t = 1, \dots, T$) contains the values of the k covariates $x_{t,j}$ ($j = 1, \dots, k$) and a shifted response value y_{t+1} . Integer-valued changepoints $\tau_1 < \tau_2 < \dots < \tau_{H-1}$ with $\tau_1 > 2$ and $\tau_{H-1} < T + 1$ can be used to divide the datapoints into H segments. Datapoint D_t belongs to segment h if and only if $\tau_{h-1} \leq t < \tau_h$, where $\tau_0 := 2$ and $\tau_H := T + 1$ are two pseudo changepoints. The segmentation yields segment-specific response vectors, design matrices, and regression coefficient vectors:

$$\mathbf{y}_h = \begin{bmatrix} y_{\tau_{h-1}+1} \\ y_{\tau_{h-1}+2} \\ \vdots \\ y_{\tau_h} \end{bmatrix}, \quad \mathbf{X}_h = \begin{bmatrix} 1 & x_{\tau_{h-1},1} & \dots & x_{\tau_{h-1},k} \\ 1 & x_{\tau_{h-1}+1,1} & \dots & x_{\tau_{h-1}+1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{\tau_h-1,1} & \dots & x_{\tau_h-1,k} \end{bmatrix}, \quad \boldsymbol{\beta}_h = \begin{bmatrix} \beta_{h,0} \\ \beta_{h,1} \\ \vdots \\ \beta_{h,k} \end{bmatrix}.$$

We have not made explicit that the segmentations of \mathbf{y} , \mathbf{X} and $\boldsymbol{\mu}$ depend on the changepoint set $\tau := \{\tau_1, \dots, \tau_{H-1}\}$. We note that \mathbf{X}_h , $\boldsymbol{\beta}_h$ and $\boldsymbol{\mu}_h$ depend on $\boldsymbol{\pi}$.

In the piecewise linear model we use the likelihood from equation (32.1) for each segment,

$$\mathbf{y}_h | (\boldsymbol{\beta}_h, \sigma^2, \boldsymbol{\pi}, \tau) \sim \mathcal{N}(\mathbf{X}_h \boldsymbol{\beta}_h, \sigma^2 \mathbf{I}), \quad (32.8)$$

where σ^2 is the noise variance parameter which is here shared among segments, and for which we again assume $\sigma^{-2} \sim \text{GAM}(a_\sigma, b_\sigma)$. The regression coefficient vectors $\boldsymbol{\beta}_h$ are segment-specific and we use a multivariate Gaussian prior for each $\boldsymbol{\beta}_h$,

$$\boldsymbol{\beta}_h | (\sigma^2, \lambda^2, \boldsymbol{\pi}, \tau) \sim \mathcal{N}(\boldsymbol{\mu}_h, \sigma^2 \lambda^2 \mathbf{I}), \quad (32.9)$$

where $\boldsymbol{\mu}_h \in \mathbb{R}^{k+1}$ is the segment-specific prior expectation vector, whose $(j+1)$ th element is the prior expectation of $\beta_{h,j}$ ($j = 0, \dots, k$), and λ^2 is a ‘signal-to-noise ratio’ parameter, which is

shared among segments. Again we impose an inverse Gamma prior, $\lambda^{-2} \sim GAM(a_\lambda, b_\lambda)$. For the posterior distribution of the parameters we then have

$$p(\beta_1, \dots, \beta_h, \sigma^2, \lambda^2 | \mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\tau}) \propto \prod_{h=1}^H p(\mathbf{y}_h | \beta_h, \sigma^2, \boldsymbol{\pi}, \boldsymbol{\tau}) \\ \cdot \prod_{h=1}^H p(\beta_h | \sigma^2, \lambda^2, \boldsymbol{\pi}, \boldsymbol{\tau}) \cdot p(\sigma^2) \cdot p(\lambda^2).$$

For the marginal likelihood, with β_1, \dots, β_H , and σ^2 integrated out, again the marginalization rule from (Bishop, 2006, Section 2.3.7) can be applied:

$$p(\mathbf{y} | \lambda^2, \boldsymbol{\pi}, \boldsymbol{\tau}) = \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-\frac{T}{2}} \cdot (2b_\sigma)^{a_\sigma}}{\prod_{h=1}^H \det(\mathbf{C}_h)^{1/2}} \\ \cdot \left(2b_\sigma + \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)^T \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h) \right)^{-\left(\frac{T}{2} + a_\sigma\right)}, \quad (32.10)$$

where $\mathbf{C}_h := \mathbf{I} + \lambda^2 \mathbf{X}_h \mathbf{X}_h^T$ depends on λ^2 . The changepoint set $\boldsymbol{\tau}$ and the covariate set $\boldsymbol{\pi}$ together imply the segment-specific values \mathbf{y}_h , \mathbf{X}_h and $\boldsymbol{\mu}_h$ on the right-hand side.

A posterior sample can be obtained by MCMC simulations. Given a fixed changepoint set $\boldsymbol{\tau}$ with $H - 1$ changepoints, we iteratively resample the parameters using the pseudocode provided in Table 32.3. We obtain as output $(\boldsymbol{\pi}_{(r)}, \{\beta_h\}_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2)_{r=1, \dots, R}$, where $\{\beta_h\}_{(r)}$ is the set of the H segment-specific regression coefficient vectors sampled in iteration r .

Table 32.3 Pseudocode. Bayesian piecewise linear regression inference (for NH-DBNs), given a fixed changepoint set $\boldsymbol{\tau}_{(r)} = \boldsymbol{\tau}$

Initialization: For example, $\boldsymbol{\pi}_{(0)} = \boldsymbol{\pi}$, $\boldsymbol{\tau}_{(0)} = \boldsymbol{\tau}$, $\beta_{h,(0)} = \mathbf{0}$ for all h , and $\sigma_{(0)}^2 = \lambda_{(0)}^2 = 1$.

Iterations: For $r = 1, \dots, R$:

- (1b) Sample $\sigma_{(r)}^{-2}$ from $GAM\left(\alpha_\sigma + \frac{T}{2}, \beta_\sigma + \frac{1}{2} \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)^T \left(\mathbf{I} + \lambda_{(r-1)}^2 \mathbf{X}_h \mathbf{X}_h^T\right)^{-1} (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)\right)$. Invert $\sigma_{(r)}^{-2}$ to obtain $\sigma_{(r)}^2$.
- (2b) For $h = 1, \dots, H$, sample $\beta_{h,(r)}$ from $\mathcal{N}\left([\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1} (\lambda_{(r-1)}^{-2} \boldsymbol{\mu}_h + \mathbf{X}_h^T \mathbf{y}_h), \sigma_{(r)}^2 [\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1}\right)$,
- (3b) Sample $\lambda_{(r)}^{-2}$ from $GAM\left(\alpha_\lambda + \frac{(|\boldsymbol{\pi}_{(r-1)}|+1)}{2}, \beta_\lambda + \frac{1}{2} \sigma_{(r)}^{-2} \sum_{h=1}^H (\beta_{h,(r)} - \boldsymbol{\mu}_h)^T (\beta_{h,(r)} - \boldsymbol{\mu}_h)\right)$. Invert $\lambda_{(r)}^{-2}$ to obtain $\lambda_{(r)}^2$.
- (4b) Propose a new covariate set.
 - Randomly choose the move type (D, A or E) and the new set $\boldsymbol{\pi}_*$, as described in Section 32.2.2.
 - Propose to move from $\boldsymbol{\pi}_{(r-1)}$ to $\boldsymbol{\pi}_*$. Accept $\boldsymbol{\pi}_*$ with probability

$$A(\boldsymbol{\pi}_{(r-1)}, \boldsymbol{\pi}_*) = \min \left\{ 1, \frac{p(\mathbf{y} | \lambda_{(r)}^2, \boldsymbol{\pi}_*, \boldsymbol{\tau}_{(r-1)})}{p(\mathbf{y} | \lambda_{(r)}^2, \boldsymbol{\pi}_{(r-1)}, \boldsymbol{\tau}_{(r-1)})} \cdot \frac{p(\boldsymbol{\pi}_*)}{p(\boldsymbol{\pi}_{(r-1)})} \cdot HR \right\},$$
 where HR is defined in equation (32.6).
 - Draw a random number $u \in [0, 1]$. If $u < A(\boldsymbol{\pi}_{(r-1)}, \boldsymbol{\pi}_*)$, set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_*$. Otherwise leave the covariate set unchanged, i.e. set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_{(r-1)}$.
 - If the covariate set has changed, adapt the columns of the design matrices \mathbf{X}_h and prior expectation vectors $\boldsymbol{\mu}_h$, correspondingly.
- (5b) Keep the changepoint set $\boldsymbol{\tau}$ fixed, i.e. set $\boldsymbol{\tau}_{(r)} = \boldsymbol{\tau}$. Then the segmentations of \mathbf{y} , \mathbf{X} and the prior expectations $\boldsymbol{\mu}_h$ ($h = 1, \dots, H$) do not change.

If the number of changepoints and their locations are unknown, the changepoint set can be inferred from the data. We assume *a priori* that the distances between changepoints are geometrically distributed with hyperparameter $p \in (0, 1)$ and that there cannot be more than $H = 10$ segments. This implies for a changepoint set $\tau = \{\tau_1, \dots, \tau_{H-1}\}$ the prior probability

$$p(\tau|p) = \begin{cases} (1-p)^{\tau_H-\tau_{H-1}} \cdot \prod_{h=1}^{H-1} p \cdot (1-p)^{\tau_h-\tau_{h-1}-1}, & \text{if } |\tau| \leq 9 \text{ (i.e. } H \leq 10\text{),} \\ 0, & \text{if } |\tau| > 9 \text{ (i.e. } H > 10\text{).} \end{cases} \quad (32.11)$$

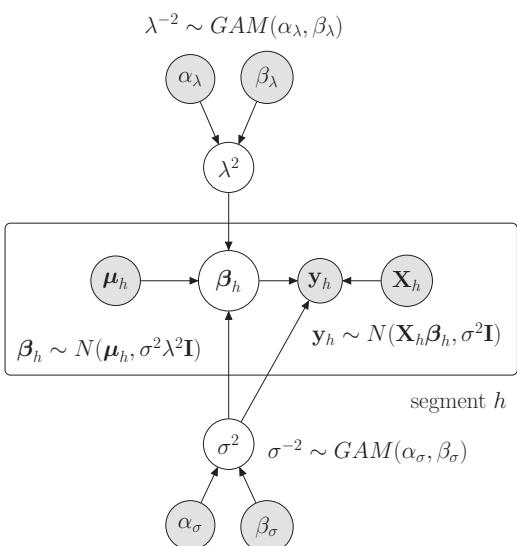
A drawback is that the hyperparameter p is fixed in advance, although it can have a strong effect on the number of inferred changepoints. The higher p , the more changepoints will be inferred. In the absence of any genuine prior knowledge about the number of changepoints, it is difficult to specify p appropriately. In practice, users often apply their models with different values for p and cross-compare the results. We do that as well in Section 32.3.2. We note that another solution is to impose a hyperprior distribution on p , so as to infer the best value from the data. The natural conjugate prior is a beta distribution.

A graphical model representation of the NH-DBN is provided in Figure 32.2. Using the marginal likelihood from equation (32.10), the posterior of the NH-DBN takes the form

$$p(\boldsymbol{\pi}, \tau, \lambda^2 | \mathbf{y}) \propto p(\mathbf{y} | \lambda^2, \boldsymbol{\pi}, \tau) \cdot p(\boldsymbol{\pi}) \cdot p(\tau) \cdot p(\lambda^2). \quad (32.12)$$

For the changepoint set inference typically three Metropolis–Hastings moves are used. The moves propose to replace the current changepoint set τ by a new changepoint set τ^* , where

Given the covariate set $\boldsymbol{\pi}$ and the changepoint set τ :



The covariate set $\boldsymbol{\pi}$ implies the columns of \mathbf{X}_h and so to which covariates the regression coefficients in $\boldsymbol{\beta}_h$ and their prior expectations in $\boldsymbol{\mu}_h$ refer.

All sets with up to \mathcal{F} covariates have equal prior probability:
 $p(\boldsymbol{\pi}) = \frac{1}{c}$ if $|\boldsymbol{\pi}| \leq \mathcal{F}$.

The changepoint set $\tau = \{\tau_1, \dots, \tau_{H-1}\}$ implies the segmentation of the design matrix \mathbf{X} into $\mathbf{X}_1, \dots, \mathbf{X}_H$ and the response vector \mathbf{y} into $\mathbf{y}_1, \dots, \mathbf{y}_H$ with segment-specific regression coefficient vectors $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_H$ having segment-specific prior expectation vectors $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_H$.

The prior for the distance $d_h := \tau_h - \tau_{h-1}$ between changepoints is: $d_h \sim GEO(p)$

Figure 32.2 Graphical model representation of the piecewise linear regression model (for NH-DBNs). The graph on the left shows the probabilistic relationships between the parameters and the data, given a fixed covariate set $\boldsymbol{\pi}$ and a fixed changepoint set τ . Parameters that have to be inferred are represented by white circles. The data and the fixed parameters are represented by grey circles. All nodes within the plate are segment-specific. The box on the right summarizes what the covariate set $\boldsymbol{\pi}$ and the changepoint set τ imply for the model.

Table 32.4 Pseudocode. Changepoint inference: new step (5b) for the pseudocode in Table 32.3

(5b) New step (5b) to infer the changepoint set:

- Randomly choose the move type (B, D or R) and the new set τ_* , as described in the main text.
- Propose to move from $\tau_{(r-1)}$ to τ_* . Accept τ_* with the acceptance probability given in equation (32.13), using $\lambda^2 = \lambda_{(r)}^2$, $\pi = \pi_{(r)}$ and $\tau = \tau_{(r-1)}$.
- Draw a random number $u \in [0, 1]$. If $u < A(\tau_{(r-1)}, \tau_*)$, set $\tau_{(r)} = \tau_*$, or otherwise leave the changepoint set unchanged, i.e. set $\tau_{(r)} = \tau_{(r-1)}$.
- If the changepoint set has changed, update the segmentations of the response vector \mathbf{y} and the design matrix \mathbf{X} , and use updated prior expectation vectors μ_1, \dots, μ_{H^*} .

τ^* implies a new data segmentation. The birth move (B) proposes to set a new changepoint at a randomly selected location. The new changepoint set τ^* then contains $H^* = H + 1$ changepoints. The new changepoint is located in a segment h and divides it into two sub-segments. The death move (D) randomly selects one changepoint $\tau \in \tau$ and deletes it. The new changepoint set τ^* then contains $H^* = H - 1$ changepoints. Removing a changepoint yields two adjacent segments that are merged into one single segment. The reallocation move (R) randomly selects one changepoint $\tau \in \tau$ and proposes to reallocate it to a randomly selected position in between the two surrounding changepoints. The reallocated changepoint yields new bounds for two consecutive segments.

When the type of move is randomly selected and the new candidate changepoint set τ_* is chosen as described above, the acceptance probability for the move from τ with segmentation $\mathbf{y}_1, \dots, \mathbf{y}_H$ to τ^* with segmentation $\mathbf{y}_1^*, \dots, \mathbf{y}_{H^*}^*$ is

$$A(\tau, \tau^*) = \min \left\{ 1, \frac{p(\mathbf{y} | \lambda^2, \pi, \tau^*)}{p(\mathbf{y} | \lambda^2, \pi, \tau)} \cdot \frac{p(\tau^*)}{p(\tau)} \cdot HR \right\}, \quad \text{where } HR = \begin{cases} \frac{T-1-|\tau^*|}{|\tau|}, & \text{for B,} \\ \frac{|\tau^*|}{T-1-|\tau|}, & \text{for D,} \\ 1, & \text{for R.} \end{cases} \quad (32.13)$$

T is the number of datapoints, and $|\cdot|$ denotes the cardinality.

If the segmentation has to be inferred from the data, step (5b) of the MCMC algorithm is replaced by a Metropolis–Hastings step on the covariate set; see the pseudocode in Table 32.4.

The output is a sample $(\pi_{(r)}, \tau_{(r)}, \{\beta_h\}_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2)_{r=1, \dots, R}$, where $\{\beta_h\}_{(r)}$ is the set of the $H_{(r)} := |\tau_{(r)}| + 1$ regression coefficient vectors sampled in iteration r . In a network domain, the marginal posterior probability that there is an edge from Z_j to Z_i is estimated by

$$\hat{p}_{j,i} := \hat{p}(Z_j \rightarrow Z_i | \mathcal{D}) = \frac{1}{R} \sum_{r=1}^R I_{\{Z_j \in \pi_{i,(r)}\}}(\pi_{i,(r)}),$$

where $\pi_{i,(r)}$ is the r th sampled covariate set for response Z_i .

32.2.4 Bayesian Piecewise Linear Regression with Coupled Regression Coefficients (Coupled NH-DBNs)

In many applications in systems biology the available time series are rather short (i.e. T is small), and the NH-DBN model in Section 32.2.3 tends to infer segmentations with short segments,

containing few datapoints only. Learning the regression coefficient vectors for each segment separately can then lead to inflated inference uncertainties. To address this bottleneck, coupling mechanisms between the segment-specific regression coefficient vectors have been proposed. For example, in Grzegorczyk and Husmeier (2012) it was proposed to sequentially couple the segment-specific regression coefficients, and in Grzegorczyk and Husmeier (2013) it was proposed to globally couple the segment-specific regression coefficients.

The sequential coupling mechanism is based on the assumption that the regression coefficients at any segment stay similar to those at the previous segment, that is, there is coupling between adjacent time segments. The globally coupled NH-DBN model, on the other hand, treats the individual segments as interchangeable units and couples all segment-specific regression coefficient vectors simultaneously. From this perspective, sequential coupling is more suited for time series where the regression coefficients are assumed to evolve gradually over time, while the global coupling concept is more flexible and can even be applied to couple a set of unordered time series (each representing a segment). The segments could then, for example, refer to data from experiments that have been performed under different experimental conditions. A principal shortcoming of the sequential coupling mechanism is that it imposes a systematic dependence between coupling strength and instability. The reason for this is that the sequential coupling mechanism is of the form of a Bayesian filter and corresponds to a diffusion process; see equations (32.16) and (32.17) below. Unlike the global coupling mechanism, the uncoupled NH-DBN in Section 32.2.3 is not the limiting case of the sequentially coupled model. As the strength of the sequential coupling decreases, the instability (diffusion) increases. For more detailed explanations we refer to Grzegorczyk and Husmeier (2013). For recently proposed improved sequential coupling mechanisms we refer to Shafiee Kamalabad and Grzegorczyk (2018).

32.2.4.1 NH-DBNs with Sequentially Coupled Regression Coefficients

In Grzegorczyk and Husmeier (2012) it was proposed to model dynamic networks with a piecewise linear model having sequentially coupled regression coefficient vectors. The key idea is that the posterior expectation $\tilde{\beta}_{h-1}$ of the regression coefficient vector β_{h-1} for segment $h-1$ is used as prior expectation μ_h for the next vector β_h . The coupling strength among the segment-specific vectors is regulated by a coupling parameter $\delta^2 > 0$. The likelihood from the NH-DBN in Section 32.2.3 can be re-employed,

$$\mathbf{y}_h | (\beta_h, \sigma^2, \pi, \tau) \sim \mathcal{N}(\mathbf{X}_h \beta_h, \sigma^2 \mathbf{I}), \quad (32.14)$$

but the prior for the segment-specific regression coefficient vectors is replaced by

$$\beta_h | (\sigma^2, \lambda^2, \delta^2, \pi, \tau) \sim \begin{cases} \mathcal{N}(\mathbf{0}, \sigma^2 \lambda^2 \mathbf{I}), & \text{if } h = 1, \\ \mathcal{N}(\tilde{\beta}_{h-1}, \sigma^2 \delta^2 \mathbf{I}), & \text{if } h > 1 \end{cases} \quad (h = 1, \dots, H), \quad (32.15)$$

where $\mathbf{0}$ is the zero vector and $\tilde{\beta}_{h-1}$ is the posterior expectation of β_{h-1} . That is, only the first segment $h = 1$ gets an uninformative prior with mean $\mu_1 := \mathbf{0}$ and covariance matrix $\Sigma_1 := \sigma^2 \lambda^2 \mathbf{I}$, while the subsequent segments $h > 1$ obtain informative priors with mean $\mu_h := \tilde{\beta}_{h-1}$ and covariance matrices $\Sigma_h := \sigma^2 \delta^2 \mathbf{I}$. We refer to the new parameter δ^2 as the coupling parameter. A low ‘signal-to-noise ratio’ parameter λ^2 yields a peaked prior for $h = 1$ in equation (32.15), and thus the distribution of β_1 is peaked around the zero vector. A low coupling parameter δ^2 yields a peaked prior for $h > 1$ in equation (32.15), and thus a strong coupling of β_h to the posterior expectation $\tilde{\beta}_{h-1}$ from the preceding segment. The vectors β_h and β_{h+1} will then tend to be similar.

The posterior distribution of β_h ($h = 1, \dots, H$) can be computed in closed form (cf. Grzegorczyk and Husmeier, 2012):

$$\beta_h | (\mathbf{y}_h, \sigma^2, \lambda^2, \delta^2, \boldsymbol{\pi}, \boldsymbol{\tau}) \sim \begin{cases} \mathcal{N}(\tilde{\Sigma}_1 \mathbf{X}_1^T \mathbf{y}_1, \sigma^2 \tilde{\Sigma}_1), & \text{if } h = 1, \\ \mathcal{N}(\tilde{\Sigma}_h (\delta^{-2} \tilde{\beta}_{h-1} + \mathbf{X}_h^T \mathbf{y}_h), \sigma^2 \tilde{\Sigma}_h), & \text{if } h \geq 2, \end{cases} \quad (32.16)$$

where $\tilde{\Sigma}_1 := [\lambda^{-2} \mathbf{I} + \mathbf{X}_1^T \mathbf{X}_1]^{-1}$ and, for $h \geq 2$, $\tilde{\Sigma}_h := [\delta^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1}$.

The posterior expectations in equation (32.16) are the prior expectations used in equation (32.15):

$$\tilde{\beta}_{h-1} := \begin{cases} [\lambda^{-2} \mathbf{I} + \mathbf{X}_1^T \mathbf{X}_1]^{-1} (\mathbf{X}_1^T \mathbf{y}_1), & \text{if } h = 2, \\ [\delta^{-2} \mathbf{I} + \mathbf{X}_{h-1}^T \mathbf{X}_{h-1}]^{-1} (\delta^{-2} \tilde{\beta}_{h-2} + \mathbf{X}_{h-1}^T \mathbf{y}_{h-1}), & \text{if } h \geq 3. \end{cases} \quad (32.17)$$

With an inverse gamma prior on the coupling parameter $\delta^2, \delta^{-2} \sim GAM(\alpha_\delta, \beta_\delta)$, the sequentially coupled NH-DBN is fully specified. A graphical model is provided in Figure 32.3.

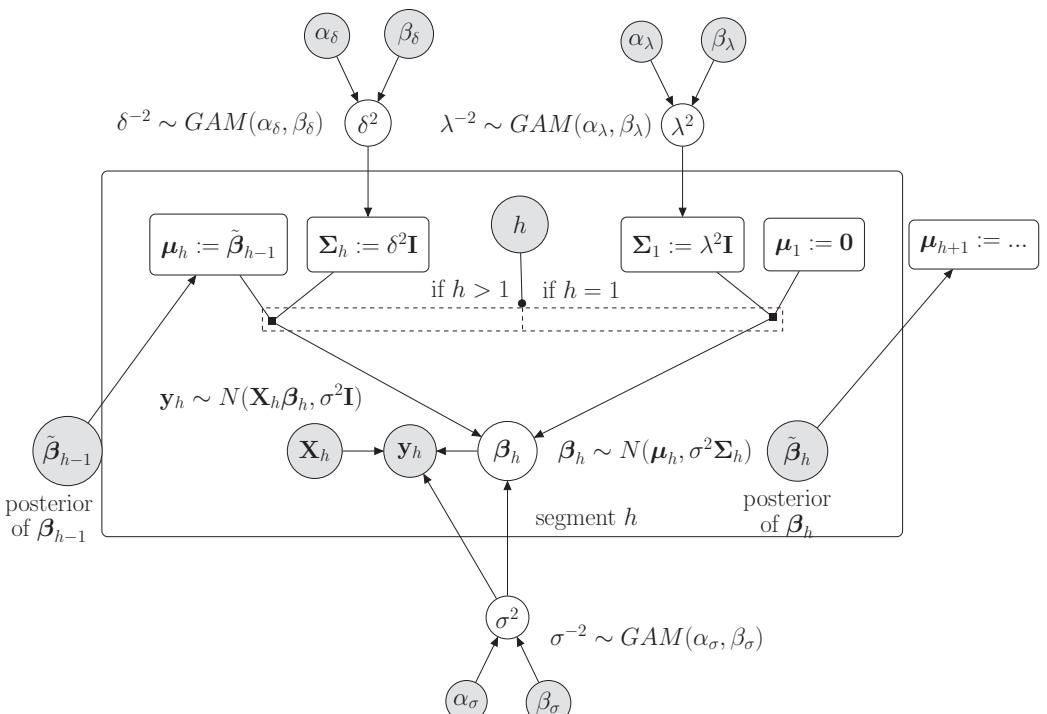


Figure 32.3 Graphical model of the sequentially coupled NH-DBN. Parameters that have to be inferred are in white circles. The data and the fixed parameters are in grey circles. The rectangles contain definitions which deterministically depend on the parent nodes. Circles within the plate are segment-specific. The prior for β_h depends on the segment h . The segment-specific instantiations of $\mathbf{y}_h, \mathbf{X}_h, \beta_h$ and $\tilde{\beta}_h$ depend on the covariate set $\boldsymbol{\pi}$ and the changepoint set $\boldsymbol{\tau}$ (not shown). The key idea is that the posterior expectation $\tilde{\beta}_{h-1}$ of the regression coefficient vector for segment $h - 1$ is used as prior expectation μ_h for the next vector β_h . The regression coefficients evolve gradually among the segments.

We have for the posterior,

$$\begin{aligned} p(\beta_1, \dots, \beta_H, \sigma^2, \lambda^2, \delta^2, \pi, \tau | \mathbf{y}) &\propto \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \beta_h, \pi, \tau) \cdot p(\beta_1 | \sigma^2, \lambda^2, \pi, \tau) \\ &\quad \cdot \prod_{h=2}^H p(\beta_h | \sigma^2, \delta^2, \pi, \tau) \cdot p(\sigma^2) \cdot p(\lambda^2) \cdot p(\delta^2). \end{aligned} \quad (32.18)$$

By application of the marginalization rule we obtain

$$\begin{aligned} p(\mathbf{y} | \lambda^2, \delta^2, \pi, \tau) &= \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-T/2} (2b_\sigma)^{a_\sigma}}{\left(\prod_{h=1}^H \det(\mathbf{C}_h)\right)^{1/2}} \\ &\quad \cdot \left(2b_\sigma + \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \tilde{\beta}_{h-1})^T \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \tilde{\beta}_{h-1})\right)^{-\left(\frac{T}{2} + a_\sigma\right)}, \end{aligned} \quad (32.19)$$

where $\tilde{\beta}_0 := \mathbf{0}$, $\tilde{\beta}_1, \dots, \tilde{\beta}_{H-1}$ are defined in equation (32.17) and

$$\mathbf{C}_h := \begin{cases} \mathbf{I} + \lambda^2 \mathbf{X}_1 \mathbf{X}_1^T, & \text{if } h = 1, \\ \mathbf{I} + \delta^2 \mathbf{X}_h \mathbf{X}_h^T, & \text{if } h > 1. \end{cases}$$

A sample $(\tau_{(r)}, \pi_{(r)}, \{\beta_h\}_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2, \delta_{(r)}^2)_{r=1, \dots, R}$ from the posterior distribution can be generated by MCMC simulations; see the pseudocode in Table 32.5. In a network domain, the marginal posterior probability that there is an edge from Z_j to Z_i is estimated by

$$\hat{p}_{j,i} := \hat{p}(Z_j \rightarrow Z_i | \mathcal{D}) = \frac{1}{R} \sum_{r=1}^R I_{\{Z_j \in \pi_{i,(r)}\}}(\pi_{i,(r)}),$$

where $\pi_{i,(r)}$ is the r th sampled covariate set for response Z_i .

32.2.4.2 NH-DBNs with Globally Coupled Regression Coefficients

In Grzegorczyk and Husmeier (2013) a global coupling scheme for the regression coefficient vectors was proposed. The key idea is to treat the segments $h = 1, \dots, H$ as interchangeable units and to impose a shared hyperprior onto the prior expectations of the segment-specific regression coefficient vectors. This is a straightforward extension of the NH-DBN in Section 32.2.3. In equation (32.9) we set $\mu_h = \mu$,

$$\beta_h | (\sigma^2, \lambda^2, \pi, \tau) \sim \mathcal{N}(\mu, \sigma^2 \lambda^2 \mathbf{I}), \quad (32.20)$$

and instead of fixing μ we make μ a free hyperparameter vector with hyperprior

$$\mu \sim \mathcal{N}(\mu^\dagger, \Sigma^\dagger). \quad (32.21)$$

For the posterior distribution of the parameters we then have

$$\begin{aligned} p(\beta_1, \dots, \beta_H, \sigma^2, \lambda^2, \mu | \mathbf{y}, \pi, \tau) &\propto \prod_{h=1}^H p(\mathbf{y}_h | \beta_h, \sigma^2, \pi, \tau) \\ &\quad \cdot \prod_{h=1}^H p(\beta_h | \sigma^2, \lambda^2, \mu, \pi, \tau) \cdot p(\sigma^2) \cdot p(\lambda^2) \cdot p(\mu), \end{aligned}$$

Table 32.5 Pseudocode. NH-DBN with sequentially coupled regression coefficients

Initialization: For example, $\boldsymbol{\pi}_{(0)} = \boldsymbol{\pi}$, $\boldsymbol{\tau}_{(0)} = \boldsymbol{\tau}$, $\boldsymbol{\beta}_{h,(0)} = \mathbf{0}$ for all h , and $\sigma_{(0)}^2 = \lambda_{(0)}^2 = \delta_{(0)}^2 = 1$.

Iterations: For $r = 1, \dots, R$:

- (1c) Sample $\sigma_{(r)}^{-2}$ from $GAM\left(\alpha_\sigma + \frac{1}{2} \cdot T, \beta_\sigma + \frac{1}{2} \cdot \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1})^T \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \tilde{\boldsymbol{\beta}}_{h-1})\right)$. Invert $\sigma_{(r)}^{-2}$ to obtain $\sigma_{(r)}^2$.
- (2c) Sample $\boldsymbol{\beta}_{1,(r)}$ from $\mathcal{N}\left([\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1} \mathbf{X}_h^T \mathbf{y}_h, \sigma_{(r)}^2 [\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1}\right)$. For $h = 2, \dots, H$, sample $\boldsymbol{\beta}_{h,(r)}$ from $\mathcal{N}\left([\delta_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1} (\delta_{(r-1)}^{-2} \tilde{\boldsymbol{\beta}}_{h-1} + \mathbf{X}_h^T \mathbf{y}_h), \sigma_{(r)}^2 [\delta_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1}\right)$.
- (3c) Sample $\lambda_{(r)}^{-2}$ from $GAM\left(\alpha_\lambda + \frac{|\boldsymbol{\pi}_{(r-1)}|+1}{2}, \beta_\lambda + \frac{1}{2} \cdot \sigma_{(r)}^{-2} \cdot \boldsymbol{\beta}_{1,(r)}^T \boldsymbol{\beta}_{1,(r)}\right)$ and $\delta_{(r)}^{-2}$ from $GAM\left(\alpha_\delta + \frac{(H-1)\cdot(|\boldsymbol{\pi}_{r-1}|+1)}{2}, \beta_\delta + \frac{1}{2} \cdot \sigma_{(r)}^{-2} \cdot \sum_{h=2}^H (\boldsymbol{\beta}_{h,(r)} - \tilde{\boldsymbol{\beta}}_{h-1})^T (\boldsymbol{\beta}_{h,(r)} - \tilde{\boldsymbol{\beta}}_{h-1})\right)$. Invert $\lambda_{(r)}^{-2}$ and $\delta_{(r)}^{-2}$ to obtain $\lambda_{(r)}^2$ and $\delta_{(r)}^2$.
- (4c) Propose a new covariate set:
 - Randomly choose the type of move (D, A or E) and the new set $\boldsymbol{\pi}_*$, as described in Section 32.2.2.
 - Propose to move from $\boldsymbol{\pi}_{(r-1)}$ to $\boldsymbol{\pi}_*$. Accept $\boldsymbol{\pi}_*$ with probability

$$A(\boldsymbol{\pi}_{(r-1)}, \boldsymbol{\pi}_*) = \min \left\{ 1, \frac{p(\mathbf{y} | \lambda_{(r)}^2, \delta_{(r)}^2, \boldsymbol{\pi}_*, \boldsymbol{\tau}_{(r-1)})}{p(\mathbf{y} | \lambda_{(r)}^2, \delta_{(r)}^2, \boldsymbol{\pi}_{(r-1)}, \boldsymbol{\tau}_{(r-1)})} \cdot \frac{p(\boldsymbol{\pi}_*)}{p(\boldsymbol{\pi}_{(r-1)})} \cdot HR \right\}$$
, where HR was defined in equation (32.6).
 - Draw a random number $u \in [0, 1]$. If $u < A(\boldsymbol{\pi}_{(r-1)}, \boldsymbol{\pi}_*)$, set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_*$. Otherwise leave the covariate set unchanged, i.e. set $\boldsymbol{\pi}_{(r)} = \boldsymbol{\pi}_{(r-1)}$.
 - If the covariate set has changed, adapt the columns of the design matrices \mathbf{X}_h and prior expectation vectors $\boldsymbol{\mu}_h$ correspondingly.
- (5c) Propose a new changepoint set:
 - Randomly choose the type of move (B, D or R) and the new set $\boldsymbol{\tau}_*$, as described in Section 32.2.3.
 - Propose to move from $\boldsymbol{\tau}_{(r-1)}$ to $\boldsymbol{\tau}_*$. Accept $\boldsymbol{\tau}_*$ with probability

$$A(\boldsymbol{\tau}_{(r-1)}, \boldsymbol{\tau}_*) = \min \left\{ 1, \frac{p(\mathbf{y} | \lambda_{(r)}^2, \delta_{(r)}^2, \boldsymbol{\pi}_{(r)}, \boldsymbol{\tau}_*)}{p(\mathbf{y} | \lambda_{(r)}^2, \delta_{(r)}^2, \boldsymbol{\pi}_{(r)}, \boldsymbol{\tau}_{(r-1)})} \cdot \frac{p(\boldsymbol{\tau}_*)}{p(\boldsymbol{\tau}_{(r-1)})} \cdot HR \right\}$$
, where HR was defined in equation (32.13).
 - Draw a random number $u \in [0, 1]$. If $u < A(\boldsymbol{\tau}_{(r-1)}, \boldsymbol{\tau}_*)$, set $\boldsymbol{\tau}_{(r)} = \boldsymbol{\tau}_*$, or otherwise leave the changepoint set unchanged, i.e. set $\boldsymbol{\tau}_{(r)} = \boldsymbol{\tau}_{(r-1)}$.
 - If the changepoint set has changed, update the segmentations of the response vector \mathbf{y} , the design matrix \mathbf{X} , and prior expectation vectors $\boldsymbol{\mu}$ accordingly.

and for the marginal likelihood it follows from equation (32.10) that

$$p(\mathbf{y} | \lambda^2, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\tau}) = \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-\frac{T}{2}} \cdot (2b_\sigma)^{a_\sigma}}{\prod_{h=1}^H \det(\mathbf{C}_h)^{1/2}} \cdot \left(2b_\sigma + \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu})^T \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu})\right)^{-\left(\frac{T}{2} + a_\sigma\right)}, \quad (32.22)$$

where $\mathbf{C}_h := \mathbf{I} + \lambda^2 \mathbf{X}_h \mathbf{X}_h^T$. A graphical model representation of the NH-DBN model with globally coupled regression coefficients is provided in Figure 32.4.

As shown in Grzegorczyk and Husmeier (2013), $\boldsymbol{\mu}$ can be efficiently sampled with a collapsed Gibbs sampling step. The full conditional distribution with $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_H$ integrated out is

$$p(\boldsymbol{\mu} | \sigma^2, \lambda^2, \boldsymbol{\pi}, \boldsymbol{\tau}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}^\ddagger, \boldsymbol{\Sigma}^\ddagger), \quad (32.23)$$

where

$$\boldsymbol{\mu}^\ddagger := \boldsymbol{\Sigma}^\ddagger \left(\sum_{h=1}^H \mathbf{X}_h^T [\sigma^2 \mathbf{I} + \sigma^2 \lambda^2 \mathbf{X}_h \mathbf{X}_h^T]^{-1} \mathbf{y}_h + [\boldsymbol{\Sigma}^\ddagger]^{-1} \boldsymbol{\mu}^\dagger \right),$$

$$\boldsymbol{\Sigma}^\ddagger := \left(\sum_{h=1}^H \mathbf{X}_h^T [\sigma^2 \mathbf{I} + \sigma^2 \lambda^2 \mathbf{X}_h \mathbf{X}_h^T]^{-1} \mathbf{X}_h + [\boldsymbol{\Sigma}^\ddagger]^{-1} \right)^{-1},$$

the changepoint set τ implies the segmentation of \mathbf{y} and \mathbf{X} , and the structure of \mathbf{X}_h depends on the covariate set π .

For model inference Grzegorczyk and Husmeier (2013) proposed to apply the concept of ‘blocking’. Blocking is a sampling technique by which correlated variables are merged into blocks and sampled together. We follow Grzegorczyk and Husmeier (2013) and form two blocks. We group the covariate set π with μ , and we group the changepoint set τ with μ , and proceed as follows.

When a move on π is performed, a new covariate set π_* is proposed, and for π_* directly a new μ_* is sampled from $p(\mu|\sigma^2, \lambda^2, \pi_*, \tau, \mathbf{y})$. The move proposes to replace the current block $[\pi, \mu]$ by the new block $[\pi_*, \mu_*]$, that is, to update π and μ together. The acceptance probability is

$$A([\pi, \mu], [\pi_*, \mu_*]) = \min \left\{ 1, \frac{p(\mathbf{y}|\lambda^2, \mu_*, \pi_*, \tau)}{p(\mathbf{y}|\lambda^2, \mu, \pi, \tau)} \cdot \frac{p(\pi_*)}{p(\pi)} \cdot \frac{p(\mu_*)}{p(\mu)} \cdot \frac{p(\mu|\sigma^2, \lambda^2, \pi, \tau, \mathbf{y})}{p(\mu_*|\sigma^2, \lambda^2, \pi_*, \tau, \mathbf{y})} \cdot HR \right\}, \quad (32.24)$$

where HR is defined in equation (32.6).

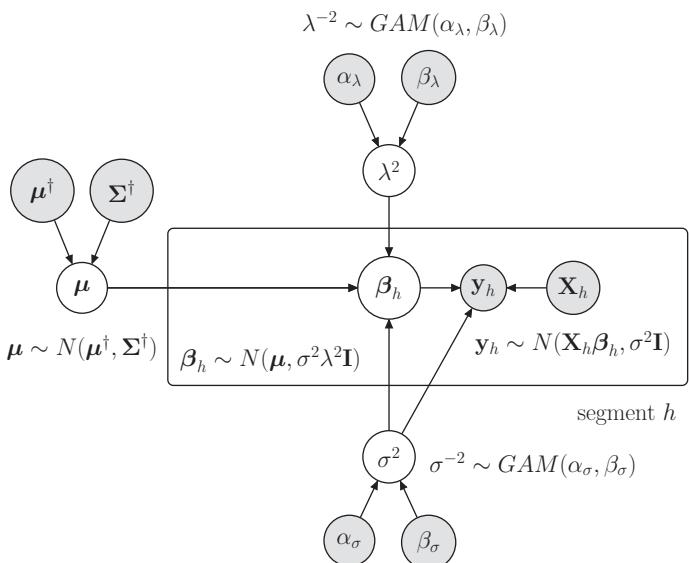


Figure 32.4 Graphical model of the globally coupled NH-DBN. Parameters that have to be inferred are in white circles. The data and the fixed parameters are in grey circles. Circles within the plate are segment-specific. The segment-specific instantiations of \mathbf{y}_h , \mathbf{X}_h , β_h and $\tilde{\beta}_h$ depend on the covariate set π and the changepoint set τ (not shown). Unlike for the sequentially coupled NH-DBN in Figure 32.3, the segments are treated as interchangeable units, and all segment-specific regression coefficient vectors β_1, \dots, β_H are coupled simultaneously.

Correspondingly, when a move on τ is performed, for the new covariate set τ_* a new μ_* is sampled from $p(\mu|\sigma^2, \lambda^2, \pi, \tau_*, \mathbf{y})$. The move proposes to replace $[\tau, \mu]$ by $[\tau_*, \mu_*]$. The acceptance probability is

$$A([\tau, \mu], [\tau_*, \mu_*]) = \min \left\{ 1, \frac{p(\mathbf{y}|\lambda^2, \mu_*, \pi, \tau_*)}{p(\mathbf{y}|\lambda^2, \mu, \pi, \tau)} \cdot \frac{p(\tau_*)}{p(\tau)} \cdot \frac{p(\mu_*)}{p(\mu)} \cdot \frac{p(\mu|\sigma^2, \lambda^2, \pi, \tau, \mathbf{y})}{p(\mu_*|\sigma^2, \lambda^2, \pi, \tau_*, \mathbf{y})} \cdot HR \right\}, \quad (32.25)$$

where HR is defined in equation (32.13).

A sample $(\tau_{(r)}, \pi_{(r)}, \{\beta_h\}_{(r)}, \sigma_{(r)}^2, \lambda_{(r)}^2, \mu_{(r)})_{r=1,\dots,R}$ from the posterior can be generated by MCMC simulations; see the pseudocode given in Table 32.6. In a network domain, again the marginal posterior probability that there is an edge from Z_j to Z_i is estimated by

$$\hat{p}_{j,i} := \hat{p}(Z_j \rightarrow Z_i | \mathcal{D}) = \frac{1}{R} \sum_{r=1}^R I_{\{Z_j \in \pi_{i,(r)}\}}(\pi_{i,(r)}),$$

where $\pi_{i,(r)}$ is the r th sampled covariate set for response Z_i .

Table 32.6 Pseudo code. NH-DBN with globally coupled regression coefficients

Initialization: $\pi_{(0)} = \pi$, $\tau_{(0)} = \tau$, $\beta_{h,(0)} = \mathbf{0}$ for all h , $\sigma_{(0)}^2 = \lambda_{(0)}^2 = 1$, and $\mu = \mathbf{0}$.

Iterations: For $r = 1, \dots, R$:

- (1d) Sample $\sigma_{(r)}^{-2}$ from $GAM\left(\alpha_\sigma + \frac{T}{2}, \beta_\sigma + \frac{1}{2} \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \mu_{(r-1)})^T \left(\mathbf{I} + \lambda_{(r-1)}^2 \mathbf{X}_h \mathbf{X}_h^T\right)^{-1} (\mathbf{y}_h - \mathbf{X}_h \mu_{(r-1)})\right)$. Invert $\sigma_{(r)}^{-2}$ to obtain $\sigma_{(r)}^2$.
- (2d) For each segment h implied by $\tau_{(r-1)}$, sample $\beta_{h,(r)}$ from $\mathcal{N}\left([\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1} (\lambda_{(r-1)}^{-2} \mu_{(r-1)} + \mathbf{X}_h^T \mathbf{y}_h), \sigma_{(r)}^2 [\lambda_{(r-1)}^{-2} \mathbf{I} + \mathbf{X}_h^T \mathbf{X}_h]^{-1}\right)$.
- (3d) Sample $\lambda_{(r)}^{-2}$ from $GAM\left(\alpha_\lambda + \frac{(|\pi_{(r-1)}|+1)}{2}, \beta_\lambda + \frac{1}{2} \sigma_{(r)}^{-2} \sum_{h=1}^H (\beta_{h,(r)} - \mu_{(r-1)})^T (\beta_{h,(r)} - \mu_{(r-1)})\right)$. Invert $\lambda_{(r)}^{-2}$ to obtain $\lambda_{(r)}^2$.
- (4d) Propose a new covariate set.
 - Randomly choose the type of move (D, A or E) and the new set π_* , as described in Section 32.2.1.
 - Propose to move from $[\pi_{(r-1)}, \mu_{(r-1)}]$ to $[\pi_*, \mu_*]$, where μ_* is sampled from $p(\mu|\sigma_{(r)}^2, \lambda_{(r)}^2, \pi_*, \tau_{(r-1)}, \mathbf{y})$.
 - Compute the acceptance probability with equation (32.24) using $\pi = \pi_{(r-1)}$, $\tau = \tau_{(r-1)}$, $\mu = \mu_{(r-1)}$, $\sigma^2 = \sigma_{(r)}^2$ and $\lambda^2 = \lambda_{(r)}^2$.
 - Draw a random number $u \in [0, 1]$. If $u < A([\pi_{(r-1)}, \mu_{(r-1)}], [\pi_*, \mu_*])$, set $\pi_{(r)} = \pi_*$ and $\mu_{(r)} = \mu_*$, otherwise set $\pi_{(r)} = \pi_{(r-1)}$ and $\mu_{(r)} = \mu_{(r-1)}$.
 - If the covariate set has changed, adapt the columns of \mathbf{X}_h correspondingly.
- (5d) Propose a new changepoint set:
 - Randomly choose the type of move (B, D or R) and the new set τ_* , as described in Section 32.2.3.
 - Propose to move from $[\tau_{(r-1)}, \mu_{(r-1)}]$ to $[\tau_*, \mu_*]$, where μ_* is sampled from $p(\mu|\sigma_{(r)}^2, \lambda_{(r)}^2, \pi, \tau_*, \mathbf{y})$.
 - Compute the acceptance probability with equation (32.25) using $\pi = \pi_{(r)}$, $\tau = \tau_{(r-1)}$, $\mu = \mu_{(r-1)}$, $\sigma^2 = \sigma_{(r)}^2$ and $\lambda^2 = \lambda_{(r)}^2$.
 - Draw a random number $u \in [0, 1]$. If $u < A([\tau_{(r-1)}, \mu_{(r-1)}], [\tau_*, \mu_*])$, set $\tau_{(r)} = \tau_*$ and $\mu_{(r)} = \mu_*$, otherwise set $\tau_{(r)} = \tau_{(r-1)}$ and $\mu_{(r)} = \mu_{(r-1)}$.
 - If the changepoint set has changed, update the segmentation of \mathbf{y} and \mathbf{X} accordingly.

32.2.5 NH-DBNs with More Flexible Allocation Schemes

Up to now we have used changepoint sets $\tau = \{\tau_1, \dots, \tau_{H-1}\}$ to divide T temporally ordered datapoints D_1, \dots, D_T into H segments. The changepoint set implicitly defines a latent allocation vector V of length T whose t th element $V_t \in \{1, \dots, H\}$ is a latent variable whose value is equal to h if datapoint D_t belongs to segment h , that is, if $\tau_{h-1} \leq t < \tau_h$. With changepoints the configuration space of the allocation vector is very restricted. In the broader context of mixture models the segments refer to components and we have the condition $V_t \leq V_{t+1}$. Hence, a component once left cannot be revisited; for example, an allocation vector of the form $V = (1, 1, 1, 2, 2, 2, 1, 1)^T$ is not part of the configuration space. It was therefore also proposed in the literature to combine Bayesian networks with mixture models rather than changepoint processes (see, for example, Ko *et al.*, 2007; Grzegorczyk *et al.*, 2008). A mixture model with H components allows for an unrestricted free allocation of the datapoints D_1, \dots, D_T to the H components. This substantially increases the configuration space of the possible data segmentations. But, on the other hand, the configuration space extension comes with the disadvantage that the temporal order of the datapoints is no longer taken into account.

In a Bayesian mixture model it is usually assumed that the number of components H follows a Poisson distribution, $H \sim POI(\lambda)$. Subsequently, given H , a categorical distribution with hyperparameter vector $p = (p_1, \dots, p_H)^T$ is used for the allocation of the datapoints. p_h is the prior probability that any datapoint D_t belongs to component h , symbolically $p_h := p(V_t = h | H)$ for all $t = 1, \dots, T$. The probability of a given allocation vector V is then

$$p(V|p, H) = \prod_{h=1}^H p_h^{n_h},$$

where n_h is the number of datapoints that are allocated to component h by V , that is, the cardinality of the set $\{t : V_t = h\}$. Using a conjugate Dirichlet prior with hyperparameters $\alpha = (\alpha_1, \dots, \alpha_H)^T$ for p , the marginal probability of V is

$$p(V|H) = \int_p p(V|p, H)p(p|H) dp = \frac{\Gamma(\sum_{h=1}^H \alpha_h)}{\Gamma(\sum_{h=1}^H (\alpha_h + n_h))} \cdot \prod_{h=1}^H \frac{\Gamma(\alpha_h + n_h)}{\Gamma(\alpha_h)}.$$

The mixture model can be combined with the standard NH-DBN model in Section 32.2.3, or with the globally coupled NH-DBN in Section 32.2.4.2. The sequential coupling scheme in Section 32.2.4.1 cannot be used, as there is no natural order of the H components. When combined with the standard NH-DBN, the posterior is

$$\begin{aligned} p(\beta_1, \dots, \beta_H, \sigma^2, \lambda^2, \delta^2, \pi, V, H | y) &\propto \prod_{h=1}^H p(y_h | \sigma^2, \beta_h, \pi, V) \\ &\quad \cdot \prod_{h=1}^H p(\beta_h | \sigma^2, \delta^2, \pi, V) \cdot p(\sigma^2) \cdot p(\lambda^2) \cdot p(\delta^2) \cdot p(H), \end{aligned} \tag{32.26}$$

and the MCMC inference algorithm from Tables 32.3 and 32.4 has to be modified as follows:

1. The component-specific response vectors y_h and design matrices X_h have to be built from those datapoints D_t that are allocated to component h . Recalling that each D_t contains values of the potential covariates $x_{t,1}, \dots, x_{t,n}$ and a shifted response value y_{t+1} (e.g. y_h is the vector of those response values y_{t+1} with $V_t = h$), the segment-specific design matrix X_h has to be built accordingly from the covariate values $x_{t,1}, \dots, x_{t,n}$ with $V_t = h$.

2. Instead of the changepoint MCMC inference (step 5b), there is now need to infer the number of mixture components H and the allocation vector \mathbf{V} . For mixture models Nobile and Fearnside (2007) proposed the ‘allocation sampler’ as an alternative to computationally expensive reversible-jump Markov chain Monte Carlo sampling schemes (Green, 1995). The allocation sampler consists of different types of move. The ‘sweep’ move and the moves M1–M3 propose to leave the number of components H unchanged and to change the allocation vector only, that is, to replace the current vector \mathbf{V} by a new one, \mathbf{V}_* and to keep H unchanged. The ejection and the absorption move propose to change the number of components and the allocation vector, that is, to move from $[\mathbf{V}, H]$ to $[\mathbf{V}_*, H_*]$ where $H_* = H + 1$ (ejection move) or $H_* = H - 1$ (absorption move). For lack of space we refer to Nobile and Fearnside (2007) for the details.

A third possibility, proposed, for example, in Thorne and Stumpf (2012) and Grzegorczyk (2016), is to combine DBNs with hidden Markov models. Homogeneous HMMs and mixture models have the same unrestricted configuration space of the allocation vector \mathbf{V} . But unlike mixture models, homogeneous HMMs do not ignore the temporal order of the datapoints. With HMMs it can be taken into account that neighbouring datapoints should be more likely to belong to the same component than distant ones. In many real-world applications the allocation vector $\mathbf{V} = (1, 1, 2, 2, 2, 1, 1)^T$ should be more likely *a priori* than $\mathbf{V} = (1, 1, 2, 2, 1, 1, 2)^T$. Both allocation vectors cannot be created by changepoints. In a mixture model both vectors have the same prior probability. HMMs offer an appropriate trade-off: both vectors belong to the configurations space and the second one should be less likely. For non-homogeneous DBNs based on homogeneous HMMs we refer to Thorne and Stumpf (2012) and Grzegorczyk (2016).

32.2.6 NH-DBNs with Time-Varying Network Structures

In the NH-DBN proposed by Lèbre *et al.* (2010) even the covariate sets are allowed to vary over time. This is highly relevant to embryogenesis and morphogenesis, where the gene regulatory network structure is likely to change across different stages of an organism’s life cycle. The segment-specific regression coefficient vectors β_1, \dots, β_H stay uncoupled, as for the NH-DBN in Section 32.2.3, but for each time segment $h = 1, \dots, H$ there is now a segment-specific covariate set π_h that has to be inferred from the data. As the covariate sets vary from segment to segment, the columns of the segment-specific design matrices \mathbf{X}_h ($h = 1, \dots, H$) refer to segment-specific covariates (e.g. \mathbf{X}_h has $1 + |\pi_h|$ columns); after a first column of 1s for the intercept, there is one column for each covariate in π_h . The regression coefficients in the h th regression coefficient vector $\beta_h = (\beta_{0,h}, \dots, \beta_{|\pi_h|+1,h})^T$ refer to the covariates in π_h . $\beta_{i,h}$ is the regression coefficient for the i th covariate in π_h ($i = 1, \dots, |\pi_h|$). The posterior of the model has the form

$$p(\{\pi_h, \beta_h\}_{h=1,\dots,H}, \sigma^2, \lambda^2, \tau | \mathbf{y}) \propto \prod_{h=1}^H p(\mathbf{y}_h | \sigma^2, \beta_h, \pi_h) \cdot \prod_{h=1}^H p(\beta_h | \sigma^2, \lambda^2, \pi_h) \\ \cdot p(\tau) \cdot p(\sigma^2) \cdot p(\lambda^2) \cdot \prod_{h=1}^H p(\pi_h), \quad (32.27)$$

where π_h is the covariate set for segment h and β_h is the corresponding regression coefficient vector of length $|\pi_h| + 1$. To avoid model over-flexibility, Lèbre *et al.* (2010) use a restrictive

prior for $\boldsymbol{\pi}_h$ which penalizes covariate sets with many covariates; see Lèbre *et al.* (2010) for details. The marginal likelihood, with β_1, \dots, β_H , and σ^2 integrated out, is

$$p(\mathbf{y}|\lambda^2, \{\boldsymbol{\pi}_h\}_{h=1,\dots,H}, \boldsymbol{\tau}) = \frac{\Gamma\left(\frac{T}{2} + a_\sigma\right)}{\Gamma(a_\sigma)} \cdot \frac{\pi^{-\frac{T}{2}} \cdot (2b_\sigma)^{a_\sigma}}{\prod_{h=1}^H \det(\mathbf{C}_h)^{1/2}} \cdot \left(2b_\sigma + \sum_{h=1}^H (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h)^T \mathbf{C}_h^{-1} (\mathbf{y}_h - \mathbf{X}_h \boldsymbol{\mu}_h) \right)^{-\left(\frac{T}{2} + a_\sigma\right)}, \quad (32.28)$$

where $\mathbf{C}_h := \mathbf{I} + \lambda \mathbf{X}_h \mathbf{X}_h^T$. The design matrix \mathbf{X}_h and the prior expectation vector $\boldsymbol{\mu}_h$ both depend on the segment-specific covariate sets $\boldsymbol{\pi}_h$.

A slightly modified version of the MCMC algorithm provided in Tables 32.3 and 32.4 can be used for model inference. The two required modifications are as follows:

- As the covariate sets are now segment-specific, step (4b) has to be replaced by a segment-specific covariate set move. The new move randomly selects one segment h_\star from the current segments $h = 1, \dots, H_{(r-1)}$ and proposes to replace the current segment-specific covariate set $\boldsymbol{\pi}_{h_\star,(r-1)}$ by a new covariate set $\boldsymbol{\pi}_{h_\star,\star}$. The acceptance probability is

$$A(\boldsymbol{\pi}_{h,(r-1)}, \boldsymbol{\pi}_{h,\star}) = \min \left\{ 1, \frac{p(\mathbf{y}|\lambda_{(r)}^2, \{\boldsymbol{\pi}_{h,\star}\}_{h=1,\dots,H_{(r-1)}}, \boldsymbol{\tau}_{(r-1)})}{p(\mathbf{y}|\lambda_{(r)}^2, \{\boldsymbol{\pi}_{h,(r-1)}\}_{h=1,\dots,H_{(r-1)}}, \boldsymbol{\tau}_{(r-1)})} \cdot \frac{p(\boldsymbol{\pi}_{h,\star})}{p(\boldsymbol{\pi}_{h,(r-1)})} \cdot HR \right\},$$

where $\boldsymbol{\pi}_{h,\star} = \boldsymbol{\pi}_{h,(r-1)}$ for all $h \neq h_\star$, and the Hastings ratio HR , defined in Equation (32.6), is computed for the segment-specific covariate sets (i.e. with $\boldsymbol{\pi} = \boldsymbol{\pi}_{h,(r-1)}$ and $\boldsymbol{\pi}_\star = \boldsymbol{\pi}_{h,\star}$). If the move is accepted, we set $\boldsymbol{\pi}_{h,(r)} = \boldsymbol{\pi}_{h,(r-1)}$ for all $h \neq h_\star$ and $\boldsymbol{\pi}_{h_\star,(r)} = \boldsymbol{\pi}_{h_\star,\star}$.

- The changepoint birth and death move in step (5b) also has to be modified. In a changepoint death move a changepoint τ from the current changepoint set $\boldsymbol{\tau}_{(h-1)}$ is randomly selected and the move proposes to delete it. This yields a new candidate changepoint set $\boldsymbol{\tau}_\star$. The changepoint τ separates two neighbouring segments h and $h+1$, and with its deletion the two segments will be merged. This means that one of the two segment-specific covariate sets $\boldsymbol{\pi}_h$ and $\boldsymbol{\pi}_{h+1}$ becomes redundant. In the move designed by Lèbre *et al.* (2010) one of the two covariate sets is randomly chosen to become the covariate set of the merged segment. That is, along with the removal of a changepoint the move also proposes to replace the current set of covariate sets $\{\boldsymbol{\pi}_{h,(r)}\}_{h=1,\dots,H_{(r-1)}}$ by a new set of covariate sets $\{\boldsymbol{\pi}_{h,\star}\}_{h=1,\dots,H_\star}$, where $H_\star = H_{(r-1)} - 1$.

In the changepoint birth move a new changepoint is placed on a randomly selected location. This yields a segment h with segment-specific covariate set $\boldsymbol{\pi}_{h,(r-1)}$ that is divided into two subsegments h and $h+1$. We toss a coin to decide whether segment h or segment $h+1$ keeps the covariate set $\boldsymbol{\pi}_{h,(r-1)}$, and for the other segment we randomly sample a new covariate set $\boldsymbol{\pi}_\star$ from the covariate set prior distribution; see equation (32.4). Along with the birth of a new changepoint it is proposed to replace the current set of covariate sets $\{\boldsymbol{\pi}_{h,(r)}\}_{h=1,\dots,H_{(r-1)}}$ by a new set $\{\boldsymbol{\pi}_{h,\star}\}_{h=1,\dots,H_\star}$, where $H_\star = H_{(r-1)} + 1$. Death and the changepoint birth move are complementary to each other and the acceptance probability of both moves turn out to be (for details, see Lèbre *et al.*, 2010)

$$\begin{aligned} & A([\boldsymbol{\tau}_{h,(r-1)}, \{\boldsymbol{\pi}_{h,(r)}\}_{h=1,\dots,H_{(r)}}], [\boldsymbol{\tau}_{h,\star}, \{\boldsymbol{\pi}_{h,\star}\}_{h=1,\dots,H_\star}]) \\ &= \min \left\{ 1, \frac{p(\mathbf{y}|\lambda_{(r)}^2, \{\boldsymbol{\pi}_{h,\star}\}_{h=1,\dots,H_\star}, \boldsymbol{\tau}_\star)}{p(\mathbf{y}|\lambda_{(r)}^2, \{\boldsymbol{\pi}_{h,(r-1)}\}_{h=1,\dots,H_{(r-1)}}, \boldsymbol{\tau}_{(r-1)})} \cdot \frac{p(\boldsymbol{\tau}_\star)}{p(\boldsymbol{\tau}_{(r-1)})} \cdot HR \right\}, \end{aligned}$$

where the Hastings ratio HR can be computed with equation (32.13), using $\tau = \tau_{h,(r-1)}$. We note that the probability for sampling the new covariate set $\pi_{h,\star}$ from the prior distribution has cancelled with the covariate set prior ratio (for details, see Lèbre *et al.*, 2010).

Allowing the network structure to change between segments leads to a highly flexible model. However, this approach faces a conceptual and a practical problem. The practical problem is potential model over-flexibility. If subsequent changepoints are close together, network structures have to be inferred from short time series segments. This will almost inevitably lead to overfitting (in a maximum likelihood context) or inflated inference uncertainty (in a Bayesian context). The conceptual problem is the underlying assumption that structures associated with different segments are *a priori* independent. This is not realistic. For instance, for the evolution of a gene regulatory network during embryogenesis, we would assume that the network evolves gradually and that networks associated with adjacent time intervals are *a priori* similar.

To address these problems, Husmeier *et al.* (2010) and Dondelinger *et al.* (2013) proposed three methods of information sharing among time series segments. The first method is based on hard information coupling between the nodes, using the exponential distribution proposed in Werhli and Husmeier (2008). The second scheme is also based on hard information coupling, but uses a binomial distribution with conjugate beta prior. The third scheme is based on the same distributional assumptions as the second scheme, but replaces the hard by a soft information coupling scheme. The three schemes are briefly summarized in the Appendix, and we refer the reader to the original publications for further details.

32.2.7 Dynamic Bayesian Network Modelling

DBNs are a powerful class of models which can be employed for learning the regulatory interactions among variables from time series data. As described in more detail in Section 32.2.1, the standard situation is that N random variables Z_1, \dots, Z_N have been observed at $T + 1$ equidistant timepoints. The goal is to infer the regulatory interactions among the variables, where each regulatory interaction is subject to a lag of one timepoint. Because of the lag, the network learning task for the models discussed in Sections 32.2.2–32.2.5 can be subdivided into N independent regression tasks. In the i th regression task, $Y = Z_i$ takes the role of the response and the remaining $n = N - 1$ domain variables $\{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N\}$ are the potential covariates. Because of the time lag, the number of datapoints for the regression task reduces by 1. The objective is thus to infer the covariate set $\pi_i \subset \{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N\}$ for each Z_i from T observations. The covariate sets π_1, \dots, π_N can then be merged in the form of a network. There is the edge $Z_j \rightarrow Z_i$ in the network if and only if $Z_j \in \pi_i$ ($i, j \in \{1, \dots, N\} : i \neq j$). The network can be presented as an N -by- N adjacency matrix \mathcal{A} . The elements of \mathcal{A} are binary: $\mathcal{A}_{j,i} = 1$ indicates that there is an edge from Z_j to Z_i (i.e. that $Z_j \in \pi_i$) while $\mathcal{A}_{j,i} = 0$ indicates that there is no edge from Z_j to Z_i (i.e. that $Z_j \notin \pi_i$).

For each Z_i we can generate a posterior sample using one of the following regression models:

- the standard DBN, which applies a linear regression model to each Z_i (see Section 32.2.2);
- an NH-DBN, which is based on a piecewise linear regression model (see Section 32.2.3);
- an NH-DBN with sequentially coupled network parameters, which incorporates sequentially coupled regression coefficients into the piecewise linear regression model (see Section 32.2.4.1);
- an NH-DBN with globally coupled network parameters, which incorporates globally coupled regression coefficients into the piecewise linear regression model (see Section 32.2.4.2).

After having selected the most appropriate model (see Figure 32.1 for an overview), for each variable Z_i ($i = 1, \dots, N$) an MCMC simulation is performed to generate a sample of covariate

sets $\pi_i^{(1)}, \dots, \pi_i^{(R)}$. As it takes a while until the Markov chain converges to its stationary distribution (i.e. to the posterior distribution), the first samples are usually withdrawn ('burn-in phase'). The remaining samples ('sampling phase') are considered to be samples from the posterior distribution. Here we assume the burn-in and sampling phases to have the same length. That is, if we perform $R = 2S$ MCMC iterations, we withdraw $\pi_i^{(1)}, \dots, \pi_i^{(S)}$ and consider $\pi_i^{(S+1)}, \dots, \pi_i^{(2S)}$ to be a posterior sample of covariate sets for Z_i . From the latter covariate sets we build a sample of adjacency matrices $\mathcal{A}^{(S+1)}, \dots, \mathcal{A}^{(2S)}$ and the mean adjacency matrix $\hat{\mathcal{A}}$:

$$\hat{\mathcal{A}} := \frac{1}{S} \sum_{r=1}^S \mathcal{A}^{(S+r)}, \quad \text{where } \mathcal{A}_{j,i}^{(S+r)} = \begin{cases} 1, & \text{if } X_j \in \pi_i^{(S+r)}, \\ 0, & \text{if } X_j \notin \pi_i^{(S+r)}. \end{cases}$$

The non-diagonal elements of the matrix $\hat{\mathcal{A}}$ are estimates of the marginal edge posterior probabilities. For example, $\hat{\mathcal{A}}_{j,i} \in [0, 1]$ is an estimate for the marginal posterior probability that there is an edge from Z_j to Z_i in the network. By imposing a threshold $\psi \in [0, 1]$ on the entries of $\hat{\mathcal{A}}$ we get a binary adjacency matrix $\hat{\mathcal{A}}^\psi$ with $\hat{\mathcal{A}}_{i,j}^\psi = 1$ if $\hat{\mathcal{A}}_{i,j} > \psi$, and $\hat{\mathcal{A}}_{i,j}^\psi = 0$ otherwise. This adjacency matrix $\hat{\mathcal{A}}^\psi$ corresponds to a concrete network prediction. The predicted network possesses all edges $Z_j \rightarrow Z_i$ with $\hat{\mathcal{A}}_{j,i}^\psi = 1$.

When the true network is known, we can compare the predicted adjacency matrix $\hat{\mathcal{A}}^\psi$ with the true adjacency matrix \mathcal{A}^* . For each threshold $\psi \in [0, 1]$ we can compute the recall $\mathcal{R}(\psi)$ and the precision $\mathcal{P}(\psi)$ of the predicted network:

$$\mathcal{R}(\psi) = \frac{|\{(j, i) | j \neq i, \hat{\mathcal{A}}_{j,i}^\psi = 1, \mathcal{A}_{j,i}^* = 1\}|}{|\{(j, i) | j \neq i, \mathcal{A}_{j,i}^* = 1\}|},$$

$$\mathcal{P}(\psi) = \frac{|\{(j, i) | j \neq i, \hat{\mathcal{A}}_{j,i}^\psi = 1, \mathcal{A}_{j,i}^* = 1\}|}{|\{(j, i) | j \neq i, \hat{\mathcal{A}}_{j,i}^\psi = 1\}|}.$$

The precision corresponds to the positive predictive value ('Which fraction of the predicted edges is correct?') and the recall to the sensitivity ('Which fraction of the true edges has been correctly predicted?'). The curve $\{(\mathcal{R}(\psi), \mathcal{P}(\psi)) | 0 \leq \psi \leq 1\}$ is called the precision-recall curve (Davis and Goadrich, 2006). The area under the precision-recall curve (AUC), which can be obtained by numerical integration, is a popular measure for the network reconstruction accuracy. The higher the AUC, the higher the accuracy of the predicted network.

A popular diagnostic for monitoring the MCMC convergence is based on potential scale reduction factors (PSRFs); see, for example, Brooks and Gelman (1998) for the general theory. We perform Q independent MCMC simulations with $R = 2S$ iterations each. For each simulation $q = 1, \dots, Q$, we compute the average adjacency matrices after $2s$ iterations, where $s = 1, 2, 3, \dots, S$. For simulation q we obtain after $2s$ MCMC iterations the adjacency matrices $\mathcal{A}^{q,s+1}, \dots, \mathcal{A}^{q,2s}$ and the mean adjacency matrix $\hat{\mathcal{A}}_{j,i}^{q,s}$. $\hat{\mathcal{A}}_{j,i}^{q,s}$ is an estimate for the marginal posterior probability of the edge $Z_j \rightarrow Z_i$. For each potential edge $Z_j \rightarrow Z_i$ ($j, i \in \{1, \dots, N\} : j \neq i$) and each run length $2s$, we compute the 'between-chain' and the 'within-chain' variance,

$$\mathcal{B}_{2s}(j, i) = \frac{1}{Q-1} \sum_{q=1}^Q (\hat{\mathcal{A}}_{j,i}^{q,s} - \bar{\mathcal{A}}_{j,i}^{q,s})^2,$$

$$\mathcal{W}_{2s}(j, i) = \frac{1}{Q(s-1)} \sum_{q=1}^Q \sum_{r=1}^s (\mathcal{A}_{j,i}^{q,s+r} - \hat{\mathcal{A}}_{j,i}^{q,s})^2,$$

where $\bar{\mathcal{A}}_{j,i}^{s,s}$ is the mean of $\hat{\mathcal{A}}_{j,i}^{1,s}, \dots, \hat{\mathcal{A}}_{j,i}^{Q,s}$, and $\mathcal{A}_{j,i}^{q,s+1}, \dots, \mathcal{A}_{j,i}^{q,2s}$ are the adjacency matrices from the sampling phase of simulation q . After $2s$ iterations the PSRF of the edge $X_j \rightarrow X_i$ is

$$PSRF_{2s}(j, i) = \frac{\left(1 - \frac{1}{s}\right) \mathcal{W}_{2s}(j, i) + \left(1 + \frac{1}{Q}\right) \mathcal{B}_{2s}(j, i)}{\mathcal{W}_{2s}(j, i)}. \quad (32.29)$$

PSRFs near 1 indicate that the MCMC simulations are close to the stationary distribution. As a convergence diagnostic for networks Grzegorczyk and Husmeier (2011) proposed to monitor the fraction of edges which fulfil the conditions $PSRF < \alpha$ (e.g. $\alpha = 1.1$) against the number of MCMC iterations $2s$:

$$\Phi_{2s}^{(\alpha)} = \frac{|\{(j, i) : PSRF_{2s}(j, i) \leq \alpha\}|}{N \cdot (N - 1)}. \quad (32.30)$$

32.2.8 Computational Complexity

The NH-DBNs in Sections 32.2.2–32.2.5 share the network structure among segments, and the computational advantage is that the task of inferring a network with N nodes can be subdivided into N separate and independent regression tasks; see Section 32.2.7 for the technical details. That is, when a computer cluster is available, the N MCMC simulations can be run in parallel. For the relatively small yeast data set with $N = 5$ genes only (Section 32.3.2) the computation costs for $R = 20,000$ MCMC iterations were moderate. With our Matlab implementation each of the $N = 5$ MCMC simulations took a few minutes only. It is very difficult to make a general statement on how the MCMC inference algorithms scale up to larger networks with higher values of N . The main issue is that the convergence rate depends not only on the number of nodes N and the number of segments H , but also on the underlying posterior landscape. Peaked posterior landscapes with many locally optimal regions can hinder convergence of the Markov chains independently of the size of the network N . On the other hand, even for large N , convergence might be reached quickly when the posterior landscape is flat and does not possess locally optimal regions. One aim of our future work is to properly study and to report on how the model inference scales up to larger values of N . In this context it is worth noting that the initializations of the MCMC algorithms can also play an important role. In the pseudocodes provided we used uninformative initializations. Convergence can be speeded up when starting from more informative initializations. The latter can, for example, be obtained from prior knowledge and/or pre-analyses of the data with simpler methods. Also more informative prior distributions can speed up convergence, as they make the posterior distribution more peaked and reduce the size of the effectively accessible parameter space.

For the NH-DBNs with time-varying network structures discussed in Section 32.2.6, inference is computationally more expensive. The fundamental differences are (i) that those models have an increased configuration space (segment-specific network structures) and (ii) that the network inference task cannot be subdivided into independent regression tasks. The MCMC simulations on the *Drosophila melanogaster* data with $N = 11$ genes (Section 32.3.1) were typically run overnight, although the convergence diagnostics chosen were rather conservative (retrospectively checking that $PSRF < 1.1$ had been achieved). In practice, the MCMC simulation times can be reduced if computation costs are an issue by checking PSRF values continuously and stopping the simulations as soon as an acceptable degree of convergence has been achieved.

32.3 Application Examples

We conclude this chapter with two application examples. The first example looks at the expression time series from 11 genes involved in muscle development of *Drosophila melanogaster* (fruit fly). The ultimate objective is to infer the time-varying gene regulatory network structure during morphogenesis. However, since the true network structure and structural changes are unknown, we use the estimation of the temporal changepoints as a proxy for prediction accuracy, investigating how closely they are aligned with the natural phases during the fly's life cycle (embryo, larva, pupa, adult). In the second example, we choose an example from synthetic biology. Here the true gene regulatory network structure is indeed known, and the accuracy of network structure inference can therefore be assessed directly.

32.3.1 Morphogenesis in *Drosophila*

Arbeitman *et al.* (2002) measured transcriptional profiles for about one-third of all predicted *Drosophila melanogaster* genes throughout the fly's life cycle, from fertilization to ageing adults. Complementary DNA microarrays were used to analyse the RNA expression levels of 4028 genes in wildtype flies examined during 66 sequential time periods beginning at fertilization and spanning the embryonic, larval and pupal periods and the first 30 days of adulthood. The authors found that a high proportion of the genes were developmentally regulated. Husmeier *et al.* (2010) and Dondelinger *et al.* (2013) selected 11 of these genes, which are known to be involved in muscle development, and applied a DBN with time-varying network structure, described in Section 32.2.6, to the corresponding gene expression time series. The predicted time-varying regulatory network structure proposes new biological hypotheses for the relevant molecular processes governing embryogenesis, but discussing that in detail is beyond the remit of the present chapter. Since the true regulatory network structure is unknown, we cannot assess the accuracy of the methods at the network structure level. What we can do instead, though, is assess how well the changepoints corresponding to the known transitions between the four developmental phases (embryo, larva, pupa, adult) are predicted. Figure 32.5(b) shows the posterior probabilities of inferred changepoints for any gene using the time-varying DBN without information coupling, while Figure 32.5(c) shows the corresponding posterior probabilities obtained with the various information coupling methods discussed in Section 32.2.6. For comparison, the gene expression time series were also analysed with the sparse regression method proposed in Ahmed and Xing (2009), using the authors' software package TESLA (Figure 32.5(a)). A comparison between Figures 32.5(a) and 32.5(b) demonstrates that even without information coupling, the time-varying DBN clearly outperforms TESLA in that it recovers the changepoints corresponding to all three transitions (embryo → larva, larva → pupa, and pupa → adult), whereas for TESLA these transitions are lost in noise. A comparison between Figures 32.5(c) and 32.5(b) demonstrates that the effect of the information coupling methods is to reduce the size of the smaller and potentially spurious peaks, while keeping the three most salient peaks (corresponding to larva → pupa, pupa → adult, and an extra transition in the embryo phase). In this way, the peak-to-trough ratio is increased, as would be expected for a method that aims to achieve noise reduction. We observe that as a consequence of this noise reduction the morphological transition embryo → larva becomes less pronounced. However, this is not necessarily surprising. A complex gene regulatory network is unlikely to transition into a new morphogenic phase all at once, and some key regulatory pathways may have to undergo activational changes earlier in preparation for the morphogenic transition. As such, it is not implausible that key transitions at the gene regulatory network level have to occur *before*

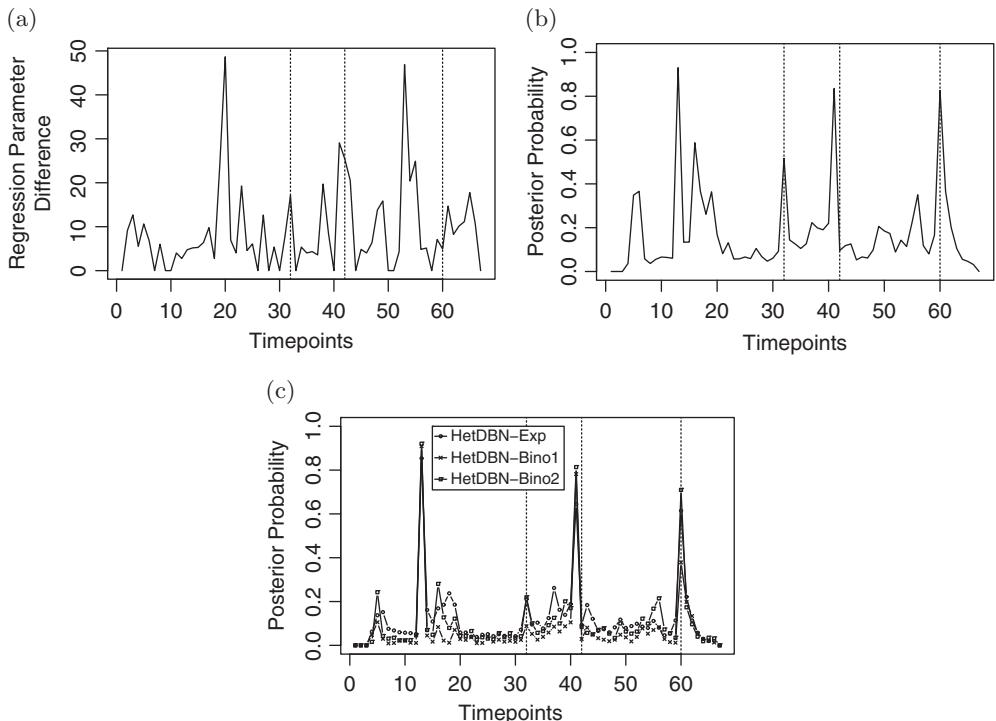


Figure 32.5 Changepoints during morphogenesis of *Drosophila melanogaster*, inferred from gene expression time series. The horizontal axes show the timepoints at which gene expression profiles were measured, and the vertical dashed lines indicate the three critical transitions during the life cycle of the fly: embryo → larva, larva → pupa, pupa → adult. The vertical axes in (b) and (c) show the marginal posterior probability of a changepoint occurring for any node at a given time plotted against time. The vertical axis in (a) shows a related surrogate score (L1 norm of the difference of the regression parameter vectors associated with two adjacent timepoints plotted against time). (a) Application of TESLA, the method proposed in Ahmed and Xing (2009). The location of the three principal life-cycle transitions is buried in noise. (b) Application of a time-varying DBN without information coupling. The three principle life-cycle transitions are much more clearly predicted than with TESLA. (c) Time-varying DBN with the three information coupling methods described in Section 32.2.6, as indicated in the legend: *HetDBN-Exp*, exponential distribution (see Appendix A.1); *HetDBN-Bino1*, binomial distribution with hard coupling (see Appendix A.2); *HetDBN-Bino2*, binomial distribution with soft coupling (see Appendix A.3). A comparison between (b) and (c) demonstrates that the information coupling leads to a suppression of spurious peaks in the signal. Figure adapted from Husmeier *et al.* (2010).

transitions at the phenotype level can take place. This explains the shift of the corresponding peak to an earlier stage of the embryo phase, around timepoint 14.

32.3.2 Synthetic Biology in Yeast

By means of synthetic biology Cantone *et al.* (2009) designed a network with $N = 5$ genes and $M = 8$ interactions in *Saccharomyces cerevisiae* (yeast); the true network structure is shown in Figure 32.10. With quantitative real-time polymerase chain reaction (RT-PCR), Cantone *et al.* (2009) then measured *in vivo* gene expression data, first under galactose and then under glucose metabolism. For both carbon sources the network structure is identical, but the strengths of the regulatory processes (i.e. the network parameters) change with the carbon source (Cantone *et al.*, 2009). In contrast to the previous example, we therefore model the system with

the NH-DBNs from Sections 32.2.2–32.2.4 where only the network parameters, but not the structure, can vary in time.

For each gene Z_i ($i = 1, \dots, 5$), 16 measurements were taken in galactose $\mathbf{D}_{i,1}, \dots, \mathbf{D}_{i,16}$ and 21 measurements were taken in glucose $\mathbf{D}_{i,1}^*, \dots, \mathbf{D}_{i,21}^*$. For both parts of the time series the initial measurements $\mathbf{D}_{i,1}$ and $\mathbf{D}_{i,1}^*$ were taken while extant glucose (galactose) was washed out and new galactose (glucose) was supplemented. We therefore withdraw the two initial observations. Subsequently, we re-merge the two parts to one time series again, and apply a gene-wise z -score standardization. For each gene Z_i ($i = 1, \dots, 5$), the remaining $(16 - 1) + (21 - 1) = 35$ observations $\mathbf{D}_{i,2}, \dots, \mathbf{D}_{i,16}, \mathbf{D}_{i,2}^*, \dots, \mathbf{D}_{i,21}^*$ then have mean 0 and variance 1.

The goal is to infer a covariate set (set of regulators) for each gene Z_i ($i = 1, \dots, 5$), with the other four genes Z_j ($j \neq i$) constituting the set of potential covariates (regulators). We make the standard assumption that all regulatory interactions are subject to a lag of one timepoint. As we have removed the initial (washing period) observations, we have to take into account that the first response observation of the ‘galactose part’, $\mathbf{D}_{i,2}^*$, is not properly related to the last observation of the ‘glucose part’, $\mathbf{D}_{j,16}^*$ ($j \neq i$). Therefore, for each response gene $Y = Z_i$ only $T = (16 - 1 - 1) + (21 - 1 - 1) = 33$ observations can be used for model inference. The response vector is given by $\mathbf{y} := (\mathbf{D}_{i,3}, \dots, \mathbf{D}_{i,16}, \mathbf{D}_{i,3}^*, \dots, \mathbf{D}_{i,21})^T$ and the corresponding shifted values of the potential covariates Z_j ($j \neq i$) are $\mathbf{D}_{j,2}, \dots, \mathbf{D}_{j,15}, \mathbf{D}_{j,2}^*, \dots, \mathbf{D}_{j,20}^*$. For example, for gene Z_2 with covariates Z_1 and Z_5 we have in a homogeneous DBN: $Y := Z_2, X_1 := Z_1, X_2 := Z_5, T = 33, k = 2, \boldsymbol{\pi} = \{X_1, X_2\}$, and

$$\mathbf{y} = \begin{bmatrix} \mathbf{D}_{2,3} \\ \vdots \\ \mathbf{D}_{2,16} \\ \mathbf{D}_{2,3}^* \\ \vdots \\ \mathbf{D}_{2,21} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & \mathbf{D}_{1,2} & \mathbf{D}_{5,2} \\ \vdots & \vdots & \vdots \\ 1 & \mathbf{D}_{1,15} & \mathbf{D}_{5,15} \\ 1 & \mathbf{D}_{1,2}^* & \mathbf{D}_{5,2}^* \\ \vdots & \vdots & \vdots \\ 1 & \mathbf{D}_{1,20}^* & \mathbf{D}_{5,20}^* \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

To be consistent with earlier studies we set the maximal cardinality of the covariate sets in equation (32.4) to $\mathcal{F} = 3$. For all inverse-gamma-distributed parameters, namely σ^2, λ^2 and δ^2 , we select the shape and rate parameters $a_\sigma = b_\sigma = 0.005, a_\lambda = 2, b_\lambda = 0.2$ and $a_\delta = b_\delta = 1$, as in Grzegorczyk and Husmeier (2012, 2013). For generating posterior samples we run the MCMC algorithms for $R = 2S = 20,000$ iterations, setting the burn-in and sampling phases both to $S = 10,000$ iterations (50% of the iterations for each phase).

In order to infer segmentations with different numbers of segments, we vary the hyperparameter p of the multiple changepoint process, that is, the hyperparameter p of the geometric distribution on the distance between changepoints in equation (32.11). We choose six hyperparameter values $p = 0.1 \cdot 2^i$, with $i \in \{-3, \dots, 2\}$. For each value of p we run $Q = 10$ independent MCMC simulations for each of the four models, namely the homogeneous DBN (see Section 32.2.2), the uncoupled NH-DBN (see Section 32.2.3), the sequentially coupled NH-DBN (see Section 32.2.4.1) and the globally coupled NH-DBN (see Section 32.2.4.2). That is, in total we infer $6 \cdot 10 \cdot 4 = 240$ networks.

To check for convergence during the $S = 10,000$ MCMC iterations of the burn-in phase, we monitored the PSRFs; see Section 32.2.7 for the technical details. To this end, we ran for each model and each hyperparameter $Q = 10$ independent MCMC simulations. We then combined the outputs from $Q = 10$ independent simulations to compute, for $s = 100, \dots, 5000$, the fraction of edges with PSRF lower than $\alpha = 1.1$, denoted by $\Phi_{2s}^{(1,1)}$ in equation (32.30). For all

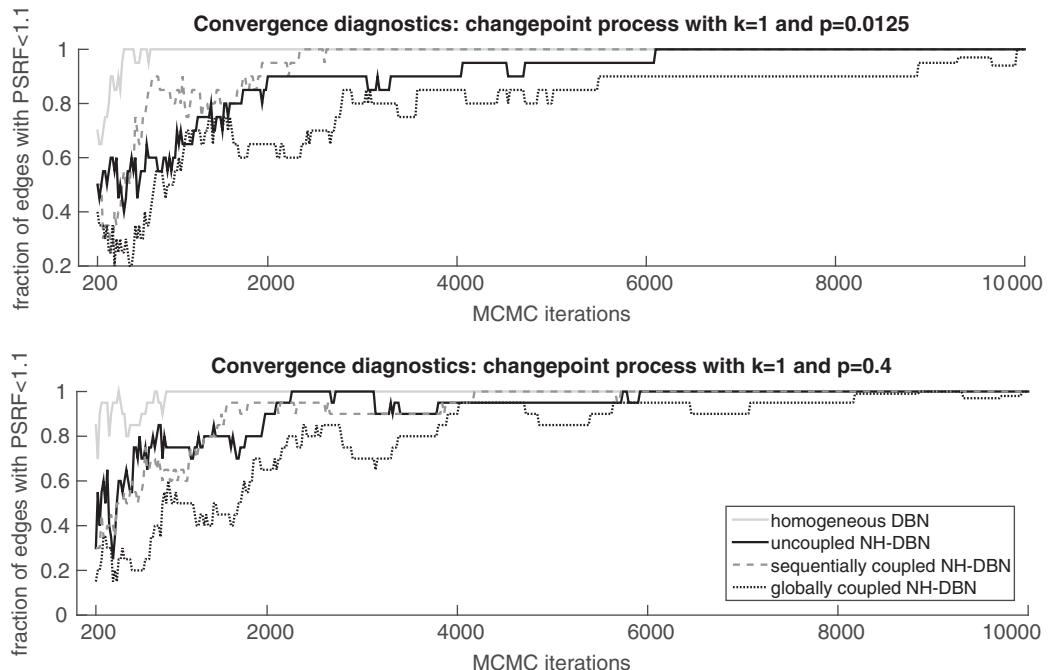


Figure 32.6 Convergence diagnostics based on potential scale reduction factors. For different changepoint hyperparameters p (see equation (32.11)), we ran for each of the four NH-DBN models $Q = 10$ independent MCMC simulations with $R = 2S = 20k$ iterations. We then monitored the model-specific fractions of edges with $PSRF < 1.1$, $\Phi_{2s}^{(1,1)}$, along the burn-in phase (for $s = 100, \dots, 5000$). The results for the lowest p (top) and the highest p (bottom) are shown.

models and all hyperparameters p we found that all PSRFs were below 1.1 at the end of the burn-in phase, which is usually taken as an indication of a sufficient degree of convergence.

Figure 32.6 shows the temporal evolution of the model-specific PSRF scores for $p = 0.1 \cdot 2^{-3} = 0.0125$ and $p = 0.1 \cdot 2^2 = 0.4$. At the end of the burn-in phase the convergence criterion is satisfied by all models. However, it can be seen that the homogeneous DBN converges faster than the NH-DBNs and that the MCMC simulations for the globally coupled NH-DBN require more iterations to converge.

To demonstrate that the hyperparameter p in equation (32.11) can have a substantial effect on the inferred data segmentations, Figure 32.7 shows the posterior averages of the number of inferred segments H . The homogeneous DBN cannot infer changepoints and keeps $H = 1$ fixed during the simulations. For the three NH-DBNs the number of inferred segments H monotonically increases in p with posterior averages ranging from $H \approx 2$ ($p = 0.0125$) to $H \approx 9$ ($p = 0.4$).

To compare the network reconstruction accuracies we computed for each of the 240 inferred networks the resulting precision–recall AUC value; see Section 32.2.7 for the technical details. Figure 32.8 shows histograms of the average precision–recall AUC values. The following trends can be found. First, the homogeneous DBN shows a consistent performance, as it does not depend on the hyperparameter p . Second, for $p \leq 0.1$ (i.e. when the number of segments H is not too large), the uncoupled and the sequentially coupled NH-DBN perform significantly better than the homogeneous DBN. Only for the highest hyperparameter ($p = 0.4$), that is, the largest number of segments, does the DBN appear to perform slightly better than the two

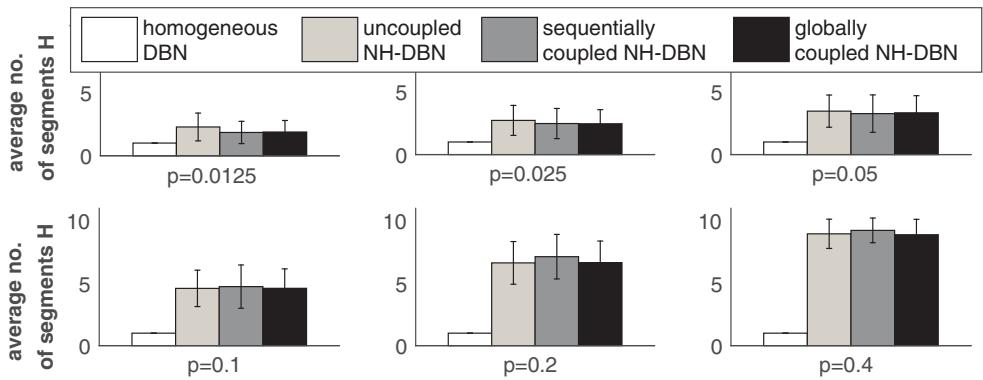


Figure 32.7 Posterior averages of the numbers of segments H . For each hyperparameter p of the changepoint prior the figure shows a histogram. There is a bar for each of the four models and the bar heights are the posterior averages of the numbers of segments H . Averages have been taken across the $N = 5$ genes and $Q = 10$ independent MCMC simulations.

NH-DBNs. Third, except for the lowest hyperparameter ($p = 0.0125$), the sequentially coupled NH-DBN consistently outperforms the homogeneous DBN (albeit by only a small amount). Fourth and finally, the globally coupled NH-DBN is the clear winner and consistently outperforms all the other models for all values of p .

For all hyperparameters p we then averaged the model-specific edge scores, across the $Q = 10$ simulations, to obtain the average model-specific precision recall curves. Figure 32.9 shows the

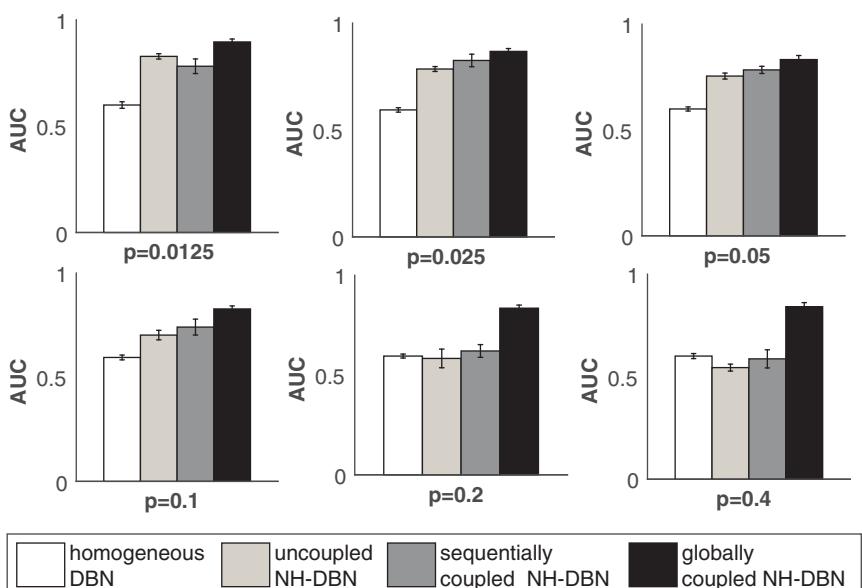


Figure 32.8 Quantification of the network reconstruction accuracies in terms of average AUC values. We analysed the yeast data with four models using six different changepoint hyperparameters p in equation (32.11). For each p there is a histogram showing the four model-specific precision-recall AUC values. The bar heights are averages over 10 AUC scores obtained from $Q = 10$ independent MCMC simulations. The error bars indicate standard deviations.

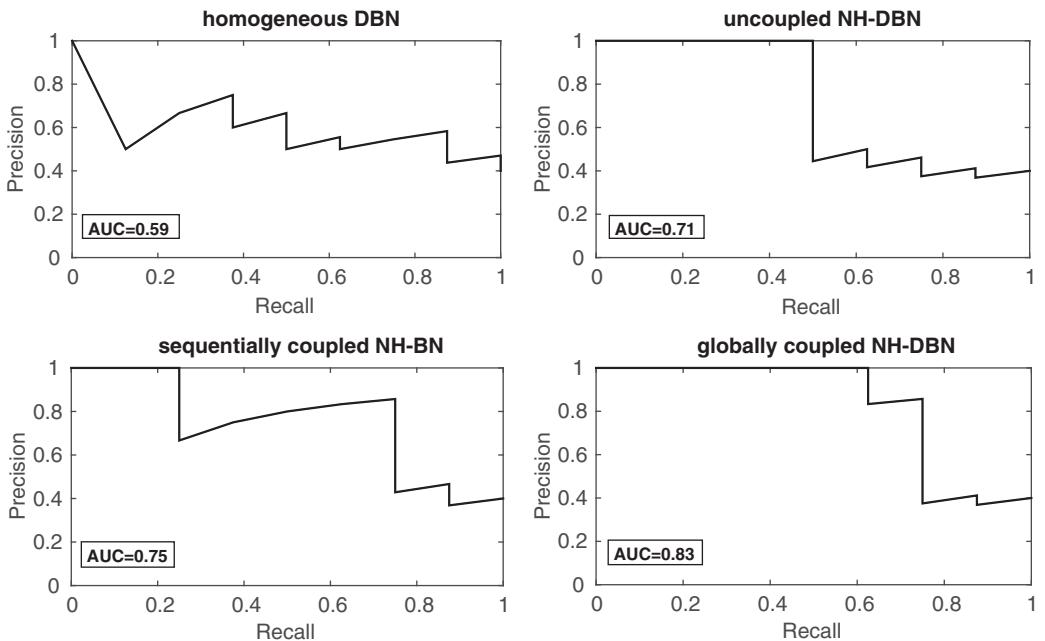


Figure 32.9 Precision–recall curves. For each changepoint hyperparameter p we averaged the model-specific edge scores across the $Q = 10$ simulations. From the average scores we then extracted the model-specific precision–recall curves. The figure shows the model-specific results for the hyperparameter $p = 0.1$. The first point $(0, 1)$ is a pseudo point (the starting point of the curve).

results for the moderate hyperparameter $p = 0.1$. The globally coupled NH-DBN model infers the highest scores for six true edges, and hence yields the highest AUC value ($AUC = 0.83$).

Finally, Figure 32.10 shows the true yeast network and two network predictions, obtained with the homogeneous DBN and the globally coupled NH-DBN. As the true network possesses eight edges, we decided to extract the eight edges with the highest average posterior probability scores. The homogeneous DBN infers four true positive (correct) and four false positive

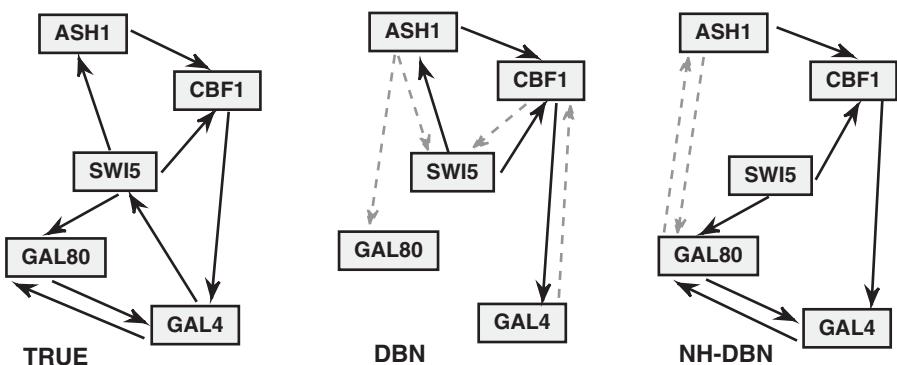


Figure 32.10 True yeast network and two inferred yeast networks. The true yeast network, designed by Cantone *et al.* (2009), is shown on the left. As the true network has eight edges, we extracted the eight edges with the highest posterior probability scores. The homogeneous DBN predicts the centre network and the globally coupled NH-DBN with $p = 0.1$ in equation (32.11) predicts the network on the right. The grey dotted edges in the predicted networks refer to false positive edges. The predicted networks show that the advanced globally coupled NH-DBN yields a substantial improvement over the standard homogeneous DBN.

(incorrect) edges. The network predicted with the homogeneous DBN leads to the point RECALL = 4/8 and PRECISION = 4/8 in the top left-hand plot of Figure 32.9. The globally coupled NH-DBN infers 6 true positive and only 2 false positive edges. The network obtained with the globally coupled NH-DBN thus yields the point RECALL = 6/8 and PRECISION = 6/8 in the bottom right-hand plot of Figure 32.9.

32.4 Summary

In this chapter we have given an overview of non-homogeneous dynamic Bayesian network models. After describing the relationship between NH-DBNs and regression models in Section 32.2.1, we reviewed the traditional homogeneous dynamic Bayesian network in Section 32.2.2. Subsequently, in Section 32.2.3, we combined this homogeneous DBN model with a multiple changepoint process, so as to obtain a piecewise homogeneous model for applications where the network parameters might be subject to temporal changes. The changepoint process divides the data into temporal segments, and only the network structure is shared among segments, while the network parameters differ from segment to segment. In Section 32.2.4 we have described a sequential and a global parameter coupling scheme for NH-DBNs. Unlike the ‘uncoupled’ NH-DBN in Section 32.2.3, the coupled NH-DBNs encourage the segment-specific network parameters to evolve gradually over time (sequential coupling) or to be similar (global coupling), so as to avoid model overflexibility. NH-DBNs with alternative allocation schemes, such as mixture and hidden Markov models as well as NH-DBNs with time-varying network structures, have been briefly reviewed in Sections 32.2.5 and 32.2.6. Figure 32.1 provides a graphical overview of those NH-DBN models and might serve as a guideline for finding the best model for a particular application.

We have concluded this chapter with an application of the methods to two data sets: a data set containing gene expression time series for 11 genes during the life cycle of *Drosophila*, and a data set containing the gene expression time series of five yeast genes in two externally controlled metabolic phases, related to glucose and galactose metabolism. In the first example, we allowed both the structure and the parameters of the DBN to vary in time. Since the true regulatory network structure is unknown, we used the prediction of the known life-cycle transitions as a proxy for assessing the accuracy of the methods. All time-varying DBNs outperformed a sparse lasso-penalized linear regression model (TESLA). The information coupling schemes described in Section 32.2.6 achieved a suppression of the smaller and presumably spurious peaks in the posterior probabilities of the changepoints, which is indicative of noise reduction. The second example is taken from synthetic biology, and the true network structure is therefore known. We compared homogeneous and non-homogeneous DBNs, where the parameters were allowed to change with time, and demonstrated that the latter can potentially improve the network reconstruction accuracy; this reflects the true non-stationarity of the process as a result of the changing metabolic conditions. We emulated the scenario of not knowing the true number of changepoints, and found that when the assumed number of changepoints substantially exceeds the true number, regularization via global information coupling is essential for the non-homogeneous DBN to show its full potential and outperform the homogenous DBN.

Appendix A: Coupling Schemes

In this appendix we briefly summarize the three coupling schemes for the NH-DBNs with time-varying network structures from Section 32.2.6. We refer the reader to the original publications for more details.

A.1 Hard Information Coupling Based on an Exponential Prior

Denote by H_i the total number of segments in the time series associated with node i , let y_i^h denote the h th time series segment associated with node i , and let $\boldsymbol{\pi}_i^h$, $1 \leq h \leq H_i$, denote the parents of node i in segment h . We impose a prior distribution $P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, \beta)$ on the structures, and the joint probability distribution factorizes according to a Markovian dependence:

$$P(y_i^1, \dots, y_i^{H_i}, \boldsymbol{\pi}_i^1, \dots, \boldsymbol{\pi}_i^{H_i}, \beta) = \prod_{h=1}^{H_i} P(y_i^h | \boldsymbol{\pi}_i^h) P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, \beta) P(\beta). \quad (32.31)$$

Similarly to Werhli and Husmeier (2008), we define

$$P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, \beta) = \frac{\exp(-\beta|\boldsymbol{\pi}_i^h - \boldsymbol{\pi}_i^{h-1}|)}{Z_i(\beta, \boldsymbol{\pi}_i^{h-1})} \quad (32.32)$$

for $h \geq 2$, where β is a hyperparameter that defines the strength of the coupling between $\boldsymbol{\pi}_i^h$ and $\boldsymbol{\pi}_i^{h-1}$, and $|\cdot|$ denotes the Hamming distance. For $h = 1$, $P(\boldsymbol{\pi}_i^h)$ is the prior distribution (e.g. a uniform distribution over all valid structures). The denominator in equation (32.32) is a normalizing constant, $Z(\beta) = \sum_{\boldsymbol{\pi}_i^h \in \mathbb{M}_i} e^{-\beta|\boldsymbol{\pi}_i^h - \boldsymbol{\pi}_i^{h-1}|}$, where \mathbb{M}_i is the set of all valid parent configurations for node i . If we ignore any fan-in restriction that might have been imposed *a priori*, then the expression for the partition function can be simplified: $Z(\beta) \approx \prod_{j=1}^N Z_j(\beta)$, where $Z_j(\beta) = \sum_{e_j^h=0}^1 e^{-\beta|e_j^h - e_j^{h-1}|} = 1 + e^{-\beta}$ and hence $Z(\beta) = (1 + e^{-\beta})^N$. Inserting this expression into equation (32.32) gives

$$P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, \beta) = \frac{\exp(-\beta|\boldsymbol{\pi}_i^h - \boldsymbol{\pi}_i^{h-1}|)}{(1 + e^{-\beta})^N}. \quad (32.33)$$

A.2 Hard Information Coupling Based on a Binomial Prior

An alternative method of information sharing among segments and nodes is by using a binomial prior,

$$P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, a, b) = a^{N_1^1[h, i]} (1-a)^{N_1^0[h, i]} b^{N_0^0[h, i]} (1-b)^{N_0^1[h, i]}, \quad (32.34)$$

where we have defined the following sufficient statistics: $N_1^1[h, i]$ is the number of edges in $\boldsymbol{\pi}_i^{h-1}$ that are matched by an edge in $\boldsymbol{\pi}_i^h$, $N_1^0[h, i]$ is the number of edges in $\boldsymbol{\pi}_i^{h-1}$ for which there is no edge in $\boldsymbol{\pi}_i^h$, $N_0^1[h, i]$ is the number of edges in $\boldsymbol{\pi}_i^h$ for which there is no edge in $\boldsymbol{\pi}_i^{h-1}$, and $N_0^0[h, i]$ is the number of coinciding non-edges in $\boldsymbol{\pi}_i^{h-1}$ and $\boldsymbol{\pi}_i^h$. Since the hyperparameters are shared, the joint distribution can be expressed as

$$P(\{\boldsymbol{\pi}_i^h\} | a, b) = \prod_{i=1}^p P(\boldsymbol{\pi}_i^1) \prod_{h=1}^{H_i} P(\boldsymbol{\pi}_i^h | \boldsymbol{\pi}_i^{h-1}, a, b) = a^{N_1^1} (1-a)^{N_1^0} b^{N_0^0} (1-b)^{N_0^1} \prod_{i=1}^N P(\boldsymbol{\pi}_i^1), \quad (32.35)$$

where we have defined $N_k^l = \sum_{i=1}^N \sum_{h=2}^{H_i} N_k^l[h, i]$, and the right-hand side follows from equation (32.34). The conjugate prior for the hyperparameters a, b is a beta distribution,

$P(a, b | \alpha, \bar{\alpha}, \gamma, \bar{\gamma}) \propto a^{(\alpha-1)}(1-a)^{(\bar{\alpha}-1)}b^{(\gamma-1)}(1-b)^{(\bar{\gamma}-1)}$, which allows the hyperparameters to be integrated out in closed form:

$$\begin{aligned} P(\{\pi_i^h\} | \alpha, \bar{\alpha}, \gamma, \bar{\gamma}) &= \int \int P(\{\pi_i^h\} | a, b) P(a, b | \alpha, \bar{\alpha}, \gamma, \bar{\gamma}) da db \\ &\propto \frac{\Gamma(\alpha + \bar{\alpha})}{\Gamma(\alpha)\Gamma(\bar{\alpha})} \frac{\Gamma(N_1^1 + \alpha)\Gamma(N_1^0 + \bar{\alpha})}{\Gamma(N_1^1 + \alpha + N_1^0 + \bar{\alpha})} \frac{\Gamma(\gamma + \bar{\gamma})}{\Gamma(\gamma)\Gamma(\bar{\gamma})} \frac{\Gamma(N_0^0 + \gamma)\Gamma(N_0^1 + \bar{\gamma})}{\Gamma(N_0^0 + \gamma + N_0^1 + \bar{\gamma})}. \end{aligned} \quad (32.36)$$

The level-2 hyperparameters $\alpha, \bar{\alpha}, \gamma, \bar{\gamma}$ are given a uniform hyperprior.

A.3 Soft Information Coupling Based on a Binomial Prior

We can relax the information sharing scheme from a hard to a soft coupling by introducing node-specific hyperparameters a_i, b_i that are softly coupled via a common level-2 hyperprior, $P(a_i, b_i | \alpha, \bar{\alpha}, \gamma, \bar{\gamma}) \propto a_i^{(\alpha-1)}(1-a_i)^{(\bar{\alpha}-1)}b_i^{(\gamma-1)}(1-b_i)^{(\bar{\gamma}-1)}$:

$$P(\pi_i^h | \pi_i^{h-1}, a_i, b_i) = (a_i)^{N_1^1[h,i]}(1-a_i)^{N_1^0[h,i]}(b_i)^{N_0^0[h,i]}(1-b_i)^{N_0^1[h,i]}. \quad (32.37)$$

This leads to a straightforward modification of equation (32.35) – replacing a, b by a_i, b_i – from which we get as an equivalent to equation (32.37), using the definition $N_k^l[i] = \sum_{h=2}^{H_i} N_k^l[h,i]$,

$$\begin{aligned} P(\pi_i^1, \dots, \pi_i^{H_i} | \alpha, \bar{\alpha}, \gamma, \bar{\gamma}) &\\ &\propto \frac{\Gamma(\alpha + \bar{\alpha})}{\Gamma(\alpha)\Gamma(\bar{\alpha})} \frac{\Gamma(N_1^1[i] + \alpha)\Gamma(N_1^0[i] + \bar{\alpha})}{\Gamma(N_1^1[i] + \alpha + N_1^0[i] + \bar{\alpha})} \frac{\Gamma(\gamma + \bar{\gamma})}{\Gamma(\gamma)\Gamma(\bar{\gamma})} \frac{\Gamma(N_0^0[i] + \gamma)\Gamma(N_0^1[i] + \bar{\gamma})}{\Gamma(N_0^0[i] + \gamma + N_0^1[i] + \bar{\gamma})}. \end{aligned} \quad (32.38)$$

In all three cases, the resulting prior distributions penalize changes in the network structure. For details of their inclusion in the MCMC scheme, see Husmeier *et al.* (2010) and Dondelinger *et al.* (2013).

References

- Ahmed, A. and Xing, E. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11878–11883.
- Arbeitman, M., Furlong, E., Imam, F., Johnson, E., Null, B., Baker, B., Krasnow, M., Scott, M., Davis, R. and White, K. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275.
- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.
- Butte, A.S. and Kohane, I.S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing* **5**, 418–429.
- Butte, A.S. and Kohane, I.S. (2003). Relevance networks: A first step toward finding genetic regulatory networks within microarray data. In G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger (eds.), *The Analysis of Gene Expression Data*. Springer, New York, pp. 428–446.

- Cantone, I., Marucci, L., Iorio, F., Ricci, M., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D. and Cosma, M. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* **137**, 172–181.
- Chen, T., He, H.L. and Church, G.M. (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* **4**, 29–40.
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, ACM, New York, pp. 233–240.
- Dondelinger, F., Lèbre, S. and Husmeier, D. (2013). Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine Learning* **90**, 191–230.
- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**, 601–620.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004). *Bayesian Data Analysis*, 2nd edition. Chapman and Hall/CRC, Boca Raton, FL.
- Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Grzegorczyk, M. (2016). A non-homogeneous dynamic Bayesian network with a hidden Markov model dependency structure among the temporal data points. *Machine Learning* **102**, 155–207.
- Grzegorczyk, M. and Husmeier, D. (2011). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning* **83**, 355–419.
- Grzegorczyk, M. and Husmeier, D. (2012). A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology. *Statistical Applications in Genetics and Molecular Biology* **11**, article 7.
- Grzegorczyk, M. and Husmeier, D. (2013). Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Machine Learning* **91**, 105–154.
- Grzegorczyk, M., Husmeier, D., Edwards, K., Ghazal, P. and Millar, A. (2008). Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. *Bioinformatics* **24**, 2071–2078.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing* **6**, 422–433.
- Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19**, 2271–2282.
- Husmeier, D., Dondelinger, F. and Lèbre, S. (2010). Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culotta (eds.), *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS)*. Curran Associates, Red Hook, NY, pp. 901–909.
- Ko, Y., Zhai, C. and Rodriguez-Zas, S. (2007). Inference of gene pathways using Gaussian mixture models. In *BIBM International Conference on Bioinformatics and Biomedicine*. IEEE Computer Society, Los Alamitos, CA, pp. 362–367.
- Lèbre, S., Becq, J., Devaux, F., Lelandais, G. and Stumpf, M. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology* **4**, 130.
- Nobile, A. and Fearnside, A. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing* **17**, 147–162.
- Robinson, J. and Hartemink, A. (2010). Learning non-stationary dynamic Bayesian networks. *Journal of Machine Learning Research* **11**, 3647–3680.

- Shafiee Kamalabad, M. and Grzegorczyk, M. (2018). Improving nonhomogeneous dynamic Bayesian networks with sequentially coupled parameters. *Statistica Neerlandica* **72**, 281–305.
- Thorne, T. and Stumpf, M.P.H. (2012). Inference of temporally varying Bayesian networks. *Bioinformatics* **28**, 3298–3305.
- Vyshevsky, V. and Girolami, M.A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics* **24**, 833–839.
- Werhli, A.V. and Husmeier, D. (2008). Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of Bioinformatics and Computational Biology* **6**, 543–572.
- Wilkinson, D. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, Boca Raton, FL.
- Zak, D.E., Doyle, F.J. and Schwaber, J.S. (2002). Local identifiability: When can genetic networks be identified from microarray data? *Proceedings of the Third International Conference on Systems Biology* pp. 236–237.

33

DNA Methylation

Kasper D. Hansen,¹ Kimberly D. Siegmund,² and Shili Lin³

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, USA

²Department of Preventive Medicine, Keck School of Medicine of USC, Los Angeles, USA

³Department of Statistics, Ohio State University, Columbus, OH, USA

Abstract

DNA methylation is associated with normal development and complex disease in humans. Our understanding of alterations is quickly extending beyond the aberrant DNA methylation of CpG islands in cancer. These discoveries are, in part, propelled by high-throughput technologies, while also limited by the accessibility of human tissue. DNA methylation is still the most commonly studied of the epigenetic marks due to its low cost and the ease of interrogation from small numbers of cells and from formalin-fixed tissue. One can interrogate specific CpG sites or, through sequencing, profile the entire genome. This chapter provides a review of statistical treatments to several problems in the analysis of DNA methylation data.

33.1 A Brief Introduction

In every cell of a human's body the genetic sequence is nearly identical. However, in order to specialize in function, each cell type expresses a characteristic subset of genes. This information is encoded by epigenetics. Epigenetics, derived from Greek meaning 'upon' genetics, refers to the transmission of information at cell division regarding expression of genes to daughter cells. Mechanisms conveying epigenetic information in humans are not fully understood, but are known to involve the interrelated processes of DNA methylation, histone tail modification and chromatin structure. Although it is the combination of these factors and others that results in gene expression or silencing, DNA methylation, a hallmark of epigenetic information, is the focus of this chapter.

Mammalian DNA methylation occurs when a methyl group is added to the cytosine residue in the context of a CpG dinucleotide (Jaenisch and Bird, 2003). In somatic cells, a large portion of genomic DNA is methylated at CpGs. Methylated cytosines contribute to hotspots for C → T mutation through spontaneous deamination to uracil followed by poor DNA repair. This results in a depletion and non-random distribution of CpG dinucleotides throughout the genome. Regions dense in CpG sites and predominantly unmethylated are called CpG islands (CGIs). One definition of CGIs is regions of at least 200 bp with a C + G content of 50% or more and a ratio of observed to expected CpG dinucleotides greater than 0.6 (Gardiner-Garden and

Frommer, 1987). Approximately half of all gene promoters occur in CGIs (Straussman *et al.*, 2009). Regions flanking CGIs and approximately 2 kb in length show cell-type specific variation in DNA methylation (Irizarry *et al.*, 2009). These regions became known as CpG island shores. Later, the geographical analogy extended from CpG islands to CpG island shelves (less than 2 kb from CGI), CpG island shelves (2–4 kb from CGI) and open sea (more than 4 kb from CGI) (Bibikova *et al.*, 2011).

Variation in DNA methylation is frequently, but not always, associated with variation in gene expression. Typically, CGI promoter methylation correlates with gene silencing. Examples of normal DNA methylation and gene silencing include X chromosome inactivation (Ross *et al.*, 2005) and genomic imprinting (Kacem and Feil, 2009). Another regulatory element that shows inverse correlations between CpG methylation and gene expression are enhancers (Yao *et al.*, 2015). In fact, enhancer methylation can be a better predictor of gene expression than promoter methylation, especially for cell-type specific regulation of genes with unmethylated promoters (Aran and Hellman, 2014). Gene bodies, on the other hand, are low in CpG density and largely methylated. Methylation of gene bodies supports transcription elongation, and positive correlations are observed between cytosine methylation and gene expression (Moen *et al.*, 2014; Yang *et al.*, 2014). Many DNA methylation variable sites show no correlation with expression. Nevertheless, in studies of complex disease, DNA methylation can provide a stable biomarker with clinical utility.

In this chapter we will focus on introducing, describing and discussing several problems and statistical treatments, related to the identification of differentially methylated regions and their use in human population-based research.

33.2 Measuring DNA Methylation

DNA methylation states are not preserved during polymerase chain reaction (PCR) nor do they appear to affect DNA hybridization. These facts make conventional approaches for measuring properties of DNA, which frequently incorporate one or both of these protocols, unusable for querying methylation states. To address this, researchers have over the years invented many different measurement methods, some of them using restriction enzymes which preferentially cut methylated or unmethylated DNA, and some using antibodies which bind to either methylated or unmethylated DNA. Laird (2010) contains a full review of these methods.

However, at the time of writing, all of these methods have largely been superseded by the use of bisulfite conversion (Clark *et al.*, 1994; Frommer *et al.*, 1992), coupled with either microarrays or high-throughput DNA sequencing. Bisulfite conversion is a chemical modification whereby unmethylated cytosines are converted to uracil and methylated cytosines are unconverted. This transforms the information encoded in the methylation mark into a difference at the DNA level. By comparing the nucleotides of the bisulfite converted DNA to those present in the sample DNA, it is possible to infer whether the DNA was methylated or unmethylated. This comparison is typically done either using DNA microarrays designed for measuring methylation or using high-throughput sequencing. Indeed, the gold-standard method for measuring DNA methylation at a specific locus is bisulfite conversion coupled with pyrosequencing. In practice, comparisons are made to a reference genome and not the sample genome. This can sometimes be problematic because the most common dinucleotide change in humans is a CG-to-TG transition, and this genetic change is read as a site of unmethylation following bisulfite conversion. In other words, (common) genetic changes between the reference genome and the sample genome can lead to bias in the inferred methylation state (Gao *et al.*, 2015; Liu *et al.*, 2012; Wulfridge *et al.*, 2018).

In mammals, most cytosines do not exist in a CpG context and are therefore unmethylated and converted by the bisulfite treatment. It follows that bisulfite converted DNA is heavily depleted of cytosines – essentially consisting of only three bases – which can cause issues for subsequent handling in the laboratory. For example, the base-calling software for the widely used Illumina HiSeq series has problems with bisulfite converted DNA. To address this, it is currently recommended to spike in an external genome such as phiX to achieve a more balanced base composition of the sample. This comes at the cost of effectively reducing sequencing depth.

DNA methylation is commonly analyzed as a continuous measure in the unit interval, representing the percentage of cells present in the sequenced sample in which a locus is methylated. To measure this quantity, an assay needs to measure two channels (methylation and unmethylation) in the same sample, which are subsequently combined to form a single methylation percentage. This process can be viewed as within-sample standardization, and could explain why DNA methylation data often have less need for normalization compared to gene expression data. Bisulfite conversion efficiency may vary between samples, but experienced laboratories are capable of producing data with consistent high conversion rates (usually estimated through spike-in of external unmethylated DNA such as the lambda phage genome).

Combining bisulfite conversion with DNA microarrays has led to an explosion in population-level studies of DNA methylation in humans through the popular Infinium HumanMethylation450K and HumanMethylationEPIC array platforms from Illumina, more commonly referred to by the names 450k and EPIC (Bibikova *et al.*, 2011). This is due to the relatively low cost of these platforms, coupled with the ability to quickly process many samples. For each site, the array needs to quantify the amount of methylated DNA and the amount of unmethylated DNA. To accomplish this, these microarrays are composed of a mixture of single-color and two-color probes, where the two channels needed to measure DNA methylation either come from two different microarray probes measured in a single channel, or from a single microarray probe measured in two color channels. This makes the array design unusually complicated (see Bibikova *et al.*, 2011, for details).

One complication with using DNA microarrays to measure bisulfite converted DNA is the case where a microarray probe overlaps a nearby CpG – different from the CpG of interest – which may or may not be converted depending on its methylation state. This is particularly relevant in areas of high CpG density such as CpG Islands. To address this, assumptions are made at the design level linking the methylation state of the CpGs in the probe to the methylation level at the CpG of interest. This explains the complicated design of the Illumina methylation arrays, as using two probes supposedly better tolerates the presence of multiple CpGs, at the cost of real estate on the array. This claim is due to Illumina, and is partly supported by published data (Bibikova *et al.*, 2011).

Combining bisulfite conversion with high-throughput sequencing is straightforward, but expensive, at the whole-genome level. The assay to do this is called whole-genome bisulfite sequencing (WGBS). One aspect of any type of bisulfite sequencing is coverage, which is the number of sequenced fragments covering a locus. Higher coverage leads to more precise estimates of the methylation percentage but comes at a higher cost. The coverage necessary for WGBS is debated in the literature, with opinions ranging from 10 \times to 100 \times (Libertini *et al.*, 2016; Ziller *et al.*, 2015). To decrease costs, two methods have been proposed to select subsets of the CpGs in the genome. The first, called reduced representation bisulfite sequencing (RRBS), uses a cocktail of restriction enzymes to select a part of the genome with high CpG density and is dramatically cheaper than WGBS (Meissner *et al.*, 2005), albeit with very uneven sample-to-sample coverage. A recent alternative is capture bisulfite sequencing, where a microarray is used to select parts of the genome, similar to how exome sequencing is performed in

measuring DNA. The design of capture bisulfite sequencing platforms is complicated by the bisulfite conversion process, evidenced by the amount of time it has taken to reliably develop these platforms (Ziller *et al.*, 2016). Following sequencing, methylation states are inferred during alignment of the reads to a genome. The alignment process is made more challenging by the fact that it takes place in bisulfite-converted space, meaning that the genome sequence depends on the methylation state. The most common approach is to computationally convert all cytosines to thymines, in both the sequencing reads and the reference genome. Bisulfite sequencing also has the potential to yield allele-specific methylation.

The advent of third-generation sequencing, or single-molecule sequencing, has the potential to have a large impact on the study of DNA methylation, using sequencing platforms such as Oxford Nanopore or PacBio SMRT sequencing. These sequencing platforms make it possible to differentiate between methylated and unmethylated cytosines (Flusberg *et al.*, 2010; Simpson *et al.*, 2017), which will make bisulfite conversion unnecessary, and routinely yield DNA methylation state every time a sample is DNA sequenced.

33.3 Differential DNA Methylation

33.3.1 Differential Methylation with Bisulfite-Sequencing Data

Traditional as well as novel statistical methods have been proposed in recent years to analyze DNA methylation data generated from bisulfite-sequencing (BS-seq) technologies. When there are no biological replicates, BS-seq data, whether generated from WGBS or RRBS, are binomial in nature at each CpG site. Therefore established traditional statistical methods have been used to analyze such data to find differentially methylated CpGs due to a biological factor of interest; Fisher's exact test is a popular tool, for example (Lister and Ecker, 2009). In studies that investigate the role of DNA methylation in human diseases, there tend to be replicates and therefore the beta-binomial distribution (or, more generally, beta and binomial distributions combined) is typically used to capture between-sample variability within a biological condition (i.e. a level of the biological factor of interest such as cases or controls of a particular disease) (Dolzhenko and Smith, 2014; Feng *et al.*, 2014; Park *et al.*, 2014). Such beta-binomial-based methods are used to detect differentially methylated cytosines (DMCs), which are essentially CpG sites that show significant differential methylations among the levels of a biological factor. When a biological condition is influenced by methylation aberration in a whole region rather than at single CpG sites, then the interest is shifted toward detecting differentially methylated regions (DMRs), which are frequently obtained by merging DMCs through setting various thresholds (Dolzhenko and Smith, 2014; Hansen *et al.*, 2012; Hebestreit *et al.*, 2013). An alternative approach is to find DMRs directly, which has the advantage of more fully taking spatial correlation of neighboring methylation signals into consideration (Park and Lin, 2018). It is also important to accommodate biological and environmental covariates, such as age, sex, and smoking status, as such variables are potential confounders of differential methylation. In the following, we restrict ourselves to the discussion of methods for detecting DMCs and DMRs when biological replicates exist; review papers are available for a more general perspective (Qin *et al.*, 2016; Robinson *et al.*, 2014a; Shafi *et al.*, 2018).

33.3.1.1 Beta-Binomial-Based Methods for Detecting DMCs and DMRs

For each CpG site and each sample, the observed BS-seq data can be viewed as coming from a binomial distribution with the number of trials being the total reads and the count of methylated reads regarded as a realization from the binomial distribution. To fix ideas, we consider a

two-group setting for finding DMCs and DMRs, where these two groups are referred to as ‘cases’ and ‘controls’ for convenience. In other words, the biological factor of interest is ‘disease status’ with two levels, affected (case) or unaffected (control), and one is interested in testing whether there is a difference in the methylation levels in these two groups. To account for between-sample variability within the same group (i.e. samples that are biological replicates), a beta distribution is typically used to model the probability of ‘success’ in the binomial distribution that is used to model the observed count of methylated reads.

Specifically, at a given CpG site, let N_{ij} and M_{ij} be, respectively, the counts of total and methylated reads for sample j ($j = 1, 2, \dots, n_i$) in group i ($i = 1, 2$). Then the above conceptual description of the model leads to the following specification:

$$M_{ij} \mid p_{ij} \sim \text{bin}(N_{ij}, p_{ij}) \quad \text{and} \quad p_{ij} \sim \text{beta}(\alpha_{ij}, \beta_{ij}),$$

where p_{ij} is the probability that a read is methylated, and $\alpha_{ij} > 0$ and $\beta_{ij} > 0$ are the natural parameters of the beta distribution. Thus, $\mu_{ij} = E(p_{ij}) = \alpha_{ij}/(\alpha_{ij} + \beta_{ij})$ and $\sigma_{ij}^2 = \text{Var}(p_{ij}) = \mu_{ij}(1 - \mu_{ij})\delta_{ij}$, where $\delta_{ij} = 1/(\alpha_{ij} + \beta_{ij} + 1)$. Integrating over the distribution of p_{ij} used to capture the between-sample variability, one can find the unconditional probability mass function of M_{ij} (a beta-binomial distribution),

$$P(M_{ij} = m_{ij}) = \binom{N_{ij}}{M_{ij}} \frac{\text{Beta}(m_{ij} + \alpha_{ij}, N_{ij} - m_{ij} + \beta_{ij})}{\text{Beta}(\alpha_{ij}, \beta_{ij})}, \quad (33.1)$$

for a realization m_{ij} of the random variable M_{ij} , where Beta is the usual beta function. Then $\text{Var}(M_{ij}) = N_{ij}\mu_{ij}(1 - \mu_{ij})(1 + \delta_{ij})$. Since $\delta_{ij} > 0$, this leads to an overdispersed setting compared to a simple binomial setting; and thus, the above-mentioned capability of equation (33.1) for accounting for between-sample variability. If covariates exist, we can use \mathbf{Z}_{ij} to denote the design vector of the sample. Then we can use a generalized linear modeling framework to model the mean of the beta distribution μ_{ij} as

$$g(\mu_{ij}) = \eta_0 + \eta_1 I(i = 2) + \mathbf{Z}_{ij} \boldsymbol{\eta}, \quad i = 1, 2, j = 1, 2, \dots, n_i,$$

where g is a link function, I is the usual indicator function, and η_0, η_1 and $\boldsymbol{\eta}$ are the intercept, the group (treatment) effect, and the coefficients of the covariates, respectively. Note that this model connects an individual sample’s mean value of DNA methylation with its covariate values. Nevertheless, regardless of whether covariates are present, the null (H_0) and alternative (H_1) hypotheses are:

$$H_0 : \eta_1 = 0 \quad \text{versus} \quad H_1 : \eta_1 \neq 0. \quad (33.2)$$

If H_0 is rejected in favor of H_1 , then we declare that the CpG site is a DMC.

For RADMeth (Dolzhenko and Smith, 2014), one of a handful of currently available methods that can account for covariates and have been compared favorably (Huh *et al.*, 2019), $\boldsymbol{\eta}$ will also be estimated to avoid confounding (and especially spurious associations), and a logistic function is typically used as g . However, we note that RADMeth can only accommodate categorical covariates. How well RADMeth can account for such, given the small number of replicates (it is common to see only three samples in each of the two groups) typically seen in BS-seq data and especially if there are multiple covariates, has yet to be carefully investigated. In fact, even though accounting for covariates is a capability of RADMeth, no demonstrations were provided (Dolzhenko and Smith, 2014). Several other methods, including MethylKit (Akalin *et al.*, 2012), MACAU (Lea *et al.*, 2015), and DSS-general (Park and Wu, 2016), can also take covariates into account, although little comparison among these methods has been made.

In the absence of covariates, this general setting reduces to several other beta-binomial based methodologies (Dolzhenko and Smith, 2014; Feng *et al.*, 2014; Hebestreit *et al.*, 2013; Park *et al.*, 2014). Under such a scenario, $\eta = \mathbf{0}$, and the test in (33.2) is equivalent to testing the two group means being the same (H_0) or different (H_1). In this case, there is no individual-specific information for the beta parameters; and thus, the α and β parameters will only need to be indexed by i (group), not j (individual sample), at each CpG site.

With the specification of this model for BS-seq methylation data, pure likelihood-based (e.g. BiSeq (Hebestreit *et al.*, 2013), MethylSig (Park *et al.*, 2014) and RADMeth (Dolzhenko and Smith, 2014)) as well as empirical Bayes methods (e.g. DSS (Feng *et al.*, 2014)) are used to estimate the parameters; Wald tests (e.g. BiSeq and DSS) or likelihood ratio tests (e.g. MethylSig and RADMeth) are then used to test the hypotheses in (33.2).

If desired, false discovery rates rather than site-wise p -values for these tests can be used to make decisions about whether a CpG site is a DMC. Recognizing the fact that methylation levels at neighboring sites are spatially correlated (Eckhardt *et al.*, 2006; Hansen *et al.*, 2012; Irizarry *et al.*, 2008), various ways of taking such a feature into consideration have also been proposed, from ‘smoothing’ the observed methylation level (e.g. BiSeq), to ‘pooling’ count data from a predefined window (e.g. MethylSig), ‘shrinkage’ estimation of parameters (e.g. DSS), and ‘combination’ of p -values across CpG sites within a predefined window (e.g. MethylSig). A related method that exploits the idea of ‘smoothing’ is BSmooth (Hansen *et al.*, 2012), which is similar to BiSeq in that the methylation level at a CpG site is estimated based not only on its own observed data, but also on the neighboring sites.

33.3.1.2 Direct Detection of DMRs

DMCs found using the methods described above can be merged to form DMRs through thresholding, which includes rules on directionality (all hypo- or hypermethylated) (Hebestreit *et al.*, 2013), minimum length (Feng *et al.*, 2014), maximum length between adjacent CpG sites (Hansen *et al.*, 2012), minimum number of DMCs (Feng *et al.*, 2014; Hansen *et al.*, 2012), and minimum percentage of DMCs among all CpG sites (Feng *et al.*, 2014). Simply merging adjacent DMCs is also practiced (Dolzhenko and Smith, 2014). As can be easily seen, the rules placed to form DMRs from DMCs are arbitrary, and different methods use different sets of rules.

To take such arbitrariness out of the equation, and more importantly, to maximally utilize the information available to find DMRs directly rather than through the merging of DMCs, one may consider an alternative approach such as BCurve (Park and Lin, 2018). To ensure that a DMR is not stretched over a region with sparse coverage of CpG sites (which can be the case with RRBS data), BCurve first finds CpG clusters so that two adjacent CpG sites (that are included in a particular study) within a cluster are close enough to each other, similarly to the procedure employed in BiSeq (Hebestreit *et al.*, 2013). Within each cluster, BCurve hypothesizes that the underlying latent methylation value at each CpG site falls on a smooth curve. Specifically, BCurve assumes that the latent methylation value follows a Gaussian distribution with the mean being the probit transformation of the probability that a read is methylated. Within-group variability is also introduced at the latent methylation value level so that each sample within a group has its own mean. Unlike BiSeq, BSmooth, and MethylSig where kernel smoothing was used, BCurve uses B-splines for constructing the smoothing function.

Let \mathbf{p}_{ij} denote the column vector of methylation probabilities in a CpG cluster for observation j in group i , with each entry representing a CpG site in the cluster. We use \mathbf{B} to denote the design matrix of the B-splines, with the number of rows being the number of CpG sites in the cluster and the number of columns being the number of basis functions, which is determined by the number of knots of the B-splines. Further, we let ξ_{ij} be a column vector of random variables, one for each CpG site in the cluster, which are assumed to follow a normal distribution with mean

zero and a common variance within a group. This leads to the specification of the following model for each group i separately ($i = 1, 2$):

$$\Phi^{-1}(\mathbf{p}_{ij}) = \mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\xi}_{ij}, \quad j = 1, 2, \dots, n_i, \quad (33.3)$$

where Φ is the cumulative distribution function of a standard normal distribution and $\boldsymbol{\gamma}$ is the coefficient of the B-spline basis functions. Although $\boldsymbol{\xi}_{ij}$ is specified generally in the above discussion, BCurve sets $\boldsymbol{\xi}_{ij} = \boldsymbol{\xi}_{ij}\mathbf{1}$, where $\mathbf{1}$ is a vector of 1s, so that each smooth curve has the same displacement from the overall mean at all CpG sites (Park and Lin, 2018). That being said, different amounts of between-sample variability in the two groups are accommodated for in this formulation; for example, if larger variability among the cases is suspected compared to that among the controls, then the variance estimates will tend to be larger for the former.

As discussed above, data from each of the two groups are analyzed separately, that is, the parameters of the models will be estimated twice, one for each of the two conditions (groups). For each group, the confidence band for the methylation levels along the CpG sites in the cluster at a desired level is constructed. The regions for which the two confidence bands (one from each) do not overlap are then taken as DMRs (Park and Lin, 2018).

33.3.2 Differential Methylation with Capture-Sequence Data

For experiments with the capture-sequence (cap-seq) based technologies, since the observed data are no longer of nucleotide resolution, the statistical methods for analyzing such data are markedly different from those for analyzing BS-seq data. Although nucleotide-resolution WGBS technology is the gold standard, due to its formidable cost, a population study aiming for whole-genome coverage on a reasonable number of samples would inevitably have to consider an alternative technology, and cap-seq is a viable alternative. However, the technology is still infrequently utilized due to several distinctive features that make the analysis of such data extremely challenging. Features that pose the greatest challenges to devise good analytical methods include the following. First, for some cap-seq technology such as MethylCap-seq (Frankhouser *et al.*, 2014), a pulldown fragment being sequenced may contain multiple CpG sites, and it is unknown which CpG(s) are methylated and responsible for the pulldown as the only requirement is that there is at least one. Second, since short-read sequencing is done from the ends of fragments, reads may not even cover any CpG sites. Third, if sequencing is done from a single end, the length of each pulldown fragment is unknown. As such, a typical window-based approach may lead to phantom DMRs (Frankhouser *et al.*, 2014). Finally, the uneven density of CpG sites throughout the genome can adversely affect the quality of data, leading to bias in DMR calls for dense CpG regions.

Standard methods for differential analysis of gene expression and DNA–protein interaction data have found their way into the analysis of cap-seq methylation data. First, there is the most basic window-based approach that uses standard *t*-tests for DMR finding (Lienhard *et al.*, 2014; Yan *et al.*, 2012). In this approach, reads in predefined windows are aggregated and compared between the two conditions. A second approach utilizes the peak detection methods developed for the study of DNA–protein binding via chromatin-immunoprecipitation sequencing (Qin *et al.*, 2016). Further discussion of these methods can be found in review papers (Qin *et al.*, 2016; Robinson *et al.*, 2014a). In a cap-seq-specific development, Frankhouser *et al.* (2014) attempt to infer ‘nucleotide level’ data from the pulldown reads to address problems encountered with cap-seq data. Exploiting the availability of the fragment sizes distribution, a probabilistic model was devised to assign each read to all CpG sites. If sample sizes are sufficiently large and the number of CpG sites is relatively small, then the Hotelling’s T^2 test would be a possible choice. In reality, the number of CpG sites can easily be larger than the sample size, begging the

use of high-dimensional counterparts to Hotelling T^2 statistics (Ayyala *et al.*, 2016; Chen and Qin, 2010).

As a final remark, it is noted that although multiple attempts have been made to better utilize and analyze cap-seq data, this area is still under-explored. As such, better statistical methods are still lacking in order to fully capitalize on the whole-genome data offered by this type of more affordable technology.

33.3.3 Differential Methylation with HumanMethylation Array Data

33.3.3.1 Normalization of HumanMethylation Array Data

Like all microarray data, HumanMethylation microarrays can benefit from normalization prior to analysis. This has been the topic of much work since the release of the 450k array platform. A special issue for normalization of HumanMethylation array data is that some studies concern comparisons where a large portion of the measured methylome changes across conditions, violating common assumptions made in the normalization literature. Examples of such studies include comparison of cancer and normal samples as well as comparisons of samples from different tissues or cell types. Good approaches for such studies include noob (Triche *et al.*, 2013) and functional normalization (Fortin *et al.*, 2014), and have been comprehensively evaluated (Fortin *et al.*, 2014; Liu and Siegmund, 2016). The noob method was later modified to be single-sample, that is, the output of the method does not depend on which other samples are processed concurrently (Fortin *et al.*, 2017). This has clear advantages for certain problems such as biomarker development. In this setting it is important to note that the absence of normalization can sometimes outperform existing normalization methods (Dedeurwaerder *et al.*, 2014).

Another large application of the HumanMethylation array platform is in epigenome-wide association studies that involve changes in only a small portion of the measured CpGs. For such studies, other normalization methods may be attractive. Quantile normalization (Bolstad *et al.*, 2003) can be successful for studies of small effect sizes in homogeneous samples, as shown by a study of prenatal tobacco smoke exposure and childhood CpG methylation in whole blood (Breton *et al.*, 2014). Subset quantile normalization (SQN) is a flexible alternative that should tolerate heterogeneous samples better. SQN equalizes the probe intensities both across and within arrays using a model sensitive to the probe design, DNA methylation state, and CpG density of the targeted genomic location (Aryee *et al.*, 2014).

Many genomic studies suffer from batch effects (Leek *et al.*, 2010) and it is prudent to use batch correction methods such as SVA (Leek and Storey, 2007, 2008) or RUVm (Maksimovic *et al.*, 2015) – an adaptation of the RUV method (Gagnon-Bartsch and Speed, 2012) to this platform. Adjusting for batch effects can be more important than normalization; some normalization methods can adjust for some types of batch effects.

33.3.3.2 Identification of DMCs and DMRs

Since the HumanMethylation arrays profile a set number of CpGs across samples, the most straightforward analysis is the identification of DMCs. For each cytosine the array targets, the estimated DNA methylation proportion, traditionally (and confusingly) called beta value, resembles a beta-distributed variable, continuous and bounded between 0 and 1. Often these are transformed using a logit transformation on the base-2 log scale (Du *et al.*, 2010) or using a $N(0, 1)$ -quantile normalization (Bell *et al.*, 2011) and analyzed using ordinary linear regression. Beta regression, a desirable approach when sample sizes are large enough to support the flexible modeling of dispersion (Triche *et al.*, 2016), is computationally intensive and has not shown sufficient advantage to be adopted in the field.

While many measured CpGs are distal to other measured CpGs, a significant fraction are close together (say, within 1 kb) making it possible to attempt to find DMRs (Aryee *et al.*, 2014). This can be done by clustering the results of a DMC analysis or by directly inferring DMRs, reviewed in Robinson *et al.* (2014b). For example, the comb-p package uses methods from meta-analysis to combine *p*-values from a DMC analysis into regions (Pedersen *et al.*, 2012). Similarly, DMRcate (which can handle both array and sequencing data) starts with a linear model at the single site level and aggregates the output of such an analysis into regions (Peters *et al.*, 2015). In contrast, the bump hunting method from minfi estimates a smoothed methylation contrast between groups and uses this to directly estimate regions without performing tests at the single cytosine level (Aryee *et al.*, 2014). Likewise, Aclust finds clusters of correlated cytosines and then tests for association between the cluster and the covariate of interest (Sofer *et al.*, 2013). Note that region-level approaches are only able to analyze a subset of the measured cytosines, those with other measured cytosines in the region.

33.4 Other Topics of Interest

The standardization of technologies for measuring DNA methylation has accelerated scientific progress. WGBS allows for a true genome-wide analysis, allowing the discovery of differentially methylated regions unmeasured by targeted approaches. However, as mentioned earlier, not all DNA methylation aberrations alter gene expression. In cancer, many DNA methylation marks are passenger events and do not drive cancer progression. Nevertheless, their cell-type specificity makes them excellent targets for monitoring treatment response and disease progression. Dead cells, including cancer cells, shed DNA fragments into circulation. The identification of cancer-specific DNA methylation in circulating cell-free DNA from plasma offers a new method for monitoring patients called liquid biopsy. DNA can also be found in urine sediment, another non-invasive tissue. For a review of DNA methylation in cancer surveillance, see Liang and Weisenberger (2017).

Low-cost microarrays have led to a wealth of research from large observational studies. Studies of cancer can measure the DNA methylation directly from resected cancer tissue. These have led to the discovery of DNA methylation-related cancer subtypes that inform therapeutic decisions. Many cancer studies also measure DNA methylation in adjacent non-cancer tissue in order to identify cancer-specific alterations. Studies of non-cancer traits are complicated by the unavailability of the target tissue, for instance lung from asthma patients, or brain from subjects with Alzheimer's disease. In such instances surrogate tissues are used, blood being a common choice.

All human tissues are a mixture of different cell types. Whole blood is a complex mixture of more than six different cell types (monocytes, neutrophils, eosinophils, natural killer cells, B-cells and T-cells of different sorts). Cancer tissue is also a mixture, with cancer cells subject to stromal and immune cell infiltration. As cell type is one of the largest determinants of DNA methylation variation, epigenome-wide association studies (EWAS) are subject to confounding by cell type (Michels *et al.*, 2013). Differences in the fractions of the cell composition can cause methylation differences between groups that are not due to differences in DNA methylation of the cell genomes.

Multiple analytic approaches exist to control for the DNA methylation variation caused by the fractional composition in EWAS. The methods are classified as either reference-based or reference-free. In reference-based approaches, a panel of measured DNA methylation profiles of the purified cell types is used to infer the cell type fractions in the mixed populations, with the fractions later treated as covariates in the EWAS. Reference-free approaches have the

advantage of not requiring a reference panel, while also adjusting for any other unwanted variation in the data (e.g. age or processing batch effects). Several in-depth evaluations comparing different methods have found that no single method performs best under all circumstances (McGregor *et al.*, 2016; Teschendorff *et al.*, 2017). A recent comparison of reference-based methods found that the best method depended on the amount of noise in the data (Teschendorff *et al.*, 2017). Under realistic levels of noise, new unconstrained approaches outperformed the most commonly used reference-based approach. However, when comparing between the unconstrained approaches, the best approach differed depending on whether the mixed cell population included epithelial cells or blood. For the reference-free methods, an improved surrogate variable analysis method, SmartSVA, was reported to outperform earlier approaches by reducing false positives and increasing statistical power (Chen *et al.*, 2017). Importantly, the field has learned that small effect sizes are difficult to study in samples of mixed cell populations due to the sensitivity of the results to the analytic method.

Another active area of research in DNA methylation is the study of aging. Age is another major determinant of DNA methylation. Globally, we see loss of DNA methylation in cells with age, however hypermethylation is observed at CpG island promoters. Horvath (2013) developed an algorithm to predict chronological age from DNA methylation measured using Illumina HumanMethylation arrays. The DNA methylation age predictor was built using elastic net regression. Elastic net regression and support vector machines were recommended for prediction modeling in DNA methylation data sets with lots of small signal strengths and effect sizes (Zhuang *et al.*, 2012). Many important studies of aging have resulted through application of these approaches. One study showed the epigenetic age is zero in embryonic and induced pluripotent stem cells while cancer shows age acceleration (Horvath, 2013, 2015). Measured in blood, DNA methylation age predicts time to death independent of chronological age and typical risk factors (Chen *et al.*, 2016). A recent genome-wide association study identified *TERT* as playing an important role in epigenetic aging (Lu *et al.*, 2018).

Finally, DNA methylation is believed to be an intermediate variable, and a possible mechanism that explains the associations between exposures and disease outcome. CpG methylation has been shown to be associated with single nucleotide polymorphisms (SNPs), termed methylation quantitative trait loci, and with environmental exposures such as air pollution, prenatal condition, and personal smoking history. By jointly considering the effects of DNA methylation and SNPs, and potentially environmental exposures, the mechanism of how these multiple factors affect a trait can start to be addressed. In particular, the effects of SNPs or environmental exposures on an outcome through DNA methylation have been studied through mediation analysis (Liu *et al.*, 2013; Tobi *et al.*, 2018). The analysis of high-dimensional mediators presents new analytical challenges that are currently an active area of research (Barfield *et al.*, 2017; Chén *et al.*, 2018).

References

- Akalin, A., Kormaksson, M., Li, S., Bakelman, F.G., Figueroa, M., Melnick, A. and Mason, C. (2012). methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**(10), R87.
- Aran, D. and Hellman, A. (2014). Unmasking risk loci: DNA methylation illuminates the biology of cancer predisposition. *BioEssays* **36**(2), 184–190.
- Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D. and Irizarry, R.A. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**(10), 1363–1369.

- Ayyala, D., Frankhouser, D., Ganbat, J., Marcucci, G., Bundschuh, R., Yan, P. and Lin, S. (2016). Statistical methods for detecting differentially methylated regions based on MethylCap-seq data. *Briefings in Bioinformatics* **17**(6), 926–937.
- Barfield, R., Shen, J., Just, A.C., Vokonas, P.S., Schwartz, J., Baccarelli, A.A., VanderWeele, T.J. and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology* **41**(8), 824–833.
- Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12**(1), R10.
- Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L., Fan, J.-B. and Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* **98**(4), 288–295.
- Bolstad, B., Irizarry, R., Astrand, M. and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193.
- Breton, C., Siegmund, K., Joubert, B., Wang, X., Qui, W., Carey, V., et al. (2014). Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. *PLoS ONE* **9**(6), e99716.
- Chen, B.H., Marioni, R.E., Colicino, E., Peters, M.J., Ward-Caviness, C.K., Tsai, P.-C., Roetker, N.S., Just, A.C., Demerath, E.W., Guan, W., Bressler, J., Fornage, M., Studenski, S., Vandiver, A.R., Moore, A.Z., Tanaka, T., Kiel, D.P., Liang, L., Vokonas, P., Schwartz, J., Lunetta, K.L., Murabito, J.M., Bandinelli, S., Hernandez, D.G., Melzer, D., Nalls, M., Pilling, L.C., Price, T.R., Singleton, A.B., Gieger, C., Holle, R., Kretschmer, A., Kronenberg, F., Kunze, S., Linseisen, J., Meisinger, C., Rathmann, W., Waldenberger, M., Visscher, P.M., Shah, S., Wray, N.R., McRae, A.F., Franco, O.H., Hofman, A., Uitterlinden, A.G., Absher, D., Assimes, T., Levine, M.E., Lu, A.T., Tsao, P.S., Hou, L., Manson, J.E., Carty, C.L., LaCroix, A.Z., Reiner, A.P., Spector, T.D., Feinberg, A.P., Levy, D., Baccarelli, A., van Meurs, J., Bell, J.T., Peters, A., Deary, I.J., Pankow, J.S., Ferrucci, L. and Horvath, S. (2016). DNA methylation-based measures of biological age: Meta-analysis predicting time to death. *Aging* **8**(9), 1844–1865.
- Chen, J., Behnam, E., Huang, J., Moffatt, M.F., Schaid, D.J., Liang, L. and Lin, X. (2017). Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics* **18**(1), 413.
- Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D. and Lindquist, M.A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**(2), 121–136.
- Chen, S.X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set. *Annals of Statistics* **38**, 808–835.
- Clark, S.J., Harrison, J., Paul, C.L. and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Research* **22**(15), 2990–2997.
- Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G. and Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics* **15**(6), 929–941.
- Dolzenko, E. and Smith, A. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15**(1), 215.
- Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**(1), 587.
- Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C.,

- Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics* **38**, 1378–1385.
- Feng, H., Conneely, K.N. and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research* **42**(8), e69.
- Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7**(6), 461–465.
- Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B.W., Hudson, T.J., Fertig, E.J., Greenwood, C.M. and Hansen, K.D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology* **15**(11), 503.
- Fortin, J.-P., Triche Jr, T. and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**(4), 558–560.
- Frankhouser, D.E., Murphy, M., Blachly, J.S., Park, J., Zoller, M.W., Ganbat, J.-O., Curfman, J., Byrd, J.C., Lin, S., Marcucci, G., Yan, P. and Bundschuh, R. (2014). PrEMeR-CG: Inferring nucleotide level DNA methylation values from MethylCap-seq data. *Bioinformatics* **30**(24), 3567–3574.
- Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America* **89**(5), 1827–1831.
- Gagnon-Bartsch, J.A. and Speed, T.P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**(3), 539–552.
- Gao, S., Zou, D., Mao, L., Liu, H., Song, P., Chen, Y., Zhao, S., Gao, C., Li, X., Gao, Z., Fang, X., Yang, H., Ørntoft, F.T., Sørensen, K.D. and Bolund, L. (2015). BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics* **31**(24), 4006–4008.
- Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology* **196**(2), 261–282.
- Hansen, K., Langmead, B. and Irizarry, R. (2012). BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13**(10), R83.
- Hebestreit, K., Dugas, M. and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29**(13), 1647–1653.
- Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology* **14**(10), 3156.
- Horvath, S. (2015). Erratum to: DNA methylation age of human tissues and cell types. *Genome Biology* **16**(1), 96.
- Huh, I., Wu, X., Park, T. and Yi, S.V. (2019). Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. *Briefings in Bioinformatics* **20**(1), 33–46.
- Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B. and Feinberg, A.P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research* **18**(5), 780–790.
- Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J., Sabuncian, S. and Feinberg, A.P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics* **41**(2), 178–186.
- Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics* **33**, 245–254.

- Kacem, S. and Feil, R. (2009). Chromatin mechanisms in genomic imprinting. *Mammalian Genome* **20**, 544–556.
- Laird, P.W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews Genetics* **11**(3), 191–203.
- Lea, A.J., Tung, J. and Zhou, X. (2015). A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genetics* **11**(11), 1–31.
- Leek, J.T. and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**(9), 1724–1735.
- Leek, J.T. and Storey, J.D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America* **105**(48), 18718–18723.
- Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**(10), 733–739.
- Liang, G. and Weisenberger, D.J. (2017). DNA methylation aberrancies as a guide for surveillance and treatment of human cancers. *Epigenetics* **12**(6), 416–432.
- Libertini, E., Heath, S.C., Hamoudi, R.A., Gut, M., Ziller, M.J., Herrero, J., Czyz, A., Ruotti, V., Stunnenberg, H.G., Frontini, M., Ouwehand, W.H., Meissner, A., Gut, I.G. and Beck, S. (2016). Saturation analysis for whole-genome bisulfite sequencing data. *Nature Biotechnology* doi: 10.1038/nbt.3524.
- Lienhard, M., Grimm, C., Morkel, M., Herwig, R. and Chavez, L. (2014). MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**(2), 284–286.
- Lister, R. and Ecker, J.R. (2009). Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research* **19**(6), 959–966.
- Liu, J. and Siegmund, K.D. (2016). An evaluation of processing methods for HumanMethylation450 BeadChip data. *BMC Genomics* **17**(1), 469.
- Liu, Y., Siegmund, K.D., Laird, P.W. and Berman, B.P. (2012). Bis-SNP: Combined DNA methylation and SNP calling for bisulfite-seq data. *Genome Biology* **13**, R61.
- Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J. and Feinberg, A.P. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology* **31**(2), 142–147.
- Lu, A.T., Xue, L., Salfati, E.L., Chen, B.H., Ferrucci, L., Levy, D., Joehanes, R., Murabito, J.M., Kiel, D.P., Tsai, P.-C., Yet, I., Bell, J.T., Mangino, M., Tanaka, T., McRae, A.F., Marioni, R.E., Visscher, P.M., Wray, N.R., Deary, I.J., Levine, M.E., Quach, A., Assimes, T., Tsao, P.S., Absher, D., Stewart, J.D., Li, Y., Reiner, A.P., Hou, L., Baccarelli, A.A., Whitsel, E.A., Aviv, A., Cardona, A., Day, F.R., Wareham, N.J., Perry, J.R.B., Ong, K.K., Raj, K., Lunetta, K.L. and Horvath, S. (2018). GWAS of epigenetic aging rates in blood reveals a critical role for *TERT*. *Nature Communications* **9**(1), 387.
- Maksimovic, J., Gagnon-Bartsch, J.A., Speed, T.P. and Oshlack, A. (2015). Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. *Nucleic Acids Research* **43**(16), e106.
- McGregor, K., Bernatsky, S., Colmegna, I., Hudson, M., Pastinen, T., Labbe, A. and Greenwood, C.M. (2016). An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biology* **17**(1), 84.
- Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research* **33**(18), 5868–5877.
- Michels, K.B., Binder, A.M., Dedeurwaerder, S., Epstein, C.B., Greally, J.M., Gut, I., Houseman, E.A., Izzi, B., Kelsey, K.T., Meissner, A., Milosavljevic, A., Siegmund, K.D., Bock, C. and

- Irizarry, R.A. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods* **10**(10), 949–955.
- Moen, E.L., Stark, A.L., Zhang, W., Dolan, M.E. and Godley, L.A. (2014). The role of gene body cytosine modifications in MGMT expression and sensitivity to temozolomide. *Molecular Cancer Therapeutics* **13**(5), 1334–1344.
- Park, J. and Lin, S. (2018). Detection of differentially methylated regions using Bayesian curve credible bands. *Statistics for Biosciences* **10**(1), 20–40.
- Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**(10), 1446–1453.
- Park, Y., Figueroa, M.E., Rozek, L.S. and Sartor, M.A. (2014). MethylSig: A whole genome DNA methylation analysis pipeline. *Bioinformatics* **30**(17), 2414–2422.
- Pedersen, B.S., Schwartz, D.A., Yang, I.V. and Kechris, K.J. (2012). Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**(22), 2986–2988.
- Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., Lord, R.V., Clark, S.J. and Molloy, P.L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin* **8**(1), 6.
- Qin, Z., Li, B., Conneely, K.N., Wu, H., Hu, M., Ayyala, D., Park, Y., Jin, V.X., Zhang, F., Zhang, H., Li, L. and Lin, S. (2016). Statistical challenges in analyzing methylation and long-range chromosomal interaction data. *Statistics in Biosciences* **18**(2), 284–309.
- Robinson, M.D., Kahraman, A., Law, C.W., Lindsay, H., Nowicka, M., Weber, L.M. and Zhou, X. (2014a). Mini review: Statistical methods for detecting differentially methylated loci and regions. Preprint, bioRxiv 007120.
- Robinson, M.D., Kahraman, A., Law, C.W., Lindsay, H., Nowicka, M., Weber, L.M. and Zhou, X. (2014b). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics* **5**, 324.
- Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., Frankish, A., Lovell, F.L., Howe, K.L., Ashurst, J.L., Fulton, R.S., Sudbrak, R., Wen, G., Jones, M.C., Hurles, M.E., Andrews, T.D., Scott, C.E., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N.P., Hunt, S.E., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M.L., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D.A., Worley, K.C., Ainscough, R., Ambrose, K.D., Ansari-Lari, M.A., Aradhya, S., Ashwell, R.I.S., Babbage, A.K., Bagguley, C.L., Ballabio, A., Banerjee, R., Barker, G.E., Barlow, K.F., Barrett, I.P., Bates, K.N., Beare, D.M., Beasley, H., Beasley, O., Beck, A., Bethel, G., Blechschmidt, K., Brady, N., Bray-Allen, S., Bridgeman, A.M., Brown, A.J., Brown, M.J., Bonnin, D., Bruford, E.A., Buhay, C., Burch, P., Burford, D., Burgess, J., Burrill, W., Burton, J., Bye, J.M., Carder, C., Carrel, L., Chako, J., Chapman, J.C., Chavez, D., Chen, E., Chen, G., Chen, Y., Chen, Z., Chinault, C., Ciccodicola, A., Clark, S.Y., Clarke, G., Cleee, C.M., Clegg, S., Clerc-Blankenburg, K., Clifford, K., Cobley, V., Cole, C.G., Conquer, J.S., Corby, N., Connor, R.E., David, R., Davies, J., Davis, C., Davis, J., Delgado, O., DeShazo, D., Dhami, P., Ding, Y., Dinh, H., Dodsworth, S., Draper, H., Dugan-Rocha, S., Dunham, A., Dunn, M., Durbin, K.J., Dutta, I., Eades, T., Ellwood, M., Emery-Cohen, A., Errington, H., Evans, K.L., Faulkner, L., Francis, F., Frankland, J., Fraser, A.E., Galgoczy, P., Gilbert, J., Gill, R., Glöckner, G., Gregory, S.G., Gribble, S., Griffiths, C., Grocock, R., Gu, Y., Gwilliam, R., Hamilton, C., Hart, E.A., Hawes, A., Heath, P.D., Heitmann, K., Hennig, S., Hernandez, J., Hinzmann, B., Ho, S., Hoffs, M., Howden, P.J., Huckle, E.J., Hume, J., Hunt, P.J., Hunt, A.R., Isherwood, J., Jacob, L., Johnson, D., Jones, S., de Jong, P.J., Joseph, S.S., Keenan, S., Kelly, S., Kershaw, J.K., Khan, Z., Kioschis, P., Klages, S., Knights, A.J., Kosiura, A., Kovar-Smith, C., Laird, G.K., Langford, C., Lawlor, S., Leversha, M., Lewis, L., Liu, W., Lloyd, C., Lloyd, D.M., Lousegod, H., Loveland, J.E., Lovell, J.D., Lozado, R., Lu, J., Lyne, R., Ma, J., Maheshwari, M.,

- Matthews, L.H., McDowall, J., McLaren, S., McMurray, A., Meidl, P., Meitinger, T., Milne, S., Miner, G., Mistry, S.L., Morgan, M., Morris, S., Müller, I., Mullikin, J.C., Nguyen, N., Nordsiek, G., Nyakatura, G., O'Dell, C.N., Okwuonu, G., Palmer, S., Pandian, R., Parker, D., Parrish, J., Pasternak, S., Patel, D., Pearce, A.V., Pearson, D.M., Pelan, S.E., Perez, L., Porter, K.M., Ramsey, Y., Reichwald, K., Rhodes, S., Ridler, K.A., Schlessinger, D., Schueler, M.G., Sehra, H.K., Shaw-Smith, C., Shen, H., Sheridan, E.M., Shownkeen, R., Skuce, C.D., Smith, M.L., Sotheran, E.C., Steingruber, H.E., Steward, C.A., Storey, R., Swann, R.M., Swarbreck, D., Tabor, P.E., Taudien, S., Taylor, T., Teague, B., Thomas, K., Thorpe, A., Timms, K., Tracey, A., Trevanion, S., Tromans, A.C., d'Urso, M., Verduzco, D., Villasana, D., Waldron, L., Wall, M., Wang, Q., Warren, J., Warry, G.L., Wei, X., West, A., Whitehead, S.L., Whiteley, M.N., Wilkinson, J.E., Willey, D.L., Williams, G., Williams, L., Williamson, A., Williamson, H., Wilming, L., Woodmansey, R.L., Wray, P.W., Yen, J., Zhang, J., Zhou, J., Zoghbi, H., Zorilla, S., Buck, D., Reinhardt, R., Poustka, A., Rosenthal, A., Lehrach, H., Meindl, A., Minx, P.J., Hillier, L.W., Willard, H.F., Wilson, R.K., Waterston, R.H., Rice, C.M., Vaudin, M., Coulson, A., Nelson, D.L., Weinstock, G., Sulston, J.E., Durbin, R., Hubbard, T., Gibbs, R.A., Beck, S., Rogers, J. and Bentley, D.R. (2005). The DNA sequence of the human X chromosome. *Nature* **434**, 325–337.
- Shafi, A., Mitrea, C., Nguyen, T. and Draghici, S. (2018). A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics* **19**(5), 737–753.
- Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J. and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods* **14**(4), 407–410.
- Sofer, T., Schifano, E.D., Hoppin, J.A., Hou, L. and Baccarelli, A.A. (2013). A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* **29**(22), 2884–2891.
- Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nature Structural & Molecular Biology* **16**, 564–571.
- Teschendorff, A.E., Breeze, C.E., Zheng, S.C. and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinformatics* **18**(1), 105.
- Tobi, E.W., Slieker, R.C., Luijk, R., Dekkers, K.F., Stein, A.D., Xu, K.M., Biobank-Based Integrative Omics Studies Consortium, Slagboom, P.E., van Zwet, E.W., Lumey, L.H. and Heijmans, B.T. (2018). DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Science Advances* **4**(1), eaao4364.
- Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. and Siegmund, K.D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research* **41**(7), e90.
- Triche, T.J., Laird, P.W. and Siegmund, K.D. (2016). Beta regression improves the detection of differential DNA methylation for epigenetic epidemiology. Preprint, bioRxiv 054643.
- Wulfridge, P., Langmead, B., Feinberg, A.P. and Hansen, K.D. (2018). Analyzing whole genome bisulfite sequencing data from highly divergent genotypes. Preprint, bioRxiv 076844.
- Yan, P., Frankhouser, D., Murphy, M., Tam, H.-H., Rodriguez, B., Curfman, J., Trimarchi, M., Geyer, S., Wu, Y.-Z., Whitman, S.P., Metzeler, K., Walker, A., Klisovic, R., Jacob, S., Grever, M.R., Byrd, J.C., Bloomfield, C.D., Garzon, R., Blum, W., Caligiuri, M.A., Bundschuh, R. and Marcucci, G. (2012). Genome-wide methylation profiling in decitabine-treated patients with acute myeloid leukemia. *Blood* **120**(12), 2466–2474.
- Yang, X., Han, H., Carvalho, D.D.D., Lay, F.D., Jones, P.A. and Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**(4), 577–590.

- Yao, L., Shen, H., Laird, P.W., Farnham, P.J. and Berman, B.P. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biology* **16**(1), 105.
- Zhuang, J., Widschwendter, M. and Teschendorff, A.E. (2012). A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics* **13**(1), 59.
- Ziller, M.J., Hansen, K.D., Meissner, A. and Aryee, M.J. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods* **12**(3), 230–232.
- Ziller, M.J., Stamenova, E.K., Gu, H., Gnirke, A. and Meissner, A. (2016). Targeted bisulfite sequencing of the dynamic DNA methylome. *Epigenetics & Chromatin* **9**(1), 55.

34

Statistical Methods in Metabolomics

Timothy M.D. Ebbels,¹ Maria De Iorio,² and David A. Stephens³

¹Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, UK

²Department of Statistical Science, University College London, London, UK

³Department of Mathematics and Statistics, McGill University, Montreal, Canada

Abstract

Metabolomics studies the levels of small molecules in living systems and can reveal much about an organism's state of health or disease. The data is complex and requires tailored data processing and modelling approaches. Here we review recent developments in the field, including data pre-processing, univariate and multivariate approaches, network and pathway analysis, and methods aiding metabolite identification.

34.1 Introduction

Metabolites are the small molecules which provide the basic building blocks and energy to sustain life. They are involved in almost all basic biological processes, from the synthesis of macromolecules (such as DNA or proteins) to signalling and energy transport. Metabolite levels form a characteristic fingerprint of the biological system, and systematically change during episodes of stress or pathology. Metabolomics (also known as metabolic profiling or metabolic phenotyping) is the study of the myriad of metabolites present in a biofluid, cell or tissue, and how they change over time and under different conditions. Since metabolites are often the end points of cellular processes, they are often considered to be closer to the phenotype than, for example, genomic or transcriptomic information. For this reason, and with the potential to discover new clinically relevant information in many areas, metabolomics as a field has greatly expanded in the last 10 years.

Metabolomics can be applied in two different modes: targeted and untargeted. In the targeted approach an assay is developed for a specific set of known metabolites, typically a few tens to low hundreds, often from the same chemical class (e.g. fatty acids). In untargeted metabolomics, one aims to profile as many metabolites across as many chemical classes as possible. Often the latter approach yields data in which not only the identity but also the number of metabolites detected is unknown initially. Statistical modelling will typically be applied to raw spectroscopic data (Figure 34.1) and there are huge challenges in dealing with noise, peak shifts, overlapping signals, and particularly the lack of annotation. In the last decade, there have been major changes in the assay technologies and study sizes, leading to concomitant developments in statistical modelling approaches. In this chapter, we do not aim at a comprehensive review

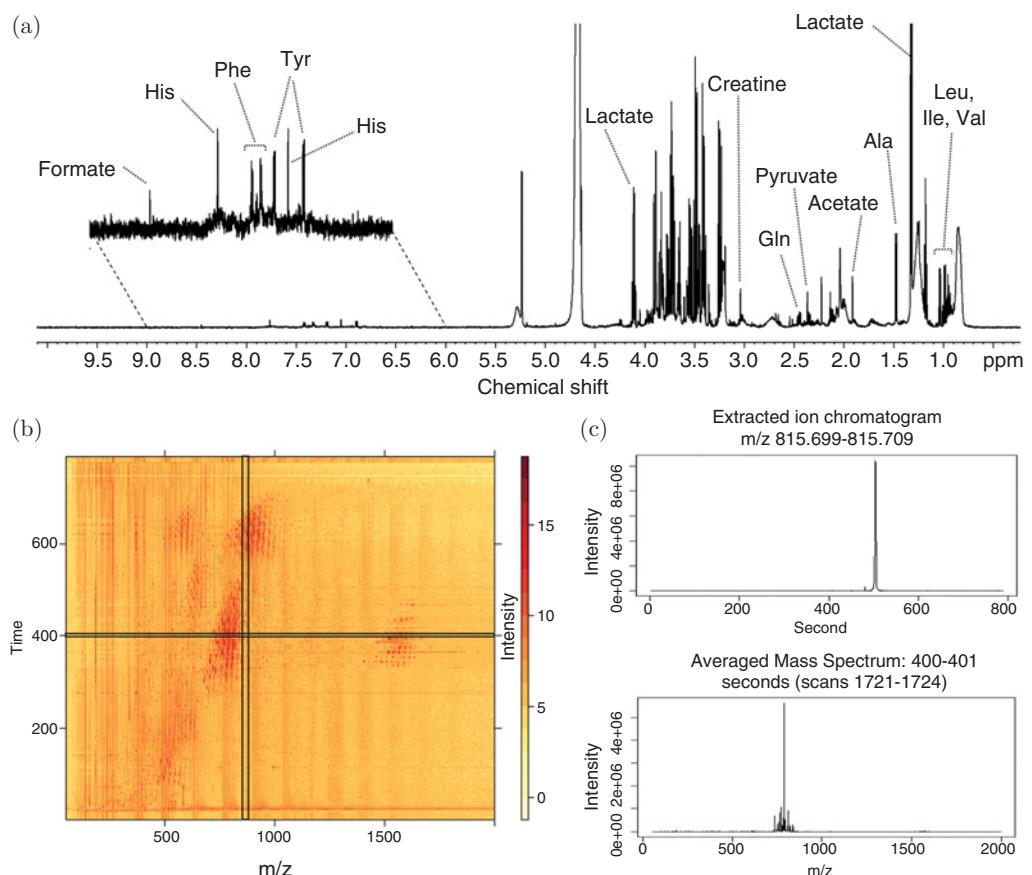


Figure 34.1 Example metabolomics data. (a) One-dimensional ^1H NMR spectrum of human serum (inset shows zoom of region between 6.0 and 9.0 ppm). (b) Heat map LC-MS trace of a single human serum sample with dimensions m/z and retention time. (c) One-dimensional slices through (b) in the retention time and m/z dimensions.

of the field. Rather we review the main developments which have taken place in recent years, focusing primarily on untargeted methods and discuss some open problems which remain to be solved.

34.2 Preprocessing and Deconvolution

The most common analytical technologies for assaying the levels of metabolites are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). To avoid the potentially damaging effects of ion suppression, MS is usually coupled with a preceding step which physically separates different molecular species before they enter the mass spectrometer. This can be achieved via either liquid or gas chromatography (LC or GC) or with capillary electrophoresis (CE), leading to the hyphenated methods LC-MS, GC-MS and CE-MS. Here we only discuss NMR and LC-MS since they are by far the most widely used methods in metabolomics.

34.2.1 Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy uses the interaction between nuclear magnetic moments and external magnetic fields to reveal information about the abundance and structure of molecules in a sample.

Each metabolite gives rise to a pattern of peaks, or resonances, which are characteristic of its molecular structure (see Figure 34.1(a)). The area under the intensity curve of each peak is proportional to the concentration of the metabolite, and is therefore the objective of data processing and statistical modelling. Raw NMR data, the free induction decay (FID), is collected in the time domain and must be Fourier transformed to obtain the frequency domain spectrum. In addition, a number of other standard preprocessing operations must be undertaken (Hoch and Stern, 1996), including phasing (compensating for phase errors in data acquisition), baseline correction and calibration of the frequency axis, known as the chemical shift. For the latter, the zero point is defined arbitrarily using the peak of an internal standard. In addition, the FID is usually multiplied by a function, typically a decreasing exponential, which introduces a certain amount of smoothing of noise in the processed spectrum.

In metabolomics, a number of NMR spectra are acquired on a set of biological samples, and these must be processed into a two-way data table where each row corresponds to one NMR spectrum and each column a variable reporting on metabolite levels. Each sample will be a complex mixture of perhaps thousands of different molecules, and this complexity introduces significant challenges to data analysis. Two major effects cause problems (Ebbels and Cavill, 2009). First, especially in ^1H NMR (the most common type), peaks from several different metabolites often overlap with each other, making it difficult to determine to which molecule the intensity at a given chemical shift relates. Second, peak positions may change slightly from sample to sample. This is due to differences in the physical conditions of each sample (e.g. pH or metal ion concentrations) which cannot be sufficiently controlled. However, the combination of peak shift and overlap means that it is often very difficult to know whether the NMR intensity at a given chemical shift reports on the same molecule in each sample.

Various approaches can be taken to address the above problems, including peak alignment, binning and peak fitting, applied in different combinations. Binning is the simplest approach, where the area under the NMR intensity curve is calculated in small chemical shift regions or bins. This integrated area is thus invariant to movement of the peak, as long as it remains within the bin. The bins could form a regular grid across the chemical shift axis, though this risks peaks moving across bin boundaries, and this approach has mostly been replaced by ‘intelligent’ – and typically manual – placement of the bin boundaries, taking account of the changes in chemical shift observed for each peak across a set of spectra.

In alignment algorithms, peaks are moved so that their positions match across the sample set (Vu and Laukens, 2013). The key difficulty here is in knowing which peaks should be matched. Since the analysis is usually untargeted, meaning that the molecular identity of each peak is not known *a priori*, the only information that can be used to match peaks is position and shape. In addition, alignment cannot separate peaks which are overlapped, or rectify position changes which swap the ordering of peaks on the chemical shift axis.

A natural way to account for changes in peak positions is to find peaks and fit them with a model peak function. In addition, this allows overlapped peaks to be deconvolved into their constituents, as well as tracking peaks which swap positions. Several approaches have been taken to peak fitting, from simply fitting individual peaks (Rubtsov and Griffin, 2007) to fitting the full peak pattern for each metabolite in a spectral library of known compounds. The latter is the approach employed by the popular commercial software NMRSuite by Chenomx (Edmonton, Alberta, Canada) (Weljie *et al.*, 2006), which allows users to search a dedicated library to annotate and deconvolve spectra. Several other methods for spectral library fitting have been published, employing a variety of algorithms to find the combination of metabolites which best fits the observed spectrum. Bayesian methods have come to the fore here, owing to their ability to take account of prior knowledge, such as that which might be used by the expert spectroscopist when doing a manual fit.

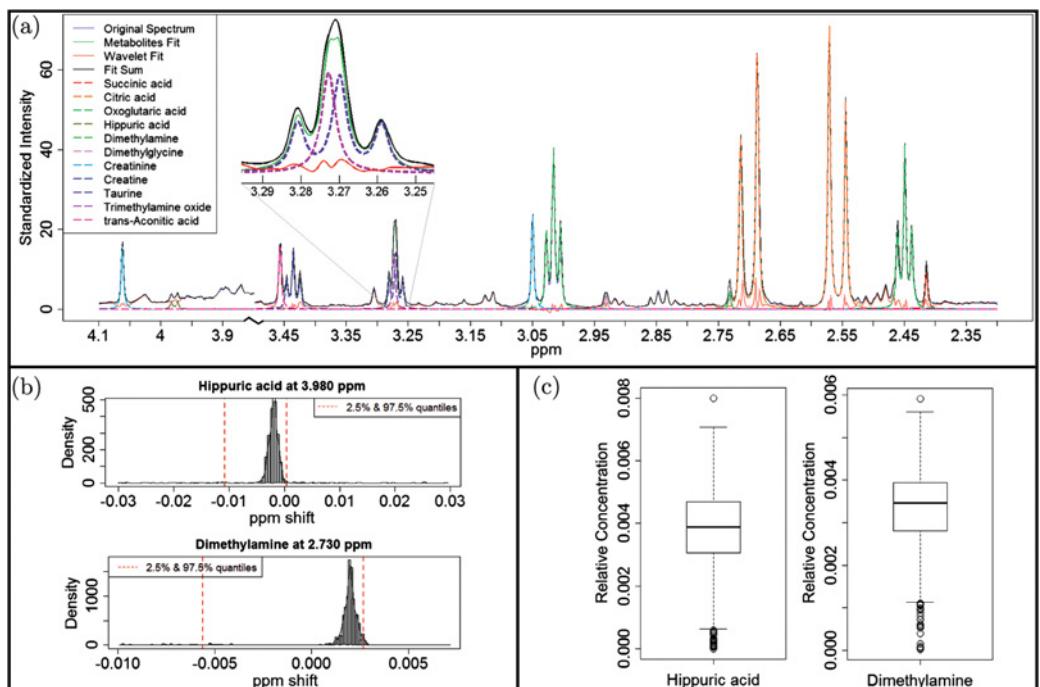


Figure 34.2 Deconvolving NMR spectra with BATMAN. (a) ^1H NMR spectrum of rat urine (black) showing mean posterior estimated fits for several metabolites (colours). (b) Posterior distribution of estimated positions for two peaks. (c) Posterior distribution of relative concentration for two metabolites. Reproduced from Hao *et al.* (2012) by permission of Oxford University Press.

One example of the Bayesian paradigm is the Bayesian automated analyzer for NMR (BATMAN; Hao *et al.*, 2012; Astle *et al.*, 2012; Hao *et al.*, 2014). BATMAN (Figure 34.2) employs a two-component model for the NMR spectra. One component models metabolites known to be present in the mixture, using a library of spectral patterns derived from the Human Metabolome Database. This library specifies the pattern for each metabolite in terms of peak positions and relative intensities. The characteristic patterns produced by many metabolites allow ambiguities in peak identity to be addressed. The second component uses wavelets to model the remaining spectral intensity not taken up by the known compounds. Wavelets prove to be a powerful way to model both broad and narrow features which are not as yet identified, and yield a parsimonious description of the spectral intensity which could then be further investigated, for example using machine learning algorithms. While Bayesian algorithms such as BATMAN have the advantage of quantifying the uncertainty in the fit (e.g. peak positions and intensities), there are also limitations, such as the computational intensity of the approach. In addition, a general difficulty which applies to most peak fitting approaches is the challenge of deciding what (small) set of metabolites might be present in the sample, from a potential set of many thousands available in databases.

34.2.2 Liquid Chromatography – Mass Spectrometry

In LC-MS, the sample is first separated on a chromatographic column, before being analysed by the mass spectrometer. This yields a three-dimensional data set with coordinates retention

time (RT), mass to charge ratio (m/z) and ion count (intensity) (Figure 34.1(b), (c)). As with NMR, several peaks can be observed for the same metabolite owing to isotopologues, adduct formation, fragmentation and dimerization, though the proportion of each component is much more variable than in NMR. Mass spectrometry is a very high-resolution and highly sensitivity technique, and this yields very large data sets (10 MB–10 GB per sample), which then require data reduction. This usually proceeds via a peak detection step, followed by RT alignment, peak matching and peak filling, described in more detail below.

Owing to the much higher resolution in the m/z dimension than the RT dimension, LC-MS peaks are much wider in RT than m/z and also usually exhibit an asymmetric, heavy-tailed shape. For this reason, the required two-dimensional peak detection and noise rejection are non-trivial. An additional difficulty is the presence of small drifts in m/z across a peak. Numerous approaches are employed, from simple thresholding to the use of wavelets (Tautenhahn *et al.*, 2008) and Kalman filters (Åberg *et al.*, 2008). Nonetheless, the peak detection step reduces the data size by a very large factor, typically resulting in 1000–10,000 peaks (tens of kilobytes).

Despite great advances made in the last decade, chromatographic retention times are still less reproducible than mass to charge ratios, so that peaks drift slightly in their RT positions across the sample set. This drift must be corrected, so that peaks from the same metabolite can be matched and compared with each other. RT correction software employs a wide variety of algorithms, including dynamic time warping and loess regression (Smith *et al.*, 2006; Pluskal *et al.*, 2010). Most approaches attempt to construct a warping function which transforms the RT values of one sample to those of another (or a reference sample). In contrast to NMR, the peak shifts are usually assumed to vary smoothly across RT, with nearby peaks of the same sample drifting by similar amounts in the same direction.

The next step is to match peaks from the same metabolite with each other across samples. The success of matching is highly dependent on having a good alignment, since poorly aligned peaks will be indistinguishable from peaks of different molecules. Again, many different approaches are used, including clustering and kernel density estimation (Smith *et al.*, 2006; Pluskal *et al.*, 2010). In some cases, the RT correction and matching steps must be iterated to improve the quality of the resulting matrix. Once peaks are matched, their intensities are estimated (usually by integrating the area under the curve or fitting a model peak). While some peaks will be detected in all samples, there will be many peaks whose variation in intensity takes them below the detection threshold in some samples. Rather than giving these zero intensity in the final matrix, a ‘peak filling’ step is usually conducted, in which the algorithm goes back to samples where the peak was not detected and attempts to estimate its intensity. This is done by integrating the intensity within the m/z –RT boundaries calculated from the samples where the peak was detected. The result of the preceding operations is a matrix of estimated intensities (typically in the range $\sim 10^3$ – 10^8) of size $n \times m$, where n is the number of samples and m is the number of matched peaks or features.

Although this matrix could be taken forward to statistical modelling, further quality control (QC) and ‘postprocessing’ steps are usually required. In particular, LC-MS data are highly variable, both across samples within one analytical run, and across batches and laboratories. Therefore, to assess reproducibility and data quality, a standard QC sample (usually consisting of a pool of the study samples) is run regularly throughout the analysis (Gika *et al.*, 2008). Since the QC samples should be of identical composition, any variation in peak positions or intensities can be attributed to the analytical procedure, including both sample preparation errors and instrumental drift. This allows errors to be both monitored and potentially corrected. Based on the QC samples, two approaches can be taken: filtering the features and/or correction of the

intensities. A very common approach is to filter out features considered to be irreproducible, for example by removing features with high coefficients of variation of intensity measured in the QC samples (e.g. greater than 30%). Another common approach is to assess linearity of each feature by including a dilution series in the analytical run design. In this approach, a typical sample (e.g. a QC) is diluted by known factors ($\times 2$, $\times 4$, $\times 8$, etc.) and these diluted samples run one or more times during the experiment. Then, for each feature, the correlation between intensity and dilution is assessed, and only features exhibiting a strong correlation (i.e. a linear response, e.g. Pearson $r > 0.8$) are kept. A further use of the QC samples is to implement a 'drift correction.' During the analytical run, the MS detector may change its response, often resulting in a decrease in intensities towards the end of the run. These changes can be monitored, by modelling the intensity–run order relationship in the QCs, one feature at a time, using a smooth nonlinear regression technique such as loess or splines. This modelled variation can then be used to correct the intensities for the feature in the biological (non-QC) samples. While the above QC sample allows monitoring and correction of study-specific variation, other standard samples (e.g. a standard reference blood sample) may be used to address between-study and between-lab variation.

A final preprocessing step which applies to both NMR and LC-MS data is that of normalization. The aim of normalization is to correct unwanted global changes in intensity which affect all features from the same sample in the same way. An example would be differences in the dilution of urine samples. These overall dilution changes are not usually biologically informative (because, for example, they may derive from different hydration states of the study participants). These can be removed by a normalization procedure where the intensities of all features for one sample are scaled by a constant factor, $x_{ik} \rightarrow x_{ik}/\alpha_i$. This normalization factor must be estimated separately for each sample. A very popular method is total intensity or constant sum normalization, where the normalization factor is simply the sum of the intensities of all features. This will result in normalized data where the total signal for all samples is the same, and has some theoretical motivation in terms of equalizing the total 'amount' of material being assayed. A disadvantage of this approach, however, is that it can result in spurious correlations between features when there are a small number of high-intensity features dominating the profile (Craig *et al.*, 2006). The probabilistic quotient normalization (PQN; Dieterle *et al.*, 2006) was developed to address this issue and is able to avoid problems with high-intensity features. In this approach, the intensity x_{ik} of the k th feature, in the sample i to be normalized, is compared to the same feature in a reference sample j , often taken as the median sample profile of the data set. The normalization factor is calculated as the median of the ratios (quotients) of these intensities:

$$\alpha_i = \underset{k}{\text{median}}(x_{ik}/x_{jk}) \quad (34.1)$$

The use of the median in the PQN method makes the assumption that at least 50% of the features in the profile are only affected by the dilution (or changes which should be normalized out). Note that this approach is similar to some normalization techniques used in other areas of omics (e.g. transcriptomics). Normalization is usually more straightforward in targeted workflows when metabolites are quantified in absolute amounts (e.g. moles per decilitre).

34.3 Univariate Methods

The aims of statistical analysis in metabolic profiling will depend on the objectives of the study. We are often interested in detecting structure and identifying important features of the data.

The main goals here are usually to cluster individuals based on their metabolic profiles and understand differences among groups of subjects. Unsupervised statistical methods are usually employed to this end. Another goal is to determine whether there is a significant difference between groups in relation to a phenotype of interest and identify which metabolites discriminate phenotype groups. In this case regression and classification methods are employed and we often work in the *small n, large p inference* problem (West, 2003), that is, each individual datum (metabolic profile) consists of a large vector of interrelated (dependent) observations, yet the number of samples in the study is relatively small. This poses substantial challenges to statistical analyses. Here, we briefly review some of the current techniques used in the statistical analysis of metabolomic data. We will denote by X the $n \times p$ matrix of metabolic profiles, where each row is a p -dimensional vector of spectral variables representing one metabolic profile from one biological sample. Usually $p \approx 10^2\text{--}10^4$. Each column of X corresponds to a single metabolic variable which plays the role of predictor variable, and we denote the n -dimensional vector of phenotypic responses by y .

The high dimensionality of metabolic spectra makes the application of advanced inferential tools computationally challenging. As mentioned above, the metabolomic workflow can follow two different strategies (Vinaixa *et al.*, 2012): targeted metabolomics, driven by a specific biochemical question or hypothesis in which a set of metabolites related to one or more pathways are defined; or untargeted metabolomics, a hypothesis-free approach in which as many metabolites as possible are measured and compared between samples. Here, we focus on the second approach, as in the first case conventional multivariate statistical methods can be employed. In particular, we consider problems where thousands of metabolic features (e.g. peaks, spectral variables), are compared across samples. As metabolic data sets contain too many correlated variables, fitting a multivariate model is challenging. Given a sample size, performance might deteriorate as the number of variables increases. Indeed, including too many variables can introduce substantial errors and reduce the predictive power of the model (e.g. due to overfitting). Therefore, using multivariate models requires intensive validation work. Excellent reviews on multivariate tools for metabolomics can be found in Liland (2011) and Hendriks *et al.* (2011).

Often in omics sciences, data analysis is approached from a univariate perspective using traditional statistical methods that consider only one variable at a time (Balding, 2006; Kalogeropoulou, 2011). There are multiple metabolomic publications that rely on univariate tests applied in parallel across all the detected metabolic features to report their main findings. This approach presents serious drawbacks. In particular, it ignores the correlations between metabolites, and/or spectral variables that report on the same metabolite. In addition, performing many univariate tests leads to an increased risks of false positive results. Often the test applied makes implicit assumptions on the distribution of the variables, which are not always satisfied in real applications and are often overlooked by researchers.

Case-control phenotype. Perhaps the most natural analysis of metabolic concentrations and case-control status is to test the null hypothesis of no difference in mean levels of metabolites between cases and controls. Users have a choice between, among others, a t test and non-parametric two-sample tests such as the Wilcoxon test. The latter does not make any specific assumptions on the distribution of the data in each sample (although they do rely on some assumptions!). These tests are implemented in common statistical software such as R. An alternative approach is offered by logistic regression: a transformation of the disease risk π_i for each i individual is modelled as a linear function of the metabolic features. Common transformations used in applications are $\text{logit}(\pi) = \log(\pi/(1 - \pi))$ and $\text{probit}(\pi) = \Phi^{-1}(\pi)$, where $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function of the standard normal. Then the transformation of disease risk is equated to βx , where β denotes the vector of regression

coefficients (effect size) and \mathbf{x} is the vector of covariates which usually includes metabolites and confounding factors. The likelihood ratio test is used to test the null hypothesis of no association.

Continuous outcomes: linear regression. The natural statistical tool for continuous (or quantitative) traits is linear regression, which assumes a linear relationship between the mean value of the response and the metabolic feature. For each metabolic feature under investigation, independent linear regression models can be fitted, possibly correcting for confounding factors. This procedure requires the response to be approximately normally distributed, with a common variance. If normality does not hold, a transformation (e.g. log or square root) of the original trait values might lead to approximate normality.

We note that the use of multivariate and univariate data analysis is not mutually exclusive, and in fact researchers often combine them to maximize the extraction of relevant information from metabolomic data sets (e.g. Goodacre *et al.*, 2007). Univariate methods are sometimes used in combination with multivariate models as a filter to retain the most 'information-rich' features, for example retaining only variables with a p -value smaller than a prespecified threshold. This strategy leads to a significant reduction in the number of variables included in a multivariate model, thereby mitigating the problems of high dimensionality and collinearity.

34.3.1 Metabolome-Wide Significance Levels

When applying univariate methods to each variable one at a time, the question of controlling the false positive rate becomes key. This is usually addressed by adjusting the significance level used to declare findings, for example using a Bonferroni correction, or a false discovery rate (FDR) approach. However, these approaches do not take the highly dependent nature of metabolic variables into account. An alternative is the metabolome-wide significance level (MWSL) defined by Chadeau-Hyam *et al.* (2010) as the threshold required to control the family-wise error rate. The evaluation of the MWSL is done through permutations, based on real-world data. ^1H NMR spectroscopic profiles of 24-hour urinary collections from the INTERMAP study were used as this study offers a unique resource providing a large-scale standardized set of urinary metabolic profiles capturing variation within and between human populations in China, Japan, the UK and USA. Case-Control outcomes were simulated from a logistic model and two-sample t tests were used for detection of associations between metabolic variables and the response. The approach adopted is similar to the evaluation of the genome-wide significance level (Hoggart *et al.*, 2008).

The MWSL accounts appropriately for the high degree of correlation in spectral data, and provides a practical threshold that can be used as a benchmark for future metabolome-wide association studies of human urine. The MWSL primarily depends on sample size and spectral resolution. A conservative estimate of the independent number of tests is approximately 35% for urine spectra, regardless of spectral resolution and sample size. This leads, for example, to an estimated MWSL of 2×10^{-5} and 4×10^{-6} for a familywise error rate of 0.05 and 0.01 respectively, at medium spectral resolution (7100 variables). The same simulation-based method for the determination of the MWSL may be applied to other metabolic profiling technologies such as LC/GC-MS.

It is well known that spectral variables from metabolic profiles exhibit a high degree of collinearity, and this is supported by the finding that the computed MWSL greatly exceeds the Bonferroni or Šidák corrected value across all three data sets. The extent of collinearity is summarized by the ratio of effective to actual number of tests, which varies between 15% and 30% across diverse spectral resolutions, sample sizes and populations. One way to interpret the

number of independent tests is an indication of the number of independent metabolic processes exhibited by the system, since each independent process might be expected to manifest itself through multiple metabolic variables. The MWSL estimates can be used to guide selection of design of metabolomic experiments in metabolome-wide association studies and the choice of the method of analysis.

34.3.2 Sample Size and Power

The number of subjects per group (i.e. sample size) is an important aspect to be determined during experimental design of a study. A low sample size may lead to a lack of power to detect differences between groups or association with responses of interest, which may fail to provide insight into the biological system under investigation. In contrast, an unnecessarily high sample size may lead to a waste of resources for minimal information gain. However, choosing the appropriate sample size for high-throughput approaches involving multivariate data is complicated. In particular, in metabolic phenotyping, there is currently no accepted approach for power and sample size determination, in large part due to the unknown nature of the expected effect. In such hypothesis-generating science, neither the number of important metabolites nor the effect size is known *a priori*. Moreover, when determining sample size, multiple testing is an issue, as usually hundreds to thousands of metabolic variables are tested for association with an outcome of interest. For these reasons, power analysis is often avoided and sample sizes are determined using *ad hoc* criteria such as cost or expert knowledge. An early approach to tackle the problem is offered by the *data driven sample size determination* (DSD; Blaise, 2013) algorithm which was developed to be used with small pilot study data and for a specific set of univariate analyses, but does not account for correlation in the data. More recently, a sample size determination module has been implemented in MetaboAnalyst 3.0 (Xia *et al.*, 2015) based on the Bioconductor package Sample Size and Power Analysis developed for genomics (Van Iterson *et al.*, 2009). This approach does not take into account correlation between variables and relies on the concept of summary average power for the data set. Unfortunately, there is no reason to expect that each variable exhibits the same power, and since for most studies no prior knowledge about important variables exists, it is preferable to set the sample size to a number where the majority of variables are expected to reach a minimum level of power.

Blaise *et al.* (2016) introduce a simulation-based approach, able to accommodate the high dimensionality of modern metabolic data and the high degree of correlation between metabolic signals. They investigate the relationship between statistical power, sample and effect size and obtain estimates of the required sample size for metabolomics studies. Figure 34.3 illustrates the workflow. On the basis of pilot data (available from previous studies or public databases), new samples are simulated with marginal distributions and correlation structure similar to the ones observed in the pilot data. These can be used to study the sensitivity of power and other metrics to sample size. Depending on the structure of the problem, the simulated data can represent metabolite concentration or spectral data. The data are generated using a multivariate log-normal distribution, whose parameters are specified based on inference performed on the pilot data. This allows us to maintain the long tails and strong correlations that are typically seen in metabolic data sets. Then the desired effect size is introduced in the simulated data set, depending on the type of outcome variables under investigation and the statistical method intended for data analysis. Multiple data sets are generated following this procedure and the outcomes of the statistical analyses are stored and used to derive empirical summary statistics, estimates and confidence intervals of performance statistics (e.g. true positive, false negative

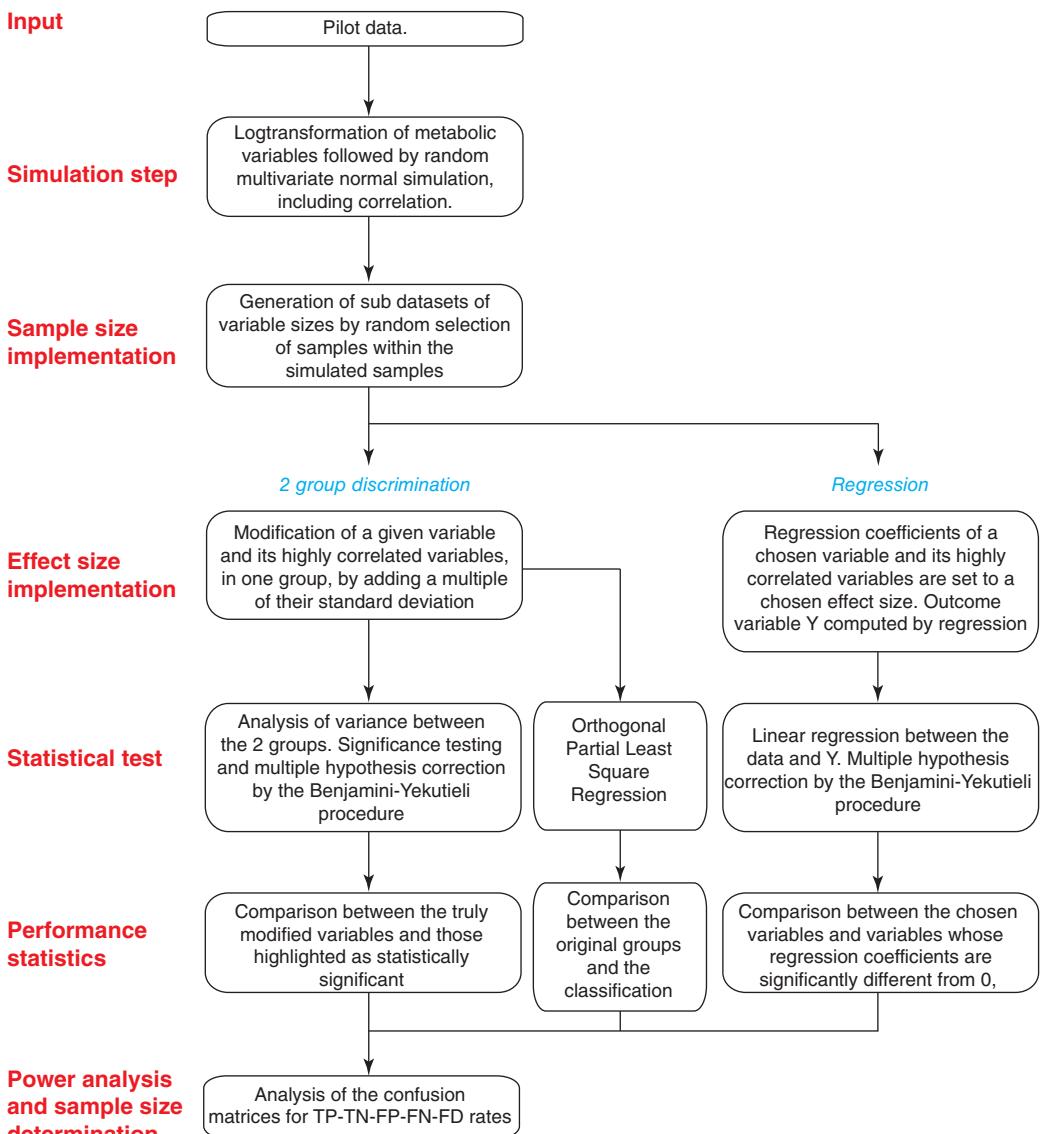


Figure 34.3 A workflow for power and sample size estimation in metabolomics using pilot data and log-normal simulations. Reprinted with permission from Blaise *et al.* (2016). © 2016 American Chemical Society.

rates), from which power and other quantities of interest may be calculated. Different multiple hypothesis-testing corrections can be used and their effect on power and efficiency (e.g. in controlling FDR) benchmarked.

34.4 Multivariate Methods and Chemometrics Techniques

Multivariate statistical methods take collections of observations on the experimental units (individuals, biological samples, etc.) and analyse them simultaneously. Two types of variables,

predictors (or *covariates*) and *outcome* variables (or *responses*), are typically considered. Predictors are collections of variables having some joint probability distribution, but where no individual variable is considered a functional consequence of the others. An outcome is regarded as a variable that results from variation in the other variables. In modern statistical applications, the number of variables considered is often very large (of the order of thousands), with the possibility of high-dimensional predictor sets and even possibly high-dimensional outcomes; as described in this chapter, metabolomic spectra can correspond to measurements on thousands of metabolites simultaneously, and this large vector could potentially be regarded as a predictor or as an outcome, depending on the context. The principal challenge is therefore one of *dimension reduction* that aims to get at the essential information in the sample while stripping out or smoothing over the non-essential components.

In the following description, we assume that n independent and identically distributed samples have been collected on p predictor variables x_1, x_2, \dots, x_p and a single (scalar) outcome variable, y . We begin with the most elementary method for studying such data, regression.

34.4.1 Linear Regression Methods

The most widely used method for handling multivariate data is the multiple linear regression model where the relationship between outcome y and predictors x is captured by the linear relationship written in vector form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (34.2)$$

where $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]^T$ is an $n \times 1$ column vector (the responses), $\mathbf{x}_i = [1, x_{i1}, x_{i2}, \dots, x_{ip}]$ a $1 \times (p+1)$ row vector, $\mathbf{X} = [1, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p]$ an $n \times (p+1)$ matrix (the design matrix), $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_p]^T$ a $(p+1) \times 1$ column vector (the parameter vector), and $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_n]^T$ an $n \times 1$ column vector (the random errors). This form of the regression model illustrates the fact that the model is *linear* in $\boldsymbol{\beta}$ (that is, the elements of $\boldsymbol{\beta}$ appear in their untransformed form). This is important as it allows particularly straightforward calculation of statistical quantities such as parameter estimates and standard errors.

The implication in this model is that there is a *structural* link between the predictors \mathbf{X} and the outcomes \mathbf{Y} , in that changing \mathbf{X} will result in a commensurate (expected) change in \mathbf{Y} . Most typically in such a model, we do not regard the measured \mathbf{X} as a random quantity, or consider its probability distribution, we simply consider the distribution of \mathbf{Y} for the given, observed \mathbf{X} values.

The model in (34.2) can be rewritten in component form as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i,$$

where coefficient β_j measures the expected change in response if continuous predictor x_j changes by one unit, or the expected change in response if a categorical (or factor) predictor changes from its baseline level, *while all other variables are held constant*. In general, there is a dependence structure among the variables: that is, the predictors \mathbf{X} are correlated, and therefore the impact of any one predictor x_j overall is not simply measured by the coefficient β_j – this is a *conditional* measure.

Using ordinary least squares (OLS), the parameter estimates from the sample of size n are obtained by minimizing the sum of squared errors quantity,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (34.3)$$

The estimates are given analytically as the solution to the linear system of equations

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}, \quad (34.4)$$

that is,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (34.5)$$

with resulting prediction vector

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

A necessary condition for the OLS estimation procedure to be viable is that the $p \times p$ matrix $\mathbf{X}^T \mathbf{X}$ is invertible. If $n \geq p$, this condition is easily satisfied, and essentially is a requirement that the columns of \mathbf{X} be linearly independent. However, if $n < p$ then the matrix is never invertible: in modern statistical settings, where huge numbers of predictor variables are often measured on relative few samples, this situation is common. We describe how OLS methods can be adapted to handle this setting in Section 34.4.2.

The only situation in which the coefficients estimated in this way can be interpreted as a marginal effect parameters is if the predictors are *orthogonal*, that is, if in the sample the column vectors $\mathbf{x}_1, \dots, \mathbf{x}_p$ are constructed such that $\mathbf{x}_j^T \mathbf{x}_k = 0$ for all j, k , in which case

$$\mathbf{X}^T \mathbf{X} = \mathbf{I}_p.$$

In most applied problems, the predictors are not naturally orthogonal. However, we will see in Section 34.5 methods that can construct orthogonal predictors using simple linear transformations.

34.4.2 Shrinkage Methods

As described above, if $p > n$ OLS methods cannot be used as the linear system in equation (34.4) has no unique solution. In effect, the system has too many parameters. One solution therefore is to amend the OLS criterion to attempt to limit the number of parameters that are included. The simplest way to do this is to remove terms from the model (i.e. columns from \mathbf{X}). However, in general it will not be clear which variables to remove, so instead we achieve the same result in an automatic fashion using *penalization*, that is, we amend the sum of squares function in equation (34.3) to be

$$S_\lambda(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + \lambda \rho(\boldsymbol{\beta}), \quad (34.6)$$

where $\rho(\boldsymbol{\beta})$ is a non-negative penalty function. By careful choice of this penalty, and the scaling (or tuning) parameter λ , the effective dimension of the solution space is reduced from p to something smaller. The main idea is to make $\rho(\boldsymbol{\beta})$ increase as $\|\boldsymbol{\beta}\|$ increases, and to have the penalty take its smallest value when $\boldsymbol{\beta} = \mathbf{0}$, so that the OLS estimates undergo *shrinkage* (i.e. the estimates of $\boldsymbol{\beta}$ are smaller in magnitude than the OLS estimates). Common choices of the penalty are as follows:

1. *Quadratic penalty*. The quadratic (or L_2) penalty takes the form

$$\rho(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2 = \boldsymbol{\beta}^T \boldsymbol{\beta}.$$

In this case, an analytic solution to the minimization problem involving (34.6) is

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{y}. \quad (34.7)$$

The penalty is termed the *ridge* penalty, and the method of estimation using this penalty is termed *ridge regression*.

2. *Absolute value penalty.* The absolute value (or L_1) penalty takes the form

$$\rho(\beta) = \sum_{j=1}^p |\beta_j|.$$

In this case, no analytic solution is available to the minimization problem involving equation (34.6), but efficient numerical procedures exist to obtain the estimates. The penalty is termed the *least absolute shrinkage and selection operator* (lasso) penalty. This penalty has the advantage that estimates can be precisely equal to zero, indicating that *variable selection* can be achieved – if the estimate is zero, then the corresponding variable is essentially omitted from the inferred model.

3. *$L_1 + L_2$ penalty.* This combined penalty takes the form

$$\rho(\beta) = \alpha \sum_{j=1}^p \beta_j^2 + (1 - \alpha) \sum_{j=1}^p |\beta_j|,$$

for some fixed parameter $0 < \alpha < 1$, and is termed the *elastic net* penalty. It combines features of the two component penalties.

4. *Group penalty.* The L_1 group penalty takes the form

$$\rho(\beta) = \|\beta\| = \sqrt{\sum_{j=1}^p \beta_j^2}.$$

Again, no analytic solution is available to the minimization problem involving equation (34.6) with this penalty, but efficient numerical procedures again exist to obtain the estimates.

5. *Oracle penalty.* An oracle penalty is a shrinkage penalty that (as the sample size increases) produces estimates that behave as if they were OLS estimates from a correctly specified model that includes precisely those variables that are in the data-generating model. There are several forms of oracle penalty, but in general the penalty must allow for *selection* (i.e. allow estimates to be precisely zero) and produce estimators that are (asymptotically) *unbiased*. In practice, the tuning parameter λ is allowed to vary with the sample size – in fact, it is allowed to diminish at a specified rate in n – so therefore the influence of the penalty term

$$\lambda_n \rho(\beta)$$

need not diminish to leave an unbiased estimator. Oracle penalties typically have two properties: they are non-differentiable at zero to allow for selection; and they become constant for β large enough (to allow for absence of bias).

34.5 Orthogonal Projection Methods

Principal components analysis (PCA) can be used to reveal the underlying variance structure of a data set and is a device for data visualization for exploratory data analysis. In its role of reducing rank of a data matrix, PCA can be used to estimate how many ‘true variates’ or sources of variability there are in a multivariate data set. *Partial least squares* (PLS) gives a similar data reduction decomposition, but also incorporates (the required) dependence between predictor and response present in a regression context.

34.5.1 Principal Components Analysis

PCA is a technique used in high-dimensional data analysis, and is used to *reduce the dimensionality* of a data set or data matrix. Broadly, it describes the data set in terms of its components of variance. Each *principal component* describes a *percentage of the total variance* of a data set, and computes *loadings* or *weights* that each variate contributes to this variance. For example, the first principal component of a data set describes the dimension which accounts for the greatest amount of variance of the data set. The *coefficients* of the principal components quantify the loading or weight of each variate to that amount of variance.

The mathematical assumptions behind PCA appeal to approximate multivariate normality of the underlying observations, although this is not a strict requirement. It is not based on a regression model formulation, as it only analyses the *predictor* variables, or, at least, treats all variables equivalently. As a technique, however, it does often contribute in regression or classification analysis because of its data-reduction properties.

We begin with a justification of the PCA method inspired by the multiple regression model. In PCA the data matrix is typically arranged with observations in rows, and different predictors in columns. In a classification context, we might wish to see how much information the predictor variables contained. Suppose that the $n \times p$ data matrix \mathbf{X} is so arranged, but also that \mathbf{X} is *, so that the mean within a column is zero – this is achieved by taking the raw predictor data matrix, and subtracting from each element in a column that column's sample mean. In linear regression, the matrix \mathbf{X} is referred to as the design matrix, and used to estimate parameters in the regression model using the formula for observed response vector \mathbf{y} .*

Note that if \mathbf{X} is the centred matrix as defined, we have that $\mathbf{S} = \mathbf{X}^T \mathbf{X} / n$ is the *sample covariance matrix*. Now using standard matrix decomposition techniques (specifically, the *singular value decomposition* as discussed below), we may (uniquely) write \mathbf{X} as

$$\mathbf{X} = \mathbf{UDV}^T, \quad (34.8)$$

where \mathbf{U} is $n \times p$ and \mathbf{V} is $p \times p$ such that

$$\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{V} \mathbf{V}^T = \mathbf{I}_p$$

for p -dimensional identity matrix \mathbf{I}_p (i.e. \mathbf{U} and \mathbf{V} are *orthogonal* matrices), and \mathbf{D} is a $p \times p$ matrix with diagonal elements $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$ and zero elements elsewhere. The representation in equation (34.8) is termed the *singular value decomposition* (SVD) of \mathbf{X} . Note that, using this form, we have

$$\mathbf{X}^T \mathbf{X} = \mathbf{V} \mathbf{D}^T \mathbf{U}^T \mathbf{U} \mathbf{D} \mathbf{V}^T = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T = \mathbf{V} \mathbf{L} \mathbf{V}^T. \quad (34.9)$$

This is called the *eigendecomposition* of $\mathbf{X}^T \mathbf{X}$. The diagonal elements of $\mathbf{L} = \mathbf{D}^2$ are

$$l_1 \geq l_2 \dots \geq l_p \geq 0;$$

these are termed the *eigenvalues* of $\mathbf{X}^T \mathbf{X}$. The columns of the matrix \mathbf{V} are termed the *eigenvectors* of $\mathbf{X}^T \mathbf{X}$, and the j th column, \mathbf{v}_j , is the eigenvector associated with eigenvalue l_j . The principal components of $\mathbf{X}^T \mathbf{X}$ are defined via the columns of \mathbf{V} , $\mathbf{v}_1, \dots, \mathbf{v}_p$. The j th principal component is \mathbf{w}_j , defined by

$$\mathbf{w}_j = \mathbf{X} \mathbf{v}_j = l_j \mathbf{u}_j$$

for normalized vector \mathbf{u}_j . We have that $\mathbf{V}^T \mathbf{V} = \mathbf{I}_p$, that is, the columns of \mathbf{V} are orthogonal. Hence the principal components $\mathbf{w}_1, \dots, \mathbf{w}_p$ are also orthogonal. The vectors $\mathbf{v}_1, \dots, \mathbf{v}_p$ are termed the *factor loadings*.

In the linear model above, we now may write

$$\mathbf{Y} = \mathbf{UDV}^T \boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{W}\boldsymbol{\beta}' + \boldsymbol{\epsilon},$$

say, where $\mathbf{W} = \mathbf{UD}$, so we retain a linear model with new orthogonal design matrix \mathbf{W} . Note from the decomposition we have that $\mathbf{W} = \mathbf{UD} = \mathbf{X}\mathbf{V}$, say, and as $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, the columns of \mathbf{V} are of unit length. This relationship confirms that we are seeking the transform of \mathbf{X} to \mathbf{W} where the columns of \mathbf{W} are the principal components. Furthermore, the first principal component \mathbf{w}_1 has largest sample variance among all normalized linear combinations of the columns of \mathbf{X} ; in general, we have that

$$\text{Var}[\mathbf{w}_j] = \frac{l_j}{n}, \quad j = 1, \dots, p.$$

The justification via the linear model reflects a common reason why PCA is used; however, as the SVD relates only to $\mathbf{X}^T\mathbf{X}$ and not to \mathbf{y} , it is evident that in fact PCA is a tool that addresses variation in the predictor set without any relation to the response. It is therefore a general approach to dimension reduction that relies on a linear transformation of the predictor set. We can equivalently regard the selection of the first principal component as the solution to the minimization problem

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \|\mathbf{w}_1\|^2 = \arg \max_{\|\mathbf{v}\|=1} \|\mathbf{X}\mathbf{v}\|^2 = \arg \max_{\|\mathbf{v}\|=1} \{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}\}, \quad (34.10)$$

which is equivalent to

$$\mathbf{v}_1 = \arg \max_{\|\mathbf{v}\|=1} \left\{ \frac{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right\},$$

to obtain the first loading, and then taking $\mathbf{w}_1 = \mathbf{X}\mathbf{v}_1$. For the subsequent component loadings $j = 2, 3, \dots, p$, we extract the influence of components up to $j - 1$ by writing

$$\mathbf{X}_j = \mathbf{X} - \sum_{k=1}^{j-1} \mathbf{X}\mathbf{v}_k \mathbf{v}_k^T$$

and then solving the equivalent minimization problem,

$$\mathbf{v}_j = \arg \max_{\|\mathbf{v}\|=1} \left\{ \frac{\mathbf{v}^T \mathbf{X}_j^T \mathbf{X}_j \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right\},$$

and finally taking $\mathbf{w}_j = \mathbf{X}\mathbf{v}_j$.

The *total variance explained* by the data is a straightforward function of the centred design matrix; it is the sum of the diagonal elements (or *trace*) of the matrix \mathbf{S} , given by

$$\text{trace}(\mathbf{S}) = \sum_{j=1}^p [\mathbf{S}]_{jj} = \frac{\text{trace}(\mathbf{X}^T \mathbf{X})}{n} = \frac{\text{trace}(\mathbf{L})}{n} = \frac{1}{n} \sum_{j=1}^p l_j,$$

and hence the j th principal component accounts for a proportion

$$\frac{l_j}{\sum_{k=1}^p l_j} \quad (34.11)$$

of the total variance. Using principal components, therefore, it is possible to find the ‘directions’ of largest variability in terms of a *linear combination* of the columns of the design matrix; a

linear combination of column vectors $\mathbf{x}_1, \dots, \mathbf{x}_p$ is a vector \mathbf{w} of the form $\mathbf{w} = \sum_{j=1}^p \pi_j \mathbf{x}_j$ for coefficients (loadings) $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$, for the first principal component, $\boldsymbol{\pi} = \mathbf{v}_1$.

The main use of principal components decomposition is in *data reduction* or *feature extraction*. It is a method for looking for the main sources of variability in the predictor variables, and the argument follows that the first few principal components contain the majority of the explanatory power in the predictor variables. Thus, instead of using the original predictor variables in the linear (regression) model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

we can use instead the principal components as predictors

$$\mathbf{Y} = \mathbf{W}\boldsymbol{\beta}' + \boldsymbol{\epsilon}.$$

The data reduction, compression or feature extraction arises if, instead of taking all p of the principal components, we take only the first k , that is, we extract the first k columns of matrix \mathbf{W} , and reduce $\boldsymbol{\beta}_Z$ to being a $(k \times 1)$ vector. Choosing k can be done by inspection of the *scree* plot of the successive scaled eigenvalues as in (34.11).

34.5.2 Partial Least Squares

The principal components analysis outlined above is an excellent method for extracting the linear combinations of the input predictors that are the largest sources of variability. But implicit in the PCA definition is the constraint that no aspect of relationship between the predictors \mathbf{X} and the response \mathbf{Y} is recognized. Hence, if the PCA is to be used as a feature extraction method for use in regression, there may well be a deficiency in the principal components as predictors themselves.

PLS is a related feature extraction procedure where the relationship between the predictors \mathbf{X} and the response \mathbf{Y} is modelled explicitly. It does this by accounting for the correlation between response and prediction under the usual linear model formulation. The PLS procedure can also be viewed as a method that depends upon a similar decomposition to PCA, but the decomposition is instead applied to a matrix based on the covariance between predictor and outcome. With centred (and standardized) matrix \mathbf{X} and outcome \mathbf{Y} , we consider

$$\mathbf{S}_{X,Y} = \frac{\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}}{n^2},$$

formed by taking the outer product of the sample covariance matrix with itself.

A more common view of PLS is as a recursive regression procedure, and an algorithm to construct the PLS components is given by Hastie *et al.* (2001):

1. Let $\mathbf{x}_j = (x_{j1}, \dots, x_{jn})^T$ be the j th column of the design matrix \mathbf{X} , appropriately centred (by subtracting the column mean \bar{x}_j) and scaled (by column sample standard deviation s_j) to have sample mean 0 and sample variance 1.
2. Set $\hat{\mathbf{y}}^{(0)} = \mathbf{1}\bar{y}$ and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$
3. For $j = 1, 2, \dots, p$,
 - $\mathbf{w}_j = \sum_{k=1}^p \hat{\phi}_{jk} \mathbf{x}_k^{(j-1)}$, where $\hat{\phi}_{jk} = \langle \mathbf{x}_j, \mathbf{x}_k^{(j-1)} \rangle$, with

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = \sum_{i=1}^n v_{1i} v_{2i}$$

being the usual dot product operator;

- $\hat{\theta}_j$ is defined by

$$\hat{\theta}_j = \frac{\langle \mathbf{w}_j, \mathbf{y} \rangle}{\langle \mathbf{w}_j, \mathbf{w}_j \rangle};$$

- $\hat{\mathbf{y}}^{(j)}$ is defined by

$$\hat{\mathbf{y}}^{(j)} = \hat{\mathbf{y}}^{(j-1)} + \hat{\theta}_j \mathbf{w}_j;$$

- for $k = 1, \dots, j-1$, $\mathbf{x}_k^{(j)}$ is defined by

$$\mathbf{x}_k^{(j)} = \mathbf{x}_k^{(j-1)} - \left[\frac{\langle \mathbf{w}_j, \mathbf{x}_k^{(j-1)} \rangle}{\langle \mathbf{w}_j, \mathbf{w}_j \rangle} \right] \mathbf{w}_j$$

so that, for each $k = 1, \dots, j-1$ $\mathbf{x}_k^{(j)}$ is orthogonal to \mathbf{w}_j .

4. Record the sequence of fitted vectors $\hat{\mathbf{y}}^{(1)}, \hat{\mathbf{y}}^{(2)}, \dots, \hat{\mathbf{y}}^{(p)}$.

5. Evaluate the PLS coefficients as

$$\hat{\beta}_j^{(\text{PLS})} = \sum_{k=1}^m \hat{\phi}_{kj} \hat{\theta}_k.$$

and the m th PLS direction is

$$\mathbf{w}_m = \sum_{k=1}^p \hat{\phi}_{mk} \mathbf{x}_k.$$

In the construction of each PLS direction \mathbf{w}_m the predictors are weighted by the strength of their univariate impact on \mathbf{y} . The algorithm first regresses \mathbf{y} on \mathbf{z}_1 , giving coefficient $\hat{\theta}_1$, then orthogonalizes $\mathbf{x}_1, \dots, \mathbf{x}_p$ to \mathbf{w}_1 , and then proceeds to regress \mathbf{y} first on \mathbf{w}_2 on these orthogonalized vectors, and so on. After $r \leq p$ steps, the vectors $\mathbf{w}_1, \dots, \mathbf{w}_r$ have been produced, and can be used as the inputs in a regression type model, to give the coefficients

$$\hat{\beta}_1^{(\text{PLS})}, \dots, \hat{\beta}_r^{(\text{PLS})}.$$

34.5.3 Orthogonal Projection onto Latent Structures

If the response \mathbf{Y} is multivariate (say, $n \times q$), then alternative orthogonal projections may be used. The method of *orthogonal signal correction* (OSC) adopts a strategy that seeks to achieve the same goal of attempting to represent \mathbf{X} using principal components, with the additional constraint that the components are orthogonal to \mathbf{Y} . Fearn (2000) derived the following approach based on a constrained version of PCA: for the first element, amending equation (34.10), we have that the first constrained orthogonal component, \mathbf{v}_1^* , solves

$$\mathbf{v}_1^* = \arg \max_{\|\mathbf{v}\|=1} \{ \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} \}, \quad \text{such that } \mathbf{v}^T \mathbf{X}^T \mathbf{Y} = \mathbf{0}. \quad (34.12)$$

As for PCA, this problem is solved using an eigendecomposition approach; we have that if

$$\mathbf{M} = \mathbf{I} - \mathbf{X}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{X}$$

then \mathbf{v}_1^* is the first eigenvector of the matrix $\mathbf{M} \mathbf{X}^T \mathbf{X}$, and the remaining eigenvectors of this matrix are the remaining constrained orthogonal components. We then have the representation

$\mathbf{W}^* = \mathbf{X}\mathbf{V}^*$. As for PCA, the constrained eigenvectors corresponding to the largest eigenvalues are selected. Simpler forms of projection matrix \mathbf{M} have also been proposed. Prediction of \mathbf{Y} is then achieved by taking a subset of the columns of \mathbf{V}^* in the usual way.

The approach based on equation (34.12) is an example of the method of *orthogonal projection to latent structures* (Trygg and Wold, 2002, 2003). An alternative orthogonal PLS-like approach uses a separate decomposition of the \mathbf{X} and \mathbf{Y} components; we have $\mathbf{W}_X = \mathbf{X}\mathbf{V}_X$ and $\mathbf{W}_Y = \mathbf{Y}\mathbf{V}_Y$, or equivalently

$$\mathbf{X} = \mathbf{W}_X \mathbf{V}_X^T, \quad \mathbf{Y} = \mathbf{W}_Y \mathbf{V}_Y^T,$$

and then a dimension-reduction step to the representation

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{E}_X, \quad \mathbf{Y} = \tilde{\mathbf{Y}} + \mathbf{E}_Y,$$

where $\tilde{\mathbf{X}} = \widetilde{\mathbf{W}}_X \widetilde{\mathbf{V}}_X^T$ is obtained by taking $\tilde{p} < p$ components from \mathbf{W}_X and \mathbf{V}_X^T , and $\tilde{\mathbf{Y}} = \widetilde{\mathbf{W}}_Y \widetilde{\mathbf{V}}_Y^T$ is obtained by taking $\tilde{q} < q$ components from \mathbf{W}_Y and \mathbf{V}_Y^T , with $\widetilde{\mathbf{W}}_X$ and $\widetilde{\mathbf{W}}_Y$ chosen to exhibit maximum covariance. In this approach, the orthogonally constrained strategy in equation (34.12) may also be used to obtain $\tilde{\mathbf{X}}$.

34.6 Network Analysis

Metabolic measurements provide a wealth of information about the biochemical status of cells, tissues or organisms and the consideration of metabolic processes has become increasingly important in the last few years, in particular, for disease monitoring/diagnosis, aetiology and to elucidate novel gene functions. Technological advances have considerably extended our ability to describe complex biological systems and it is now possible to generate quantitative profiles of thousands of molecular species (metabolites). A feature of metabolomic data is that a significant number of metabolite levels are highly interrelated. These correlations, or more generally ‘associations’, do not necessarily occur between metabolites that are neighbours in a metabolic pathway (e.g. substrates/products), but are the result of direct enzymatic conversions and of indirect cellular regulations of transcriptional and biochemical processes (Steuer, 2006). For example, two or more molecules involved in the same pathway may present high intermolecular correlation and exhibit a similar response to a stimulus. The primary interest in the analysis of metabolite associations (metabolic association networks) stems from the fact that the observed pattern provides information about the physiological state of a metabolic system and complex metabolite relationships. The working assumption is that a network of metabolic associations represents a fingerprint of the underlying biochemical network in a given biological state, which then can be related to changes between different genotypes, phenotypes or experimental conditions (Steuer *et al.*, 2003; Weckwerth, 2003). Associative patterns as identified by network techniques may represent an important set of information that can be related to molecular marker variation. Therefore, topological differences in metabolic association networks might prove useful to complement findings of subtle differences in variances and averages of metabolite concentration levels to discriminate between different groups (e.g. diseased/not diseased). At the metabolite level, it is plausible to assume that a transition to a different physiological state may not necessarily involve changes in the mean levels of the metabolite concentrations but may also arise due to changes in metabolic interactions (Steuer, 2006). It is possible that metabolite patterns as identified by network techniques might eventually provide a better understanding of molecular marker variation than marginal metabolite variation (Ursem *et al.*, 2008).

Statistical network analysis is complementary to the elucidation of biochemical reaction networks. Although the latter offer many advantages – in particular, they link more closely to functional phenotypes (i.e. fluxes) – biochemical network reconstruction is typically labour-intensive as it requires significant biochemical characterization and represents many years of accumulated experimental data (Price and Shmulevich, 2007). On the other hand, the ability to rapidly reconstruct networks from high-throughput data is an inherent advantage of inferred statistical networks, which have also been proved to be able to recover the reaction pathway (e.g. Arkin *et al.*, 1997).

When dealing with metabolic network analysis we need to distinguish two situations: (i) the underlying network is unknown and needs to be estimated from data; (ii) the network is observed and the analysis is mainly focused on characterizing network properties and explaining the network generative process. Situation (i) typically involves independent estimation of a graph under each condition under investigation, mainly based on correlation measures, and subsequent comparison of their topological properties in the case of multiple conditions, such as connectivity patterns observed in each state. The most common statistical method to infer a network from data is the Gaussian graphical model (GGM; Lauritzen, 1996) framework. GGMs have emerged as a fundamental tool for studying dependency structures among variables measured at the individual level, with statistical associations shown as edges in a graph. Suppose we have p -variate data $X = (X_1, \dots, X_p)$ (e.g. concentrations of p metabolites) arising from a Gaussian distribution with precision matrix Ω . Given an independent and identically distributed sample, we are interested in estimating Ω . Estimation of the precision matrix is equivalent to model selection among undirected GGMs. A graph $G = (V, E)$ is described by a set of nodes V and edges E , with variables X_1, \dots, X_p placed at the nodes. The edges define the global conditional independence structure of the joint distribution of X . Absence of an edge corresponds to a zero in Ω , indicating conditional independence between a pair of variables; presence of an edge points to a non-zero element in Ω . Thus, G represents the interconnections between interacting nodes (e.g. metabolic modules) in the form of subgraphs. Most of the techniques for GGM inference centre on efficient ways of estimating the precision matrix, which describes the linear associations between variables. Recent developments, in particular favour shrinkage and sparse solutions, reduce the likelihood of spurious associations and allow application in high-dimensional and noisy settings (see, for example, Friedman *et al.*, 2008; Schaefer *et al.*, 2009; Tan *et al.*, 2017).

Valcárcel *et al.* (2011) have developed the idea of differential networks, which can be used to reveal metabolite interactions that discriminate between states. Differential networks do not have biological interpretations *per se* but highlight interesting interactions, as they are formed of all interactions that significantly change between states. Figure 34.4 shows the estimated differential network between two rat brain tissues, the occipital cortex and the hippocampus, under normal physiological conditions. Every edge in the differential network represents an inter-metabolite relationship which differs between the two tissues. Metabolites that also significantly differ in mean-level concentration are highlighted in red. It appears that most statistically different interactions between tissues occur because the (partial) correlations change sign, which points to a marked change in the underlying regulation of the system and potentially reflects the existence of multiple steady states. An interesting finding is that most of the metabolites which change their mean levels between states tend to be highly connected in the differential network, that is, participate in many significantly different interactions between states. Moreover, we observe that many of these interactions correspond to known metabolic relationships. For example, they are members of the same pathway and/or have been implicated in normal and pathological brain function. Differential networks are interesting because they can lead to formulation of new hypotheses and guide further experimentation.

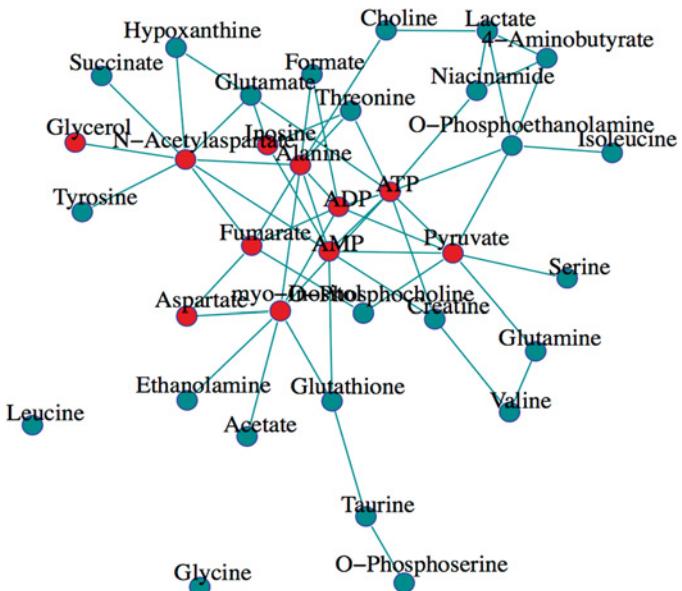


Figure 34.4 An example differential dependency network estimated from partial correlations of metabolite levels between rat hippocampus and occipital cortex under normal physiological conditions.

This type of analysis provides insightful results and offers an important starting point for future research.

The approach proposed by Valcárcel *et al.* (2011) is heuristic, as it is based on estimating the graphs under the two conditions independently and then comparing their properties. This strategy, although easy to implement, may lead to spurious network estimation because each network contains considerable noise due to limited sample sizes and does not consider the overall complexity of the system. Recently more rigorous methodological developments for joint inference of multiple graphs have been proposed, with the aim of highlighting common structures and differences (e.g. Guo *et al.*, 2011; Telesca *et al.*, 2012; Peterson *et al.*, 2015; Tan *et al.*, 2017).

In the case where the network is observed (situation (ii)), the focus of analysis is on characterizing the network, looking at its properties, such as degree distribution, clustering coefficients of each node and network cohesion (Kolaczyk, 2009). Moreover, it is often of interest to understand the data generation process and in this case random graph theory plays an important role. Also in this context, comparison of networks under different biological conditions is becoming more and more frequent. Early analyses are based primarily on empirical observations or comparison of summary statistics derived from individual networks built under different conditions. An early example of networks analysed differentially is the comparison of protein–protein interaction data across species (Matthews *et al.*, 2001). Despite the increasing number of differential networks techniques that have been proposed to date, the statistical community has only in the last few decades started to devote its interest to issues associated with multiple related networks. Differential network analysis introduces a number of statistical challenges (Kerr *et al.*, 2000; Hatfield *et al.*, 2003) as only one random network, representing a snapshot of the entire system, is observed under each condition. It is important to understand systematic and independent errors influencing interaction networks under different conditions to perform meaningful analysis. Moreover, questions about experimental design and power of

statistical methods able to capture differences in association patterns are still the object of open debate.

Various computational methodologies have been developed to compare networks, including graph matching and graph similarity algorithms (Brandes and Erlebach, 2005). Here we mention two representative approaches. Yates and Mukhopadhyay (2013) approach the problem as one of hypothesis testing and describe an inferential method for performing one- and two-sample hypothesis tests where the sampling unit is a network and the hypotheses are stated via network model(s). Testing is based on a dissimilarity measure that incorporates nearby neighbour information among nodes and determines the sampling distribution of the test statistic under the null hypothesis via resampling techniques. Ruan *et al.* (2015) propose an algorithm (dGHD algorithm) based on the generalized Hamming distance, for assessing the degree of topological difference between networks and evaluating its statistical significance.

In conclusion, the analysis of metabolic association networks still presents many open and challenging research problems and calls for the development of statistical methods and corresponding computational machinery to deal with the complexity and high dimensionality of the data and model space.

34.7 Metabolite Identification and Pathway Analysis

As mentioned above, a key problem in untargeted metabolomics is the lack of identification (or annotation) of the variables. This is due to the difficulty of unambiguously identifying features in raw spectral data without extensive work by an expert, which can ultimately be traced back to the wide array of chemistries represented by metabolites and the (relative) lack of standardization in the field. While identification of unknowns mostly depends on expert interpretation of experimental data, there are some statistical approaches which can help guide experiments. One of the simplest and most widely used is statistical correlation spectroscopy.

34.7.1 Statistical Correlation Spectroscopy

In this technique, the pattern of statistical correlations between spectral intensities is exploited to aid identification and other tasks. Often a single metabolite gives rise to several peaks in the spectral data. In LC-MS this could be due to adduct formation, isotopologues and fragmentation processes. In NMR multiple peaks are a consequence of chemical shift and spin–spin coupling. The intensities of these peaks will rise and fall with the concentration of the metabolite in the sample, and therefore become highly correlated. By examining the high correlations, it may be possible for a spectroscopist to associate an unknown peak with others in the spectrum, and this may give clues as to the identity of the underlying metabolite. For a data set \mathbf{X} with n spectra (rows) and m variables (columns, e.g. NMR intensities or LC-MS feature intensities), statistical correlation spectroscopy depends on examining the Pearson product-moment correlation between pairs of variables,

$$r_{jk} = \frac{1}{n-1} \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sigma_j \sigma_k}, \quad (34.13)$$

where \bar{x}_j and σ_j denote the mean and standard deviation of the j th variable.

Statistical total correlation spectroscopy (STOCSY) was the first statistical spectroscopy technique in metabolomics (Cloarec *et al.*, 2005). It focuses on NMR and is the most widely used approach today. As can be seen from Figure 34.5(a), a plot of the two-dimensional $m \times m$

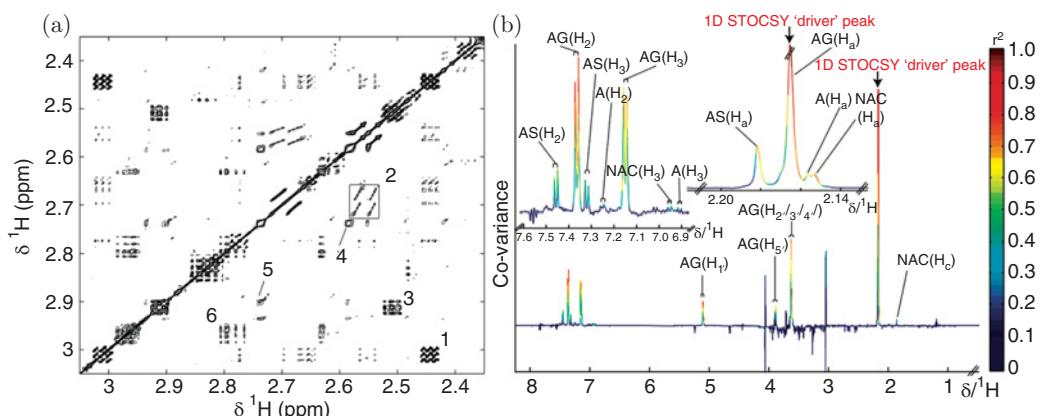


Figure 34.5 Statistical total correlation spectroscopy (STOCSY). (a) Contour map of the two-dimensional STOCSY correlation matrix derived from ^1H NMR spectra of mouse urine. Reprinted with permission from Cloarec *et al.* (2005). © 2005 American Chemical Society. (B) One-dimensional STOCSY examining metabolites of the drug acetaminophen using ^1H NMR spectra of human urine and driven from the peak at 2.17 ppm. Insets show zooms of the regions around 7.3 and 2.2 ppm. Abbreviations: A, acetaminophen; AG, acetaminophen glucuronide; AS, acetaminophen sulfate; NAC, *N*-acetyl-L-cysteine acetaminophen. Reprinted with permission from Holmes *et al.* (2007). © 2007 American Chemical Society.

matrix \mathbf{R} reveals a number of off-diagonal correlations resulting mainly from peaks related to the same metabolite. A more typical visualization is that of Figure 34.5(b) in which the correlations from one row (or column) of the matrix are mapped onto a spectrum-like plot. Here, a ‘driver peak’, usually the one whose identity we are attempting to determine, is used to select the relevant row of \mathbf{R} . In the plot, the x -axis is chemical shift, the y -axis the covariance (with respect to the driver) and the colours denote squared correlation to the driver. Using covariance as the y -coordinate leads to a spectrum-like plot which is easy for a spectroscopist to interpret, allowing peaks to be differentiated from noise, and the sign of the correlation to be visualized. Strong positive correlations occurring on a clear peak suggest that it may be related to the unknown.

Since the original STOCSY, a large number of variants have been published, aimed at exploiting statistical correlations for different purposes, or incorporating other information. For example, iterative STOCSY (iSTOCSY) thresholds the correlations and uses them as new driver peaks in a second round of STOCSY (Sands *et al.*, 2011). Further rounds of iterative thresholding and correlation calculations produce a network of strong correlations which can reveal previously hidden relationships, such as those between drugs and their metabolites. Another modification is subset optimization by reference matching (STORM) which aims to select the subset of spectra which most strongly exhibit a specific correlation pattern (Posma *et al.*, 2012). This is useful when the unknown metabolite may be present only in a subset of the individuals, often the case in large studies. Starting with a driver pattern (a set of ‘interesting’ variables), STORM uses iterative rounds of subset selection and STOCSY to find the optimal subset, and also employs statistical shrinkage to avoid spurious correlations. This can significantly enhance the correlation pattern derived via standard STOCSY. From the chemical analytic point of view, it is very helpful to use statistical correlations to link data from different instruments, thus providing complimentary information on an unknown. This is the motivation behind statistical heterospectroscopy, in which MS data are correlated to NMR data from the same samples (Crockford *et al.*, 2006). Strong correlations between these very different experimental methods can yield a wealth of information on unknowns, allowing new molecules to be identified (Crockford *et al.*, 2008).

34.7.2 Pathway and Metabolite Set Analysis

If many metabolites are identified in a data set, it becomes possible to incorporate biological questions more explicitly into statistical analysis. One approach for doing this is pathway analysis. Biological pathways are groups of molecules (e.g. genes, proteins, metabolites) which act together in a coordinated way as part of the larger biological system. In metabolomics, the relevant pathways are metabolic pathways, in which metabolites are transformed from one to another via metabolic reactions. Since each reaction is catalysed by an enzyme (a protein which can speed up the natural rate of the reaction), metabolic pathways also become a useful tool for integration of multi-omics data since metabolomics data can be combined with proteomic or transcriptomic data for enzymes (Cavill *et al.*, 2011). A further important class of proteins is the transporters which control influx and efflux of metabolites into and out of the cell.

Pathway analysis in metabolomics follows the same approach as gene set analysis in other areas of bioinformatics. The basic workflow is as follows. A data set of metabolites and their concentrations is combined with a database of known pathways. Each pathway is considered one at a time and a statistical test is made to decide whether the metabolite levels in the pathway are significantly affected by the experimental factors (e.g. treatment versus control). The output is a list of pathways which are deemed significantly altered. There are many approaches for pathway analysis, most of which can be assigned to one of three types: over-representation analysis (ORA; Tavazoie *et al.*, 1999), set enrichment analysis, and topology-based approaches. Since it is the most widely used, we only describe ORA in detail here.

Consider the situation in which m identified metabolites (the background) are quantified under a control and treatment conditions. Any valid approach (e.g. t tests) could be used to determine which of these metabolites have significantly different concentrations in each condition, using an appropriate multiple-testing strategy. This yields a set of s ‘interesting’ metabolites significantly altered in concentration between the two conditions. We compare this to the set of K metabolites known to be members of the pathway and determine the number l of metabolites in the intersection of both sets. Under the null hypothesis that the s interesting metabolites are selected at random from the set of m observable metabolites, the probability of observing l or more metabolites in the intersection is given by the cumulative hypergeometric distribution:

$$\Pr(\geq l) = 1 - \sum_{k=0}^{l-1} \frac{\binom{K}{k} \binom{m-K}{s-k}}{\binom{m}{s}}. \quad (34.14)$$

Thus, the null hypothesis can be rejected if this value is less than α (usually $\alpha = 0.05$), and we say that the pathway exhibits significant over-representation of metabolites with altered concentrations. This process can be repeated for each pathway in the database, with appropriate control for multiple testing. The same argument can of course be used for sets of metabolites selected in other ways to examine their over-representation within a pathway. For example, one could ask whether metabolites of a particular chemical class are associated with particular pathways.

There are a number of limitations to pathway analysis in general, and ORA in particular. First, pathways are arbitrary sets of metabolites which have been grouped together for subjective reasons, which may be purely historical. In reality, the collection of reactions forms a metabolic network, of which a pathway is a local connected graph. Importantly, pathways may overlap, so that tests within an ORA analysis are not independent. Further, pathway definitions often vary between different metabolic databases, so that different results may be obtained for

seemingly similar biological functions. A key problem when applying ORA to metabolomic data is that the background estimate (the set of m observable metabolites) is always much smaller than the full metabolome, since it is not yet possible to assay all metabolites in an organism. Therefore, the set of K metabolites from the pathway must be adjusted to only include those which were observable with the assay used to obtain the data set at hand. If this is not done, and the background is assumed to be the full metabolome of the organism, p -values tend to be overly optimistic, and incorrect conclusions can be drawn.

34.8 Conclusion

Metabolomics science continues to evolve as the underlying analytical technologies and applications expand and improve. There remain many open problems, such as optimal determination of peak correspondence (alignment) and integration of metabolomic data with other omics data. In addition, new generations of instruments (e.g. ion mobility mass spectrometry) bring new challenges for data analysis. While the metabolomics-focused statistical community is small, it benefits from interactions with other fields such as machine learning, chemometrics and applied statistics. Given the continued flow of ideas between these fields, as well as the rapidly expanding nature of metabolomics itself, we look forward to seeing a wide array of new statistical applications in this area in the future.

References

- Åberg, K.M., Torgrip, R.J., Kolmert, J., Schuppe-Koistinen, I. and Lindberg, J. (2008). Feature detection and alignment of hyphenated chromatographic–mass spectrometric data: Extraction of pure ion chromatograms using kalman tracking. *Journal of Chromatography A* **1192**(1), 139–146.
- Arkin, A., Shen, P. and Ross, J. (1997). A test case of correlation metric construction of a reaction pathway from measurements. *Science* **277**(5330), 1275–1279.
- Astle, W., De Iorio, M., Richardson, S., Stephens, D. and Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association* **107**(500), 1259–1271.
- Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**(10), 781.
- Blaise, B.J. (2013). Data-driven sample size determination for metabolic phenotyping studies. *Analytical Chemistry* **85**(19), 8943–8950.
- Blaise, B.J., Correia, G., Tin, A., Young, J.H., Vergnaud, A.-C., Lewis, M., Pearce, J.T., Elliott, P., Nicholson, J.K., Holmes, E., et al. (2016). Power analysis and sample size determination in metabolic phenotyping. *Analytical Chemistry* **88**(10), 5179–5188.
- Brandes, U. and T. Erlebach (eds) (2005). *Network Analysis, Lecture Notes in Computer Science* Vol. 3418. Springer, Berlin.
- Cavill, R., Kamburov, A., Ellis, J.K., Athersuch, T.J., Blagrove, M.S., Herwig, R., Ebbels, T.M. and Keun, H.C. (2011). Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Computational Biology* **7**(3), e1001113.
- Chadeau-Hyam, M., Ebbels, T.M., Brown, I.J., Chan, Q., Stamler, J., Huang, C.C., Daviglus, M.L., Ueshima, H., Zhao, L., Holmes, E., et al. (2010). Metabolic profiling and the metabolome-wide

- association study: Significance level for biomarker identification. *Journal of Proteome Research* **9**(9), 4620–4627.
- Cloarec, O., Dumas, M.E., Craig, A., Barton, R.H., Trygg, J., Hudson, J., Blancher, C., Gaugier, D., Lindon, J.C., Holmes, E. and Nicholson, J.K. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets. *Analytical Chemistry* **77**(5), 1282.
- Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K. and Lindon, J.C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry* **78**(7), 2262–2267.
- Crockford, D.J., Holmes, E., Lindon, J.C., Plumb, R.S., Zirah, S., Bruce, S.J., Rainville, P., Stumpf, C.L. and Nicholson, J.K. (2006). Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: Application in metabonomic toxicology studies. *Analytical Chemistry* **78**(2), 363–371.
- Crockford, D.J., Maher, A.D., Ahmadi, K.R., Barrett, A., Plumb, R.S., Wilson, I.D. and Nicholson, J.K. (2008). ^1H NMR and UPLC-MS^E statistical heterospectroscopy: Characterization of drug metabolites (xenometabolome) in epidemiological studies. *Analytical Chemistry* **80**(18), 6835–6844.
- Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^1H NMR metabonomics. *Analytical Chemistry* **78**(13), 4281–4290.
- Ebbels, T.M. and Cavill, R. (2009). Bioinformatic methods in NMR-based metabolic profiling. *Progress in Nuclear Magnetic Resonance Spectroscopy* **55**(4), 361–374.
- Fearn, T. (2000). On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems* **50**(1), 47–52.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441.
- Gika, H.G., Macpherson, E., Theodoridis, G.A. and Wilson, I.D. (2008). Evaluation of the repeatability of ultra-performance liquid chromatography–TOF-MS for global metabolic profiling of human urine samples. *Journal of Chromatography B* **871**(2), 299–305.
- Goodacre, R., Broadhurst, D., Smilde, A.K., Kristal, B.S., Baker, J.D., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3**(3), 231–241.
- Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**(1), 1–15.
- Hao, J., Astle, W., De Iorio, M. and Ebbels, T.M.D. (2012). Batman – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **28**(15), 2088–90.
- Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J.G. and Ebbels, T.M. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using Batman. *Nature Protocols* **9**(6), 1416.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer, New York.
- Hatfield, G., Hung, S.P. and Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. *Molecular Microbiology* **47**(4), 871–877.
- Hendriks, M.M., van Eeuwijk, F.A., Jellema, R.H., Westerhuis, J.A., Reijmers, T.H., Hoefsloot, H.C. and Smilde, A.K. (2011). Data-processing strategies for metabolomics studies. *Trends in Analytical Chemistry* **30**(10), 1685–1698.
- Hoch, J. and Stern, A. (1996). *NMR Data Processing*. Wiley, New York.

- Hoggart, C.J., Clark, T.G., De Iorio, M., Whittaker, J.C. and Balding, D.J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology* **32**(2), 179–185.
- Holmes, E., Loo, R.-L., Cloarec, O., Coen, M., Tang, H., Maibaum, E., Bruce, S., Chan, Q., Elliott, P., Stamler, J., Wilson, I.D., Lindon, J.C. and Nicholson, J.K. (2007). Detection of urinary drug metabolite (xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy. *Analytical Chemistry* **79**(7), 2629–2640.
- Kalogeropoulou, A. (2011). Pre-processing and analysis of high-dimensional plant metabolomics data. PhD thesis, University of East Anglia.
- Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**(6), 819–837.
- Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer, New York.
- Lauritzen, S.L. (1996). *Graphical Models*. Clarendon Press, Oxford.
- Liland, K.H. (2011). Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *Trends in Analytical Chemistry* **30**(6), 827–841.
- Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S. and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or ‘interologs’. *Genome Research* **11**(12), 2120–2126.
- Peterson, C., Stingo, F.C. and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association* **110**(509), 159–174.
- Pluskal, T., Castillo, S., Villar-Briones, A. and Orešić, M.; (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**(1), 395.
- Posma, J.M., Garcia-Perez, I., De Iorio, M., Lindon, J.C., Elliott, P., Holmes, E., Ebbels, T.M. and Nicholson, J.K. (2012). Subset optimization by reference matching (STORM): An optimized statistical approach for recovery of metabolic biomarker structural information from ^1H NMR spectra of biofluids. *Analytical Chemistry* **84**(24), 10694–10701.
- Price, N.D. and Shmulevich, I. (2007). Biochemical and statistical network models for systems biology. *Current Opinion in Biotechnology* **18**(4), 365–370.
- Ruan, D., Young, A. and Montana, G. (2015). Differential analysis of biological networks. *BMC Bioinformatics* **16**(1), 327.
- Rubtsov, D.V. and Griffin, J.L. (2007). Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy. *Journal of Magnetic Resonance* **188**(2), 367–379.
- Sands, C.J., Coen, M., Ebbels, T.M., Holmes, E., Lindon, J.C. and Nicholson, J.K. (2011). Data-driven approach for metabolite relationship recovery in biological ^1H NMR data sets using iterative statistical total correlation spectroscopy. *Analytical Chemistry* **83**(6), 2075–2082.
- Schaefer, J., Opgen-Rhein, R. and Strimmer, K. (2009). GeneNet: Modeling and inferring gene networks. R package version 1. <http://cran.r-project.org/package=GeneNet/index.html>.
- Smith, C.A., Want, E.J., O’Maille, G., Abagyan, R. and Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry* **78**(3), 779–787.
- Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Briefings in bioinformatics* **7**(2), 151–158.
- Steuer, R., Kurths, J., Fiehn, O. and Weckwerth, W. (2003). Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**(8), 1019–1026.
- Tan, L.S., Jasra, A., De Iorio, M., Ebbels, T.M., et al. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *Annals of Applied Statistics* **11**(4), 2222–2251.

- Tautenhahn, R., Boettcher, C. and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9**(1), 504.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**(3), 281.
- Telesca, D., Müller, P., Parmigiani, G. and Freedman, R.S. (2012). Modeling dependent gene expression. *Annals of Statistics* **6**(2), 542.
- Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **16**(3), 119–128.
- Trygg, J. and Wold, S. (2003). O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics* **17**(1), 53–64.
- Ursem, R., Tikunov, Y., Bovy, A., Van Berloo, R. and Van Eeuwijk, F. (2008). A correlation network approach to metabolic data analysis for tomato fruits. *Euphytica* **161**(1–2), 181.
- Valcárcel, B., Würtz, P., al Basatena, N.K.S., Tukiainen, T., Kangas, A. J., Soininen, P., Järvelin, M.R., Ala-Korpela, M., Ebbels, T.M. and de Iorio, M. (2011). A differential network approach to exploring differences between biological states: An application to prediabetes. *PLoS ONE* **6**(9), e24702.
- Van Iterson, M., Pedotti, P., Hooiveld, G., Den Dunnen, J., van Ommen, G., Boer, J., Menezes, R., et al. (2009). Relative power and sample size analysis on gene expression profiling data. *BMC Genomics* **10**(1), 439.
- Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J.J. and Yanes, O. (2012). A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. *Metabolites* **2**(4), 775–795.
- Vu, T.N. and Laukens, K. (2013). Getting your peaks in line: A review of alignment methods for NMR spectral data. *Metabolites* **3**(2), 259–276.
- Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology* **54**(1), 669–689.
- Weljie, A.M., Newton, J., Mercier, P., Carlson, E. and Slupsky, C.M. (2006). Targeted profiling: Quantitative analysis of ¹H NMR metabolomics data. *Analytical Chemistry* **78**(13), 4430–4442.
- West, M. (2003). Bayesian factor regression models in the ‘large *p*, small *n*’ paradigm. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West (eds), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. Clarendon Press, Oxford, pp. 733–742.
- Xia, J., Sinelnikov, I.V., Han, B. and Wishart, D.S. (2015). Metaboanalyst 3.0 – making metabolomics more meaningful. *Nucleic Acids Research* **43**(W1), W251–W257.
- Yates, P.D. and Mukhopadhyay, N.D. (2013). An inferential framework for biological network hypothesis tests. *BMC Bioinformatics* **14**(1), 94.

35

Statistical and Computational Methods in Microbiome and Metagenomics

Hongzhe Li

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

Abstract

High-throughput sequencing technologies enable individualized characterization of the microbiome composition and functions. The human microbiome, defined as the community of microbes in and on the human body, impacts human health and risk of disease by dynamically integrating signals from the host and its environment. The resulting data can potentially be used for personalized diagnostic assessment, risk stratification, disease prevention and treatment. I review the statistical and computational methods for measuring various important features of the microbiome and how these features are used as an outcome of an intervention, as a mediator of a treatment and as a covariate to be controlled for when studying disease/exposure associations. The classical statistical methods require modifications to account for covariates, sparsity and compositional nature of the data, and new statistical and computational methods are needed for integrative analysis of microbiome, small molecules and metabolomics to better understand microbiome–host interactions. I illustrate these methods using several real microbiome studies.

35.1 Microbiome in Human Health and Disease

Microbiota, the microbial organisms in and on the human body, including bacteria, archaea, protists, fungi and viruses, constitute the human microbiome, with the composition of the microbiome varying substantially between different body sites and between healthy and diseased individuals (Cho and Blaster, 2012). Our largest collection of microbes resides in the gut, which extends human metabolism and affects human health by enabling the degradation and metabolism of components of our diet and by providing many important small molecules. The microbiota and the genes that they provide play key roles in human health by affecting immunity, metabolism, development, and behavior. Dysbiosis of gut microbiome contributes to many diseases, including metabolic syndrome, inflammatory bowel disease, cardiovascular diseases, cancer and responses to cancer treatment (Gopalakrishnan *et al.*, 2017; Koeth *et al.*, 2013; Routy *et al.*, 2018; Turnbaugh *et al.*, 2009).

High-throughput sequencing (HTS) technologies make large-scale studies of microbiome possible. One approach to sequencing-based microbiome studies is to sequence specific genes (often the 16S rRNA gene) to produce a profile of diversity of bacterial taxa. Taxa are classifications of bacteria and are often arranged in a hierarchy in descending order as kingdom, phylum class, order, family, genus, species, and subspecies. Alternatively, the HTS-based sequencing strategy, also called metagenomic shotgun sequencing, the technique of sequencing the collective genome of all microorganisms from a given sample, provides further insights at the molecular level, such as species/strain quantification, gene function analysis and microbial growth dynamics. Such data also provide information for identifying new microbial species. These sequence-based analyses have expanded our knowledge of microbiomes by generating enormous new data sets that provide important insights into the composition and functional properties of a large number of microbial communities. The US Human Microbiome Project (Human Microbiome Project Consortium, 2012), together with the European project MetaHIT (Qin *et al.*, 2010), has provided preliminary understanding of the biological significance of the human microbiome. The sequencing data, coupled with functional data such as metabolomics (see **Chapter 34**), present an important approach to studies of complex phenotypes and the roles that the microbiome plays.

As an example, Lewis *et al.* (2015) conducted a longitudinal microbiome study of 90 children with Crohn's disease who were initiating treatment with either a defined formula diet or anti-TNF therapy and compared them to 26 healthy control children. They tracked symptoms, mucosal inflammation (fecal calprotectin, FCP), and changes in the gut microbiome over an 8-week study period, where the gut microbiome was quantified using shotgun metagenomic sequencing of the fecal samples. Dysbiosis was quantified as the distance of each sample's microbial composition from the centroid of the microbial compositions from healthy controls. They observed that the microbial communities were partitioned into two distinct clusters based on the bacterial composition and showed that bacterial community membership was associated independently with intestinal inflammation, antibiotic use, and therapy. By tracking the microbiota composition over the course of therapy, they found that dysbiosis was reduced in response to decreased bowel inflammation. Figure 35.1 compares the composition of the relatively common bacterial taxa at the genus level in the healthy control subjects and pediatric Crohn's disease cohort prior to initiation of therapy with clinical metadata summarized above the heat map. It shows the difference in microbial composition between normal and Crohn's disease patients and a large variability in bacterial compositions among the patients.

Driven by the recent surge in human gut microbiome research, new statistical and computational methods have been developed to answer the questions regarding the data normalization, species abundance quantifications, association between high-dimensional compositional data and clinical data and associations among the bacterial species and microbial ecology. The taxonomic diversity that is inherent in complex environmental communities, the potential systemic biases in sequencing data, and the compositional nature of the data create unique statistical challenges. This chapter reviews some recent statistical methods and computational metagenomic tools for microbiome feature estimation, for linking microbiomes with various clinical phenotypes, microbiome mediation analysis and computational tools for integrating microbiome and metabolomics data (**Chapter 34**), covering more on the computational methods for studying the metabolic functions of the microbiomes based on metagenomic sequencing data. Although this review focuses on the microbiome in human health and disease, the methods reviewed can be equally applied to microbiome research in plants, soil, oceans, and the atmosphere.

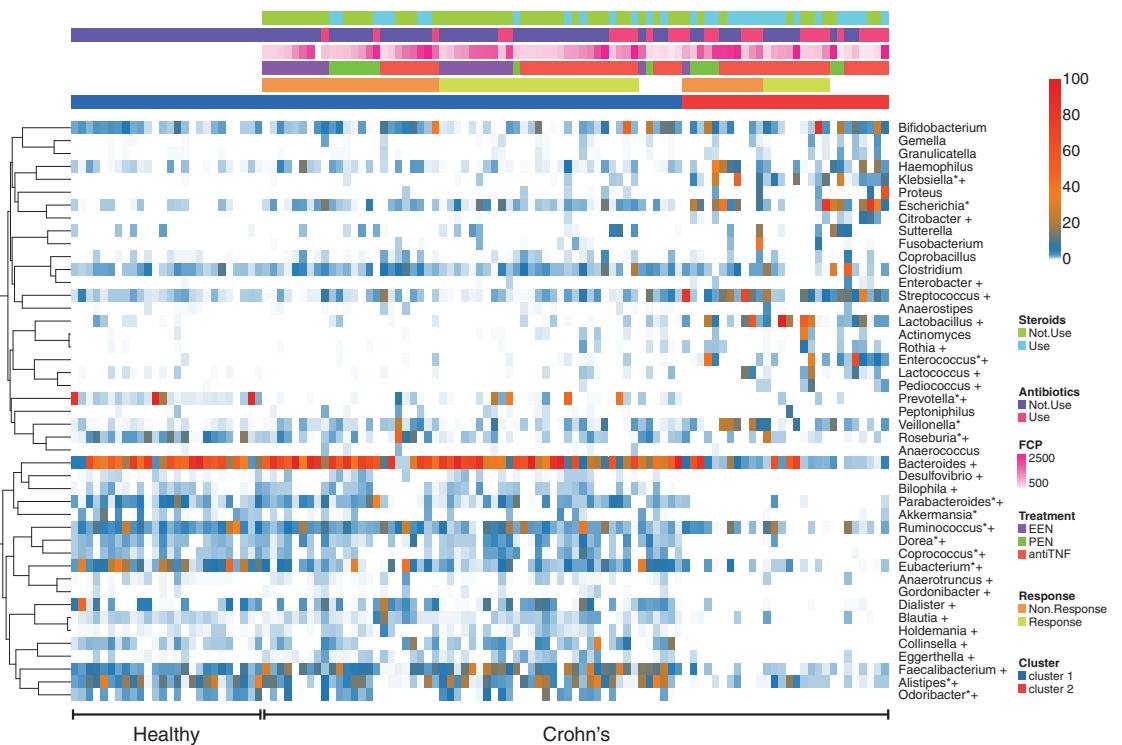


Figure 35.1 A heatmap demonstrating relative abundance of bacterial genera prior to therapy according to presence or absence of Crohn's disease, microbial cluster assignment, use of corticosteroids and antibiotics, FCP concentration, and response to therapy (Lewis *et al.*, 2015). Metadata are indicated by the color code at the top of the figure. White cells indicate missing data. Taxa that were statistically different in abundance between Crohn's disease and healthy controls are identified by *; taxa that were statistically different in abundance between the two Crohn's disease clusters are identified by + ($q < 0.05$, Wilcoxon rank-sum test). Samples were ordered healthy versus Crohn's samples, cluster 1 versus cluster 2 defined based on the microbial compositions, then other forms of metadata.

35.2 Estimation of Microbiome Features from 16S rRNA and Shotgun Metagenomic Sequencing Data

There are two sequencing approaches to quantifying the composition of the human microbiome. The first involves sequencing only specific marker genes (Tringe and Rubin, 2005), also called the amplicon-based approach. The 16S rRNA gene is often sequenced to study bacterial composition. This gene is ubiquitous in the bacterial kingdom and contains highly variable regions whose sequences can serve as unique markers for different types of bacteria. The basic idea is to isolate from all bacteria the DNA strands corresponding to some variable region of the 16S rRNA gene and to create and sequence the resulting amplicons. Amplicon sequencing can reveal the phylogenetic structure of a microbial community, which can be very useful in downstream analyses. Since rRNA makes up 80% of total bacterial RNA, 16S rRNA sequencing allows detection of rare members of the community with high sensitivity. However, 16S data do not provide any information about bacterial gene inventory and functionality. In addition, the data do not provide high sensitivity in identifying bacteria at the species or strain level.

Alternatively, shotgun metagenomic sequencing involves DNA being extracted from all cells in a community and subsequently sheared into fragments that are independently sequenced, resulting in a very large set of reads. Some of these reads will be sampled from taxonomically informative genomic loci (e.g. universal marker genes, clade-specific marker genes), and others will be sampled from coding sequences that provide insight into the biological functions such as enzymes coded. As a result, shotgun metagenomic data provide the opportunity to simultaneously quantify the microbial community composition and gene functions. Quince *et al.* (2017) present a review of current methods in metagenomics, and provide practical guidance and advice on sample processing, sequencing technology, assembly, binning, annotation, and experimental design. The data also allow the reconstruction of draft genome sequences for individual community members and the estimation of the bacterial growth dynamics. Finally, this approach makes possible the detection of new microbial species and new genes. However, in order to achieve the same level of sensitivity in detecting rare taxa as 16S sequencing, much deeper sequencing is required.

The features that can be derived from 16S rRNA sequencing include counts of reads aligned to operational taxonomic units (OTUs) or bacterial taxa, which are often normalized into proportions to account for differences in sequencing depths. Alternatively, the individual reads can also be placed into an existing phylogenetic tree, which allows us to classify the sequences in the sample as to taxonomic position on the tree (McCoy and Matsen, 2013). In contrast, for shotgun metagenomic sequencing, the important features include species/strain relative abundances, gene/pathway relative abundances, and bacterial replication rates to reflect the growth dynamics. We briefly review some of the most commonly used methods to estimate these features from sequencing data.

35.2.1 Estimation of Microbiome Features in 16S rRNA Data

To obtain the taxonomic classification, gene marker amplicon sequences, like certain hyper-variable regions of the bacterial 16S rRNA gene, are clustered by two main approaches. First, sequences can be clustered into phylotypes according to their similarity to previously annotated sequences in a reference database. Second, OTUs can be constructed by clustering sequences *de novo*, purely based on the sequence similarity. In all cases, an arbitrary similarity threshold is used to differentiate clusters, where the similarity between two sequencing reads can be quantified by the normalized Hamming distance. The 99% similarity threshold is generally

accepted as a good proxy for species, and the 97% similarity threshold for genera. However, the former threshold is often insufficient to discriminate between closely related species. The most widely used software package, QIIME (<https://qiime2.org>), provides three high-level protocols for binning the reads into OTUs, including *de novo*, closed-reference, and open-reference OTU binning. However, these naive methods ignore the possible sequencing errors in read data.

To account for possible sequencing errors in Illumina reads in OTU construction, Callahan *et al.* (2016) developed an algorithm that models errors as occurring independently within a read, and independently between reads. Under this error model, the probability that an amplicon read with sequence i is produced from sample sequence j can be calculated. Let sequence i with abundance a_i be in partition j containing n_j reads. For such a sequence i , the abundance p -value is defined as the probability of seeing a_i or more identical reads of type j , where the null hypothesis is that the a_i reads of sequence i are from n_j reads of sample sequence j due to random errors. Therefore, a small p -value indicates that there are more reads of sequence i than can be explained by errors introduced during the amplification and sequencing of n_j copies of sample sequence j (Callahan *et al.*, 2016). These p -values are then used to assess support for whether the reads come from the same or different clusters.

Callahan *et al.* (2016) further developed a divisive partitioning algorithm where sequencing reads with the same sequence are grouped into unique sequences with an associated abundance and consensus quality profile. The divisive partitioning algorithm is initialized by placing all unique sequences into a single set, and assigning the most abundant sequence as the representative of that set. All unique sequences are then compared to the representative of their partition, and error rates at which an amplicon read with sequence i is produced from sample sequence j for all (i,j) read pairs are calculated and stored. The abundance p -value is then calculated for each unique read. If the smallest p -value, after Bonferroni correction, falls below a threshold, a new partition is formed with the unique sequence with the smallest p -value as its representative, and all unique sequences are compared to the representative of that new partition. This procedure has been implemented as the DATA2 pipeline (<https://benjneb.github.io/dada2/tutorial.html>), which outputs an amplicon sequence variant (ASV) table, a higher-resolution analog of the traditional OTU table, which records the number of times each ASV is observed in each sample. These ASVs can be used for downstream data analysis such as calculating the distance between two microbiome samples.

35.2.2 Estimation of Microbial Composition in Shotgun Metagenomic Data

Many bioinformatics and computational tools have been developed to quantify the relative abundances of microbial species, including approaches that use clade-specific marker genes (Mende *et al.*, 2013; Sunagawa *et al.*, 2013) and approaches based on lowest common ancestor (LCA) positioning. These computational tools for quantifying the microbiome relative abundances are based on the genomes of known microbial species/strains, which are also called reference-based approaches. For the former, a gene marker catalog is pre-computed from previously sequenced bacterial genomes and sequences are taxonomically classified by aligning the reads to these marker genes. The marker genes used can be universal or clade-specific. Segata *et al.* (2012) proposed to estimate the taxon abundances using sets of clade-specific marker genes. Chen *et al.* (2017) further developed a multi-sample approach to account for marker-to-marker variability. For the LCA approach, pre-aligned sequences are hierarchically classified on a taxonomy tree using a placement algorithm. Sequences that surpass a dissimilarity threshold (bit-score) are progressively placed on higher taxonomy levels.

Alternatively, reference-free methods have also been developed, including methods based on *de novo* assembly with subsequent analysis of long contigs and composition-based methods

including k -mer spectrum analysis and Markov models (Wood and Salzberg, 2014). To quantify known and unknown microorganisms at species-level resolution using shotgun sequencing data, Sunagawa *et al.* (2013) developed a method that establishes metagenomic OTUs based on single-copy phylogenetic marker genes.

35.2.3 Estimation of Microbial Gene/Pathway Abundance in Shotgun Metagenomic Data

Shotgun metagenomic sequencing also provides information to estimate the abundance of all the microbial genes in a given microbial community. Along with library size or sequencing depth and gene length, average genome size (AGS) should be accounted for in comparative metagenomic studies to identify the genomic differences between microbial communities. The AGS can be estimated using universal marker genes (Frank and Sorensen, 2011). The rationale is that the AGS of a community is inversely proportional to the relative abundance, R , of an essential single-copy gene in that community: $AGS \propto R^{-1}$ (Nayfach and Pollard, 2015; Raes *et al.*, 2007). Using this idea, Nayfach and Pollard (2015) developed Microbe-Census to rapidly and accurately estimate AGS from shotgun metagenomic data based on sequencing read data from 30 universal marker genes and used these estimated average genome lengths to normalize the relative abundances of the microbial gene families.

35.2.4 Quantification of Bacterial Growth Dynamics

Korem *et al.* (2015) showed that the pattern of metagenomic sequencing read coverage for different microbial genomes contains a single trough and a single peak, the latter coinciding with the bacterial origin of replication. Furthermore, the ratio of sequencing coverage between the peak and trough (PTR) provides a quantitative measure of a species' growth rate. They further observed that the growth rates were different between Crohn's disease patients and normal individuals for some species. Figure 35.2 describes how a single bacterium replicates its genome in two directions from a single origin of replication that can result in non-uniform read coverage along the genome. Korem *et al.* (2015) presented a method for estimating PTR for species that have a complete, closed, circular reference genome. For such species, after aligning the sequencing reads to the complete bacterial genome, one can obtain the position-specific coverage values that can be used to estimate the PTRs.

However, for metagenomic data, since reads are sampled from a mixture of many different genomes of different abundances, the problem of estimating such growth dynamics becomes more challenging. Brown *et al.* (2016) presented a method for measurement of bacterial replication rates in microbial communities that include species with only draft genomes constructed

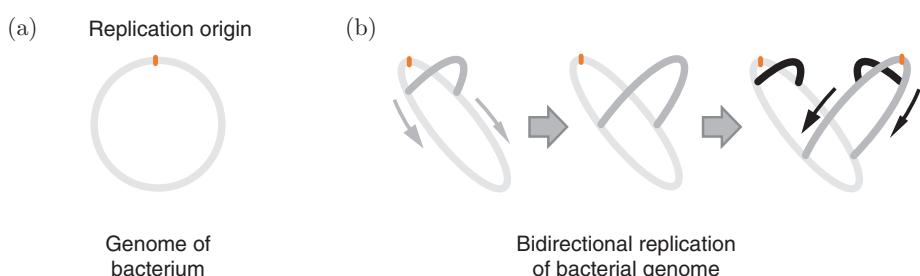


Figure 35.2 Read coverages from bidirectional replication from a single origin provide information on bacterial growth dynamics. (a) The circular bacterial genome has a replication origin marked in red. (b) Bidirectional replications create uneven read coverages, where the darker arcs represent new DNA strands.

based on the metagenomic sequencing reads. This approach first constructs the draft genomes or contigs and then maps the reads to the collection of assembled contigs that represent a draft genome. The challenge is that the genomic ordering of these constructed contigs is unknown. Coverage is then calculated across overlapping segments of genome fragments, where extreme high and low coverage windows are excluded (exceeding 8-fold compared to the median), as they may correlate with highly conserved regions, strain variation, or integrated phage. After the GC bias correction, the average coverage values for each window are ordered from lowest to highest to assess the coverage trend across the genome. Because coverage values for each window are rearranged, the order of the fragments in the complete genome is usually not known. The overall slope of coverage across the genome versus the coverage rank is used to calculate *iRep*, a measure of the average genome copy number across a population of cells, which can be used to measure on average the percentage of the microbial cells that are replicating.

35.2.5 Microbial Diversity Index

At the microbial community level, OTU or ASV groupings from 16S data or species composition from metagenomic sequencing provide one way to measure alpha (within-sample) diversity of the taxa. A diversity index provides a quantitative measure that reflects how many different taxa there are in a community (richness), and simultaneously takes into account how evenly these taxa are distributed in a community (evenness). Such diversity analyses are frequently applied to microbiome data.

There are several measures of diversity, including Chao 1, Simpson's diversity and the Shannon index, all of which reduce microbial compositional data to a single value that takes into account both taxon richness and evenness (Chao, 1984; DeJong, 1975). In general, an increase in the number of taxa, or a more even distribution in their abundances, results in a greater diversity score (DeJong, 1975; Li *et al.*, 2012). For example, Chao 1 estimates total species richness as

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{n_1^2}{2n_2},$$

where S_{obs} is the number of observed species, n_1 is the number of species captured once, and n_2 is the number of species captured twice (Chao, 1984). However, these commonly used microbial diversity indices are not sensitive to rare taxa, which are often observed in microbiome studies. Li *et al.* (2012) introduced a tail statistic τ that is the standard deviation of the rank abundance curve,

$$\tau = \sqrt{\sum_{i=2}^n p_{(i)}(i-1)^2},$$

where n is the number of taxa discovered in the sample and $p_{(i)}$ is the proportion of the i th most abundant taxon. Due to its greater sensitivity to low abundant taxa, this statistic captures the microbiome diversity better than the commonly used Simpson or Shannon index.

35.3 Methods for Analysis of Microbiome as an Outcome of an Intervention or Exposure

In 16S rRNA sequencing-based microbiome studies, we obtain the read counts at a given taxonomic level. These counts are often normalized into proportions in order to account for the

different sequencing depths. For shotgun metagenomic sequencing, most software only outputs the relative abundances of bacterial species. In many applications, the microbiome is treated as an outcome of an intervention or exposure. Examples include assessing the genetic effects on microbiomes (Bonder *et al.*, 2016) or comparing the microbiome difference between normal and diseased individuals (Lewis *et al.*, 2015). Such analyses can be done at each individual taxon level to identify the bacterial taxa that are associated with disease, or at the overall microbial community level to understand how intervention/exposure perturbs microbial communities.

35.3.1 Modeling Multivariate Sparse Count Data as the Response Variable

For 16S rRNA sequencing, one can obtain taxonomic counts as an outcome. Such multivariate count data are often sparse (with lots of zeros) and over-dispersed. Parametric models of such data include the multinomial distribution, Dirichlet multinomial (Chen and Li, 2013b; La Rosa *et al.*, 2012) and Dirichlet-multinomial mixture models (O'Brien *et al.*, 2016). If the analysis is done at the higher taxonomic levels such as class, family and phylum, these models fit the data reasonably well. For repeatedly measured multivariate count data, Shi and Li (2017) developed a model for paired-multinomial data and a test for changes of the underlying proportion parameters. However, if the counts are given at the genus or species level, there are often zeros, especially for rare species, in which case the zero-inflated Dirichlet-multinomial distribution should be used. To account for such over-dispersion, Xia *et al.* (2013) propose an additive logistic normal multinomial regression model to associate the covariates to bacterial composition. The model naturally accounts for sampling variabilities and zero observations and also allows for a flexible covariance structure among the bacterial taxa.

Methods have also been developed at each individual taxon level, including use of the zero-inflated negative binomial distribution. Tang *et al.* (2017) developed a general framework to perform robust association tests to assess the overall association of the microbial community with the covariates that allow for arbitrary inter-taxon dependencies. Unlike existing methods for microbiome association analysis, this framework does not make any distributional assumptions on the microbiome data. It allows adjustment for confounding variables and accommodates many zero observations.

35.3.2 Modeling High-Dimensional Compositional Response Data in Microbiome Studies

If only the relative abundances are available, the resulting data are compositional with a unit-sum constraint. The problem of compositional data is well known in the field of ecology but was ignored in the early days of microbiome data analysis, and only in recent years have we seen appropriate methods developed for compositional data in microbiome studies. While classical compositional data analysis methods work well for low-dimensional data (Aitchison, 1982), they are not appropriate for the high-dimensional and sparse compositional data seen in microbiome studies. At each individual taxon level, nonparametric tests or zero-inflated beta regression are often used to identify differentially abundant taxa (Chen and Li, 2016). It is important to understand the limitation of such data in drawing biological conclusions. The changes in relative abundances do not imply changes in true abundance unless the total bacterial counts remain relatively constant. If microbial load varies substantially between samples, relative abundances will hamper attempts to link microbiome features to metabolic concentrations or physiological parameters. In the extreme case, a change in the true abundance of one taxon can lead to changes in the relative abundances of all other taxa, although their absolute abundances remain the same. A recent study has demonstrated this limitation and possible misinterpretation of the results (Vandepitte *et al.*, 2017). One way to overcome this limitation is to adjust

microbiome profiles for differences in microbial load by spike-in bacteria (Stammmer *et al.*, 2017). However, such spike-in experiments are difficult to perform.

Although 16S rRNA or metagenomic sequencing does not provide information about the absolute abundances of the bacterial taxa, ideas from the compositional data analysis literature can be useful. Cao *et al.* (2018) clarified that with compositional data, one can only test a constant shift of the true abundances of all bacterial taxa and developed a test for testing such a shift. Denote by $\mathbf{X}^{(k)} = (\mathbf{X}_1^{(k)}, \dots, \mathbf{X}_{n_k}^{(k)})^T$ the observed $n_k \times p$ data matrices for group k ($k = 1, 2$), where $\mathbf{X}_i^{(k)}$ represent compositions that lie in the $(p - 1)$ -dimensional simplex $S^{p-1} = \{(x_1, \dots, x_p) : x_j > 0 (j = 1, \dots, p), \sum_{j=1}^p x_j = 1\}$. We assume that the compositional variables arise from a vector of latent variables, called the basis, which may refer to the true abundances of bacterial taxa in a microbial community. Denote by $\mathbf{W}^{(k)} = (\mathbf{W}_1^{(k)}, \dots, \mathbf{W}_{n_k}^{(k)})^T$ the $n_k \times p$ matrices of unobserved bases, which generate the observed compositional data via the normalization

$$X_{ij}^{(k)} = \frac{W_{ij}^{(k)}}{\sum_{\ell=1}^p W_{i\ell}^{(k)}} \quad (i = 1, \dots, n_k; j = 1, \dots, p; k = 1, 2), \quad (35.1)$$

where $X_{ij}^{(k)}$ and $W_{ij}^{(k)} > 0$ are the j th components of $\mathbf{X}_i^{(k)}$ and $\mathbf{W}_i^{(k)}$, respectively.

Denote by $\mathbf{Z}_i^{(k)} = (Z_{i1}^{(k)}, \dots, Z_{ip}^{(k)})^T$ the log-basis vectors, where $Z_{ij}^{(k)} = \log W_{ij}^{(k)}$. Suppose that $\mathbf{Z}_1^{(k)}, \dots, \mathbf{Z}_{n_k}^{(k)}$ ($k = 1, 2$) are two independent samples, each from a distribution with mean $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kp})^T$ and common covariance matrix $\boldsymbol{\Omega} = (\omega_{ij})$. Instead of testing the hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, Cao *et al.* (2018) proposed to test

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 + c\mathbf{1}_p \text{ for some } c \in \mathbb{R} \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 + c\mathbf{1}_p \text{ for any } c \in \mathbb{R}, \quad (35.2)$$

which is testable using only the observed compositional data. Cao *et al.* (2018) developed an overall test for (35.2) based on centered log-ratio transformation under the sparse alternative where only a few taxa have differential abundances. Rejecting the null hypothesis implies a shift in abundance (at the log-scale).

35.4 Methods for Analysis of Microbiome as a Covariate

Microbiome can also be treated as a covariate when studying the disease/exposure associations. One may want to adjust for the gut microbiome composition when evaluating the treatment effects of a drug on the clinical outcome. For example, Routy *et al.* (2018) showed that metagenomics of patient stool samples at diagnosis revealed correlations between clinical responses to immune checkpoint inhibitors and the relative abundance of *Akkermansia muciniphila*. Such analyses can be performed at each individual taxon level or at the overall microbial community level. The compositional nature of the microbiome data makes some of the standard regression methods not directly applicable (Aitchison, 1982; Lin *et al.*, 2014; Shi *et al.*, 2016).

35.4.1 Regression Analysis with Compositional Covariates

To deal with the compositional nature of the data and to select taxa from a list of high-dimensional taxon composition, Lin *et al.* (2014) and Shi *et al.* (2016) introduced compositional data regression and developed methods for parameter estimation and statistical inferences. The linear log-contrast model (Aitchison and Bacon-Shone, 1984) has been proposed for

compositional data regression. Specifically, suppose an $n \times p$ matrix \mathbf{X} consists of n samples of the composition of a mixture with p components, and suppose Y is a response variable depending on \mathbf{X} . The nature of composition makes each row of \mathbf{X} lie in a $(p - 1)$ -dimensional positive simplex $S^{p-1} = \{(x_1, \dots, x_p) : x_j > 0, j = 1, \dots, p \text{ and } \sum_{j=1}^p x_j = 1\}$. Lin *et al.* (2014) formulate a regression problem with a linear constraint on the coefficients by letting $\beta_p = -\sum_{j=1}^{p-1} \beta_j$,

$$Y = \mathbf{Z}\beta + \epsilon, \quad \mathbf{1}_p^T \beta = 0, \quad (35.3)$$

where $\mathbf{1}_p = (1, \dots, 1)^T \in \mathbb{R}^p$, $\mathbf{Z} = (z_1, \dots, z_p) = (\log x_{ij}) \in \mathbb{R}^{n \times p}$, and $\beta = (\beta_1, \dots, \beta_p)^T$. The linear constraint $\mathbf{1}_p^T \beta = 0$ makes the regression coefficients more interpretable when the covariates are in the p -dimensional simplex. Shi *et al.* (2016) extended this model to include multiple linear constraints that can be used to include the relative abundances of the taxa at different taxonomic levels. They further developed a penalized estimation method and a de-biased procedure for statistical inferences.

As an example, model (35.3) was applied to the data set of a microbiome study to link microbial composition with body mass index (BMI) (Wu *et al.*, 2011) and identified the subcomposition of four bacterial genera that are associated with BMI, including *Alistipes*, *Clostridium*, *Acidaminococcus* and *Allisonella*. The ternary plots of three of these four genera in Figure 35.3 clearly show different compositions between normal and overweight individuals.

Lu *et al.* (2018) developed generalized linear models with linear constraints for microbiome compositional data. Bates and Tibshirani (2017) presented a link between the linear model with constraints and the model that includes as covariates the log of all pairwise ratios, which provides another interpretation of model (35.3).

35.4.2 Kernel-Based Regression in Microbiome Studies

Based on the microbiome compositions, methods based on microbial distances are widely applied in microbiome studies. Commonly used distances include UniFrac and its various extensions (Chen *et al.*, 2012; Lozupone *et al.*, 2007; Lozupone and Knight, 2005) and Bray–Curtis distance (Bray and Curtis, 1957). With such a distance matrix available, distance-based

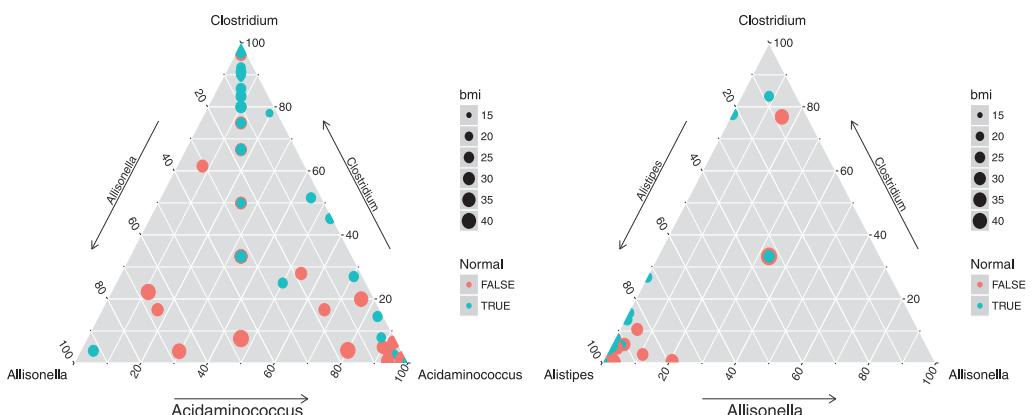


Figure 35.3 Ternary plots of three of the four bacterial genera that are associated with BMI identified using the compositional data regression. Each dot represents a sample, with color used to indicate normal versus overweight and area/radius corresponding to BMI.

tests such as PERMANOVA can be used. Alternatively, Chen and Li (2013a) and Wu *et al.* (2011) developed a kernel regression approach to test for the association between an outcome and the overall microbiome compositions.

We assume that n samples have been collected and that their microbiomes have been measured. For the i th subject, let y_i denote the outcome variable, $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ denote the abundances of all OTUs for individual i where p is the total number of OTUs, and $\mathbf{U}_i = (U_{i1}, U_{i2}, \dots, U_{im})'$ be the vector of the covariates. The goal is to test for association between the outcome and microbial profiles while adjusting for covariates \mathbf{U} . Specifically, for a continuous outcome variable y , one can use the linear kernel machine model

$$y_i = \beta_0 + \boldsymbol{\beta}' \mathbf{U}_i + f(\mathbf{X}_i) + \epsilon_i, \quad (35.4)$$

and for a dichotomous outcome variable, one can use the logistic kernel machine model

$$\text{logit}(P(y_i = 1)) = \beta_0 + \boldsymbol{\beta}' \mathbf{U}_i + f(\mathbf{X}_i), \quad (35.5)$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)'$ is the vector of regression coefficients for the m covariates, and ϵ_i is an error term with mean 0 and variance σ^2 for continuous phenotypes.

Under the kernel machine regression framework, $f(\mathbf{X}_i)$ is assumed to be from a reproducing kernel Hilbert space generated from a positive definite kernel function, $K(\cdot, \cdot)$ such that $f(\mathbf{X}_i) = \sum_{j=1}^n \alpha_j K(\mathbf{X}_i, \mathbf{X}_j)$ for some $\alpha_1, \dots, \alpha_n$. One can construct such a kernel matrix based on the pairwise distances calculated from the observed compositional vector \mathbf{X}_i . The score statistic for testing $H_0 : f(\mathbf{X}) = 0$ is computed as

$$Q = \frac{1}{2\phi} (y - \hat{y}_0)' K (y - \hat{y}_0),$$

where \hat{y}_0 is the predicted mean of y under H_0 , $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are estimated under the null model by regression of y on only the covariates \mathbf{U} , and ϕ is the dispersion parameter. For the linear kernel machine regression, $\phi = \hat{\sigma}_0^2$, where $\hat{\sigma}_0^2$ is the estimated residual variance under the null model. In the logistic kernel machine regression, $\phi = 1$ (see Wu *et al.*, 2011, for details).

In practice, since there are several different ways of defining the pairwise distances between the samples based on the microbiome compositions and therefore different ways of specifying the kernel matrix, Wu *et al.* (2011) suggested that one can apply the score test for each of these possible distances and take the smallest p -value. Its significance can be assessed using permutations.

35.5 Methods for Analysis of Microbiome as a Mediator

The recent literature has demonstrated that the microbiome often serves as a mediator in linking treatment to the outcomes. Wu *et al.* (2017) showed that metformin alters the gut microbiome of individuals with treatment-naïve type 2 diabetes, contributing to the therapeutic effects of the drug. This introduces the notion that altered gut microbiota mediates some of metformin's antidiabetic effects. Similarly, Zackular *et al.* (2016) showed that dietary zinc alters the microbiota and decreases resistance to *Clostridium difficile* infection (CDI) and excess dietary zinc substantially alters the gut microbiota and, in turn, reduces the minimum amount of antibiotics needed to confer susceptibility to CDI. For both examples, it is important to quantify the microbial mediation effect and to identify the bacterial taxa that mediate the treatment effects.

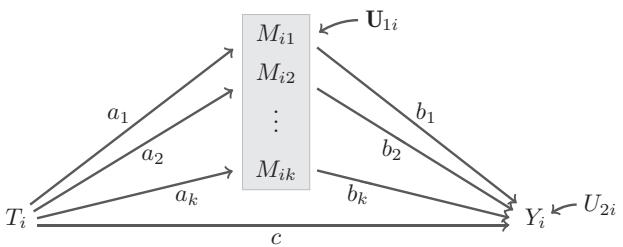


Figure 35.4 A compositional mediation model to link treatment T_i to the continuous outcome Y_i , where a_j, b_j and c are path coefficients, $j = 1, \dots, k$; U_{1i} and U_{2i} are disturbance variables for k compositional mediators $\mathbf{M}_i = (M_{i1}, \dots, M_{ik})$ with $\sum_{j=1}^k M_{ij} = 1$ and an outcome Y_i , respectively.

Methods for microbiome mediation analysis are limited. Different from most of the published work on mediation analysis methods, microbiome data are compositional and high-dimensional. If we assume that the treatment affects the microbiome composition as a whole, the problem is then closer to univariate mediation analysis than the mediation analysis with multiple mediators, where exposure–mediator interactions and, to a certain extent, mediator–mediator interactions need to be considered (VanderWeele and Vansteelandt, 2014).

Suppose we have a random sample of size n from a population where for each unit i we observe Y_i , T_i , and \mathbf{M}_i , where Y_i is the response for the i th individual and T_i is the treatment assignment. For simplicity of notation, we assume $\mathbf{X}_i = X_i$ is a one-dimensional covariate in the model formulation. The composition vector $\mathbf{M}_i \in S^{k-1}$ for all i , that is, $\mathbf{M}_i = \{(M_{i1}, \dots, M_{ik}) : M_{ij} > 0, j = 1, \dots, k, \sum_{j=1}^k M_{ij} = 1\}$, is the microbiome mediator. Figure 35.4 illustrates the effect of T_i on Y_i mediated through \mathbf{M}_i .

The compositional mediation model introduced by Sohn and Li (2017) is based on use of the two compositional operators as in Aitchison (1982). For two compositions $\boldsymbol{\eta}, \boldsymbol{\zeta} \in S^{k-1}$, the perturbation operator is defined by

$$\boldsymbol{\eta} \oplus \boldsymbol{\zeta} = \left(\frac{\eta_1 \zeta_1}{\sum_{j=1}^k \eta_j \zeta_j}, \frac{\eta_2 \zeta_2}{\sum_{j=1}^k \eta_j \zeta_j}, \dots, \frac{\eta_k \zeta_k}{\sum_{j=1}^k \eta_j \zeta_j} \right)^T,$$

and the power transformation for a composition $\boldsymbol{\eta}$ by a scalar α by

$$\boldsymbol{\eta}^\alpha = \left(\frac{\eta_1^\alpha}{\sum_{j=1}^k \eta_j^\alpha}, \frac{\eta_2^\alpha}{\sum_{j=1}^k \eta_j^\alpha}, \dots, \frac{\eta_k^\alpha}{\sum_{j=1}^k \eta_j^\alpha} \right)^T.$$

With these operators, Sohn and Li (2017) proposed the following compositional mediation model:

$$\mathbf{M}_i = (\mathbf{m}_0 \oplus \mathbf{a}^{T_i} \oplus \mathbf{h}^{X_i}) \oplus \mathbf{U}_{1i}, \quad (35.6)$$

$$Y_i = c_0 + c T_i + (\log \mathbf{M}_i)^T \mathbf{b} + g X_i + U_{2i}, \quad \text{subject to } \mathbf{b}^T \mathbf{1}_k = 0, \quad (35.7)$$

where \mathbf{m}_0 is the baseline composition (i.e. when $T_i = \mathbb{E}(T_i)$); similarly, c_0 is the baseline for Y_i ; \mathbf{a}, \mathbf{b} and c are path coefficients; \mathbf{h} and g are nuisance coefficients corresponding to the covariate X ; $\mathbf{1}_k$ is a k -vector of 1s. The distribution of the disturbance variable \mathbf{U}_{1i} need not be specified, but the disturbance U_{2i} is assumed to follow an $N(0, \sigma^2)$ distribution. Model (35.6) describes how a treatment perturbs a composition from the baseline composition as measured by the composition parameter \mathbf{a} . With the compositional operators, all the calculations are within the simplex space and the parameter \mathbf{a} is directly interpretable as a composition. For

those individuals with $T_i = 1$, the new composition is the baseline composition (\mathbf{m}_0) perturbed by \mathbf{a} .

Sohn and Li (2017) showed that the causal total indirect effect $\delta(t)$ for the compositional mediation model is identifiable and given by

$$\begin{aligned}\zeta(\tau) &\equiv \mathbb{E}[Y_i(t, \log \mathbf{M}_i(\tau)) - Y_i(t_0, \log \mathbf{M}_i(\tau))] \\ &= c(t - t_0),\end{aligned}\quad (35.8)$$

$$\begin{aligned}\delta(\tau) &\equiv \mathbb{E}[Y_i(\tau, \log \mathbf{M}_i(t)) - Y_i(\tau, \log \mathbf{M}_i(t_0))] \\ &= (\log \mathbf{a})^T \mathbf{b}(t - t_0),\end{aligned}\quad (35.9)$$

where t is the observed treatment for unit i , t_0 a reference value for the treatment and $\tau = t$ or t_0 .

Statistical inference of model (35.7), including estimation of the mediation effects and the confidence intervals, is performed by a combination of methods for compositional data and high-dimensional inferences. In addition, a bootstrap method was developed to obtain the confidence intervals of the individual mediation effect of each taxon considered (Sohn and Li, 2017). Due to the high dimensionality of the mediators, simulations indicate that a large sample size is required in order to have adequate power to identify the relevant microbial mediators. Since microbiomes are usually measured over time to assess treatment effects, extension of such models to longitudinal data is needed.

35.6 Integrative Analysis of Microbiome, Small Molecules and Metabolomics Data

Besides metagenomic sequencing, technological advances in RNA-sequencing provide us with an ability to gain insight into the genes that are actively expressed in complex bacterial communities (Bashiardes *et al.*, 2016). In addition, the gut microbiome is responsible for bile acid biotransformation and production of important metabolites such as short-chain fatty acid. Mass spectrometry has become a powerful tool for metabolomics studies due to its wide dynamic range and its ability to analyze samples of significant molecular complexity (Wikoff *et al.*, 2009). To better understand the function of the microbiome, research interest has focused on combining metagenomics, metatranscriptomics, and metabolomics data in order to address the physiological role of the human microbiome in health and disease in relation to the end products of primary metabolism and secondary metabolites (Magnúsdóttir *et al.*, 2017; Sharon *et al.*, 2014). It has become common that microbiome studies collect both microbiome and metabolomics data. Methods for integrative analysis of such data are still being developed in order to understand how the microbiome contributes to host metabolism and how metabolites change microbial community dynamics. Some simple correlation analysis has been applied to associate microbiome and metabolites. For example, Ni *et al.* (2017) identified an association between disease severity, gut dysbiosis, and bacterial production of free amino acids by calculating correlations between the taxonomic abundances and fecal metabolites and by identifying differentially abundant taxa and differentially expressed metabolites.

35.6.1 Computational Analysis of Small Molecules from the Human Microbiota

The microbes living in the human gut produce thousands of small molecules with low molecular weight, including lipids, monosaccharides and secondary metabolites. Concentrations of these small molecules can vary dramatically from person to person, which can lead to

differences in disease risks and treatment outcomes. It is therefore important to understand what small molecules a given microbe species can produce. The thousands of prokaryotic genomes in sequence databases provide an opportunity to identify the biosynthetic gene clusters (BGCs): physically clustered groups of two or more genes along the genome that encode the biosynthetic enzymes for a natural product pathway (Cimermancic *et al.*, 2014). These BGCs have been discovered for hundreds of bacterial metabolites, although our knowledge of their diversity remains limited. The minimum information about a biosynthetic gene cluster data standard has recently been proposed (Medema *et al.*, 2015) and will facilitate the analysis of the BGCs.

Cimermancic *et al.* (2014) developed a hidden Markov model-based algorithm, ClusterFinder, that aims to identify gene clusters of both known and unknown biosynthetic classes. ClusterFinder is based on a training set of 732 BGCs with known small molecule products compiled and manually curated. To scan a genome for BGCs, it converts a nucleotide sequence into a string of contiguous Pfam domains using Glimmer (Salzberg *et al.*, 1999) and HMMER (Finn *et al.*, 2015), where Glimmer uses interpolated Markov models to identify the coding regions and distinguish them from noncoding DNA, and HMMER is used to search a sequence database for homologs of a protein family of interest. A hidden Markov model is then fitted to assign each domain a probability of being part of a gene cluster, based on the frequencies at which these domains occur in the BGC and non-BGC training sets. These posterior probabilities are then used to identify the neighboring domains.

ClusterFinder detected more than 14,000 BGCs with an average of six gene clusters per genome based on 2430 reference genomes of the human microbiota for a broad range of small molecule classes, including saccharides, non-ribosomal peptides (NRPS), polyketides (PKS), ribosomally encoded and posttranslationally modified peptides (RiPPPs), and NRPS-independent siderophores (Donia *et al.*, 2014). As an example, Figure 35.5 shows four of the 3118 BGCs identified by Donia *et al.* (2014) in the genomes of human-associated bacteria in the 752 metagenomic samples from the NIH Human Microbiome Project.

In order to quantify the abundance of the BGCs in a given metagenomic sample, one can apply the mBLAST package (Davis *et al.*, 2015) to align the metagenomic sequencing reads to the genes in the BGCs and take the average abundance over the genes as the abundances of the BGCs (Donia *et al.*, 2014). It would be interesting to associate the abundances of these BGCs with metabolomics and clinical phenotypes.

35.6.2 Metabolic Modeling in Microbiome

Bacterial genome sequences provide important functional information about metabolism. Besides identifying the BGCs, genome-scale metabolic models (GEMs), mathematical

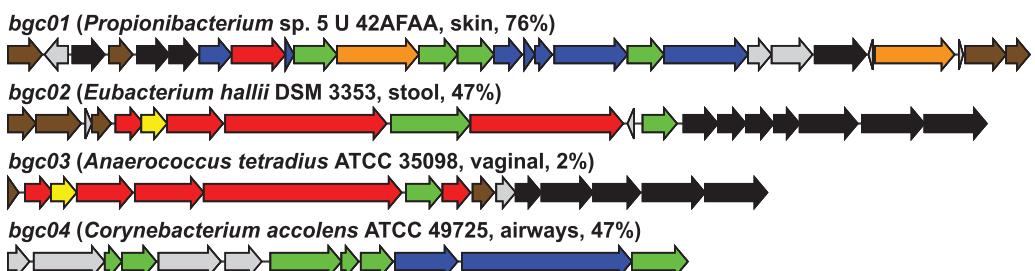


Figure 35.5 Examples of the BGCs identified in Donia *et al.* (2014), where colors represent BGC types (PKS, blue; NRPS, red; RiPPPs, orange; terpenes, pink; NI siderophores, tan; saccharides, purple – for detailed descriptions, see Donia *et al.*, 2014), and arrows represent protein-coding genes.

representations of the knowledge on an organism's metabolic capacity, provide another way of studying the bacterial systems and their functions (Santos *et al.*, 2011). These models have been previously applied in bacterial systems for phenotypic characterizations, metabolic engineering, drug discovery and for studying interspecies interactions through metabolic exchanges (Oberhardt *et al.*, 2009; Feist and Palsson, 2008; van der Ark *et al.*, 2017).

Such a GEM is often constructed at the strain level of the species. GEM construction can be performed semi-automatically using the genome annotation of the microbe of interest since this predicts the enzymes a microbe encodes. This list of enzymes, together with known metabolic reactions in various databases, provides a list of possible chemical reactions the microbe can perform. However, the known metabolic/chemical reactions are far from complete, including missing reactions due to incorrect, missing, or low-quality annotations. These missing reactions or gaps lead to parts of the metabolic network being not connected, which makes the analysis of metabolic networks difficult. Certain gap-filling algorithms can be used to predict the presence of additional reactions from reaction databases such as KEGG or Metacyc and to connect disconnected parts of the network (Thiele *et al.*, 2014; van der Ark *et al.*, 2017).

As a useful resource, Magnúsdóttir *et al.* (2017) generated genome-scale metabolic reconstructions of 773 semi-automatically curated GEMs of gut microbes. An interesting question is how to integrate these GEMs with metagenomic sequencing data sets and metabolomics data sets in order to infer the metabolic diversity of microbial communities and to study the role that particular microbes play in disease etiology or treatment outcome. By aligning the metagenomic reads to the genomes of the strains with known GEMs, one can quantify the metabolic diversity of GEM reactions for a given sample.

35.7 Discussion and Future Directions

The roles played by the microbiome in human health and diseases have been intensively investigated in recent years. The future will see microbiome-wide association studies (MWAS) to link dynamic microbial consortia to diseases in large population cohorts (Gilbert *et al.*, 2016). While MWAS have some similarity to genome-wide association studies (GWAS), they have their own features and challenges, including unobserved confounding due to population stratification, temporal variability of microbiome and many uncharacterized microorganisms. Such associations can be at the individual taxon level, at the whole community level or at sub-community levels. Incorporation of biological knowledge and metabolic information into such analyses will provide important insights into how the microbiome affects disease onset and progression. Details on bacterial GWAS are described in **Chapter 36**.

The future will also see single-cell genomics of microbial cells, which promises to provide more accurate microbial community quantification and overcome the limitations caused by the uncertainty in the grouping of sequence reads according to the strain of origin (Blainey, 2013; Lasken, 2012; Tolonen and Xavier, 2017). Similar to analysis of single-cell RNA-sequencing data (**Chapter 26**), new computational methods for dealing with such single-cell genomics in microbiome studies are much needed.

Acknowledgements

This research was supported by NIH grants CA127334 and GM123056.

References

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B* **44**(2), 139–177.
- Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–330.
- Bashiardes, S., Zilberman-Schapira, G. and Elinav, E. (2016). Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights* **10**, 19–25.
- Bates, S. and Tibshirani, R. (2017). Log-ratio Lasso: Scalable, sparse estimation for log-ratio models. Preprint, arXiv:1709.01139v1.
- Blainey, P. (2013). The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews* **37**, 407–427.
- Bonder, M.J., Kurilshikov, A., Tigchelaar, E.F., Mujagic, Z., Imhann, F., Vila, A.V., Deelen, P., Vatanen, T., Schirmer, M., Smekens, S.P., Zhernakova, D.V., Jankipersadsing, S.A., Jaeger, M., Oosting, M., Cenit, M.C., Masclee, A.A.M., Swertz, M.A., Li, Y., Kumar, V., Joosten, L., Harmsen, H., Weersma, R.K., Franke, L., Hofker, M.H., Xavier, R.J., Jonkers, D., Netea, M.G., Wijmenga, C., Fu, J. and Zhernakova, A. (2016). The effect of host genetics on the gut microbiome. *Nature Genetics* **48**, 1407–1412.
- Bray, J.R. and Curtis, J.T. (1957). An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* **27**, 325–349.
- Brown, C.T., Olm, M.R., Thomas, B.C. and Banfield, J.F. (2016). Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology* **34**, 1256–1263.
- Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P. (2016). Dada2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**, 581–583.
- Cao, Y., Lin, W. and Li, H. (2018). Two-sample tests of high dimensional means for compositional data. *Biometrika* **105**, 115–132.
- Chao, A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 265–270.
- Chen, E., Bushman, F. and Li, H. (2017). A model-based approach for species abundance quantification based on shotgun metagenomic data. *Statistics in Biosciences* **9**, 13–27.
- Chen, E.Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32**(17), 2611–2617.
- Chen, J., Bittinger, K., Charlson, E., Hoffmann, C., Lewis, J., Wu, G., Collman, R., Bushman, F. and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**(16), 2106–2113.
- Chen, J. and Li, H. (2013a). Kernel methods for regression analysis of microbiome compositional data. In Y. Liu, M. Hu, and J. Lin (eds), *Topics in Applied Statistics: 2012 ICSA Applied Statistics Symposium Proceedings*. Springer, New York, pp. 191–201.
- Chen, J. and Li, H. (2013b). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Annals of Applied Statistics* **7**, 418–442.
- Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., Birren, B.W., Takano, E., Sali, A., Linington, R.G. and Fischbach, M.A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421.
- Davis, C., Kota, K., Baldhandapani, V., Gong, W., Abubucker, S., Becker, E., Martin, J., Wylie, K.M., Khetani, R., Hudson, M.E., Weinstock, G.M. and Mitreva, M. (2015). mBLAST: Keeping up with the sequencing explosion for (meta)genome analysis. *Journal of Data Mining in Genomics & Proteomics* **4**, 135.

- DeJong, T. (1975). A comparison of three diversity indices based on their components of richness and evenness. *Oikos* **26**, 222–227.
- Donia, M., Cimermancic, P., Schulze, C., Laura, C., Wieland, B., Martin, J., Mitreva, M., Clardy, J., Linington, R.G. and Fischbach, M. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414.
- Feist, A. and Palsson, B. (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature Biotechnology* **26**, 659–667.
- Finn, R., Clements, J., Arndt, W., Miller, B., Wheeler, T., Schreiber, F., Bateman, A. and Eddy, S. (2015). HMMER web server: 2015 update. *Nucleic Acids Research* **43**, W30–W38.
- Frank, J. and Sorensen, S. (2011). Quantitative metagenomic analyses based on average genome size normalization. *Applied and Environmental Microbiology* **77**, 2513–2521.
- Gilbert, J.A., Quinn, R.A., Debelius, J., Xu, Z.Z., Morton, J., Garg, N., Jansson, J.K., Dorrestein, P.C. and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103.
- Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., Prieto, P.A., Vicente, D., Hoffman, K., Wei, S.C., Cogdill, A.P., Zhao, L., Hudgens, C.W., Hutchinson, D.S., Manzo, T., Petaccia de Macedo, M., Cotechini, T., Kumar, T., Chen, W.S., Reddy, S.M., Sloane, R.S., Galloway-Pena, J., Jiang, H., Chen, P.L., Shpall, E.J., Rezvani, K., Alousi, A.M., Chemaly, R.F., Shelburne, S., Vence, L. M., Okhuyzen, P.C., Jensen, V.B., Swennes, A.G., McAllister, F., Sanchez, E.M.R., Zhang, Y., Le Chatelier, E., Zitzvogel, L., Pons, N., Austin-Breneman, J.L., Haydu, L.E., Burton, E.M., Gardner, J.M., Sirmans, E., Hu, J., Lazar, A.J., Tsujikawa, T., Diab, A., Tawbi, H., Glitza, I.C., Hwu, W.J., Patel, S.P., Woodman, S.E., Amaria, R.N., Davies, M.A., Gershenwald, J.E., Hwu, P., Lee, J.E., Zhang, J., Coussens, L.M., Cooper, Z.A., Futreal, P.A., Daniel, C.R., Ajami, N.J., Petrosino, J.F., Tetzlaff, M.T., Sharma, P., Allison, J.P., Jenq, R.R. and Wargo, J.A. (2017). Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**(6371), 97–103.
- Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214.
- Koeth, R., Wang, Z., Levison, B., Buffa, J., E., O., et al. (2013). Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature Medicine* **19**, 576–585.
- Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E., Jona, G., Harmelin, A., Cohen, N., Sirota-Madi, A., Thaiss, C., Pevsner-Fischer, M., Sorek, R., Xavier, R., Elinav, E. and Segal, E. (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106.
- La Rosa, P., Brooks, J., Deych, E., Boone, E., Edwards, D., et al. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* **7**, e52078.
- Lasken, R. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology* **10**, 631–640.
- Lewis, J.D., Chen, E.Z., Baldassano, R.N., Otley, A.R., Griffiths, A.M., Lee, D., Bittinger, K., Bailey, A., Friedman, E.S., Hoffmann, C., Albenberg, L., Sinha, R., Compher, C., Gilroy, E., Nessel, L., Grant, A., Chehoud, C., Li, H., Wu7, G.D. and Bushman, F.D. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host & Microbe* **18**, 489–500.
- Li, K., Bihan, M., Yooseph, S. and Methé, B. (2012). Analyses of the microbial diversity across the human microbiome. *PLoS ONE* **7**(6), e32118.
- Lin, W., Shi, P., Feng, R. and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–797.

- Lozupone, C., Hamady, M., Kelley, S. and Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**(5), 1576–1585.
- Lozupone, C. and Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**(12), 8228–8835.
- Lu, J., Shi, P. and Li, H. (2018). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* doi: 10.1111/biom.12956.
- Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R.M.T. and Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology* **35**, 81–89.
- McCoy, C. and Matsen, F. (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ* **1**, e157.
- Medema, M.H., Kottmann, R., Yilmaz, P., et al. (2015). Minimum information about a biosynthetic gene cluster. *Nature Chemical Biology* **11**, 625–631.
- Mende, D., Sunagawa, S., Zeller, G. and Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nature Methods* **10**, 881–884.
- Nayfach, S. and Pollard, K.S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology* **16**(1), 51.
- Ni, J., Shen, T.-C.D., Chen, E.Z., Bittinger, K., Bailey, A., Roggiani, M., Sirota-Madi, A., Friedman, E.S., Chau, L., Lin, A., Nissim, I., Scott, J., Lauder, A., Hoffmann, C., Rivas, G., Albenberg, L., Baldassano, R.N., Braun, J., Xavier, R.J., Clish, C.B., Yudkoff, M., Li, H., Goulian, M., Bushman, F.D., Lewis, J.D. and Wu, G.D. (2017). A role for bacterial urease in gut dysbiosis and Crohn's disease. *Science Translational Medicine* **9**(416), eaah6888.
- Oberhardt, M., Palsson, B. and Papin, J. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology* **5**, 320.
- O'Brien, J., Record, N. and Countway, P. (2016). The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. Preprint, bioRxiv 045468.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Consortium, M., Bork, P., Ehrlich, S.D. and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–70.
- Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* **35**, 833–844.
- Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C. and Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biology* **8**(1), R10.
- Routy, B., Le Chatelier, E., Derosa, L., Duong, C.P.M., Alou, M.T., Daillère, R., Fluckiger, A., Messaoudene, M., Rauber, C., Roberti, M.P., Fidelle, M., Flament, C., Poirier-Colame, V., Opolon, P., Klein, C., Iribarren, K., Mondragón, L., Jacquemet, N., Qu, B., Ferrere, G., Clémenson, C., Mezquita, L., Masip, J.R., Naltet, C., Brosseau, S., Kaderbhai, C., Richard, C., Rizvi, H., Levenez, F., Galleron, N., Quinquis, B., Pons, N., Ryffel, B., Minard-Colin, V., Gonin, P., Soria, J.-C., Deutsch, E., Loriot, Y., Ghiringhelli, F., Zalcman, G., Goldwasser, F., Escudier, B., Hellmann, M.D., Eggermont, A., Raoult, D., Albiges, L., Kroemer, G. and Zitvogel, L. (2018). Gut

- microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**(6371), 91–97.
- Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. *Genomics* **59**, 24–31.
- Santos, F., Boele, J. and Teusink, B. (2011). A practical guide to genome-scale metabolic models and their analysis. In D. Jameson, M. Verma, and H.V. Westerhoff, (eds), *Methods in Systems Biology, Methods in Enzymology*. Vol. **500**, Academic Press, San Diego, CA, pp. 509–532.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods* **9**(8), 811–814.
- Sharon, G., Garg, N., Debelius, J., Knight, R., Dorrestein, P.C. and Mazmanian, S.K. (2014). Specialized metabolites from the microbiome in health and disease. *Cell Metabolism* **20**(5), 719–730.
- Shi, P. and Li, H. (2017). A model for paired-multinomial data and its application to analysis of data on a taxonomic tree. *Biometrics* **73**(4), 1266–1278.
- Shi, P., Zhang, A. and Li, H. (2016). Regression analysis for microbiome compositional data. *Annals of Applied Statistics* **10**(2), 1019–1040.
- Sohn, M. and Li, H. (2017). Compositional mediation analysis for microbiome studies. Preprint, bioRxiv 149419.
- Stammler, F., Glasner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P., Gessner, A. and Spang, R. (2017). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, 28.
- Sunagawa, S., Mende, D., Zeller, G., Izquierdo-Carrasco, F., Berger, S., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* **10**, 1196–1199.
- Tang, Z., Chen, G., Alekseyenko, A. and Li, H. (2017). A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics* **33**, 1278–1285.
- Thiele, I., Vlassis, N. and Fleming, R. (2014). fastGapFill: Efficient gap filling in metabolic networks. *Bioinformatics* **30**, 2529–2531.
- Tolonen, A.C. and Xavier, R.J. (2017). Dissecting the human microbiome with single-cell genomics. *Genome Medicine* **9**(1), 56.
- Tringe, S. and Rubin, E. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* **6**(11), 805–814.
- Turnbaugh, P., Hamady, M., Yatsunenko, T., Cantarel, B., Ley, R., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**(7228), 480–484.
- van der Ark, K.C.H., van Heck, R.G.A., Martins Dos Santos, V.A.P., Belzer, C. and de Vos, W.M. (2017). More than just a gut feeling: Constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes. *Microbiome* **5**(1), 78.
- Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G. and Raes, J. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511.
- VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods* **2**, 95–115.
- Wikoff, W.R., Anfora, A.T., Liu, J., Schultz, P.G., Lesley, S.A., Peters, E.C. and Siuzdak, G. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences of the United States of America* **106**(10), 3698–3703.
- Wood, D.E. and Salzberg, S.L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46.

- Wu, G., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y., *et al.* (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**(6052), 105–108.
- Wu, H., Esteve, E., Tremaroli, V., Khan, M.T., Caesar, R., Mannerøas-Holm, L., Støahlman, M., Olsson, L.M., Serino, M., Planas-Fèlix, M., Xifra, G., Mercader, J.M., Torrents, D., Burcelin, R., Ricart, W., Perkins, R., Fernàndez-Real, J. and Bäckhed, F. (2017). Metformin alters the gut microbiome of individuals with treatment-naïve type 2 diabetes, contributing to the therapeutic effects of the drug. *Nature Medicine* **23**, 850–858.
- Xia, F., Chen, J., Fung, W. and Li, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69**, 121–139.
- Zackular, J.P., Moore, J.L., Jordan, A.T., Juttukonda, L.J., Noto, M.J., Nicholson, M.R., Crews, J.D., Semler, M.W., Zhang, Y., Ware, L.B., Washington, M.K., Chazin, W.J., Caprioli, R.M. and Skaar, E.P. (2016). Dietary zinc alters the microbiota and decreases resistance to *Clostridium difficile* infection. *Nature Medicine* **22**, 1330–1334.

36

Bacterial Population Genomics

Jukka Corander,^{1,2} Nicholas J. Croucher,³ Simon R. Harris,⁴ John A. Lees,⁵ and Gerry Tonkin-Hill⁴

¹Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

²Department of Biostatistics, University of Oslo, Oslo, Norway, and Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

³Department of Infectious Disease Epidemiology, Imperial College London, London, UK

⁴Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

⁵Department of Microbiology, School of Medicine, New York University, New York, USA

Abstract

We provide an in-depth review of the popular and emerging statistical methods for bacterial population genomics, covering the major aspects relevant from a population-based perspective, including population structure, phylogenetic trees, recombination analysis, transmission modeling, genome-wide association analysis and genome-wide epistasis analysis. We describe challenges specific to the analysis of bacterial genomes, including accounting for the pangenome, clonality and horizontal gene transfer as well as interactions between bacteria and their host or environment. The chapter focuses on the computational scalability of methods to cope with the increasing affordability of sequencing.

36.1 Introduction

Statistics has never played a more central role in the study of bacteria than it does today, a situation that is unlikely to change soon, with further technological advances expected to bring more complex data. Here we provide an in-depth review of the popular and emerging statistical methods for bacterial population genomics, covering the major aspects relevant from the population-based perspective, including population structure, phylogenetic trees, recombination analysis, transmission modeling, genome-wide association studies (GWASs) and genome-wide epistasis analysis. As sequencing technology has continued to advance over the past decade, the new possibilities offered by affordable sequencing have required the developers of statistical methods to focus on the computational scalability of their solutions.

Bacteria have been intensively studied and manipulated ever since van Leeuwenhoek's discovery of single-celled micro-organisms during the seventeenth century. However, it was not until 2010 that it became possible to study bacterial evolution and diversity in detail across whole genomes, with the arrival of high-throughput, short-read next-generation sequencing (NGS) techniques (Harris, 2010). This pioneering study rapidly gave rise to many further investigations into genome-wide variation in pathogenic bacterial populations. Hand-in-hand with

this explosion of data came new biological questions requiring the development of new statistical tools.

Bacterial genomes typically consist of a small number of circular replicons, comprising 10^5 to 10^7 base pairs in total. Essential replicons are referred to as chromosomes; most species have a single chromosome, although some pathogen species (e.g. *Vibrio cholerae* and *Burkholderia cenocepacia*) have two. In terms of heredity, the vast majority of bacteria are haploid, although cells will often contain multiple copies of the chromosome. Non-essential replicons are referred to as plasmids. Bacteria may harbor multiple plasmids, each of which can be present at varying copy number in the cell. Examples of linear bacterial chromosomes and plasmids are known. The set of genes that are present in all individuals of a collection of bacterial isolates is known as the core genome, while those genes that are present in only a subset of the collection belong to the accessory genome. The pan-genome refers to the combined set of all non-orthologous genes present in a defined taxonomic group; this is different to the meta-genome, which refers to the genes found in an environment, regardless of the relatedness between the individuals from which they originated. As well as varying in their gene content, bacteria also exhibit diversity as a consequence of both point mutations and rearrangement, similar to eukaryotes. However, unlike sexually reproducing organisms, bacteria do not exchange genetic material through recombination at every generation. Some bacteria are almost entirely clonal, whereas others experience acquisitions of new genetic material through horizontal gene transfer.

36.2 Genetic Population Structure and Clustering of Genotypes

36.2.1 Background

Genetic population structure has a long history, dating back to the works of Ronald Fisher and Sewall Wright in the 1920s and 1930s. While various indices of the level of population structuring, such as the F -statistics, were a popular concept for decades, the most practically successful ways of studying population structure often involve statistical clustering models, or computer simulations depicting the relatedness of individuals in gene pools backward or forward over time across generations. These methods are often computationally expensive and have led to the development of fast and easy-to-apply approaches based on pairwise distance metrics, but these are heavily dependent on both the distance measure and graphical representation chosen for the clustering. Likelihood-based methods resolve these issues and have the advantage of being interpretable in the context of a genetic model. Consequently, likelihood-based methods are the most popular tools for genetic mixture and admixture analyses, the former corresponding to identification of subpopulations (i.e. statistical clustering of individuals on the basis of observed genomic variation), and the latter extending the clustering to the level of individual alleles such that their ancestral origin is identified. Most of the methods frequently used for mixture and admixture analyses of bacterial population data, such as ADMIXTURE, BAPS, fineSTRUCTURE, hierBAPS and STRUCTURE, have their origin in the study of eukaryotic organisms (see Chapter 8) and were adopted for bacterial genomics when NGS finally enabled large-scale studies. The above list of software packages is by no means exhaustive; there are more than 10 other likelihood-based and Bayesian methods for genetic population structure analysis that have been used in molecular ecology and human genetics. For a review of some of these methods, see Excoffier and Heckel (2006).

36.2.2 Model-Based Clustering

A genetic mixture model describes a population in terms of a finite number of subpopulations, each with their own set of allele frequencies at the polymorphic sites. They are similar to

statistical clustering models encountered in other domains such as text document analysis by topic models (Blei *et al.*, 2003). Each individual is either assumed to originate entirely from one subpopulation (the ‘no admixture model’) or it inherits its genome from different subpopulations (the ‘admixture model’). In the ‘no-admixture model’ individuals are assumed to be drawn from K subpopulations, each with a different set of allele frequencies across L loci. Given a set of genotyped individuals X , the aim is then to identify the underlying subpopulation they originate from and the frequencies of alleles in those sub-populations. That is, one aims to find the vectors \mathbf{z} and \mathbf{p} where z_i indicates which of the K populations sample i belongs to and p_{klj} indicates the frequency of allele j at locus l in population k . The ‘admixture’ model generalizes to allow for individuals to inherit their genomes from more than one sub-population. The fraction q_k of an individual’s genome that comes from population k must then be estimated with the restriction that all loci must be assigned as originating from one subpopulation ($\sum_k q_k = 1$). The actual likelihood of the sampled haplotypes, or alternatively their marginal data distribution under the Bayesian approach, can be formulated in several different ways, using for example a latent finite mixture model (Pritchard *et al.*, 2000), a product-partition model (Corander and Marttinen, 2006) or a Dirichlet process partition model (Huelskenbeck and Andolfatto, 2007).

The Structure approach of (Pritchard *et al.*, 2000) uses a Markov chain Monte Carlo (MCMC) algorithm based on Gibbs sampling to estimate both the population frequencies of alleles and the allocation of individuals to populations. One of the first applications of the algorithm to bacterial data sets investigated the population structure of *Helicobacter pylori*, taking advantage of its higher mutation rate to investigate human migrations (Falush *et al.*, 2003b). Difficulties in using the STRUCTURE algorithm include the symmetry present in the model: two valid solutions can give the same likelihood, leading to issues with non-identifiability. This can lead further to mixing issues, with the sampler often failing to explore different symmetric modes, and can lead to problems when combining runs with different starting points (Rosenberg *et al.*, 2002). Methods have been developed to post-process the results of such runs, including CLUMPAK which uses a Markov clustering algorithm to group together similar modes and match up the cluster labels from different runs (Kopelman *et al.*, 2015). The initial choice of the number of clusters K is also a challenging problem. A number of methods have been described to choose K , with most involving running the algorithm for different values of K and choosing based on model selection criteria such as the Akaike information criterion.

The BAPS algorithm (Corander *et al.*, 2003; 2004) takes a different approach and focuses on estimating the partition among individuals, directly and analytically integrating over the allele frequency parameters for each sub-population. This has the advantage of providing a more natural method for selecting the number of clusters K . In addition, the hierBAPS extension of the method allows for the exploration of clusters at different resolutions, which is useful when there is more than one valid clustering in a data set. By integrating out the sub-population frequency parameters, the approach also reduces inaccuracies introduced in the simulation of many parameters. To cope with the increasing size of data sets, a later version of BAPS swapped to using a greedy stochastic search method to find the clustering that locally maximizes the marginal likelihood. This avoided the computationally expensive MCMC stage of previous methods (Corander *et al.*, 2006) and allowed the algorithm to be applied to data sets comprised of thousands of whole-genome sequences (WGSs) such as the analysis of population structure in over 3000 WGSs of *Streptococcus pneumoniae* from a refugee camp (Chewapreecha *et al.*, 2014). There is an optional subsequent step to investigate admixed populations (Corander and Marttinen, 2006).

The importance of computationally efficient methods, such as the stochastic search method in BAPS, is growing as pathogen sequence data sets start to number in the tens of thousands of isolates. hierBAPS (Cheng *et al.*, 2013) iteratively uses the BAPS clustering approach on each inferred cluster to allow for investigation of a data set at multiple resolutions. This increases

the statistical power to detect separate sub-lineages. A version of hierBAPS using the R language (R Core Team, 2013) allows for ease of use and plotting of results (<https://github.com/gtonkinhill/rhierbaps/>).

Another program that addresses similar performance problems to hierBAPS is a maximum likelihood algorithm, snapclust (Beugin *et al.*, 2018). Like hierBAPS, snapclust uses a geometric clustering approach to provide an initial clustering before using an iterative search operation to locally maximize the likelihood. However, unlike hierBAPS, snapclust requires the number of latent clusters K to be specified beforehand and makes use of an EM algorithm to optimize the likelihood rather than using a greedy stochastic search.

36.2.3 Linkage Disequilibrium

An important consideration when investigating genetic admixture is the impact of linkage disequilibrium (LD). LD refers to the dependence of alleles at different locations on the same genome (Chapter 2). It is usually the result of genetic recombination having less of an impact on loci that are in close proximity. The impact of LD on clustering can vary depending on the rate of recombination in the organism and the time-scale over which the samples have been taken. For very clonal organisms such as *Mycobacterium tuberculosis*, nearly all the sites in the genome will be in high LD. In this case a phylogenetic tree-building approach may be more suitable, such as maximum parsimony (Camin and Sokal, 1965), maximum likelihood (Stamatakis, 2014) or Bayesian methods (Drummond and Rambaut, 2007). Some bacteria, including *S. pneumoniae* and *Neisseria gonorrhoeae*, have much higher rates of recombination, and consequently constructing accurate phylogenies can be more difficult. The impact of LD on these more recombinant organisms is highly dependent on the time interval over which the samples were taken. If there are considerable gaps between sample dates there may be enough recombination for the impact of LD to be safely ignored, in which case methods such as Structure (Pritchard *et al.*, 2000) and BAPS (Corander and Marttinen, 2006) can be used. For more densely sampled isolates clustering methods that account for recombination such as Structure version 2 (Falush *et al.*, 2003a; Tang *et al.*, 2006) and fineSTRUCTURE (Lawson *et al.*, 2012) are more suitable.

The Structure (v2) model of Falush *et al.* (2003a) adapts the original unlinked model to allow for ‘chunks’ of linked markers to be derived from the estimated ancestral populations. That is, it allows for dependence between loci through the use of a Markov chain linking adjacent loci. It is suitable for data where the loci are linked, but not so closely that recent recombinations are likely to be evident. A recent adaptation of this model makes use of hierarchical Dirichlet processes to infer the number of clusters and their composition simultaneously (De Iorio *et al.*, 2015). For very densely sampled populations where recent recombination is present, approaches such as BratNextGen (Marttinen *et al.*, 2012) and fineSTRUCTURE (Lawson *et al.*, 2012) are more powerful. FineSTRUCTURE models each sample genome as the recipient of a series of chunks of DNA from other ‘donor’ genomes in the sample. It makes use of another hidden Markov Model introduced by Li and Stephens (2003) to generate a coancestry matrix between every genome in the data set. Both of these approaches reduce to the unlinked case as the recombination rate approaches infinity. A recent paper made use of fineSTRUCTURE to infer the population history of *H. pylori* and consequently its human host from a sample derived from the stomach of the ‘Iceman’, a copper-age glacier mummy (Maixner *et al.*, 2016).

36.2.4 Distance-Based Methods

Pairwise distance-based approaches have gained appeal with the ever-increasing size of genome sequence data sets. The speed of some modern distance-based methods has

allowed for the clustering of otherwise unfeasibly large data sets. A notable example was the clustering of all 54,118 NCBI RefSeq genomes (Ondov *et al.*, 2016). The lack of assumptions on the type of population structure can be a further advantage of distance-based methods. Discriminant analysis of principal components (DAPC; Jombart *et al.*, 2010) takes advantage of the speed of principal component analysis to reduce the dimensionality of large sequencing data sets before using discriminant analysis (Lachenbruch and Goldstein, 1979) to choose a clustering. If priors are known for the clusters, *K*-means clustering can be used in place of discriminant analysis. While DAPC is considerably faster than model-based approaches, it still relies on a genotype matrix that can require considerable time to compute. With large genomes and high sequencing depth the computation required to produce a consensus sequence alignment or call single nucleotide polymorphisms (SNPs) can become prohibitive when many samples are involved. By adapting the MinHash dimensionality reduction technique (Indyk and Motwani, 1998) the Mash algorithm is able to avoid the alignment stage and allows for the rapid calculation of a distance matrix (Ondov *et al.*, 2016). This matrix can then be used to cluster samples using traditional graph clustering methods such as *K*-means.

While distance-based methods are faster, they do not offer the interpretability that an explicit genetic model provides. Methods such as snapclust (Beugn *et al.*, 2018) and hierBAPS (Cheng *et al.*, 2013) provide faster model-based alternatives but currently do not scale as well as distance based approaches. There remains a need for model-based methods that can scale to future bacterial data sets in excess of a hundred thousand genomes (Tonkin-Hill *et al.*, 2018).

36.3 Phylogenetics and Dating Analysis

A phylogenetic tree can be used to represent the evolutionary relationships among bacteria sampled from a population. The shape, or topology, of the tree shows the relationships among bacteria based on their vertical evolution: the tips are observed sequences, nodes are ancestors of the observed sequences, the root (if present) is the most recent common ancestor of the population. The branch lengths give the amount of genetic change between samples, usually in units of nucleotide substitutions per site. Phylogenetic trees can only represent vertical relationships, and horizontal events such as recombination cannot be shown. Sites involved in horizontal events should be removed before constructing a phylogeny, as they cannot be correctly modeled (for an example of this, see Figure 36.1(a)). Convergent evolution, resulting in homoplasy, where the same allele appears and gets selected multiple times at a locus across a population, can also be a challenge for reconstruction. Phylogenetic relationships can be used to inform outbreak tracing, as an additional data source for epidemiology and to date the emergence of sequences (which is discussed in more detail below). The shape of the tree is affected by population dynamics, spatial structure of the population, and the strength of selection (Grenfell *et al.*, 2004; Volz *et al.*, 2013). Phylogenetic inference can therefore be informative of population dynamics, either through summary statistics or by including a model of the population dynamics during the inference.

One of the simplest algorithms for phylogenetic reconstruction is the neighbor joining (NJ) approach which, taking a distance matrix as input, hierarchically clusters samples starting with the lowest distances while iteratively recalculating the distance matrix between clusters and remaining samples (Saitou and Nei, 1987). While fast for small sample numbers, the algorithmic complexity is $O(N^3)$ so it is unsuitable for very large populations. With the use of heuristics to speed up many of the calculations, as implemented in software such as RapidNJ, it has become feasible to apply NJ algorithms to tens of thousands of sequences (Simonsen *et al.*, 2008).

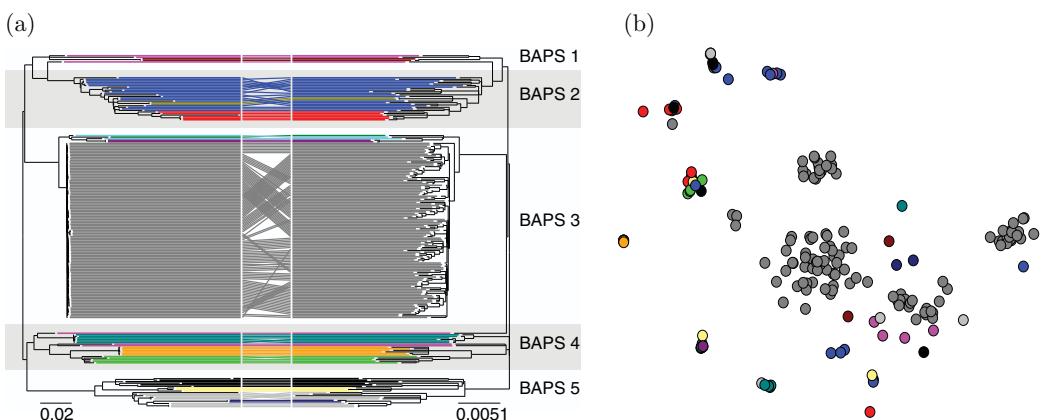


Figure 36.1 Illustration of phylogenomics and gene content analysis. (a) Maximum likelihood trees for the PMEN2 population (Croucher *et al.*, 2014). The right-hand side shows the estimated phylogeny prior to exclusion of recombination events. The left-hand side shows the estimated phylogeny after exclusion of significant recombination events based on Gubbins (Croucher *et al.*, 2015). Branch coloring refers to the secondary-level hierBAPS clusters and the labels next to the phylogeny depict first-level clustering. (b) PANINI network of the same isolates shown in the phylogenetic tree based on their gene content. Each isolate is colored using the same labeling as for the tree branches. For details see Section 36.7.

The NJ algorithm does not model the mechanisms of evolution; more sophisticated methods use aligned sequences as input, and NJ or maximum parsimony to create a starting topology which is then improved. Each site in a multiple sequence alignment is a hypothesis of homology, which is assumed by the phylogenetic model to have the same function in each sample, and to have evolved under the model selected. The alignment can be of nucleotides, amino acids or binary characters (such as a gene presence/absence). In each case a model of relative transition rates between character states is needed, as well as an overall transition rate. With simple models and a sufficient sample of polymorphic sites, these parameters can be inferred from the alignment itself, such as with the general time-reversible (GTR) model for nucleotide evolution (Tavaré, 1986; **Chapter 13**). Otherwise previously fitted rates can be used, for example the LG matrix for amino acids (Le and Gascuel, 2008). Two common modifications are made to these rate assumptions. The first is to allow rate heterogeneity between sites. For example, some positions may have a conserved function and therefore little variation, while other sites may be under diversifying selection. This is normally modeled with a rate that is a discrete approximation of the gamma distribution, with each alignment site assigned to a category, of which there are typically four. The second modification introduces a class of completely invariant sites when using only variable sites in an alignment, which can lead to an ascertainment bias that results in overestimation of branch lengths (Lewis, 2001). Selection of the most appropriate approach can be aided by approaches such as jModelTest (Posada, 2008), which calculate how well these substitution models describe the input alignment. Optimizing the rate parameters and topology efficiently under a given model represents a further challenge, though RAxML (Stamatakis, 2014) and IQ-TREE (Nguyen *et al.*, 2015) are two methods that have been shown to cope with whole-genome alignments of thousands of bacteria. Comparisons of the accuracy of tree topology from different models and software implementations have shown these approaches to be best, though computationally intensive (Lees *et al.*, 2018b).

An evolutionary model and aligned data allows inference of a posterior distribution of phylogenies using MCMC sampling (Nascimento *et al.*, 2017), for example with MrBayes (Ronquist and Huelsenbeck, 2003; **Chapter 6**), or more quickly with maximum likelihood.

Probability intervals, often termed credible intervals, are implicit in the Bayesian approach, while bootstrapping can be used to estimate the confidence in node placement of maximum likelihood trees (Minh *et al.*, 2013).

As well as reconstructing a microbe's evolutionary history, tree structure can be highly informative about its demographic history through the application of phylodynamic methods. There are two distinct approaches to such analyses. The first, coalescent theory (Kingman, 1982; Chapter 5), operates backwards in time. Analyses start with the observed sequences at the tips of the tree and reconstruct their history until they coalesce at their ancestral node. The second, birth–death models (Kendall, 1948), operates forwards in time. Populations diverge from a common ancestor, with nodes representing points at which lineages are formed and branch off from one another. A further important assumption distinguishing these approaches is that coalescent theory assumes that a small fraction of the evolving population is present in the sample, whereas birth–death models explicitly model the proportion of the population being sampled, and how this changes over time. Hence birth–death approaches tend to be more complex and require systematic sampling. Although originally devised based on homochronous sampling, in which all samples are regarded as being isolated contemporaneously, both have been adapted to work on heterochronous samples of measurably evolving populations. This allows the divergence detected between early and late isolates to be exploited to calibrate the substitution model. In turn, this allows a 'timed' phylogeny to be generated, such that the time at which divergences pre-dating the first isolate can be estimated.

By combining a substitution model with a tree model incorporating the times of observations (i.e. sampling dates), the dates of ancestral nodes can be estimated. This can be used to infer the time a species, or one of its lineages, emerged, as well as the effective population size through time (Drummond *et al.*, 2005). The tree model consists of a model of population dynamics affecting tree shape, usually either a birth–death or coalescent model, and a molecular clock model that relates the rate of molecular changes to observed times. A 'strict clock' requires that the substitution rate per unit time is equal across all branches of the tree, which has been observed to fit to sequences from species such as *Acinetobacter baumannii*, whereas a 'relaxed clock' allows the rate to vary across branches, which is observed to give a better fit to most species.

Inference using these combined models is usually performed by MCMC with BEAST (Drummond *et al.*, 2012). A successful analysis requires appropriate priors to be set for each component of the model, and the choice of a model that describes the data well. For the former, understanding the relation between each prior and its predicted phylodynamic outcome, as well as a comparison of the posterior samples from realized chains to the prior distribution, can be used to inform this choice (Bromham *et al.*, 2018). For the latter, after ensuring convergence of MCMC, as well as comparing likelihoods of each model, the relative predictive power of the model posterior can be used to compare the adequacy of different model choices while accounting for uncertainty in the model parameters (Bollback, 2002; Duchêne *et al.*, 2015).

BEAST was designed for analysis of viral dynamics before it was known that the substitution rates in bacterial species were sufficient to be informative for dating. Analysis of whole-genome alignments from bacterial populations is therefore computationally expensive and requires long runs of serial chains to reach convergence. For dating ancestors in large populations, even in cases where the molecular clock (as defined above) is violated, a fast alternative is least-squares dating (To *et al.*, 2016). This uses a linear relationship between time and branch length, assuming a constant substitution rate and Poisson-distributed errors. Dates are given by minimizing the least-squares distance between model expected values and the data, which can be achieved with linear complexity $O(N)$. Nonparametric bootstraps can be used to estimate

confidence intervals on the inferred node dates. Another recent and promising computationally scalable method is TreeTime, which uses maximum likelihood phylodynamics and offers a considerable range of inference options (Sagulenko *et al.*, 2018).

36.4 Transmission Modeling

The tracking of infectious disease through a population has historically been dominated by epidemiological methods such as contact tracing and case counting. One of the best-known examples is from the nineteenth century when John Snow mapped the location of cholera cases to identify the source of an outbreak as contaminated water from the Broad Street pump in the Soho district of London. More recently, by leveraging pathogen genetic sequence data, it is possible to build a more detailed picture of transmission and to disentangle transmission patterns that are obscured when only epidemiological data are available.

Previously, transmission networks were inferred from detailed contact tracing information. While this approach is still highly informative, person to person contact information is not always readily available. Additionally, contact information alone cannot always differentiate between transmission patterns such as those seen in person to person transmission versus zoonotic transmission. Past genotyping methods such as multilocus sequence typing generally do not provide enough resolution at the genetic level to be helpful in determining recent transmission. As whole-genome sequencing becomes clinically affordable, near-real-time analysis of pathogen transmission will be feasible. One of the first examples of the use of whole-genome sequencing in studying bacterial transmission at a clinically relevant time-scale involved the analysis of methicillin-resistant *Staphylococcus aureus* (MRSA) in a neonatal intensive care unit (Köser *et al.*, 2012). Here whole-genome sequencing allowed for the identification of direct transmission, including one previously unidentified transmission event. Additionally, the WGS clearly distinguished those cases involved in the outbreak from similar infections occurring through alternative transmission routes.

To date, the increased resolution provided by genomic data has often been more valuable in ruling out suspected routes through which infections have spread than in being able to definitively identify transmission chains. Although there undoubtedly exist new opportunities for unraveling past and future transmission networks, there are still many methodological challenges plaguing who-infected-whom studies. Biased sequence sampling, population structure, genetic recombination, within-host evolution and multiple infection can all affect the accuracy of inferred transmission networks (Frost *et al.*, 2015).

36.4.1 Challenges

36.4.1.1 Within-Host Diversity

Short-term evolution rates in bacteria over months or years have been shown to be higher than previous estimates from long-term evolution studies. As most studies using bacterial whole-genome sequencing make use of a single genome per infection, this higher evolutionary rate and resulting high diversity within a host can both facilitate and hamper efforts at accurately inferring transmission links. For example, a genome sampled in one host may be closer to that sampled from a second host, than to the host from which both were infected. Conversely, a higher mutation rate allows for the resolution of transmission links at shorter time-scales than was previously thought possible. A more thorough review of within-host diversity in bacterial populations is given in Didelot *et al.* (2016).

36.4.1.2 Population Structure and Sampling

Population structure and uneven sampling can lead to biases in transmission inference (Dearlove *et al.*, 2017). Oversampling a sub-population can lead to artificially high estimates for transmission rates within that population. Sampling may be focused on high-risk groups as a consequence of public health policies. Differences in access to health care and thus uneven diagnosis in different host populations can also bias sampling efforts. Frequently, sampling practices will have changed over time, or differ between countries, leading to a trade-off in the scope and consistency of isolate collections. These biases can become a significant problem in the identification of super-spreaders, hosts that give rise to many more subsequent infections than average and may lead to the misallocation of outbreak response resources. Some work has been done to distinguish outbreak clusters from those due to sampling (McCloskey and Poon, 2017; Dearlove *et al.*, 2017), but these approaches do not attempt to infer the full transmission network.

Missing samples can have a large impact on the success of transmission inference methods. Many approaches currently assume that all infected individuals involved in an outbreak are sampled. This is likely to be true only for the minority of cases where dense testing is present, the pathogen rapidly becomes acutely symptomatic, the infecting species is easily detected and the outbreak is relatively small.

36.4.1.3 Mixed Infections

The close genetic distances that allow for the identification of transmission events can be obscured when hosts are infected with multiple distinct lineages or mixed infections. This occurs when a host is infected by multiple strains and can lead to larger genetic distances between hosts than would be seen when comparing individual genetic lineages. One approach to dealing with mixed infections is to try and reconstruct the individual haplotypes within each host. This has seen some success, with Eyre *et al.* (2013) implementing a two-step maximum likelihood approach to first identify mixed infections before making use of reference strains to infer the individual haplotypes. Other approaches such as Phyloscanner (Wymant *et al.*, 2017) identify mixed infections but do not explicitly model them in transmission.

36.4.1.4 Recombination and Horizontal Gene Transfer

As with standard phylogenetic inference, failing to adequately account for recombination can lead to erroneous conclusions. Although some bacterial species are essentially clonal and do not undergo recombination, in many cases recombination plays a significant role in the evolution of a species. *S. pneumoniae* is an example of such a highly recombinant species, with a recombination rate much greater than its point mutation rate. If recombination is not adequately dealt with it can lead to incorrect phylogenies, resulting in errors in transmission inference. Many of the tools used to account for recombination can also be utilized in analyzing transmission. Methods such as Gubbins (Croucher *et al.*, 2015) and ClonalFrame (Didelot and Wilson, 2015; Didelot and Falush, 2007) can be used to generate phylogenetic trees that account for recombination in recently diverged populations (see Figure 36.1(a)), which can then be used to infer transmission.

36.4.1.5 Inference Methods for Transmission Analysis

The number of methods for inferring transmission networks from WGS data is growing quickly, with each attempting to tackle many of the challenges already discussed. The ‘best’ method is likely to vary depending upon the data set considered, and care should be taken to ensure that the assumptions of a method fit the problem.

36.4.1.6 Pairwise Distance-Based Methods

Some of the simplest and fastest methods for inferring transmission from WGS data involve comparing pairwise distances, with pairwise SNP distance being a common choice. This strategy involves aligning sequencing reads to a reference genome and comparing the number of SNPs that differ between every sample. A threshold is usually chosen to identify direct transmission where one sampled person is believed to have directly infected another. While approaches based on SNP distances are simple and often effective, they fail to account for many of the previously mentioned challenges. A particular problem is encountered when there exists substantial variation in the degree of diversity between donor and recipients in transmission chains. This may arise due to high within-host diversity accumulating during chronic infections, microbes diversifying through recombination, or differences in mutation rate, as observed in 'hypermutator' strains. These can lead to direct transmission having larger SNP distances, resulting in a transmission being incorrectly classified as indirect.

An early approach went beyond the simple threshold criterion and integrated temporal and spatial separation between infected sites to generate a combined likelihood of transmission (Ypma *et al.*, 2012). Available implementations of analogous approaches include one based on the distribution of pairwise genetic distances (Worby *et al.*, 2014) and Outbreaker (Jombart *et al.*, 2011). Both of these methods assume complete sampling, a constant mutation rate, no mixed infections and no significant population structure. The Worby *et al.* method develops an approximation to the distribution of SNPs which allows for a fixed bottleneck at discrete transmission times. It then attempts to find the transmission network that maximizes the likelihood given this distribution. The Jombart *et al.* approach instead attempts to find the most parsimonious network given a pairwise distance matrix. Ties in parsimony are broken by an alternative proximity measure such as geographic distance.

36.4.1.7 Inference of Transmission from Phylogeny

While transmission cannot be directly inferred from a phylogenetic tree, a transmission tree must be consistent with the underlying phylogeny. A number of methods take advantage of this and make use of an initial dated phylogeny from which transmission is inferred. This approach has the added advantage of incorporating both sampling times and a molecular clock that considers branch lengths and can account for varying rates of evolution.

TransPhylo (Didelot *et al.* 2014, 2017) takes a dated phylogeny as input and attempts to paint the branches of the tree with a unique color for each host, indicating the host in which each branch evolved. Changes in color then represent transmission events. The initial algorithm (Didelot *et al.* 2014) assumed that an outbreak was completely sampled and had finished. An updated algorithm allows for the case of an ongoing outbreak as well as missing samples by incorporating reversible jumps into the MCMC moves. The approach models within-host diversity by assuming a fixed within-host population size and fixed neutral coalescence rate. The model also assumes a complete transmission bottleneck which requires that only a single genotype can be transmitted between hosts. Transmission is inferred using an MCMC approach which simulates within-host diversity by first independently simulating the genealogy within each host. These subtrees are then knitted together to form a transmission chain.

Phyloscanner (Wymant *et al.*, 2017) is also based on the idea of inferring subtrees and linking them to form transmission networks. Phyloscanner takes advantage of sequence data from multiple genotypes within a host to improve transmission inference. In doing so, Phyloscanner is able to identify multiple infections, account for contamination, infer within- and between-host phylogenies and identify crossover recombination breakpoints. Phyloscanner splits the

genome into windows before inferring subtrees within each window. By investigating the subgraphs of each host, within each window, it is possible to identify probable transmission paths. Multiple infections can be identified as they result in multiple subgraphs within a window for a host. A disadvantage of the Phyloscanner approach over those based on phylogenetic and epidemiological models is that it does not provide a likelihood for an estimate and consequently the uncertainty within the transmission classification is not easily quantified. Phyloscanner also sacrifices accounting for epidemiological information such as sampling time for the added flexibility of the approach.

36.4.1.8 Dual Inference of Both Phylogenetic and Transmission Trees

The most flexible model-based approaches estimate both the phylogenetic and transmission trees concurrently. By including estimation of the phylogeny in a Bayesian MCMC approach it is possible to incorporate uncertainty in the phylogenetic tree into transmission inference. Outbreaker (Jombart *et al.*, 2014) utilizes a discrete-time stochastic model that incorporates both genetic sequence data and sampling dates. Outbreaker assumes that transmission events occur as branching events in the phylogeny, and it does not account for within-host diversity. Similar to Transphylo, it requires estimates of both the generation time of the pathogen and sampling times. As Outbreaker does not account for within-host diversity, it is better suited to analyzing the transmission of pathogens with short generation times and little opportunity to evolve within the host. The flexibility of simultaneously inferring both the phylogenetic tree and transmission tree allows for many epidemiological considerations to be incorporated into the method. Ongoing work to make these adjustments easier is available in the Outbreaker2 package (Campbell *et al.*, 2018), part of the RECON project to integrate epidemiological tools into workflows within R.

Phybreak (Klinkenberg *et al.*, 2017) is similar to Outbreaker in that it jointly infers the transmission and phylogenetic trees using an MCMC algorithm. Significantly, however, it accounts for within-host diversity using a model similar to Transphylo. Unlike Outbreaker, phybreak does not account for missing samples. This restricts the applicability of phybreak to very densely sampled outbreaks.

Some of the most flexible approaches to inferring transmission use Bayesian transmission networks. One approach to cope with the difficulty in scaling methods based on MCMC sampling is to first subset the network into likely clusters of transmission using methods such as the MMPP clustering method (McCloskey and Poon, 2017).

A method that takes advantage of the BEAST framework to jointly infer both the phylogenetic and transmission trees is that of Hall *et al.* (2015). Here, candidate phylogenetic trees are sampled and a transmission tree is inferred by looking at partitions of the tree into hosts, with the restriction that all tips corresponding to one sample are taken from only one host. This approach assumes that all infections have been sampled and does not consider mixed infections. SCOTTI (De Maio *et al.*, 2016) is another method based on the BEAST framework. Unlike the other methods discussed so far, SCOTTI makes use of an approximate version of the structured coalescent (De Maio *et al.*, 2015), treating each host as a distinct population and modeling transmission as migration events between populations. This accounts for within-host diversity and is able to take advantage of multiple samples per individual. The approach is also flexible enough to account for exposure times of individuals and allows for missing infections. Care should be taken in the interpretation of the number of missing infections inferred by the method, because an inferred missing host can be infected many times as it is assumed to have an infinite exposure interval.

MCMC-based approaches rely on being able to rapidly calculate the likelihood. If calculating the likelihood itself is intractable, approximate Bayesian computation (ABC) based approaches

can be used (Beaumont *et al.*, 2002). A successful application of ABC to transmission inference focused on the transmission dynamics of *M. tuberculosis* in San Francisco (Lintusaari *et al.*, 2019).

36.4.1.9 Endemic Disease and Colonization

All the transmission analysis techniques discussed so far have been designed for, and mainly applied to, outbreaks and epidemics. That is, they often rely on the assumption of a single introduction into an uninfected population. In an endemic disease where the pathogen is maintained stably over a long time the situation is likely to be more complex. When a pathogen is endemic all cases may be related in some way at the broader scale of global transmission. This can hamper efforts to infer transmission accurately at the local scale. An approach developed by Mollentze *et al.* (2014) attempts to deal with this issue by modeling two types of unsampled hosts. The first type of missing host includes those which have been infected either directly or indirectly by a sampled host and corresponds with those used by other methods. The second type of unsampled host is one which has no ancestors within the samples. The method is also one of the few that incorporates geographic distance in their likelihood calculation. Another type of approach to modeling transmission in an endemic case was introduced by Numminen *et al.* (2014). Their method is based on a two-stage ABC inference about feasible transmission trees, which can incorporate various observation models and prior assumptions. Future development of methods that are suited to endemic diseases and well scalable to large data sets is crucial to understand some of the most significant endemic pathogens.

36.5 Genome-Wide Association Studies in Bacteria

36.5.1 Background

Mapping variation in a bacterial phenotype, such as antimicrobial drug resistance or virulence, to the causal genotypic variation is a task well addressed by the application of GWAS methodology. Classically, the relation of a change in phenotype to a causal genetic locus would have to be determined by wet-lab experiments. The technical complexities of these microbiological methods aside, there are three major limitations of this approach. Firstly, it usually necessitates a hypothesis-driven approach, where a single candidate gene, chosen using prior biological knowledge, is knocked out in an isogenic mutant strain, and is then checked for a difference in phenotype compared to the wildtype strain. Secondly, often only variants with large effect sizes can be confidently identified by these approaches, owing to the small sample sizes attainable by expensive or time-consuming assays, such as *in vivo* models. Finally, this technique can only be applied in well-characterized model systems (both *in vitro* and *in vivo*) rather than in the natural population, which also restricts the phenotypes tested to those which can be measured in the lab.

High-throughput sequencing has made available many sequences of isolates taken from their natural host. By calculating the strength of association between a phenotype and all genome-wide sequence variation, a GWAS has the potential to overcome the above limitations, and therefore complements wet-lab investigations into bacterial genetics. However, direct application of GWAS techniques developed for diploid eukaryotes such as humans is usually inappropriate for two main reasons: strong population structure and the existence of a more extensive pan-genome, that is, the collection of all genes present in at least one haplotype in a population of a bacterial species.

Most human genetic variation is due to small variants, which can be detected by resequencing and mapping to a reference from a single population. Though some variation is missed by

considering a single reference, the amount of pan-genomic variation is low, ~1% of the overall sequence (Li *et al.*, 2010). In bacteria, short variants in core genes are undoubtedly important but the presence of an accessory genome not covered by simple SNP mapping is a significant source of variation. An accessory gene is any coding sequence that is variably present in individual haplotypes in a population of a bacterial species. An alignment-free method of variant detection is therefore ideal, as the computational burden of multiple reference mappings, the bias of available references and the issue of varying levels of missing calls across the genome make alignment generally less suitable than in human genomes.

Sequence words of length k , called k -mers, have been used for many tasks in bioinformatics, including genome assembly, seeding alignment matches and estimation of evolutionary distance between sequences. Bacterial GWASs can use these k -mers as a generalized variant which covers common variation of all sizes from SNP to genes. Typically 31-mers are used as they have a good sensitivity–specificity trade-off; however, using all k -mers between 9 and 100 bases long have been shown to increase discovery power. The total number of k -mers is large ($\sim 6 \times 10^7$ for a typical data set with 10^3 samples), so requires an efficient association method, automated post-association annotation and careful consideration of the multiple testing issue.

Bacteria have strong population structure because they do not generate variation in their genomes through crossing over every generation. In the case of completely non-recombinogenic species such as *M. tuberculosis*, vertical inheritance of point mutations leads to most of the genome being in LD. If mutations are introduced *de novo* over time, one of which is causal for the phenotype of interest, a naive association will find the entire set of variants to be associated with the phenotype (the causal variant, and the genetic background). While this is locally true around causal variants in the human genome, the rapid decay of LD allows the association to be mapped to a small region. However, in these bacteria LD extends across the entire chromosome, so the set of associations will also be genome-wide, preventing fine mapping of the causal variant. In species with high rates of homologous recombination such as *S. pneumoniae*, the strength of LD is lower, but still extends across the whole genome.

Variants correlated with a specific genetic background and the phenotype are known as ‘lineage’ associations. The best that bacterial GWAS can reasonably hope to achieve with such associations is to identify them as such, and prioritize sets of associated variants for study by other means. It is also possible for variants to be associated with a phenotype independent of genetic background. These ‘locus’ variants are homoplasic, that is, the causal alleles have appeared independently in multiple genetic backgrounds. They can therefore be mapped to a more specific region of the genome, and are currently the main focus of bacterial GWASs. This is not because they are more important than lineage variants (both types of variant may in theory explain any amount of the total heritability), but they are easier to interpret. Locus variants arise from horizontal inheritance of the alleles due to recombination or due to convergent evolution. An example illustrating these aspects is given in Figure 36.2.

As explained above, it is critical to account for population structure in a bacterial GWAS, and this can be done using either a phylogeny or a regression-based approach.

36.5.2 Phylogenetic Methods

The first phylogenetic methods were designed for testing the correlation of single traits between mammals, using the species tree to control for their genetic relatedness. Methods for both binary and continuous traits have been developed, and can be applied between genomic variants and phenotypes to perform a GWAS. Independent contrasts, motivated by a Brownian motion model of trait evolution on the tree, use the difference in phenotype between phylogenetic sister isolates and their branch lengths to adjust for expected correlations between species

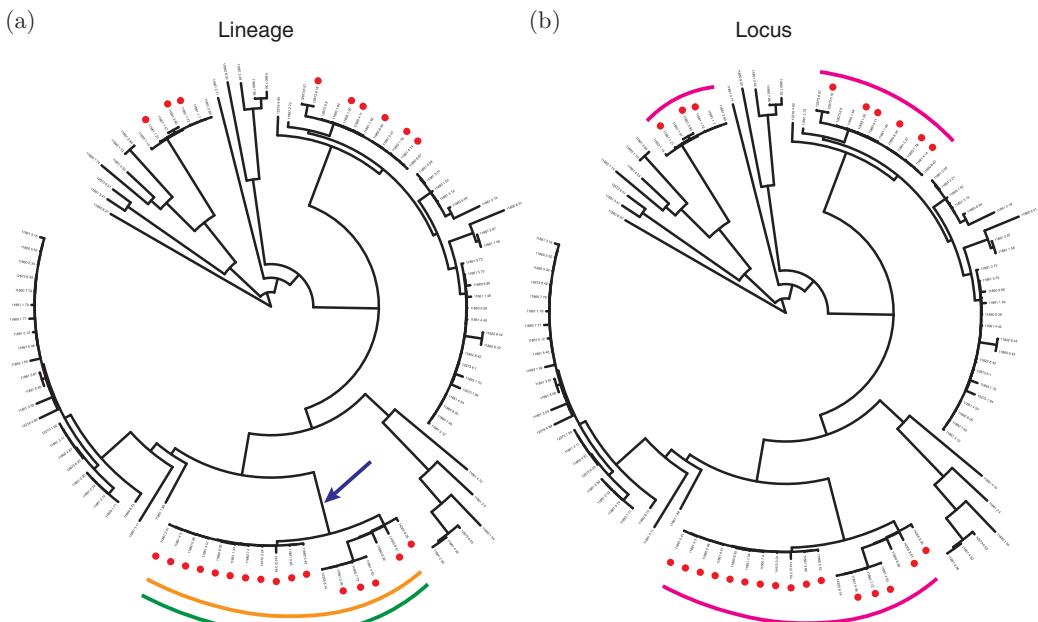


Figure 36.2 Phylogenetic illustration of lineage and locus variants. Depicted is an example phylogeny, with cases identified by red dots at the tips, and controls without dots. Colored arcs indicate the presence of a variant of interest. (a) The orange-colored variant is a causal lineage variant, and will be associated with the phenotype in a naive analysis. The green variant, present in the same clade, is not causal but will also appear associated at the same level of significance. Indeed, any mutation that has occurred on the branch indicated by the arrow will appear associated, hindering association mapping. (b) The magenta-colored variant has arisen independently in three separate clades containing cases, making it independent of genetic background and so it provides stronger evidence for association with the phenotype. A population-adjusted association should have a higher p -value and slightly lower odds ratio than the green and orange variants due to the reduced penetrance observed.

(Felsenstein, 1985). An alternative approach is to use least squares regression, but instead of assuming uncorrelated error terms for each sample, the shared distance from the root of the phylogeny between each pair of samples is used as the covariances between error terms in the model (Garland and Ives, 2000). This can be solved by changing the basis of the resulting matrices such that the errors are uncorrelated, and applying ordinary least squares regression. The computational complexity of this method increases with the cube of the number of samples, so it cannot be applied to large data sets.

Many of these models have been implemented in BayesTraits (Meade and Pagel, 2016), though not all available models will scale to genome-wide analysis, and they require manual formatting of input files. Care should also be taken to ensure that correct practice is followed to adjust the output for multiple testing. For binary traits, Scoary has been developed specifically for bacterial GWAS (Brynildsrød *et al.*, 2016). It reports a χ^2 test p -value for every cluster of orthologous gene sequences (COG), unadjusted for population structure, for association between COG presence/absence and phenotype. Pairwise comparisons (Read and Nee, 1995) on the phylogeny are used to estimate the number of times the trait has evolved independently. However, this model does not offer a way of combining the test of evolutionary convergence with phenotypic association.

It is possible to simulate the null distribution of test statistics accounting for phylogenetic correlations by using Monte Carlo simulations with a model of character evolution (Martins

and Garland, 1991). The observed test statistic for each variant can be tested against the generated null distribution to determine an association *p*-value, though the large number of multiple tests necessitates a large number of simulations (ten times the number of variants being tested, as a rule of thumb). The first test statistic used in this way was the correlation of genotype and phenotype at the tips of the phylogeny, much like a χ^2 test though with the modified null distribution to account for population structure (Sheppard *et al.*, 2013). A proposed extension, the R package treeWAS, calculates two further test statistics using ancestral state reconstruction, which may be better at finding associations that have persisted across evolutionary history (Collins and Didelot, 2017). The first captures variants with correlated evolution with the phenotype by counting simultaneous changes of variant and phenotype at nodes, and the second by summing the length of branches at which variant and phenotype are correlated.

These methods offer precise control of Type I error rate when accounting for population structure, but rely on having a trusted phylogeny, not tainted by recombination, with good branch support. This may be possible for small, closely related collections of isolates where recombinants can be removed, or in species without extensive recombination, but is not feasible across an entire recombinogenic species. For some methods a posterior tree distribution can be used as input rather than a single tree, which can partly account for poorly supported branch splits but at the expense of a greater computational burden. The total computational burden of these methods is generally high, especially if they use Monte Carlo simulations, and they may not scale to the millions of tests needed to test variation across the entire pan-genome. Hence application has mostly been limited to analysis of accessory COGs, or species/strains with limited levels of SNP variation.

36.5.3 Regression-Based Methods

Regression-based methods are more similar to approaches used for GWASs in sexually reproducing species such as humans. A linear model is fitted independently at each variant, using the *p*-value of the slope of the regression to test for association with the phenotype. Additional fixed and/or random effects are added as predictors to account for population structure, and if necessary the error distribution and link function can be modified to suit binary phenotypes.

The software SEER uses this approach to be to perform an association that scales linearly with numbers of samples and variants, and can therefore test all variable-length *k*-mers across the pan-genome (Lees *et al.*, 2016). The approach has the advantage that it is not dependent on an alignment or phylogeny. If the input data are separable, as occurs with high effect size variants frequently causal for antimicrobial resistance phenotypes, standard logistic regression will report a highly conservative *p*-value due to the artificially inflated standard error of the fit. In SEER this is avoided by automatically applying Firth's correction to the likelihood in these cases (Heinze and Schemper, 2002). To account for population structure, SEER adds fixed effects to each regression based on a pairwise distance matrix constructed using the shared *k*-mer content between all samples. This matrix is projected into a lower number of dimensions (selected by the user, based on a scree plot) using metric multidimensional scaling, with the positions in this space used as fixed effects, analogous to the use of principal components as covariates in human GWASs.

In some cases, the association result can be sensitive to the choice of these fixed effects, which can lead to highly inflated test statistics. By instead using a kinship matrix to model the variance of a random effect (a linear mixed model (LMM)), this includes the genetic relationships between all samples rather than selecting a proportion of the population structure, and has been shown to give good control of Type I error without loss of power (Lippert *et al.*, 2011). Trans-ethnic human GWASs have driven the development of efficient LMMs to perform this

analysis with a complexity which is linear in the numbers of samples and variants (Lippert *et al.*, 2011; Zhou and Stephens, 2012). The computational tricks used with these models are similar to the change of basis used in phylogenetically adjusted regression, and they allow for estimation of the ratio of genetic variation to environmental variation (giving the narrow-sense heritability h^2).

The R package bugwas implements the human GWAS software GEMMA for use with bacterial data, including k -mers. By applying this LMM to their top variants from a naive association test, the authors were able to find locus variants affecting antibiotic resistance while controlling Type I error from population structure (Earle *et al.*, 2016). Within the same model, they were also able to deconstruct the random effects to further identify potential lineage associations with both the phenotype and the population structure components, albeit with greatly reduced power.

In contrast to phylogenetic methods, regression-based methods are fast, do not require an accurate phylogeny, and may also be alignment-free. They are therefore more scalable with the large sample sizes needed for high-powered GWASs, and the large number of variants which must be tested across the pan-genome. Compared to well-calibrated phylogenetic methods, fixed-effect methods may have an elevated Type I error rate. LMMs have a similar control of the Type I error rate and can be run across all k -mers (Lees *et al.*, 2017), but they are generally restricted to the discovery of locus variants, and can only test association at the tips of the tree rather than over the evolutionary history of the bacteria. Picking a significance threshold remains a challenge for these methods, but using the number of unique patterns as the number of tests in a Bonferroni correction has proven to be a good starting point.

The recent publication of Pyseer has attempted to harmonize the various regression approaches, by allowing association of any form of bacterial variation under either a fixed effects or mixed effects linear model with software that is specific to microbial GWASs (Lees *et al.*, 2018a). It contains all the features of SEER, bugwas and Scoary, plus enhanced processing of significant k -mers and the ability to perform burden testing of rare variation.

36.6 Genome-Wide Epistasis Analysis

Genome-wide epistasis analysis is an emerging field of bacterial population genomics where the purpose is to identify sets of SNP loci under co-selective pressure, by seeking significant deviations of the observed haplotype distribution from the expectation based on LD. The first genome-wide epistasis analysis of this kind focused on using pairwise statistical correlation analysis to successfully reveal significant coevolutionary patterns across the genome for 51 *Vibrio parahaemolyticus* isolates (Cui *et al.*, 2015). Later, Skwark *et al.* (2017) showed that statistical genome-wide modeling of joint SNP variation using direct coupling analysis (DCA) can efficiently uncover valuable information about coevolutionary pressures from large-scale population genomic data. DCA emerged less than a decade ago and has opened up a new direction of biological research by demonstrating that large population-based protein sequence analysis can be leveraged to make accurate predictions about protein structure (Weigt *et al.*, 2009; Morcos *et al.*, 2011, 2014; Feinauer *et al.*, 2014; Ovchinnikov *et al.*, 2014, 2017, Söding, 2017).

Maximum likelihood inference for the Potts models employed in DCA is intractable due to the form of the normalizing constant of the model distribution, hence various weaker criteria or approximations have been used to derive estimators of the model parameters. Notably, maximum pseudolikelihood is a statistically consistent inference method which has typically outperformed variational methods (Wainwright and Jordan, 2008), such as the mean-field estimator (Ekeberg *et al.*, 2013).

To enable use of DCA for large numbers of polymorphisms in a bacterial genome, Skwark *et al.* (2017) stratified a genome into non-overlapping windows and sampled randomly a single SNP from each window to form haplotypes of approximately 1500 sequence positions, on which the plmDCA implementation by Ekeberg *et al.* (2014) could be directly applied. They then used a large number of repeated random sampling of positions from the stratified genome to aggregate information about interactions between polymorphisms across the genome. While this approach was demonstrated to successfully capture both known and novel interactions, it remains very computationally intensive and may still leave important interactions undiscovered as only a fraction of all possible combinations of interactions will be covered even when using large numbers of repeated samples.

More recently, Puranen *et al.* (2018) built upon this initial observation to develop DCA into a powerful tool that is applicable to a majority of the existing bacterial population genome data sets in a computationally scalable manner. These advances are based on a new computational architecture exploiting efficient parallelization and optimization to achieve scalability for up to 10^5 polymorphisms. In addition to being significantly faster with modest computational resources, they also show that the global inference with SuperDCA allows the discovery of previously undetected epistatic interactions that inform our understanding of bacterial biology related to survival of the pneumococcus at lower temperatures. SuperDCA is freely available from <https://github.com/santeripuranen/SuperDCA>.

As an alternative to computationally expensive model-based methods for detecting significant co-variation of SNP beyond local LD, pairwise methods (e.g. Cui *et al.*, 2015) offer better scalability, which has been the main motivation for them in the general field of statistical graphical model learning. However, a recent simulation study on high-dimensional structure learning of synthetic network models showed that a family of pairwise methods based on mutual information (MI) may be as accurate as and even outperform model-based methods in the small sample regime, which is particularly relevant to bacterial population genomics (Pensar *et al.*, 2019a, <https://arxiv.org/abs/1901.04345>). To employ this advantage, a recent GWES software SpydrPick performs an MI scan over all pairs of SNPs present in an alignment and discards online those pairs that do not show signal beyond global significance limits. Simulation study with neutral models showed that the method is able to maintain a good control of the rate of false positive findings. Examples with both core and pangenomic population variation in major human pathogen species demonstrated that the method offers considerable potential to drive molecular discoveries, even in the absence of phenotypic data (Pensar *et al.*, 2019b, <https://www.biorxiv.org/content/10.1101/523407v1>). Since the GWES approaches did emerge only relatively recently, we expect that there will be substantial further development in this field.

36.7 Gene Content Analysis

In less than a decade, bacterial population genomics has progressed from sequencing of dozens to thousands of strains. Phylogenetic trees are the main framework utilized for visualization and exploration of population genomic data, both in terms of the level of relatedness of strains and for mapping relevant metadata such as geographic locations and host characteristics. While trees are highly useful, they are in general estimated using only core genomic variation, that is, those regions of the genome common to all members of a species/sample, which may represent only a fraction of the relevant differences present in genomes across the study population. Several recent studies highlight the importance of considering variation in gene content when investigating the ecological and evolutionary processes leading to the observed data (Corander *et al.*, 2017).

The rapidly increasing size of population genomic data sets calls for efficient visualization methods to explore patterns of relatedness based on core genomic polymorphisms, accessory gene content, epidemiological, geographical and other metadata. Abudahab *et al.* (2018) introduced a framework (PANINI) of identifying neighbors of strains using gene content sharing analysis that integrates within the web application Microreact (Argimón *et al.*, 2016), by utilizing a popular unsupervised machine learning technique for big data to infer neighbors of bacterial strains from accessory gene content data and to efficiently visualize the resulting relationships. The machine learning method, called t-SNE, has already gained widespread popularity for exploring image, video and textual data, but was not previously utilized for bacterial population genomics. When applied to population-wide genomic data sets, the algorithm was clearly able to identify distinct lineages within both dense and diverse collections. This analysis could highlight which clusters, defined using the core genome, could be sensibly subdivided, and which small groups of unclustered isolates could justifiably be regarded as new clusters. Within lineages, the same congruence between core and accessory genomes across clades was not observed. Instead, clusters were distinguished by rapidly occurring, homoplasic alterations, such as phage infection. An example is shown in Figure 36.1(b), in which PANINI is applied to the PMEN2 lineage (Croucher *et al.*, 2014b), the core genome of which was analyzed using hierBAPS. The clusters defined based on the polymorphisms in shared genes show little correlation with the distribution inferred from variation in gene content. This is in spite of the central grey BAPS3 clade having lost the ability to acquire exogenous DNA for homologous recombination through its competence system (Croucher *et al.*, 2014a); instead, these rapid short-term changes tend to reflect the rapid movement of phage and other fast-transmitting mobile genetic elements. In this context, PANINI provides an intuitive way in which to understand the distribution of rapidly evolving aspects of the genome, which are difficult to analyze in a conventional phylogenetic framework. PANINI is therefore a promising platform through which biologically important changes in bacterial gene content can be uncovered at all levels of evolutionary, ecological and epidemiological analyses.

References

- Abudahab, K., Prada, J.M., Yang, Z. and Bentley, S.D. (2018). PANINI: Pangenome Neighbor Identification for Bacterial Populations. *Microbial Genomics* 2018;4 DOI 10.1099/mgen.0.000220.
- Argimón, S., Abudahab, K., Goater, R.J.E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E.J., *et al.* (2016). Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics* 2(11), e000093.
- Beaumont, M.A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics* 162(4), 2025–2035.
- Beugn, M.-P., Gayet, T., Pontier, D., Devillard, S. and Jombart, T. (2018). A fast likelihood solution to the genetic clustering problem. *Methods in Ecology and Evolution* 9(4), 1006–1016.
- Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.
- Bollback, J.P. (2002). Bayesian Model adequacy and choice in phylogenetics. *Molecular Biology and Evolution* 19(7), 1171–1180.
- Bromham, L., Duchêne, S., Hua, X., Ritchie, A.M., Duchêne, D.A. and Ho, S.Y.W. (2018). Bayesian molecular dating: Opening up the black box. *Biological Reviews of the Cambridge Philosophical Society* 93(2), 1165–1191.

- Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology* **17**, 238.
- Camin, J.H. and Sokal, R.R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19**(3), 311–326.
- Campbell, F., Didelot, X., Fitzjohn, R., Ferguson, N., Cori, A., and Jombart, T. (2018). “outbreaker2: A modular platform for outbreak reconstruction.” *BMC Bioinformatics* **19** (Suppl 11), 363.
- Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M. and Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution* **30**(5), 1224–1228.
- Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., et al. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics* **46**(3), 305–309.
- Collins, C. and Didelot, X. (2017). A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. Preprint, bioRxiv 140798.
- Corander, J. and Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**(10), 2833–2843.
- Corander, J., Waldmann, P. and Sillanpää, M.J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**(1), 367–374.
- Corander, J., Waldmann, P., Marttinen, P. and Sillanpää, M.J. (2004). BAPS 2: Enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**(15), 2363–2369.
- Corander, J., Marttinen, P. and Mäntyniemi, S. (2006). A Bayesian method for identification of stock mixtures from molecular marker data. *Fishery Bulletin* **104**(4), 550–558.
- Corander, J., Fraser, C., Gutmann, M.U., Arnold, B., Hanage, W.P., Bentley, S.D., Lipsitch, M. and Croucher, N.J. (2017). Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nature Ecology & Evolution* **1**, 1950–1960.
- Croucher, N., Hanage, W., Harris, S., McGee, L., van der Linden, M., de Lencastre, H., Sá-Leão, R., Song, J.-H., et al. (2014). Variable recombination dynamics during the emergence, transmission and ‘disarming’ of a multidrug-resistant pneumococcal clone. *BMC Biology* **12**, 49, <https://doi.org/10.1186/1741-7007-12-49>.
- Croucher, N.J., Chewapreecha, C., Hanage, W.P., Harris, S.R., McGee, L., van der Linden, M., Song, J.-H., Ko, K.S., de Lencastre, H., Turner, C., Yang, F., Sá-Leão, R., Beall, B., Klugman, K.P., Parkhill, J., Turner, P. and Bentley, S.D. (2014a). Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome Biology and Evolution* **6**(7), 1589–1602.
- Croucher, N.J., Hanage, W.P., Harris, S.R., McGee, L., van der Linden, M., de Lencastre, H., Sá-Leão, R., Song, J.-H., Ko, K.S., Beall, B., Klugman, K.P., Parkhill, J., Tomasz, A., Kristinsson, K.G. and Bentley, S.D. (2014b). Variable recombination dynamics during the emergence, transmission and ‘disarming’ of a multidrug-resistant pneumococcal clone. *BMC Biology* **12**, 49.
- Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J. and Harris, S.R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* **43**(3), e15.
- Cui, Y., Yang, X., Didelot, X., Guo, C., Li, D., Yan, Y., Zhang, Y., et al. (2015). Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen *Vibrio parahaemolyticus*. *Molecular Biology and Evolution* **32**(6), 1396–1410.
- Dearlove, B.L., Xiang, F. and Frost, S.D.W. (2017). Biased phylodynamic inferences from analysing clusters of viral sequences. *Virus Evolution* **3**(2), vex020.

- De Iorio, M., Elliott, L.T., Favaro, S., Adhikari, K. and Teh, Y.W. (2015). Modeling population structure under hierarchical dirichlet processes. Preprint.
- De Maio, N., Wu, C.-H., O'Reilly, K.M. and Wilson, D. (2015). New routes to phylogeography: A Bayesian structured coalescent approximation. *PLoS Genetics* **11**(8), e1005421.
- De Maio, N., Wu, C.-H. and Wilson, D.J. (2016). SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Computational Biology* **12**(9), e1005130.
- Didelot, X. and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**(3), 1251–1266.
- Didelot, X. and Wilson, D.J. (2015). ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Computational Biology* **11**(2), e1004041.
- Didelot, X., Gardy, J. and Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution* **31**(7), 1869–1879.
- Didelot, X., Walker, A.S., Peto, T.E., Crook, D.W. and Wilson, D.J. (2016). Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology* **14**(3), 150–162.
- Didelot, X., Fraser, C., Gardy, J. and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution* **34**(4), 997–1007.
- Drummond, A.J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.
- Drummond, A.J., Rambaut, A., Shapiro, B. and Pybus, O.G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* **22**(5), 1185–1192.
- Drummond, A.J., Suchard, M.A., Xie, D. and Rambaut, A. (2012). Bayesian phylogenetics with BEAUTi and the BEAST 1.7. *Molecular Biology and Evolution* **29**(8), 1969–1973.
- Duchêne, D.A., Duchêne, S., Holmes, E.C. and Ho, S.Y.W. (2015). Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Molecular Biology and Evolution* **32**(11), 2986–2995.
- Earle, S.G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N.C., Walker, T.M., Spencer, C.C.A., et al. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* **1**, 16041.
- Ekeberg, M., Lökvist, C., Lan, Y., Weigt, M. and Aurell, E. (2013). Improved contact prediction in proteins using pseudolikelihoods to infer Potts models. *Physical Review E* **87**(1), 012707.
- Ekeberg, M., Hartonen, T. and Aurell, E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics* **276**, 341–356.
- Excoffier, L. and Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews Genetics* **7**(10), 745–758.
- Eyre, D.W., Cule, M.L., Griffiths, D., Crook, D.W., Peto, T.E.A., Walker, A.S. and Wilson, D.J. (2013). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. *PLoS Computational Biology* **9**(5), e1003059.
- Falush, D., Stephens, M. and Pritchard, J.K. (2003a). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**(4), 1567–1587.
- Falush, D., Wirth, T., Linz, B., Pritchard, J.K., Stephens, M., Kidd, M., Blaser, M.J., et al. (2003b). Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**(5612), 1582–1585.
- Feinauer, C., Skwark, M.J., Pagnani, A. and Aurell, E. (2014). Improving contact prediction along three dimensions. *PLoS Computational Biology* **10**(10), e1003847.
- Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* **125**(1), 1–15.

- Frost, S.D.W., Pybus, O.G., Gog, J.R., Viboud, C., Bonhoeffer, S. and Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics* **10**, 88–92.
- Garland, T., Jr and Ives, A.R. (2000). Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist* **155**(3), 346–364.
- Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A. and Holmes, E.C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**(5656), 327–332.
- Hall, M., Woolhouse, M. and Rambaut, A. (2015). Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set. *PLoS Computational Biology* **11**(12), e1004613.
- Harris, S.R., Feil, E.J., Holden, M.T.G., Quail, M.A., Nickerson, E.K., Chantratita, N., Gardete, S., et al. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**(5964), 469–474.
- Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**(16), 2409–2419.
- Huelsenbeck, J.P. and Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175**(4), 1787–1802.
- Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 604–613.
- Jombart, T., Devillard, S. and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.
- Jombart, T., Eggo, R.M., Dodd, P.J. and Balloux, F. (2011). Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity* **106**(2), 383–390.
- Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C. and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology* **10**(1), e1003457.
- Kendal, D.G. (1948). On the generalized ‘birth-and-death’ process. *Annals of Mathematical Statistics* **19**(1), 1–15.
- Kingman, J.F.C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13**(3), 235–248.
- Klinkenberg, D., Backer, J.A., Didelot, X., Colijn, C. and Wallinga, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Computational Biology* **13**(5), e1005495.
- Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. and Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**(5), 1179–1191.
- Köser, C.U., Holden, M.T.G., Ellington, M.J., Cartwright, E.J.P., Brown, N.M., Ogilvy-Stuart, A.L., Hsu, L.Y., et al. (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine* **366**(24), 2267–2275.
- Lachenbruch, P.A. and Goldstein, M. (1979). Discriminant analysis. *Biometrics* **35**(1), 69–85.
- Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics* **8**(1), e1002453.
- Le, S.Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**(7), 1307–1320.
- Lees, J.A., Vehkala, M., Välimäki, N., Harris, S.R., Chewapreecha, C., Croucher, N.J., Marttinen, P., et al. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications* **7**, 12797.

- Lees, J.A., Croucher, N.J., Goldblatt, D., Nosten, F., Parkhill, J., Turner, C., Turner, P. and Bentley, S.D. (2017). Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife* **6**, e26255.
- Lees, J.A., Galardini, M., Bentley, S.D., Weiser, J.N. and Corander, J. (2018a). Pyseer: A comprehensive tool for microbial pan-genome-wide association studies. *Bioinformatics* **34**(24), 4310–4312.
- Lees, J.A., Kendall, M., Parkhill, J., Colijn, C., Bentley, S.D. and Harris, S.R. (2018b). Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: A simulation based study. *Wellcome Open Research*, **3**, 33.
- Lewis, P.O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**(6), 913–925.
- Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**(4), 2213–2233.
- Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Qian, W., et al. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology* **28**(1), 57–63.
- Lintusaari, J., Blomstedt, P., Sivula, T., Gutmann, M.U., Kaski, S. and Corander, J. (2019). Resolving outbreak dynamics using approximate bayesian computation for stochastic birth-death models. *Wellcome Open Res*, **4**, 14, <https://doi.org/10.12688/wellcomeopenres.15048.1>.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**(10), 833–835.
- Maixner, F., Krause-Kyora, B., Turaev, D., Herbig, A., Hoopmann, M.R., Hallows, J.L., Kusebauch, U., et al. (2016). The 5300-year-old *Helicobacter pylori* genome of the Iceman. *Science* **351**(6269), 162–165.
- Martins, E.P. and Garland, T. (1991). Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. *Evolution* **45**(3), 534–557.
- Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D. and Corander, J. (2012). Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Research* **40**(1), e6.
- McCloskey, R.M. and Poon, A.F.Y. (2017). A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *PLoS Computational Biology* **13**(11), e1005868.
- Meade, A. and Pagel, M. (2016). BayesTraits. <http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/> BayesTraitsV3.0.1.html.
- Minh, B.Q., Nguyen, M.A.T. and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* **30**(5), 1188–1195.
- Mollentze, N., Nel, L.H., Townsend, S., le Roux, K., Hampson, K., Haydon, D.T. and Soubeyrand, S. (2014). A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings. Biological Sciences* **281** (1782), 20133251.
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108**(49), E1293–E1301.
- Morcos, F., Hwa, T., Onuchic, J.N. and Weigt, M. (2014). Direct coupling analysis for protein contact prediction. *Methods in Molecular Biology* **1137**, 55–70.
- Nascimento, F.F., dos Reis, M. and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution* **1**(10), 1446.
- Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**(1), 268–274.

- Numminen, E., Chewapreecha, C., Siren, J., Turner, C., Turner, P., Bentley, S.D. and Corander, J. (2014). Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society, Series B* **281**(1794), 20141324.
- Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**(1), 132.
- Ovchinnikov, S., Kamisetty, H. and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030.
- Ovchinnikov, S., Park, H., Varghese, N., Huang, P.-S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyripides, N.C. and Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science* **355**(6322), 294–298.
- Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, **25**(7), 1253–1256, <https://doi.org/10.1093/molbev/msn083>.
- Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**(2), 945–959.
- Puranen, S., Pesonen, M., Pensar, J., Xu, Y.Y., Lees, J.A., Bentley, S.D., Croucher, N.J. and Corander, J. (2018). SuperDCA for genome-wide epistasis analysis. *Microbial Genomics*. doi: [10.1099/mgen.0.000184](https://doi.org/10.1099/mgen.0.000184).
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- Read, A.F. and Nee, S. (1995). Inference from binary comparative data. *Journal of Theoretical Biology* **173**(1), 99–108.
- Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12), 1572–1574.
- Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. and Schooler, E. (2002). SIP: Session Initiation Protocol. <http://www.rfc-editor.org/info/rfc3261>.
- Sagulenko, P., Puller, V. and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**(1), vex042.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**(4), 406–425.
- Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., Bentley, S.D., Maiden, M.C.J., Parkhill, J. and Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences* **110**(29), 11923–11927.
- Simonsen, M., Mailund, T. and Pedersen, C.N.S. (2008). Rapid neighbour-joining. In K.A. Crandall and J. Lagergren (eds.), *Algorithms in Bioinformatics*. Springer, Berlin, pp. 113–122.
- Skwark, M.J., Croucher, N.J., Puranen, S., Chewapreecha, C., Pesonen, M., Xu, Y.Y., Turner, P., et al. (2017). Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genetics* **13**(2), e1006508.
- Söding, J. (2017). Big-data approaches to protein structure prediction. *Science* **355**(6322), 248–249.
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313.
- Tang, H., Coram, M., Wang, P., Zhu, X. and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *American Journal of Human Genetics* **79**(1), 1–12.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**(2), 57–86.
- To, T.H., Jung, M., Lycett, S. and Gascuel, O. (2016). Fast dating using least-squares criteria and algorithms. *Systematic Biology* **65**(1), 82–97.

- Tonkin-Hill, G., Lees, J., Bentley, S., Frost, S. and Corander, J. (2018). Fast hierarchical Bayesian analysis of population structure. *bioRxiv* 454355; doi: <https://doi.org/10.1101/454355>.
- Volz, E.M., Koelle, K. and Bedford, T. (2013). Viral phylodynamics. *PLoS Computational Biology* 9(3), e1002947.
- Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1(1–2), 1–305.
- Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* 106(1), 67–72.
- Worby, C.J., Chang, H.-H., Hanage, W.P. and Lipsitch, M. (2014). The distribution of pairwise genetic distances: A tool for investigating disease transmission. *Genetics* 198(4), 1395–1404.
- Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, C., Fraser, C., STOP-HCV Consortium, Maela Pneumococcal Collaboration, and BEEHIVE Collaboration (2017). PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. *Molecular Biology and Evolution* 35(3), 719–733.
- Ypma, R.J.F., Bataille, A.M.A., Stegeman, A., Koch, G., Wallinga, J. and van Ballegooijen, W.M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings. Biological Sciences* 279(1728), 444–450.
- Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44(7), 821–824.

Reference Author Index

a

Aanensen, D.M. 1015

Abagyan, R. 974

Abbott, J. 789

Abdellaoui, A. 728, 732

Abdel-Rehim, A. 345

Abdi, K. 314

Abdo, H. 759

Abe, K. 80

Abecasis, G.R. 81, 109–111, 317, 593, 624, 628, 630, 649–650, 788, 793

Abel, L. 788

Aberg, K.M. 972

Abney, M. 594

Abraham, G. 833

Abrahamsson, S. 875

Abrams, K.R. 649

Absher, D. 943, 945

Absher, D.M. 271

Abubucker, S. 992

Abudahab, K. 1014

Açan, S.C. 317

Acevedo, N. 945

Acharya, C. 527, 529

Achaz, G. 415

Ackerman, H.C. 419

Acuna-Hidalgo, R. 788

Adachi, J. 361–362

Adai, A.T. 728

Adaktylou, F. 48

Adami, H.-O. 834

Adamian, M. 791

Adams, A.M. 268

Adams, D.R. 797

Adams, M.D. 416, 729

Adamson, B. 753–754

Adélaïde, J. 790

Adey, A. 754

Adhikari, K. 1016

Adusumalli, S. 797

Adzhubei, I. 361, 788

Aerts, J. 753, 796

Aerts, S. 753, 755

Aeschbacher, S. 47, 291

Affourtit, J. 291, 316

Afonso, L. 674

Agarwala, V. 627

Agostini, M. 793

Agrawal, A.F. 454

Aguadé, M. 417

Aguiar, D. 869

Aguilar, I. 810, 812

Ahituv, N. 796

Ahlquist, P. 874

Ahlström, T. 314

Ahmadi, K.R. 973

Ahmed, S. 650

Ahsan, H. 674

Aibar, S. 753, 755

Ainscough, R. 946

Airoldi, E.M. 872

Aitchison, J. 992

Aitman, T.J. 730, 873–874

Aittomäki, K. 674

Ajami, N.J. 993

Akaike, H. 213

Akalin, A. 942

Akashi, H. 393

Akbarian, S. 726

Akers, N.K. 726

Akey, J.M. 85, 290–293, 314, 320, 323, 629

Akhunov, E.D. 521

- Akhunova, A.R. 521
Akylbekova, E.L. 269
Ala-Korpela, M. 975
Albenberg, L. 993–994
Alberdi, M.T. 322
Albers, C.A. 77
Albert, J. 869
Alberts, R. 725
Albiges, L. 994
Albrecht, E. 674
Albrechtsen, A. 49, 76, 82, 272–273, 292, 314, 318, 320–323, 415, 492, 496–498, 594, 627, 629
Aldrich, M.C. 269
Alekseyenko, A. 995
Alesh, I. 674
Alessi, J. 320
Alexander, D.H. 268, 314, 492
Alexandre, M.J.J. 343
Alfaro, M.E. 215
Alfaro-Almagro, F. 110
Alföldi, J. 792
Alfredsson, L. 945
Ali, F. 729
Aliloo, H. 810
Alkan, C. 291–292, 316, 319, 321
Allaby, R.G. 522
Allan, W.P. 549
Allcock, R.J. 79
Allen, A.S. 788, 790–791, 794
Allen, D.R. 729
Allen, L.J.S. 47
Allen, N. 84, 629, 677
Allen, R. 319
Allendorf, F.W. 496
Allentoft, M.E. 273, 314–315, 318, 321
Allin, K. 836
Allison, D.B. 876
Allison, J.P. 993
Allman, E.S. 242
Almasy, L. 593
Al-Meeri, A. 270
Almeida, J. 79
Al Olama, A.A. 623
Alou, M.T. 994
Alousi, A.M. 993
Alpar, D. 754
Al-Ramahi, I. 731
Al-Rasheid, K.A. 322
Alt, K.W. 316, 319
Altemose, N. 77
Altmann, A. 625
Altmann, T. 528
Altmüller, J. 623
Altschuh, D. 342
Altschul, S.F. 342
Altschuler, S.J. 731
Altshuler, D. 79, 419, 593, 628, 676, 729, 788, 794
Alves, I. 319
Amadeu, R.R. 522
Amador, C. 492
Amanatides, P.G. 729
Amaral, L.A. 417
Amaria, R.N. 993
Amberger, J.S. 788
Ambrose, K.D. 946
Amendola, L.M. 788, 790
Amenga-Etego, L.N. 77
Amin, V. 757
Amir, E.-A.D. 754
Amit, I. 755
Ammerman, A.J. 314, 320
Amos, C. 269
Amoussou, J. 524
An, D. 529
An, H.J. 729
An, P. 792
Anbazhagan, R. 874
Ancrenaz, M. 495
Andau, P. 495
Anders, A. 315
Anders, S. 753–754, 872
Andersen, M.M. 548
Andersen, P.K. 319
Andersen, S.W. 674
Anderson, C.A. 623–624
Anderson, C.N.K. 314
Anderson, E.C. 76, 492–493
Anderson, J.A. 521
Anderson, M. 793
Anderson, O.D. 521
Andiappan, A.K. 732
Andolfatto, P. 171, 270, 1017–1018
Andreassen, O. 833
Andrés, A.M. 292, 319
Andrew, T. 595
Andrews, H. 794

- Andrews, M.C. 993
Andrews, T.D. 946
Andrews, T.S. 753
Andridge, R. 839
Andrieu, C. 869
Andrulis, I.L. 674
Andújar, C. 493
Ané, C. 244–245
Aneas, I. 76
Anex, D.S. 549
Anfora, A.T. 995
Angelis, K. 391
Angibault, J.M.A. 494
Angiuoli, S.V. 84
Angrist, J. 671
Anishchenko, I. 342
Anisimova, M. 367, 391
Anker, S.D. 676
Annicchiarico, P. 522
Ansari-Lari, M.A. 946
Ansielo, G. 365
Antaki, D. 788
Antanaitis-Jacobs, I. 314
Antao, T. 500
Anthony, D. 316, 319
Antolin, M.F. 495
Antonarakis, S. 77
Anton-Culver, H. 674
Antoniak, C.E. 213
Antonio, M. 293
Antoniou, A.C. 623, 647, 833–838
Anttila, V. 672
Anway, M.D. 725
Anzick, S.L. 273, 321
Aoi, T. 788
Apel, J. 323
Apetrei, C. 293
Apinjoh, T. 77
Aquadro, C.F. 171, 173–174
Aquino-Michaels, K. 726, 732
Aradhya, S. 946
Aran, D. 942
Arauna, L.R. 76
Archard, D. 568
Archer, D. 273
Archer, N. 753
Ardanaz, E. 676
Ardeshirdavani, A. 796
Arenas, M. 361
Arens, P. 523
Argmann, C. 730
Aris-Brosou, S. 213
Arjas, E. 869
Arkin, A. 972
Armagan, A. 869
Armitage, P. 623
Armour, C.D. 728, 731
Arnason, U. 213
Arndt, V. 674
Arndt, W. 342, 993
Arno, G. 788
Arnold, B. 1015
Arnold, S.J. 452–454
Aros, M.C. 80
Arriaza, B. 273
Arribas, A. 529
Arribas, P. 493
Arriola, L. 676
Arsuaga, J.L. 316, 319, 321
Arumugam, M. 994
Aryee, M.J. 942, 945, 948
Asai, K. 362
Aschard, H. 834
Ascher, D.B. 794, 796
Ashander, J. 80
Ashcroft, R.E. 568–570
Ashurst, J.L. 946
Ashwell, R.I.S. 946
Asimenos, G. 362, 789
Asimit, J.L. 626
Assimes, T.L. 729, 945
Ast, G. 419
Astle, W. 972–973
Astrand, M. 943
Atak, Z.K. 753
Athanasiadis, G. 319
Athersuch, T.J. 972
Atherton, R.A. 367
Atkinson, E.G. 420
Atlin, G.N. 522
Attie, A.D. 728, 872
Attwood, J. 943
Atzmon, G. 271
Auch, A. 320
Aud, D. 728
Auer, P. 837
Auinger, H.J. 522
Aulagnier, S. 494

- Aulchenko, Y.S. 625–626, 836
Aurell, E. 342, 1016
Ausiello, G. 344
Austin, J. 314
Auton, A. 76–77, 79, 82, 269, 272, 418, 788
Averof, M. 362
Avila, P.C. 268
Avila-Arcos, M.C. 272, 322
Avila-Campillo, I. 731
Aviv, A. 945
Avnit-Sagi, T. 993
Awadalla, P. 81
Aximu-Petri, A. 316, 319
Axton, R. 797
Aylward, W. 320
Ayme, S. 795
Ayodo, G. 318
Ayres, D.L. 216
Ayub, Q. 272
Ayyagari, R. 834
Ayyala, D. 943, 946
Azais, J.M. 522
Azeloglu, E.U. 725
Azencott, C.A. 625
Aziz, N. 795
Azizi, E. 757
- b**
- Baake, E. 140
Baake, M. 140
Babbage, A.K. 946
Babbitt, C.C. 84
Babiker, H.A. 318
Babron, M.C. 110
Babtie, A.C. 754
Babu, M. 346
Bacanu, S.-A. 111
Baccarelli, A.A. 943, 945, 947
Bach, K. 756
Bacharach, E. 393
Bacher, R. 753
Backer, J.A. 1018
Bäckhed, F. 996
Bacon-Shone, J. 992
Bae, K. 869
Baeriswyl, S. 84
Baetscher, D.S. 493
Bafna, V. 419
Bagger, F.O. 755
Baggerly, K. 945
Bagguley, C.L. 946
Baginska, J. 994
Bahlo, M. 171, 755
Bailey, A. 993–994
Bailey, L.R. 628
Bailey, R.A. 522
Bailey, S. 794
Bailliet, G. 318
Bainbridge, M. 790
Baird, D. 317
Baird, L. 549
Baird, S.J.E. 140
Bakelman, F.G. 942
Baker, D. 342–344, 1019
Baker R.J. 522
Bakken, T.E. 292, 320
Bakker, P.I. 498
Bakshi, A. 728, 733, 813
Balakrishnan, S. 342
Balanovska, E. 273, 318, 321
Balanovsky, O. 268, 273, 318, 321
Balascakova, M. 270
Balasubramanian, S. 793
Baldassano, R.N. 993–994
Baldassi, C. 342–343
Baldhandapani, V. 992
Baldi, P. 869, 973
Balding, D.J. 47, 49, 76, 144, 268, 272, 415, 493–494, 529, 548–549, 629, 676, 796, 813, 839, 972, 1014
Baldwin, N.E. 726
Baldwin-Pitts, D. 729
Bale, S. 795
Balkau, B. 676
Ball, E. V. 796
Ball, M. 317
Ball, P. 788
Ball, S.L. 495
Ballabio, A. 946
Ballantyne, C. 676
Ballarini, A. 995
Ballinger, M. 840
Balloux, F. 273, 321, 493, 495, 1017–1018
Ballouz, S. 754
Bancelis, C. 593
Band, G. 77, 79, 109–110, 269
Bandera, E.V. 269
Bandinelli, S. 943

- Bandyopadhyay, S. 758
Banerjee, R. 946
Banerjee, S. 869, 876
Banfield, J.F. 992
Banks, E. 77
Bansal, N. 677
Bao, L. 756
Bao, W. 834
Bapst, B. 811
Barabaschi, D. 522
Barabasi, A.L. 725, 730
Baragatti, M. 869
Barahona, M. 756
Baran, Y. 268
Baranski, M. 812
Barbacioru, C. 758
Barbato, M. 493
Barbe, V. 84
Barbeira, A.N. 725
Barber, D. 47
Barber, T. 522
Bard, M. 728
Bardel, C. 791
Barfield, R. 943
Bar-Joseph, Z. 756
Barker, G.E. 946
Bar-Lev, T.H. 756
Barlow, K.F. 946
Barnes, B. 943
Barnes, I. 273, 321
Barnes, K.C. 49–50, 83, 272
Barnett, R. 314, 319
Barnstead, M. 730
Baron, J. 314
Barracough, T.G. 497
Barrantes, R. 318
Barratt, B.J. 80
Barratt, E.M. 493
Barrdahl, M. 837, 839
Barrett, A. 973
Barrett, I.P. 946
Barrett, J.C. 629, 649, 835–836
Barrett, S.C.H. 522
Berrick, J.E. 417
Barroso, I. 730
Barrowdale, D. 836
Barsh, G.S. 271
Bartha, I. 789
Bartlett, G.J. 342
Barton, N.H. 140–144, 171, 173, 416
Barton, R.H. 973
Baryawno, N. 757
Bar-Yosef, O. 270, 317
Bashiardes, S. 992
Bason, J. 346
Bassas, L. 593
Bastarache, L. 293
Bastolla, U. 361–362
Basu, A. 756
Bataille, A.M.A. 1020
Bataillon, T. 530
Bateman, A. 342, 993
Bates, K.N. 946
Bates, P.A. 366
Bates, S. 992
Batista, S. 732
Batley, J. 530
Batto, J.-M. 994
Batzoglou, S. 362, 759, 789
Baudat, F. 76, 78
Bauer, E. 522, 528, 994
Bauer, M.W. 568
Bauer, V.L. 173
Baum, L.E. 47
Baur, M.P. 625
Bayes, T. 47, 213
Baying, B. 754
Baylink, D.J. 876
Bayzid, M.S. 244
Beall, B. 1015
Beare, D.M. 790, 946
Beasley, H. 946
Beasley, O. 946
Beaty, T.H. 49–50, 83, 272
Beauchamp, J. 873
Beaujard, P. 76
Beaulaurier, J. 727
Beaumont, M.A. 47, 76, 415, 493–494, 498, 1014
Beavis, W.D. 810
Beazley, C. 730
Bech-Hebelstrup, I. 324
Beck, A. 946
Beck, S. 944–945, 947
Becker, C. 270
Becker, D.M. 50
Becker, E. 992
Becker, L.C. 50

- Becker, T. 625, 629
Beckmann, L. 836
Beckmann, M.W. 674
Becskei, A. 753
Beddows, I. 290
Bedford, F. 522
Bedford, T. 1017–1018, 1020
Bedoya, G. 318
Beeby, M. 345
Beecham, G.W. 548
Beeghly-Fadiel, A. 674
Beerli, P. 213, 242, 453
Beeson, K.Y. 729
Beger, R. 973
Begun, D.J. 171
Behar, D.M. 268
Behnam, E. 943
Bejenaru, L. 270, 318
Bejerano, G. 795
Bekada, A. 76
Bekele, E. 272, 274
Belfer-Cohen, A. 270, 317
Belgrader, P. 759
Belkhir, K. 494
Bell, E.T. 213
Bell, G.W. 945
Bell, J.T. 943, 945
Bellemain, E. 493, 497
Bellenguez, C. 110
Bellows, S.T. 788
Bellwood, P. 315
Below, J.E. 629, 676
Belrose, J.L. 549
Belyeu, J.R. 788
Belzer, C. 995
Ben-Ami, H. 318
Benazzi, S. 323
Bendall, S.C. 754
Bender, C.A. 392
Bender, D. 83, 498, 595
Bendixen, C. 319
Bene, J. 318
Benedet, A.L. 675
Benes, V. 754
Bengio, Y. 343
Ben-Haim, N. 756
Benhamamouch, S. 76
Benitez, J. 674
Benito-Vicente, A. 789
Benjamini, Y. 623, 725
Benn, M. 671
Benn, P. 568
Benner, C. 109, 671
Benner, S.A. 363
Benner, T. 726
Bennett, C.A. 788
Bennett, D.A. 726, 732
Bennett, H.A. 728
Bennett, J.H. 141
Bennett, L.E. 812
Benson, M. 792
Bent, Z.W. 759
Bentley, A.R. 522, 524
Bentley, D.R. 82, 143, 947
Bentley, S.D. 1014–1015, 1019
Benvenisty, N. 947
Benyamin, B. 596, 630, 813, 840
Beral, V. 84, 677
Bercovitch, F.B. 493
Berestycki, N. 141
Berg, D.A. 758
Berg, J.J. 320, 415, 419
Berg, J.S. 791, 795
Berg, P. 530
Bergen, S. 112
Berger, B. 111, 271–272, 318, 345
Berger, J.O. 873, 875
Berger, J.P. 732
Berger, R.L. 47
Berger, S. 995
Bergman, N.H. 1019
Bergman, S. 272, 418
Bergström, A. 273
Berkovic, S.F. 788, 796
Berlin, K. 944
Berliner, J.A. 727
Berman, B.P. 945, 948
Bermúdez de Castro, J.M. 316
Bernal-Vasquez, A.M. 522
Bernard, A. 726
Bernard, V. 324
Bernardinelli, L. 672
Bernardo, R. 522–523
Bernatsky, S. 945
Berndt, S.I. 269, 594, 623, 626, 649, 836, 839
Berno, A. 595
Bernstein, B.E. 623
Bernstein, L. 269

- Béroud, C. 789
Berrada, F. 318
Berrigan, D. 453
Berriz, G.F. 727
Berry, D.A. 548
Berry, P.S.M. 493
Bersaglieri, T. 76
Bertalan, M. 994
Berthier, P. 493
Bertin, N. 727–728
Bertolino, A. 344
Bertranpetti, J. 269–270, 322
Berzi, P. 811
Berzuini, C. 672
Besansky, N.J. 419
Besenbacher, S. 79
Beskow, L.M. 571
Bessant, C. 973
Best, N. 496
Bethel, G. 946
Bethke, P.C. 529
Betsholtz, C. 726, 729
Beugin, M.-P. 1014
Bevan, A.P. 790
Beyene, Y. 523
Beyer, R. 315
Beyleveld, D. 568
Bhadra, A. 869
Bhai, J. 1014
Bhalala, O. 833
Bhangale, T. 77, 269
Bharadwaj, R. 760
Bhaskar, A. 268, 323
Bhatia, G. 112, 269, 813
Bhattacharjee, M. 869, 875
Bhattacharya, A. 869
Bhattacharya, K. 623
Bhutani, K. 730
Bi, Y. 869
Bialas, A.R. 756
Bianba, Z. 317
Bianchi, N.O. 394
Bibikova, M. 943
Biçakçı, E. 317
Bickhart, D.M. 813
Bicknell, L.S. 792
Biddick, K. 729
Bidet, P. 84
Biegert, A. 344
Bielas, J.H. 760
Bielawski, J.P. 363, 391
Biernacka, J. 839
Biesecker, B.B. 796
Biesecker, L.G. 788, 796
Bigdeli, T.B. 111
Biggs, D. 77
Biggs, P.J. 367
Biggs, W.H. 796
Biglari, F. 314
Bihan, M. 993
Bijma, P. 522
Bild, A. 873
Billamboz, A. 324
Billard, C. 324
Billaud, Y. 324
Binder, A.M. 945
Bindoff, L.A. 270
Bingen, E. 84
Bink, M.C. 813
Bird, A.C. 797, 944
Bird, C.P. 946
Birney, E. 291, 316
Birren, B.W. 992
Bishop, S.C. 494, 812
Biswas, S. 837
Bitbol, A.-F. 342
Bittinger, K. 992–994, 996
Bizet, M. 943
Bjorkegren, J.L. 726, 729
Björklund, A.K. 757
Blachly, J.S. 944
Black, L.J. 792
Blackburne, B. 789
Blackford, A. 837
Blackshaw, J. 676–677
Blackwell, T. 626
Blagrove, M.S. 972
Blainey, P. 992
Blaise, B.J. 972
Blake, J.A. 796
Blake, N. 521
Blanc, G. 522
Blanche, H. 292, 320
Blanche, P. 835
Blancher, C. 973
Blangero, J. 593
Blankenberg, S. 870
Blanquart, S. 213, 215, 362

- Blaser, M.J. 1016
Blattner, F. 874
Blechschmidt, K. 946
Blei, D.M. 869, 1014
Blencowe, B.J. 731
Blin, N. 317
Blischak, J.D. 759
Blöcher, J. 48, 314
Blockley, E. 529
Blondel, V.D. 754
Bloom, J.D. 362
Bloomer, A.C. 342
Bloomfield, C.D. 947
Bloomquist, E. 244
Blot, W.J. 269
Blows, M.W. 452, 454
Blüher, M. 270, 318
Blum, B. 947
Blum, M.G.B. 47, 322
Blum, S. 82
Blum, W. 947
Blumenstiel, B. 78
Blundell, T.L. 365, 367, 794, 797
Bo, T. 788
Bobo, D.M.D. 48
Bobrow, M. 838
Bocchini, C.A. 788
Bochkina, N. 869
Bock, C.H. 269, 754, 945
Bocquentin, F. 270, 318
Bodea, L.G.
Bodeau, J. 758
Boden, S.A. 530
Bodmer, W.F. 522
Bodo, J.-M. 269, 291
Boehm, C. 77
Boehnke, M. 317, 595, 626, 628, 630, 649
Boele, J. 995
Boer, J. 975
Boer, M.P. 523, 527
Boerwinkle, E. 269, 793, 840
Boettcher, C. 975
Bogdanova, N.V. 674
Bogunov, Y. 273
Boguski, M.S. 731
Bohlender, R.J. 321
Bohlin, J. 1015
Boichard, D. 810
Boisguerin, V. 317
Boitard, S. 271, 496
Bojesen, S.E. 674
Bolanos, R. 729
Bolla, M.K. 674
Bollback, J.P. 314, 415, 493, 1014
Bolliger, M. 324
Bollongino, R. 314–315, 318
Bolormaa, S. 810
Bolstad, B. 943
Bolund, L. 944
Bonacorsi, S. 84
Bonaguidi, M.A. 758
Bonamour, S. 452
Bonanni, B. 674
Bonasio, R. 755
Bonatto, S.L. 315
Bonazzola, R. 725
Bondarev, A.A. 290
Bondell, H.D. 872
Bonder, M.J. 992
Bone, W.P. 788
Bonhoeffer, S. 391, 1017–1018
Bonin, A. 493, 497
Bonin, M. 272
Bonne-Tamir, B. 268
Bonnin, D. 946
Bonomi, L. 797
Bonsall, D. 1020
Bontempi, G. 943
Bontrop, R.E. 82
Boocock, J. 726–727
Boomsma, D.I. 728, 732
Boone, E. 993
Boontheung, P. 731
Bordner, A.J. 362
Borglykke, A. 836
Borgwardt, K.M. 625
Bork, P. 994
Borlaug, N.E. 523
Borsboom, G. 836
Borsting, C. 548
Borstnik, B. 366
Bos, K.I. 318
Bosch, M.D. 323
Boskamp, B. 525
Boskova, V. 367
Botas, J. 731
Botha, G. 363
Botigué, L.R. 269

- Botstein, D. 417, 594
Bottinger, E. 293
Bottolo, L. 82, 111, 869–870, 873–874
Bottoms, C. 527
Bottou, L. 343
Botvinnik, O.B. 758
Boucher, K.M. 796
Boucher, L. 789
Bouchier, C. 84
Bouckaert, R.R. 242, 244
Boumertit, A. 81, 271
Bourke, P.M. 523
Boussau, B. 213, 215, 364
Boutselakis, H. 790
Bouvet, O. 84
Bovy, A. 975
Bowden, J. 648, 672, 675–676
Bowden, R. 76, 82
Bowler, E. 626, 793
Bowman, P.J. 525, 811–813
Boyd, A. 568
Boyko, A.R. 272, 416, 418
Boyle, E.A. 172, 416
Boyle, K.V. 321
Bradburd, G. 495
Bradbury, P. 527
Bradford, Y.M. 791
Bradley, D.G. 48, 77, 270, 314–315, 317
Bradley, J.R. 677
Bradman, N. 272, 274
Bradshaw, C.J. 494, 498
Brady, N. 946
Brajkovic, D. 291, 316
Bramanti, B. 314
Brand, J. 674
Brandariz, S.P. 523
Brand-Arpon, V. 84
Brandenburg, S.A. 944
Brandes, U. 972
Brandström, M. 319
Brandt, G. 316
Brandt, M.M. 795
Brant, S.R. 624
Bras, J.M. 270
Brauch, H. 674
Braun, J. 994
Braund, P.S. 648
Braverman, J.M. 416
Bravi, C.M. 318
Bravo, H.C. 945
Bravo González-Blas, C. 755
Brawley, O. 839
Bray, J.R. 992
Bray, N.L. 754
Bray-Allen, S. 946
Breeze, C.E. 947
Brekke, P. 499
Brem, R.B. 623, 725, 733
Brennand, K.J. 726
Brennecke, P. 754
Brenner, C.H. 548–549
Brenner, H. 674, 836
Brenner, J. 732
Breslow, J.L. 79, 593
Bressler, J. 943
Breton, C. 943
Brewster, D. 837
Brickman, A.M. 794
Bridgeman, A.M. 946
Briggs, A.W. 243, 291–292, 314, 316, 319, 321
Bright, J.A. 549–550
Brinkmann, B. 321
Brinkmann, H. 363, 366
Brinton, L. 835
Brion, M. 834
Brisbin, A. 268–269
Brisighelli, F. 318
Britton, A. 270
Broadhurst, D. 973
Brochmann, C. 493
Brockschmidt, F.F. 625
Brodmerkel, C. 730
Brody, J. 732
Brody, L.C. 78
Broet, P. 870
Bromham, L. 1014
Brondum, R.F. 810
Bronken Eidesen, P. 493
Brook, B.W. 494, 499
Brook, M. 837
Brooks, A.I. 732
Brooks, J. 993
Brooks, L.D. 795
Brookshaw, A. 529
Brookstein, F.L. 452
Broquet, T. 493–494
Brosseau, S. 994
Brotherton, P. 314

- Broushaki, F. 314
Brown, A.A. 730
Brown, A.J. 946
Brown, C.T. 992
Brown, D.J. 319, 810
Brown, I.J. 972
Brown, L. 497
Brown, M.A. 650
Brown, M.D. 593
Brown, M.J. 946
Brown, N.M. 1017–1018
Brown, P.J. 527, 870–871, 875
Brown, S.D.M. 788
Browne, A. 726
Browning, B.L. 76, 109, 269, 290, 314, 493, 593
Browning, S.R. 76, 109, 269, 290, 314, 493,
593, 626
Brownsworth, R. 568
Broxholme, J. 76
Broyden, C.G. 47
Brozell, A. 727
Brozynska, M. 727
Børresen-Dale, A.L. 674
Brucato, N. 76
Bruce, S.J. 973
Bruford, E.A. 946
Bruford, M.W. 493, 495
Brunak, S. 273, 314, 321, 994
Brüning, T. 674
Brunner, H. 493
Bruno, V. 870
Bruno, W.J. 362, 364
Bryan, G.J. 525
Bryant, D. 216, 242
Bryant, J. 835
Bryc, K. 268–269, 272, 292, 319, 418
Brynildsrød, O. 1015
Bryson, B.D. 755
Buard, J. 76
Bucan, M. 270
Buchan, D.W.A. 342–343
Buchanan, J.A. 626
Büchse, A. 528
Buchwald, M. 626
Buckler, E.S. 529–530
Buckleton, J.S. 548–550
Buckley, C. 875
Buckley, M.J. 946
Budowle, B. 549–550
Buechner, J. 793
Buenrostro, J.D. 315
Buetow, K.H. 77
Buettner, F. 754
Buffa, J.E.O. 993
Buhay, C. 946
Bühler, M. 756
Bühlmann, P. 416
Buigues, B. 320
Buil, A. 727, 730
Bulayeva, K. 76
Bulik-Sullivan, B.K. 111, 626, 672
Buljan, M. 345
Bull, J.J. 215
Bull, S.B. 648
Bullaugh, K. 174
Bulmer, M.G. 141, 452, 523
Bult, C.J. 796
Bumgarner, R.E. 733, 871
Bunce, M. 291, 321
Bundock, P. 524
Bundschuh, R. 943–944, 947
Bundy, J.G. 973
Buratto, B. 732
Burbano, H.A. 291, 293, 316, 318, 324
Burbrink, F.T. 243
Burcelin, R. 996
Burch, P. 946
Burchard, E.G. 268, 628
Burchard, J. 731
Burdick, J.T. 726
Burford, D. 946
Burgdorf, K.S. 994
Burge, C.B. 798, 872
Burger, J. 48, 314–315, 318
Burger, L. 342
Burger, M. 943
Bürger, R. 141, 452
Burgess, J. 946
Burgess, S. 628, 672–673, 675–677
Burgueño, J. 523, 526
Burke, M. 677, 835
Burks, T. 754
Burnell, M. 837
Burnett, J.E. 759
Burnham, K.P. 500
Burns, E. 271
Burren, A. 493
Burren, O.S. 674

- Burrill, W. 946
 Burrows, C. 946
 Burton, E.M. 993
 Burton, J. 943, 946
 Burton, P. 84, 677
 Burton, P.R. 568
 Burwinkel, B. 674
 Busby, G.B.J. 76–77, 79, 110, 269, 318
 Bush, R.M. 392
 Bush, W.S. 293, 623
 Bushman, F.D. 993–994
 Buske, O.J. 788
 Bustamante, C.D. 48, 268–269, 271–273, 317,
 321, 418, 420, 494, 627
 Bustos-Korts, D.V. 523, 527, 530
 Butler, A. 754–755
 Butler, C. 732
 Butler, D.G. 523
 Butler, J.M. 548
 Butte, A.J. 725
 Butterworth, A.S. 672, 676–677
 Butthof, A. 291, 316
 Buttner, M. 754
 Buxbaum, J.D. 726
 Buxbaum, S.G. 269
 Buyse, G. 791
 Buyx, A. 570
 Buzbas, E.O. 78, 417
 Byar, D. 835
 Bycroft, C. 77, 109
 Bye, J.M. 946
 Byers, P.H. 794
 Byrd, J.C. 944, 947
 Byrne, E.H. 83, 648
 Byrne, R.P. 77
 Byrnes, J.K. 268–269, 627
 Bystrykh, L.V. 725
- C**
 Caballero, A. 419, 493, 499
 Cabrera, V.M. 317
 Cacchiarelli, D. 759
 Caceres, R.M. 731
 Caesar, R. 996
 Caffo, B.S. 943
 Cai, G. 872
 Cai, N. 676, 813
 Cai, Q. 269
 Cai, S. 729
 Cai, T.T. 675
 Caldejon, S. 726
 Caleshu, C. 788, 792
 Cali, F. 318
 Caliebe, A. 270, 548
 Califano, A. 725
 Caligiuri, M.A. 947
 Callahan, B.J. 992
 Callahan, H.S. 454
 Calonne, E. 943
 Caltabiano, G. 366
 Calus, M.P.L. 811
 Calvert, W. 497
 Calzone, K. 835
 Camacho-González, J.M. 523
 Camarillo, G. 1019
 Cambien, F. 870
 Cambisano, N. 811
 Cameron, E.C. 498
 Camin, J.H. 1015
 Campbell, A. 270, 318
 Campbell, A.V. 568
 Campbell, F. 1015
 Campbell, H. 112, 837
 Campbell, K.R. 757
 Campbell, M.J. 975
 Campbell, R.H. 84
 Campbell, S.J. 419
 Campillo, M. 366
 Campion, G. 791
 Campos, P.F. 273, 321
 Cann, H.M. 270–271, 273
 Cannarozzi, G.M. 366
 Cannings, C. 141, 593
 Cannoodt, R. 754
 Cantarel, B. 995
 Cantor, C.R. 215, 364, 392
 Cantor, R.M. 731
 Canzian, F. 836
 Cao, J. 754, 994
 Cao, Y. 362, 992
 Capelli, C. 79, 110, 269, 318
 Caporaso, N. 269
 Capra, J.A. 83
 Caprioli, R.M. 996
 Capuani, G. 973
 Carbonell, E. 316, 319
 Carbonell, J. 342
 Carbonetto, P. 813, 841, 870, 877

- Carder, C. 946
Cardin, N.J. 82, 143, 173, 271, 418
Cardon, L.R. 109, 593, 623–624, 627, 630, 793,
 812
Cardona, A. 945
Cardoso, A. 497
Careau, V. 452
Carey, V. 943
Cargnelutti, B. 494
Carlson, C.S. 77, 623
Carlson, E. 975
Carlson, J. 796
Carlson, M.R.J. 317
Carlson, S. 725
Carlsson, E. 729
Carmel, L. 292
Carmi, S. 77, 85, 144
Carmo-Fonseca, M. 797
Carneiro, M.O. 797
Carneiro, M.S. 524
Carnes, M. 729
Carninci, P. 792
Caron, F. 870
Carpenter, J. 676
Carpenter, M.L. 315
Carr, A.J. 759
Carracedo, A. 549
Carrel, L. 946
Carrell, D.S. 293
Carreras-Carbonell, J. 493
Carretero, J.M. 316, 323
Carrick, M.J. 812
Carrigan, M. 77
Carroll, D. 673
Carroll, R.J. 726, 870
Carroll, S.B. 244
Carss, K.J. 788
Carstensen, T. 791
Carter, H. 788
Carter, N.P. 790, 946
Carter, P.A. 452
Cartwright, E.J.P. 1017–1018
Cartwright, R.A. 526
Carty, C.L. 943
Carvalho, B. 944
Carvalho, C.M. 870
Carvalho-Silva, D. 792
Carver, A. 729
Casadio, R. 342
Casale, F.P. 754
Casals, T. 593
Casane, D. 365
Casanova, J.-L. 788
Casas, J. 677
Casella, G. 47–49, 874
Casey, G. 269, 674
Caskey, C.T. 795
Caspari, R. 324
Casparino, A. 726
Cassidy, L.M. 48, 77, 270, 314–315, 317, 319
Castellani, C. 789
Castellano, S. 318–319
Castellanos, R. 727
Castellini, L.W. 725
Castillo, S. 974
Castle, J. 728, 731
Castle, K. 792
Casto, A.M. 271
Castro, V. 727
Cattivelli, L. 522
Cattuto, C. 84
Cauchemez, S. 1018
Caulk, P.M. 729
Cavalleri, G. 270, 318
Cavalli-Sforza, L.L. 213, 271, 314
Cavanagh, C.R. 529
Cavet, G. 731
Cavill, R. 973
Cayton, L. 625
Cedar, H. 947
Celli, J. 790
Cellon, C. 522
Center, A. 729
Cerami, E. 789
Cerezo, M. 626, 793
Cervino, A.C. 725
Cezairliyan, B.O. 394
Chabane, K. 791
Chadeau-Hyam, M. 869, 972
Chaix, R. 81
Chako, J. 946
Chakraborty, R. 85, 269
Chakraburty, K. 728
Chakravarti, A. 77, 626, 812
Chalifa-Caspi, V. 795
Challis, C.J. 362, 364
Chamaillard, M. 625
Chamberlain, A.J. 525, 812–813

- Chambers, J.C. 789
Chambert, K. 839
Chan, D. 872
Chan, K. 84
Chan, Q. 972, 974
Chan, T.E. 754
Chan, Y.L. 314
Chancerel, E. 324
Chandra, T. 756
Chang, B.S. 391, 394
Chang, E. 834
Chang, H.-H. 1020
Chang, J. 549, 870
Chang, M.A. 391
Chang, R. 725
Chang, S.C. 677
Chang, W.H. 676
Chang-Claude, J. 674
Chanock, S.J. 270
Chantratita, N. 1018
Chao, A. 992
Chao, K.R. 797
Chapela, R. 268
Chapman, J.C. 946
Chapman, J.M. 77, 109
Chapman, S. 523
Charcosset, A. 522, 526
Charleston, M.A. 213
Charlesworth, B. 141, 171, 174, 452
Charlesworth, D. 141, 171, 174, 416, 523
Charlesworth, J. 1016
Charlson, E. 992
Charmantier, A. 452
Chasman, D.I. 111, 624
Chatr-Aryamontri, A. 789
Chatterjee, N. 834, 838–840
Chaturvedi, A. 834
Chaturvedi, K. 730
Chau, L. 994
Chaubey, G. 268, 271
Chaudhuri, R. 83
Chavez, D. 946
Chavez, L. 945
Chawla, R. 758
Chazin, W.J. 996
Che, N. 731
Check, D. 837
Chehoud, C. 993
Chemaly, R.F. 993
Chen, A. 794
Chen, B.H. 943, 945
Chen, E.Z. 992–994
Chen, F. 111
Chen, G.-B. 834
Chen, G.K. 269
Chen, H. 80, 416, 623, 728, 755
Chen, J. 732, 754, 834, 943, 992, 996
Chen, K. 529
Chen, L. 624, 627, 729
Chen, M. 728, 872
Chén, O.Y. 943
Chen, P.L. 993
Chen, Q. 729
Chen, R. 731, 946
Chen, S.T. 525
Chen, S.X. 943
Chen, T.H. 732
Chen, T.J. 754
Chen, W.S. 993
Chen, Y. 530, 729, 753, 756, 794, 944, 946, 996
Chen, Y.H. 729
Chen, Z. 527, 946
Chenevix-Trench, G. 674
Cheng, C. 730
Cheng, H. 811
Cheng, J.Y. 319
Cheng, L. 1015
Cheng, L.-F. 869
Cheng, M.L. 729
Cherny, S.S. 109, 593
Cherrington, J.M. 726
Chesler, E.J. 726
Chess, A. 726
Chesser, R. 497
Cheung, V.G. 726, 729
Chevalet, C. 496
Cheverud, J.M. 452
Chevin, L.M. 452
Chew, E.Y. 626
Chewapreecha, C. 1015–1019
Chi, A. 677
Chia, J.M. 526
Chiang, C.C. 729
Chiang, C.W.K. 175
Chiaroni, J. 317
Chib, S. 869
Chibnik, L. 834
Chifman, J. 243

- Chikhi, L. 271, 315, 495, 497
Chikina, M. 362
Childe, V.G. 315
Chillán, M. 593
Chin, S.H. 729
Chinault, C. 946
Chines, P.S. 628
Chintalapati, M. 292
Chipman, H. 870
Chisholm, R.L. 293
Chiurugwi, T. 525
Chivall, D. 314
Cho, J.H. 648
Cho, M.Y. 753
Cho, R.J. 975
Cho, Y.S. 315, 673, 675
Choi, S.C. 362, 367
Chong, K.W.Y. 548
Chopin, L.K. 795
Chopin, N. 875
Chothia, C. 362
Choudhry, S. 628
Choudhury, P. 834
Christenfeld, N. 673
Christensen, O.F. 362, 810, 812–813
Christian, K.M. 758
Christiansen, F.B. 141, 392
Christiansen, M.W. 732
Christodoulou, J. 792
Chu, A.Y. 813
Chu, H. 673
Chu, L.-F. 753, 756
Chudin, E. 727, 731
Chun, S. 726, 789
Chung, D. 834
Chung, L.M. 870
Chung, S. 728
Chung, W. 732
Church, G.M. 728, 975
Churchhouse, C. 269
Churchill, G.A. 174, 362, 526, 731, 876, 974
Churnosov, M. 318
Chuzhanova, N. 77
Ciampi, A. 675
Ciccodicola, A. 797, 946
Cicek, A.E. 726
Cieslewicz, M.J. 84
Cimermancic, P. 992–993
Cipollini, G. 317
Cirulli, E.T. 789
Ciullo, M. 110
Ciurlionis, R. 732
Civit, S. 320
Claesens, J. 992
Clardy, J. 992–993
Clark, A.G. 418, 624
Clark, H.G. 291
Clark, M.J. 727
Clark, N.L. 362
Clark, R.T. 529
Clark, S.J. 943, 946
Clark, S.Y. 946
Clark, T.A. 944
Clark, T.G. 112, 974
Clarke, A.J. 797
Clarke, G.M. 273, 623–624, 626
Clarke, M. 675
Clarke, R. 322, 871
Clary, J. 322
Claustres, M. 593, 789
Clayton, D.G. 77, 109, 624
Clayton, S. 790
Clee, C.M. 946
Clegg, M.T. 521
Clegg, S. 946
Clémenson, C. 994
Clemento, A.J. 493
Clements, J. 993
Clements, M.N. 455
Clemments, J. 342
Clerc-Blankenburg, K. 946
Clerget-Darpoux, F. 594
Clerkson, B. 757
Clifford, K. 946
Clinton, R. 725
Clish, C.B. 994
Cloarec, O. 973–974
Clurman, B. 732
Clutton-Brock, T.H. 452
Clyde, M.A. 873
Cobelli, C. 83
Coble, M.D. 548
Cobley, V. 946
Cocco, S. 342
Cochran, W.G. 523, 530, 648
Cockerham, C.C. 493, 500, 548
Cockram, J. 522
Cocks, B.G. 810

- Codeluppi, S. 757
Coelho, M.J. 319
Coen, M. 974
Coffey, A.J. 946
Coffey, E. 728
Cogdill, A.P. 993
Coggill, P. 79
Cohain, A. 726, 729
Cohen, M.A. 363
Cohen, N. 993
Cohen, T. 676
Coifman, R.R. 758
Coin, L.J. 876
Cokkinides, V. 839
Cole, C.G. 946
Cole, D.E.C. 318
Cole, J.B. 813
Cole, M.B. 754
Cole, S.R. 673
Coleman-Derr, D. 521
Colicino, E. 943
Colijn, C. 1016–1018
Colinayo, V. 731
Collins, C. 362, 1015
Collins, F.S. 625, 789, 812
Collins, M. 273, 321
Collins, R. 792
Collis, C.M. 944
Collman, R. 992
Collod-Béroud, G. 789
Colmegna, I. 945
Colwell, L. 342, 344
Colwell, L.J. 343
Comadran, J. 527
Comai, L. 523, 527
Comas, D. 76, 268–270, 318
Combes, V. 526
Comeron, J.M. 391
Compher, C. 993
Comstock, K. 85, 500
Comstock, R.E. 523
Conesa, A. 754
Conley, E.J. 521
Conley, M.E. 788
Conlin, T. 791
Conneely, K.N. 944, 946
Connell, S. 270, 315, 317–318
Connelly, A. 82
Connelly, J. 836
Conner, J.K. 452
Connor, R.E. 946
Connor, S. 973
Connor, T.R. 758, 1015
Conomos, M.P. 623–624
Conquer, J.S. 946
Conrad, D.F. 269, 789, 795, 797
Consortium, M. 994
Constantin, S. 315
Contestabile, A. 875
Conway, M.E. 788
Cook, J.P. 624
Cook, R. 548
Cook, S.A. 730, 870, 874
Cook-Deegan, R. 796
Cookson, W.O. 80, 109, 593
Coombes, N. 523
Coombs, J.J. 529
Coop, G. 76–77, 83, 143, 269, 291, 320,
 415–420
Cooper, A. 314–319, 322
Cooper, D.N. 77, 788, 793, 796
Cooper, G.M. 362, 789, 792
Cooper, J. 839
Cooper, J.D. 77, 109, 649
Cooper, M. 523, 527
Cooper, R. 419
Cooper, Z.A. 993
Copel, B.R. 791
Coppieters, W. 811
Coppola, G. 293, 731
Corach, D. 318
Coram, M. 273, 1019
Corander, J. 493, 1015, 1019–1020
Corbin, L.J. 494
Corby, N. 946
Cordell, H.J. 80, 624
Corder, E.H. 624
Corri, A. 1015, 1017–1018
Corle, D. 835
Cornejo, O.E. 273, 321, 418
Cornelis, M.C. 677
Cornuet, J.M. 493–494, 496
Corpas, M. 790
Corrada-Bravo, H. 942
Correia, G. 972
Cortes, C. 416
Cortese, R. 943
Cosson, J.F. 494–495

- Costa, E.A. 524
 Costantino, J. 835
 Costa-Silva, J. 754
 Costello, J.F. 623
 Cotechini, T. 993
 Cotsapas, C. 83, 676, 726
 Couch, F.J. 625, 674
 Coulson, A. 947
 Coupland, P. 758
 Coussens, L.M. 993
 Coutinho, A. 322
 Couvet, D. 499
 Covarrubias-Pazaran, G. 523
 Cowlishaw, G. 499
 Cox, A.J. 110
 Cox, C. 77
 Cox, G.M. 523
 Cox, M.P. 76
 Cox, N.J. 392, 630, 725–726, 732, 812
 Cox, R.D. 834
 Cox, T.K. 626
 Cox, T.V. 943
 Coyne, M.D. 729
 Cozzetto, D. 343
 Crabtree, J. 83–84
 Craddock, N. 627
 Craig, A. 973
 Craig, O.E. 314
 Crainiceanu, C. 943
 Craiu, R.V. 650
 Crampton-Platt, A. 493
 Crandall, K.A. 391
 Crawford, D.C. 77, 269
 Crawford, J.E. 319
 Crawford, M. 273
 Cree, A. 946
 Creel, S. 494
 Crespi, B.J. 452
 Crews, J.D. 996
 Crochet, P.A. 452
 Crock, B.N. 571
 Crockford, D.J. 973
 Crook, D.W. 1016
 Cross, S.S. 674
 Crossa, J. 523, 526
 Crosslin, D.R. 293
 Crossman, C.C. 521
 Croteau-Chonka, D.C. 629, 726
 Croucher, N.J. 1015–1019
 Croutsch, C. 324
 Crow, J.F. 142, 453, 494
 Crow, M. 754
 Crowder, M. 729
 Crowell, S.L. 173
 Crowson, D. 522
 Croyle, R. 835
 Cuddy, W.S. 530
 Cuevas, J. 523
 Cui, H. 944
 Cui, W. 870
 Cui, Y. 1015
 Cule, M.L. 1016
 Cullen, M. 419
 Culleton, B.J. 317
 Cullis, B.R. 523, 529
 Culver, J. 839
 Cummings, B.B. 789
 Cundiff, P. 729
 Cunliffe, B. 81, 271
 Cunningham, A. 834, 836
 Cuo, Z.X.P. 420
 Cuomo, P.J. 728
 Cupp, A.S. 725
 Cupples, L.A. 269
 Curfman, J. 944, 947
 Curnow, R.N. 524
 Curran, J.M. 548
 Currat, M. 48, 270, 314, 317, 323
 Curtis, J.T. 992
 Cusanovich, D.A. 754
 Cushman, B.J. 790
 Cusick, M.E. 727
 Custádio, N. 797
 Cutler, D.J. 630
 Cuzick, J. 840
 Cvejic, A. 758
 Cybulski, J. 273
 Cywinska, A. 495
 Czabarka, E. 271
 Czedik-Eysenberg, A. 528
 Czene, K. 834
 Czyz, A. 945
- d**
- Dabney, J. 292, 319
 Dabrowski, P. 314
 Daetwyler, H.D. 492, 524, 527, 810–811, 813
 Dagan, T. 391

- Dagg, A. 493
 Dagnachew, B.S. 530
 Dagtas, N.D. 317
 Dahlke, C.E. 729
 Dahmer, N. 524
 D'hoe, K. 995
 Dai, H. 728, 731
 Dalen, L. 319
 Dalley, R.A. 726
 Dalmasso, C. 870
 Dalton, J.C. 797
 Daly, J.M. 1017–1018
 Daly, M.J. 79, 83, 112, 498, 593–595, 625, 628,
 672, 729, 794–795, 798
 Damba, L. 318
 D'Ambrosio, D. 732
 Damesa, T. 524
 Damgaard, P.D.B. 273
 Damine, P. 876
 Danaher, S. 729
 Danecek, P. 77, 110–111
 Danek, A. 77
 Danesh, J. 84, 676–677
 Dang, C.C. 362, 365
 Dang, K.K. 726
 Daniel, C.R. 993
 Daniels, M.J. 493
 Danko, C.G. 269, 316
 Dannemann, M. 290, 292–293, 320–323
 Darabi, H. 834
 Darden, T. 142, 173
 Darvasi, A. 497, 595
 Darwin, C. 213
 Darzi, Y. 995
 Das, A. 730
 Das, S. 110–111, 627, 649
 Daskalaki, E.A. 316–317
 Dastani, Z. 626
 Date, S.V. 728
 Datlinger, P. 754
 Datta, J. 869
 Datto, M. 792
 Daubin, V. 213
 Daucourt, J. 526
 Daughdrill, G. 345
 Daura, J. 320
 Davenport, L.B. 729
 Davey, J.W. 291
 Davey Smith, G. 624, 648, 672–673, 675–677
 David, E. 755
 David, J.L. 530
 David, M. 947
 David, O. 524
 David, R. 946
 Davidson, R.I. 626
 Davidson, W.S. 812
 Davies, B. 77
 Davies, J. 946
 Davies, M.A. 993
 Davies, N.M. 672, 675
 Davies, R.W. 77
 Daviglus, M.L. 972
 Davilla-Velderrain, J. 730
 Davis, B.P. 974
 Davis, C. 946, 992
 Davis, J. 794, 946
 Davis, J.R. 789
 Davis, K.L. 754
 Davis, L.G. 273
 Davison, A.C. 47, 875
 Davison, D. 81, 271
 Davoudi, H. 314
 Davuluri, R.V. 869
 Davydov, E.V. 789
 Dawson, E.T. 78
 Dawson, J.A. 873
 Dawson, K.J. 494
 Day, F. 674
 Day, F.R. 672–673, 945
 Day, J.W. 797
 Day, N. 834
 Day, T. 81, 271
 Dayhoff, M.O. 362–363, 391
 Day-Williams, A.G. 593, 626
 Daza, R. 754
 Deacon, F. 493
 Deadman, R. 946
 Deal, K.R. 521
 Deal, M. 273
 Dean, A.M. 392
 Dearlove, B.L. 1015
 Deary, I.J. 943, 945
 De Baets, B. 527
 De Bakker, P.I. 83
 Debelius, J. 993, 995
 DeCandia, T. 837
 de Castro, J.M.B. 319
 de Cesare, M. 1020

- De Commer, L. 995
Dedeurwaerder, S. 943, 945
Dee, N. 726
Deelen, P. 992
DeFelice, M.
Defesche, J.C. 789
de Filippo, C. 290–292, 318–320
De Finetti, B. 47
Defrance, M. 943
Degenhardt, J. 268
Degenhardt, J.D. 416
de Geus, E.J. 728, 732
DeGiorgio, M. 272–273, 321
Degnan, J.H. 242–245, 270, 324, 494
Degner, J.F. 730, 943
de Haan, G. 725
de Haan, J.R. 525
Dehghan, A. 840
De Iorio, M. 973–974, 1016
De Jager, P.L. 726
DeJong, T. 993
De Jongh, M. 273, 322
Dekkers, J.C.M. 525
Dekkers, K.F. 947
DeLacy, I.H. 530
Delaneau, O. 77, 82, 109–112, 624, 628, 730
Delaney, A.J. 1015
Delano, D. 943
de la Rasilla, M. 291, 316, 318
De La Vega, F.M. 321
Delcher, A.L. 729, 995
de Lencastre, H. 1015
De Leonardis, E. 342
Delgado, O. 946
Del Greco, F. 628, 672
Del Greco, M.F. 672
Della Casa, P. 314
Dellaportas, P. 871
Delmans, M. 754
Del Moral, P. 870
Deloukas, P. 82, 112, 143, 624
Delport, W. 363, 393
Delsate, D. 318
Delsuc, F. 363, 366
De Maio, N. 1016
DeMassy, B. 76, 78
Demerath, E.W. 943
Demeshchenko, S. 321
Deming, S.L. 269
Demirkale, C.Y. 728
De Moor, B. 794, 796
Dempster, A. 47
Den Dunnen, J. 975
Deng, J. 291, 755
Deng, M. 85, 876
Deng, Q. 754, 757
Deng, Z. 729
Dennis, J. 649
Denniston, C. 494
Denny, J.C. 83, 726
Deorowicz, S. 77
De Paoli, E. 527
Depaulis, F. 140, 416
DePrimo, S.E. 726
DePristo, M.A. 77, 794
Derenko, M. 273
Derevianko, A.P. 292, 319–323
Dermitzakis, E.T. 729–730
Dermitzakis, M. 728
De Roos, A.P.W. 811
Derosa, L. 994
Derry, J.M. 726
Der Sarkissian, C. 322
DerSimonian, R. 648
Desai, M.M. 417
Desai, N. 758
Deschenes, M. 569
Deschepper, C.F. 875
DeShazo, D. 946
Desiderio, F. 522
Desilets, R. 729
Desmet, F.O. 789
DeTomaso, D. 754
Deusch, O. 367
Deutsch, E. 994
Devilee, P. 674
Devillard, S. 495, 1014, 1017–1018
de Vlaming, R. 813
Devlin, B. 77, 548, 624, 726
Devlin, J.L. 729
Devos, J. 324
de Vos, W.M. 995
Dew, I.M. 729
deWaard, J.R. 495
Dewar, K. 293
Dewey, C.N. 756
Deych, E. 993
Dhami, P. 946

- Dhillon, B.S. 529
Dhindsa, R.S. 791
Dhombres, F. 795
Diab, A. 993
Diaconis, P. 47
Diamond, J. 315
Diaz, R. 77
Dib, C. 524
Di Camillo, B. 83, 755
Dickens, N.J. 730
Dickinson, S.P. 725
Didelez, V. 674–675
Didelot, X. 362, 1015–1019
Diekhans, M. 793
Diekmann, Y. 48, 314, 317
Dieterle, F. 973
Dieters, M.J. 530
Dietz, S.M. 729
Díez-Del-Molino, D. 48, 314, 317
Di Genova, G. 80
Di Lena, P. 342
Dilthey, A. 77, 82
DiMaio, F. 344
Dimas, A.S. 730
Dimitromanakis, A. 648
Dimmic, M.W. 363
Dimmock, D.P. 793
Di Narzo, A. 726, 730
Ding, J. 81, 111
Ding, Q. 291
Ding, S.L. 726
Ding, Y. 946
Dinh, H. 946
Dionne, D. 754
Di Renzo, A. 416
Di Taranto, M.D. 789
Diver, W.R. 269
Divers, J. 269
Dixit, A. 753–754
Dixon, L.E. 530
Dixon-Woods, M. 569
Do, C. 500
Do, K. 871
Do, R. 291, 319, 798, 813
Dobbyn, A. 726
Dobney, K. 319, 321
Dobra, A. 871
Dobrin, R. 727, 730, 732
Dodd, P.J. 1017–1018
Dodgson, J. 394
Dodson, K.L. 729
Dodsworth, S. 946
Doebeli, M. 141
Doebley, J. 524
Doebley, J.F. 524
Dogrusoz, U. 789
Doheny, K.F. 317
Doherty, J.B. 493
Dolan, M.E. 946
Dolbeare, T.A. 726
Dolzhenko, E. 943
Dombek, K.M. 733
Domboróczki, L. 315
Domchek, S. 835
Domenici, E. 726
Domingues, D. 754
Domoney, C. 530
Donati, C. 84
Dönertas, H.M. 317
Dong B. 528
Donia, M. 993
Donnelly, K.P. 593
Donnelly, M.T. 729
Donnelly, P. 49, 78–82, 110–112, 171–174,
271–272, 416, 497, 548, 627–629, 838, 873,
1019
Donnelly, P.J. 143
Donoghue, M.J. 391
Doré, J. 994
Dörk, T. 674
Doron-Faigenboim, A. 363, 393
Doronichev, V.B. 291–292, 316, 320
Dorrestein, P.C. 993, 995
Dorschner, M.O. 790
Dortch, J. 321
Doss, H. 216
Doss, S. 726–727
Dos Santos, H.G. 361
Douaud, G. 110
Doucet, A. 870
Douches, D.S. 529
Dougherty, E. 872
Douka, K. 290
Doup, L. 729
Douville, C. 788
Douze, M. 755
Dove, E.S. 568
Dow, M. 797

- Down, T.A. 943
Downey, P. 84, 677
Doyle, S.M. 273, 321
Draghici, S. 947
Dragoni, I. 875
Drake, J.A. 76
Drake, T.A. 726–729, 731
Draper, H. 946
Drees, B. 733
Dreher, B.P. 500
Dreisigacker, S. 523
Drenos, F. 625
Drineas, P. 628
Drobek, T. 526
Drögemüller, C. 493
Drouot, N. 524
Drummond, A.J. 213, 243–244, 1016
Dryomov, S. 318–319
Du, L. 316
Du, P. 943
Duan, J. 649, 839
Duarte, S.T.S.T. 792
Dubcovsky, J. 521
Dubreuil, P. 526
Dubuc, A.M. 791
Duceman, B.W. 549
Duchêne, D.A. 1014, 1016
Duchêne, S. 1014, 1016
Du Cheyron, P. 526
Dudbridge, F. 80, 624, 648, 650, 672, 835,
 839
Dudley, A.M. 728
Dudley, J.T. 726
Dudoit, S. 754, 758
Dueck, H. 755, 759
Duerr, R.H. 624
Duffy, D.L. 273
Duffy, P.R. 314
Duffy, S. 840
Dufour, J.M. 792
Dufresne, F. 524
Dugan-Rocha, S. 946
Dugas, M. 944
Duggal, G. 757
Duggan, A.T. 321
Dugoujon, J.-M. 318
Dullaghan, P. 729
Dumas, F. 526
Dumas, M.E. 973
Dumitrascu, B. 869
Dumont, M. 674
Duncan, L. 672
Duncavage, E.J. 792
Dunham, A. 946
Dunham, K.K. 492
Dunham, M.J. 796
Dunker, A. 345
Dunn, K.A. 363, 391, 788
Dunn, M. 316, 946
Dunn, S. 342
Dunning, A. 674
Dunning, A.M. 648
Dunson, D. 838
Dunson, D.B. 869
Dunstan, N. 110
Dunstone, N. 529
Duong, C.P.M. 994
Dupanloup, I. 47, 269, 315, 323
Dupont, W.D. 628
Dupuis, J. 728
Dupuy, D. 727
Duran, D.P. 497
Duran, J. 975
Durand, E.Y. 243, 269, 290–292, 315–316,
 321, 324
Durbin, K.J. 946
Durbin, R. 78, 81–83, 110, 112, 271, 273, 293,
 318, 322, 342, 418, 731, 875, 947
Durbin, R.M. 173
Durbin-Johnson, B.P. 549
Duret, L. 78, 213
Durheim, M.T. 794
Durkin, A.S. 84
Durrant, C. 526, 624
Durrett, R. 141
Dursi, L.J. 947
Dussex, N. 494
Dutkowski, G.W. 526
Dutta, I. 946
Dwyer, R. 342
Dybbs, M. 728
Dyer, K.A. 392
Dyer, T.D. 593
Dyke, S.O.M. 795

e

- Eade, R. 529
Eades, T. 946

- Earle, S.G. 1016
 Easton, D. 833–838
 Easton, D.F. 623, 647–648, 674
 Eaves, I.A. 80
 Ebbels, T.M. 972–975
 Ebel, A.V. 314
 Eberle, M.A. 77, 623
 Ebersberger, I. 243
 Ebrahim, S. 673, 677
 Echave, J. 363, 365
 Eck, R.V. 362, 391
 Ecker, J.R. 945
 Eckhardt, F. 943
 Economou, C. 316, 318
 Eddy, S. 342, 993
 Edlund, H. 316, 322
 Edmondson, A. 840
 Edmondson, R. 524
 Edsgård, D. 757
 Edwards, A.W.F. 213
 Edwards, D. 526, 530, 993
 Edwards, J.H. 593
 Edwards, M. 524
 Edwards, S. 243, 725, 729, 731
 Edwards, S.W. 731–732
 Edwards, T. 725
 Eckman, F.H. 795
 Effgen, S. 525
 Efron, B. 47, 213, 871
 Efron, M.J. 84
 Efstratiadis, A. 394
 Egeland, T. 498, 548
 Egger, M. 648
 Eggermont, A. 994
 Eggert, J.A. 494
 Eggert, L.S. 494
 Eggo, R.M. 1018
 Egholm, M. 291, 316
 Ehrenreich, I.M. 526
 Ehret, G.B. 624
 Ehrlich, S.D. 994
 Eichler, E.E. 291–292, 316, 318–321,
 788
 Eilbeck, K. 790
 Eisen, A. 835
 Eisen, E.J. 876
 Eisenberg, D. 365
 Eisenmenger, F. 367
 Eizenga, J.M. 78
 Ek, W. 837
 Ekeberg, M. 342, 344, 1016
 Ekong, R. 272
 Ekström, T.J. 945
 Ek, W. 837
 Elbers, C.C. 291
 Eldholm, V. 1015
 Eldon, B. 141
 Elefante, S. 836
 Elinav, E. 992–993
 Ellard, S. 790
 Ellington, M.J. 1017–1018
 Elliott, L.T. 77, 109–110, 1016
 Elliott, P. 84, 677, 972, 974
 Ellis, J.K. 972
 Ellis, S. 676
 Ellison, G.T.H. 569–570
 Ellwood, M. 946
 Elmohamed, M.A.S. 82
 Elo, L.L. 754
 Elofsson, A. 344–345
 Elowitz, M.B. 754
 Elsen, J.-M. 811
 Elshire, R.J. 525–526, 529
 Elston, R.C. 593–594
 Emery, J. 840
 Emery-Cohen, A. 946
 Emili, A. 728
 Emilsson, V. 732
 Enard, W. 760
 Encinas, J.M.V. 322
 Endelman, J.B. 529
 Endicott, P. 314
 Endo, T.R. 524
 Eng, C. 268
 Engel, J. 755
 Engelhardt, B.E. 730, 869
 Engelstad, M. 792
 Engen, S. 453
 England, P.R. 500
 English, G. 813
 Enikolopov, G. 758
 Enimil, A. 77
 Ennis, S. 78
 Enns, G.M. 790
 Enquobahrie, D.A. 732
 Epimakhov, A. 314
 Epstein, C.B. 945
 Erban, R. 758

- Erbe, M. 811, 835
 Erdal, Y.S. 317
 Erdmann, J. 648
 Eriksson, A. 270–273, 315, 317, 321
 Eriksson, G. 316
 Eriksson, K.F. 729
 Eriksson S. 570
 Ermel, R. 726
 Ermini, L. 322
 Ernfors, P. 759
 Ernst, D. 417
 Errington, H. 946
 Ersoz, E.S. 525
 Erwood, M. 788
 Escalante, L.E. 758
 Escobar, M. 871
 Escudier, B. 994
 Eshed, V. 270, 318
 Eskin, E. 81, 111, 594, 625–626, 648, 675,
 727
 Esko, T. 630, 728, 732, 813
 Esposito, L. 80
 Esteve, E. 996
 Estivill, X. 593
 Estoup, A. 494–495
 Estrada, K. 594, 813, 836
 Etheridge, A.M. 80, 140–142, 171, 317
 Ethier, S.N. 171
 Etter, P.D. 497
 Etxebarria, A. 789
 Evangelista, C.C. 729
 Evanno, G. 494
 Evans, B. 835
 Evans, C.A. 729
 Evans, D. 834
 Evans, K.L. 946
 Evans, S.N. 319, 322, 418–419
 Everett, B.M. 676
 Evett, I.W. 548–549
 Ewart, S.L. 728
 Ewens, K.G. 726, 729
 Ewens, W.J. 78, 172, 416, 453, 494, 593
 Ewing, G.B. 418
 Excoffier, L. 47, 49–50, 110, 269, 315, 624,
 1016
 Eyler, A.E. 726
 Eynaud, F. 324
 Eyre, D.W. 1016
 Eyre-Walker, A. 394, 416

f

- Facer, B.A. 726
 Fadhloui-Zid, K. 76, 269
 Fadista, J. 790
 Faggart, M. 78
 Faghri, F. 84
 Fagundes, N.J.R. 315
 Fairbairn, A. 317
 Fairfax, B.P. 732
 Faith, J.J. 730
 Falciani, F. 875
 Falconer, D.S. 453
 Fall, T. 672
 Fallin, M.D. 945
 Falony, G. 995
 Falush, D. 79, 81, 110, 269–271, 291, 316, 318,
 494–495, 1016–1019
 Fan, J.-B. 943
 Fan, Y. 49, 217
 Fang, H. 625
 Fang, X. 944
 Fanous, A.H. 111
 Farbmacher, H. 677
 Faridani, O.R. 757
 Fariselli, P. 342
 Farnham, P.J. 948
 Farnir, F. 811
 Farrell, J.A. 758
 Fasching, P.A. 674
 Fast, N.M. 364
 Fasulo, D. 729
 Faubet, P. 494
 Faulkner, L. 946
 Fauré, S. 524
 Fay, J.C. 78, 391, 416, 789
 Faye, L.L. 648
 Fear, J.M. 873
 Fearn, T. 870, 973
 Fearhead, P. 47, 78, 416, 496
 Feder, A.F. 315, 416
 Federer, W.T. 524
 Fedorova, S.A. 318, 321
 Fedosejev, A. 1014
 Fefferman, C. 272
 Feil, E.J. 1014, 1017
 Feil, R. 945
 Feinauer, C. 342, 1016
 Feinberg, A.P. 942–947
 Feingold, E. 873

- Feist, A. 993
Feldman, M.W. 273
Feller, W. 141
Felsenstein, J. 47, 80, 141, 213–215, 242–243,
321, 363, 391, 1016
Feng, D. 877
Feng, H. 944
Feng, L. 523
Feng, Q. 291
Feng, R. 993
Feng, S. 624
Feoktistov, V. 524
Feolo, M.L. 728
Férec, C. 77, 593
Fereday, D. 529
Ference, B.A. 674
Ferguson, J.P. 648, 870
Ferguson, N. 1015, 1017–1018
Ferguson, T.S. 214
Ferguson-Smith, A. 759
Fernandes, D. 270, 318–319, 418
Fernández, A. 498
Fernández, J. 492, 498
Fernandez, M.F. 366
Fernández, R. 322
Fernando, R.L. 525, 811, 836
Ferreira, M.A. 83, 498
Ferreira, T. 624–625, 630, 676, 813, 840
Ferrer-Admetlla, A. 47, 416
Ferrere, G. 994
Ferriera, S. 729
Ferrucci, L. 732, 943, 945
Ferry, M. 319
Ferrýao, L.F.V. 525
Fertig, E.J. 944
Ferwerda, B. 291
Fidelis, K. 365
Fidelle, M. 994
Fiehn, O. 974
Field, C. 363
Field, Y. 172, 416
Fiévet, J. 526
Figueroa, J. 674
Figueroa, M.E. 946
Filbin, M.G. 758
Fimmers, R. 548
Finak, G. 755
Finlay, K.W. 524
Finn, R. 342, 993
Finotello, F. 755
Finucane, H.K. 81, 111, 672, 674, 813, 835, 840
Firth, D. 49
Firth, H.V. 790, 797
Fischbach, M.A. 992
Fischer, A.P. 420
Fischer, G. 623
Fischer, J. 730
Fischer, K. 626, 648, 732
Fischl, B. 726
Fishelson, M. 593
Fisher, J.M. 758
Fisher, R.A. 47, 141, 172, 214, 416, 453, 524
Fisher, R.N. 496
Fitch, W.M. 214, 363, 392
Fitzjohn, R. 1015
Fizames, C. 524
Flachs, P. 78
Flack, J.M. 674
Flament, C. 994
Flanagan, S.E. 790
Flanigan, M.J. 729
Fledel-Alon, A. 76, 420
Fleiss, J.L. 648
Fleming, R.M.T. 994
Flesch-Janys, D. 674
Fletcher, O. 648, 674
Fletcher, R. 47
Flickinger, M. 317
Flint-Garcia, S. 527
Flournoy, N. 838
Fluckiger, A. 994
Fluder, E. 732
Flury, C. 493
Flusberg, B.A. 944
Flutre, T. 871
Flygare, S. 791
Flyger, H. 674
Fodor, A. 343
Foley, C.N. 673
Foley, R.A. 293, 321–322
Foll, M. 47–48, 315, 418, 494
Folland, C.K. 529
Fonseca, F. 676
Forbes, S.A. 790
Ford, C. 811
Ford, J.G. 268
Forejt, J. 78
Forer, L. 81, 110

- Fornage, M. 269, 943
Fornander, E. 318
Fornasari, M.S. 363
Forrest, A. 792
Forsberg, R. 392
Forster, J.J. 871
Forster, M. 317
Forster, P. 314
Fort, J. 315, 320
Fortea, J. 291, 316
Fortin, J.-P. 944
Fortun, M. 569
Fortunato, G. 789
Fortune, M.D. 674
Fortune, S. 755
Fosler, C.R. 729
Fost, N.C. 570
Foster, D.P. 871
Fostira, F. 674
Fourment, M. 393
Fowler, B. 757
Fowler, D.M. 796
Fox, C.S. 677
Fox, D. 797
Fox, G. 342
Fox, J. 453
Francescatto, M. 876
Franceschini, A. 796
Franciolini, L.C. 790, 792
Francis, F. 946
Francis, J.M. 758
Francken, M. 316, 318
Franco, O.H. 943
François, O. 47
Frank, J. 993
Franke, A. 317, 835
Franke, L. 674, 727, 732, 992
Frankham, R.R. 499
Frankhouser, D.E. 944, 947
Frankish, A. 793, 946
Frankland, J. 946
Franzen, O. 726, 729
Fraser, A.E. 946
Fraser, C. 1015–1018, 1020
Fraser, D.J. 500
Fraser, M. 316
Frayling, T.M. 732
Frazer, K.A. 629
Freedman, A. 835
Freedman, J.E. 728
Freedman, M.L. 624
Freedman, R.S. 975
Freeman, C. 77, 82, 109, 111, 271
Frei, K.M. 314, 318
Freimer, N.B. 50, 293, 594, 625
Freitag, D. 673, 676
Freitag-Wolf, S. 270
French, C.E. 795
Frere, J.M. 811
Freytag, S. 755
Fricke, W.F. 83
Friederich, S. 316
Friedlaender, J.S. 293
Friedman, C. 84
Friedman, E.S. 993–994
Friedman, J.M. 79, 593, 811, 973
Friedman, J.R. 730
Friedman, N. 48, 754
Friend, A.D. 273
Friend, S.H. 728, 731
Fries, A. 875
Frigge, M.L. 80, 110–111
Frikke-Schmidt, R. 671
Frisch, M. 529
Fritz, M.H.-Y. 291, 316
Froese, D. 497
Froguel, P. 172, 270, 318, 323, 416
Froment, A. 269, 291
Frontini, M. 945
Frost, S.D.W. 393, 1015, 1017–1018
Fruhwirth-Schnatter, S. 871
Fry, B. 83, 112, 419
Fu, J. 732, 992
Fu, L. 343
Fu, Q. 270, 290–293, 315–316, 318–321
Fu, W. 291
Fu, Y.-X. 78, 172, 416
Fuchsberger, C. 81, 110, 627
Fujita, M.K. 243
Fuks, F. 943
Fukui, Y. 730
Fulco, C.P. 754
Fullard, J.E. 726
Fuller, B.T. 273
Fuller, C.K. 727
Fuller, D.Q. 522
Fulton, R.S. 946
Fumagalli, M. 292

- Fung, H.C. 270
Fung, W. 996
Furey, T.S. 625
Furlan, A. 759
Furlan, S.N. 754
Furmanek, M. 314
Furumichi, M. 728
Furusawa, C. 757
Fusi, N. 871
Futema, M. 790
Futreal, P.A. 993
Futschik, A. 47, 418
Fuxreiter, M. 345
- g**
Gabai-Kapara, E. 835
Gabel, H.W. 728
Gabo, K. 944
Gabriel, S.B. 78, 83, 419
Gabrielian, A.E. 730
Gabrielson, E. 874
Gachotte, D. 728
Gaffney, D.J. 754, 943
Gaggiotti, O.E. 494
Gagliano Taliun, S. 796
Gagneur, J. 872
Gagnon-Bartsch, J.A. 944–945
Gail, M. 835, 838, 840
Gaiteri, C. 732
Gajer, P. 83
Galan, M. 494
Galgoczy, P. 243, 946
Galinsky, K.J. 269
Gall, A. 1020
Gallacher, J. 84, 629, 677
Gallager, R.G. 214
Gallais, A. 522, 526
Gallego-Llorente, M. 270, 315
Gallego Romero, I. 272, 321
Galleron, N. 994
Gallins, P. 732
Gallone, G. 795
Galloway-Pena, J. 993
Galtier, N. 78, 214, 363, 392, 416
Gamarra, B. 270, 318
Gamazon, E.R. 726
Gamba, C. 270, 315, 317, 319, 418
Gambaro, G. 110
Gammon, M. 674
Gan, L.M. 726
Gan, W. 729
Ganbat, J.-O. 944
Gandrillon, O. 757
Ganna, A. 790, 835
Gansauge, M.-T. 291, 319–322
Gao, C. 944
Gao, H. 494
Gao, J. 789
Gao, S. 944
Gao, X. 756
Gao, Z. 172, 416, 944
Garber, A. 570
Garcia, A.A.F. 522, 524–525, 529
Garcia, M.O. 524
Garcia, T. 725, 732
Garçia Borja, P. 320
Garcia-Closas, M. 625, 674, 836, 840
Garcia-Perez, I. 974
Garçia-Portugués, E. 363
Garçia-Ruiz, A, Cole, J.B. 811
Gardete, S. 1017–1018
Gardiner-Garden, M. 944
Gardner, E.J. 796
Gardner, J.M. 993
Gardner, M.J. 995
Gardy, J. 1016
Garey, M.R. 727
Garg, A. 549
Garg, N. 993, 995
Gargalovic, P.S. 727
Garimella, K. 112
Garland, T. 452, 1018
Garner, C. 648
Garrels, J. 974
Garrett, E. 874
Garrett-Engele, P.W. 731
Garrick, D.J. 811, 836
Garrigan, D. 315
Garrison, E. 78, 82
Garry, D. 756
Gartner, J.J. 791
Garud, N.R. 78, 417
Garza, J.C. 493
Garzon, R. 947
Gasch, A.P. 753
Gascuel, O. 362, 365, 1017–1019
Gaskell, G. 568
Gaskell, P.C. 624

- Gasparian, B. 270, 318
Gatesy, J. 243
Gattepaille, L.M. 273, 322
Gauch, H.G. 524, 529
Gaudet, R. 83
Gaugier, D. 973
Gaulton, K.J. 172, 416, 625, 676
Gaut, B.S. 216, 365, 394, 418, 524
Gayet, T. 1014
Gaynor, C.R. 524
Gazaffi, R. 524
Gazina, E.V. 793
Ge, H. 974
Ge, W. 625
Gebregziabher, M. 649
Geiger, D. 593
Geiger, H.H. 522
Gelernter, J. 834, 837
Gelfman, S. 790
Gelfond, J.A. 871
Gelman, A. 214, 495, 648, 871
Gelsi-Boyer, V. 790
Geman, D. 945
Genin, E. 594
Gennert, D. 758
Genovese, G. 790, 813
Genschoreck, T. 272, 292, 320
Gentile, M. 789
George, A.W. 525
George, E.I. 870–871, 873
George, R.J. 317
George, V. 876
Georges, M. 811
Georgiev, S. 269
Gerard, D. 525
Gerbault, P. 80
Gerds, T.A. 835
Gerety, S.S. 790
Germer, S. 728
Gerrish, P.J. 141
Gerritsen, F. 48, 270, 318–319
Gershenwald, J.E. 993
Gerstein, M. 728, 759
Gersuk, V. 755
Gervasio, F. 345
Geschwind, D.H. 729, 731
Gessner, A. 995
Gether, U. 270
Geurts, P. 417, 753
Geyer, S. 947
Ghahramani, Z. 872
Ghalichi, A. 317
Ghandour, G. 595
Gharavi, N.M. 727
Gharib, S.A. 732
Ghazalpour, A. 727
Ghiringhelli, F. 994
Ghirotto, S. 496
Gho, C. 523, 527
Ghosh, A. 648
Ghosh, D. 870
Ghosh, S.K. 872
Giacobbe, C. 789
Giagtzoglou, N. 757
Giambartolomei, C. 674, 726–727
Giancarlo, R. 78
Giannarelli, C. 726
Gianola, D. 727
Gibbons, S.M.C. 569
Gibbs, H.L. 244
Gibbs, R.A. 112, 947
Gibbs, R.J. 270
Gibson, G. 728
Gibson, R. 79
Gidding, S.S. 789
Gieger, C. 270, 943
Giemsch, L. 315
Gierahn, T.M. 755
Gignoux, C. 268, 324, 837
Gignoux, C.R. 627
Gika, H.G. 973
Gilad, S. 756
Gilad, Y. 730, 759, 873, 943
Gilbert, E. 78, 270, 318
Gilbert, J.A. 993
Gilbert, J.G. 79
Gilbert, K.J. 495
Gilbert, L.A. 753
Gilbert, M.T.P. 315, 319–323
Gilbert, P. 453
Gilbert, W. 394
Gilchrist, A.S. 498
Giles, G.G. 674
Gilissen, C. 788
Gilks, W.R. 214
Gill, B.S. 524
Gill, P. 548–549
Gill, R. 946

- Gillanders, E.M. 269
Gillespie, J.H. 392
Gillingwater, T.H. 321
Gillis, J. 754
Gillis, T. 732
Gilmour, A. R. 525
Gilmour, S.G. 523, 527
Gilroy, E. 993
Ginolhac, A. 315, 317, 322, 497
Giraud, B. 343
Girdhar, K. 726
Gire, H.C. 729
Gittelman, R.M. 290, 293, 323
Gittelson, S. 549
Giugliano, R.P. 795
Gkatzionis, A. 673
Glasner, C. 1014
Glasner, J. 995
Glaubitz, J.C. 526, 529
Glaus, P. 871
Glazier, A. 873
Glazko, G. 216
Glazner, C.G. 593
Glessner, J. 85, 113, 269
Glicksberg, B.S. 726
Gliedt, T.P. 628
Glimelius, B. 834
Glitza, I.C. 993
Glöckner, G. 946
Glodek, A. 729
Gloor, G. 342
Gluecksmann, A. 729
Glymour, M.M. 677
Gnirke, A. 945, 948
Go, M.J. 676
Goater, R.J.E. 1014
Gobel, U. 343
Gocayne, J.D. 730
Godambe, V.P. 48
Goddard, M.E. 271, 273, 495, 525, 596, 626,
 648, 733, 810–813, 874
Godfrey, P.A. 992
Godley, L.A. 946
Godzik, A. 343
Goebel, T. 273, 293, 322
Goel, A. 649
Goemans, N.M. 791
Goff, L. 875
Gog, J.R. 1017–1018
Gogele, M. 628
Gojobori, T. 392, 394–395, 418
Gokhman, D. 292
Golan, D. 172, 416, 835
Gold, J.R. 495
Goldberg, A. 316
Goldberg, D.S. 727
Golden, B.L. 811
Golden, M. 363
Goldfarb, D. 48
Golding, G.B. 364, 392
Goldman, M. 756
Goldman, N. 214, 216–217, 344, 363–367,
 392–395, 417, 420
Goldstein, D.B. 789, 791–794, 812
Goldstein, M. 1017–1018
Goldstein, R.A. 363–364, 392–393
Goldwasser, F. 994
Golovanova, L.V. 291–292, 316
Golub, T.R. 729
Golubchik, T. 1020
Gomez-Cabrero, D. 754
Gomez-Zurita, J. 497
Gonçalves, A. 876
Gong, L. 530
Gong, W. 756, 992
Gonin, P. 994
Gonnet, G.H. 363, 366
Gonzalez, A. 366
González-Blas, C.B. 753
Gonzalez-Fortes, G. 270, 315–318
González-Pérez, A. 791
González-Recio, O. 810, 839
Good, B.H. 417
Good, J.M. 291–292, 316, 321
Goodacre, R. 973
Goode, D.L. 789
Goodman, A.H. 569
Goossens, B. 495
Gopalakrishnan, V. 993
Gopalan, S. 48
Gordeev, V. 837
Gordillo, A. 522
Gordon, M. 529
Gordon, N.C. 1016
Gordon, S. 596, 813, 840
Gore, M.A. 525
Goremykin, V.V. 367
Gorjanc, G. 524–525

- Gorman, D. 673
Gort, G. 530
Gosman, N. 522
Gosset, W.S. 525
Goswami, S. 811, 835
Götherström, A. 316–320, 323
Gottardo, R. 755, 871, 875
Gottelli, D. 493
Göttgens, B. 756
Gouda, N. 757
Goudet, J. 493–494, 548, 550
Gough, J. 345, 876
Gould, B.A. 324
Gould, M.N. 873
Goulian, M. 994
Gourna, E.G. 571
Gouy, M. 213, 364
Govind, R. 797
Grabowski, S. 77
Gracia, A. 319
Graefen, A. 317
Graf, E. 795
Graf, K.E. 320
Graffen, A. 453
Graham, D.V. 946
Graham, C. 549
Graham, J. 549
Gralak, T. 314
Grallert, H. 732
Grandke, F. 525
Grant, A. 993
Grantham, R. 363
Grarup, N. 627
Graubard, B. 834–835
Graur, D. 214, 391
Gravel, G. 269
Gravel, S. 269, 271, 291, 315, 627
Gravenor, M.B. 363
Gray, A.N. 753
Gray, J. 730
Gray, M. 731
Gray, R.J. 872
Graybeal, A. 214
Graze, R.M. 873
Greally, J.M. 945
Grealy, A. 291
Greco, M. 674
Gredler, B. 811
Greely, H.T. 569
Green, A.R. 756
Green, E.D. 362, 789
Green, E.J. 291
Green, J. 84, 677
Green, P. 364, 875
Green, P.J. 78, 214, 495, 874
Green, R.C. 791
Green, R.E. 243, 291–292, 314, 316, 319–321
Greenawalt, D.M. 727
Greenbaum, D. 728
Greenblatt, J.F. 728
Greenblatt, M.S. 796
Greene, D. 791
Greene, E.A. 527
Greene, M. 838
Greenhalgh, K. 994
Greenland, S. 628, 673–674
Greenwood, C.M.T. 626, 630, 675
Gregg, A. 794
Gregg, M. 270, 318
Gregorová, S. 78
Gregory, M.T. 759
Gregory, S.G. 946
Greiner, E.R. 731
Greminger, M.P. 497
Grenfell, B.T. 1018
Grever, M.R. 947
Grey, C. 76
Grgicak, C.M. 549
Gribble, S. 946
Gribkova, S. 758
Grieder, C. 528
Griffin, J.E. 871–872
Griffin, J.L. 974
Griffin, S. 840
Griffiths, A.M. 993
Griffiths, C. 946
Griffiths, D. 1016
Griffiths, R.C. 49, 78, 82, 171–172, 416–417,
593
Grigg, G.W. 944
Grigorieff, N. 346
Grills, G.S. 525
Grimes, V. 273
Grimm, C. 945
Grimmett, G.R. 392
Grimsby, J. 759
Grisart, B. 811
Grishin, N.V. 343, 363

- Grobler, J.P. 293
Grocock, R. 946
Grody, W.W. 791
Gromov, A. 314
Gronau, I. 83, 269, 316, 318, 419
Gronkiewicz, S. 314
Groop, L.C. 729
Gropman, B. 729
Gross, C.A. 753
Gross, T. 346
Grosse, S. 835, 839
Grossen, C. 497
Grosshennig, A. 648
Grossman, M. 811
Grote, S. 292–293, 323
Groussin, M. 364
Grove, M.E. 788
Gruber, M. 345
Gruber, S. 836
Grün, D. 755
Grupe, A. 728
Grupe, G. 314
Grusea, S. 271
Gsponer, J. 345
Gu, H. 948
Gu, J. 726, 871
Gu, W. 79, 798
Gu, X. 392
Gu, Y. 758, 946
Gu, Z. 729
Guan, P. 729
Guan, W. 943
Guan, Y. 269, 291, 871
Guardamagna, O. 789
Guarner, F. 994
Gubina, M. 273, 318
Gudbjartsson, D.F. 79–80, 110–111
Gudjonsson, S.A. 79
Gudmundsdóttir, V. 273, 321
Gudnason, V. 732
Guénel, P. 674
Guengerich, F.P. 731
Guerra, M.A.R. 319
Guerreiro, L. 526
Guerreiro, R. 270
Gueudré, T. 343
Guhathakurta, D. 731
Guichoux, E. 324
Guidon, N. 273
Guigo, R. 729
Guill, K.E. 526
Guillaume, J.-L. 754
Guillaumet-Adkins, A. 760
Guillemaud, T. 494
Guillén, S. 315
Guillermin, Y. 791
Guillot, G. 494–495
Guillozet-Bongaarts, A. 726
Guindon, S. 392
Guinet, J.-M. 318
Guinovart, J.J. 975
Guldbrandsen, B. 810
Guldener, U. 727
Gulko, B. 269, 316
Gullberg, A. 213
Gumus, Z.H. 726
Gunaratne, P. 946
Gunawan, R. 757
Gunderson, K.L. 943
Günther, T. 316, 318, 320–323
Guo, C. 1015
Guo, G. 757
Guo, H. 672, 674
Guo, J. 757, 973
Guo, M. 755
Guo, N. 365
Guo, S.T. 525
Guo, X. 292, 321
Guo, Y. 674
Guo, Z. 675
Gupta, M. 871
Gupta, N. 112
Gupta, R. 273, 321
Gur, R.E. 726
Gur, T. 630
Gurdasani, D. 111, 628, 791, 794
Gurwitz, D. 268
Guschanski, K. 322
Gusev, A. 79, 112, 593, 672, 813, 835, 840
Gusfield, D. 83
Gusic, I. 316
Gusmao, L. 548
Gusnanto, A. 624, 648
Gussow, A.B. 791, 794
Gustafsson, S. 813
Gut, I.G. 945
Gut, M. 945
Gutenkunst, R.N. 48, 269, 291

- Guthrie, S. 82
 Gutierrez-Achury, J. 839
 Gutierrez-Arcelus, M. 729
 Gutmann, M.U. 1015
 Guttmacher, A.E. 625
 Guttman, B. 548
 Guy, D.R. 812
 Gwilliam, R. 112, 946
 Gylfason, A. 79–80, 111
- h***
- Ha, M.J. 732
 Haag, J.D. 873
 Haak, W. 314, 316–319, 322
 Haas, D.W. 112
 Haas, J. 317
 Habier, D. 525, 811, 836
 Hackett, C.A. 525
 Hadfield, J.D. 453–454
 Hadly, E.A. 314
 Haeckel, E. 214
 Haefliger, C. 943
 Haeggström, J. 759
 Haesaert, G. 527
 Haffner, P. 343
 Hager, J. 875
 Haghverdi, L. 755
 Hahn, C.G. 726
 Hahn, L.W. 628
 Hahn, M.W. 419
 Haidle, M.N. 314
 Haigh, E. 549
 Haigh, J. 81, 143, 175, 419
 Haile-Mariam, M. 812
 Haiman, C.A. 270, 674
 Haimes, E. 569
 Haines, J.L. 624
 Hajdinjak, M. 292, 315
 Hajdu, T. 314
 Hakonarson, H. 85, 113, 270
 Halakivi-Clarke, L. 871
 Haldane, J.B.S. 141, 594
 Haley, C.S. 144, 812
 Hall, I. 82, 112
 Hall, J. 836
 Hall, M. 1017–1018, 1020
 Hall, P. 323, 674
 Hall, T.O. 549
 Hallauer, A.R. 525
- Halldorsson, B.V. 79
 Hallgren, F. 316
 Halls, K. 79
 Halperin, E. 81, 111, 273, 419, 630
 Halpern, A. 364, 729
 Halpin, C. 527
 Halstead, P. 48
 Halvorsen, M. 794
 Hamady, M. 994–995
 Hamamsy, T. 726
 Hamann, U. 674
 Hamelryck, T. 363
 Hamilton, A. 812
 Hamilton, C. 946
 Hamilton, R.S. 345
 Hamilton, W. 840
 Hammer, M.F. 268, 291–292, 315, 319, 324, 550
 Hamosh, A. 788, 796
 Hamoudi, R.A. 945
 Hamroun, D. 789
 Han, A.W. 992
 Han, B. 525, 648, 975
 Han, C. 675
 Han, F. 529
 Han, H. 947
 Han, J.D. 727
 Han, L. 594
 Han, Y. 794
 Hanage, W.P. 1015, 1020
 Handa, K. 362
 Handley, M. 1019
 Handsaker, R.E. 77
 Haned, H. 549
 Hanghoj, K. 316–317, 321
 Hanihara, T. 321
 Hans, C. 871
 Hansen, K.D. 942, 944, 947–948
 Hansen, N.F. 291, 316, 791
 Hansen, T.V. 82
 Hansson, M.G. 569
 Hansson, O. 790
 Hansson M.G. 570
 Hao, J. 973
 Hao, K. 726, 730–731
 Hao, L. 271
 Hao, T. 727
 Haque, A. 755
 Harada-Shiba, M. 789

- Harbord, R.M. 624, 676
Hardcastle, T.J. 872
Hardiman, O. 77
Harding, R.M. 78
Hardy, G.H. 48
Hardy, J.A. 270, 324
Hardy, O.J. 495
Hardy, R. 834
Harkins, T.T. 317
Harmelin, A. 993
Harney, E. 270, 316, 318, 322, 418
Harold, D. 837
Haroutunian, V. 726
Harrington, E.A. 788
Harris, D. 796
Harris, J. 569
Harris, K. 79, 83, 142, 144, 272–273, 316, 495
Harris, S. 1015
Harris, S.R. 758, 1015–1018
Harrison, J. 943
Harrison, S.M. 788, 796
Harrow, J.L. 79
Hart, A.A. 731
Hart, A.H. 757
Hart, B. 730
Hart, E.A. 946
Hart, J. 497
Hart, K.W. 731
Hartemink, A.J. 759
Hartge, P. 834–835, 840
Hartigan, J.A. 392
Hartl, C. 797
Hartl, D.L. 291, 419
Hartonen, T. 1016
Hartwell, B. 315
Hartwig, F.P. 675
Harvig, L. 314
Harwood, W. 530
Hasegawa, M. 214, 361–362, 367, 392, 395
Haseman, J.K. 594
Hasin, Y. 791
Hastie, D.I. 869
Hastie, T. 811, 973
Hastings, A. 142
Hastings, W.K. 48, 214
Hatcher, S. 810
Hatfield, G. 973
Hatoum, I.J. 727
Hatrack, K. 345
Hatta, M.A.M. 530
Hatton, E. 77
Hauberg, M. 727
Haubold, B. 419
Hauser, A. 344
Hauser, S.L. 269
Hausser, J. 493
Haussler, D. 82, 366, 625, 795
Havlak, P. 946
Havrilla, J.M. 788
Havulinna, A.S. 109, 671
Hawes, A. 946
Hawks, J. 324
Hawrylycz, M.J. 726
Hayamizu, S. 362
Hayashi, T. 757
Hayashizaki, Y. 792
Haycock, P.C. 672
Hayden, M. 796
Hayden, M.J. 524
Haydu, L.E. 993
Hayes, B.J. 273, 492, 495–496, 524–527,
 810–813, 874
Hayes, P.M. 529
Hayette, S. 791
Haynes, C. 729
Haynes, J. 730
Hayward, C. 76
Hazell, S. 497
Hazes, B. 362
He, A. 727
He, L. 730
He, M. 317
He, X. 727
He, Y.D. 728, 731
Heard-Costa, N. 728
Hearne, S. 529
Heath, A.C. 596, 813
Heath, P.D. 946
Heath, S.C. 594, 945
Heath, T.A. 215
Hebbring, S.J. 293
Hebenstreit, D. 753
Hebert, C. 758
Hebert, P.D.N. 495
Hébert, R. 836
Hebestreit, K. 944
Hechter, E. 798, 841
Heckel, G. 1016

- Heckerman, D. 626
Hedgecock, D. 498
Hedin, U. 729
Hedrick, P.W. 79, 495–496
Heeney, C. 569
Heffner, E.L. 525
Hegele, R.A. 789
Heger, A. 758
Heijmans, B.T. 947
Heiman, T.J. 729
Hein, J. 85, 144, 172, 175, 363–364
Hein, M.Y. 753
Heinig, M. 874
Heinken, A. 994
Heintzman, P.D. 273, 321
Heinze, A. 292, 320
Heinze, G. 1017–1018
Heinzen, E.L. 794
Heisler, M.G. 754
Heitmann, K. 946
Held, L. 48, 872
Heled, J. 243
Helfer, J. 48
Helgason, A. 730
Helgason, H.H. 569
Hellenthal, G. 48, 76–79, 81, 110, 269–271,
 274, 314, 318, 1018
Heller, R. 497
Hellman, A. 942
Hellmann, I. 292, 320, 760
Hellmann, M.D. 994
Helmer-Citterich, M. 344, 365
Hemani, G. 624, 813, 839
Hemberg, M. 753–754, 756, 759
Hemby, S.E. 726
Hemphill, S.E. 790
Henders, A.K. 596, 813
Henderson, B.E. 270
Henderson, C.R. 453, 594
Hendriks, M.M. 973
Henikoff, S. 365, 527
Henn, B. 317
Henn, B.M. 269, 318, 418, 420
Hennig, S. 946
Hennig, W. 215
Henrich, S. 454
Henrickson, C. 273
Henriksson, J. 756
Henry, R.J. 524
Herbig, A. 293
Herbots, H.M. 172
Herder, C. 732
Herman, J.L. 364
Hermanson, L. 529
Hermisson, J. 142, 417
Hernán, M.A. 675, 677
Hernandez, D.G. 270, 732, 943
Hernandez, J. 946
Hernandez, R.D. 48, 416, 594
Herold, C. 625
Herrero, J. 945
Herrington, D. 627
Hershberg, R. 417
Hervig, T. 318
Herwig, R. 945, 972
Herzig, A.F. 110
Heslot, N. 524–525
Hetrick, K.N. 317
Heussner, K.U. 324
Heutink, P. 876
Heward, J. 80
Hewison, A.J.M. 494
Hey, J. 79, 172, 243
Heyer, E. 273
Heymsfield, S.B. 727
Heyn, H. 760
Heywood, S. 796
Hickerson, M.J. 495
Hickey, G. 78, 82
Hickey, J.M. 110, 524–525, 811
Hickey, L.T. 530
Hickman, M.J. 417
Hiergeist, A. 995
Higgins, D.G. 342, 345
Higgins, J.M. 83, 419
Higgins, J.P. 648
Higgs, P.G. 367
Higham, T.F.G. 290, 315, 317
Hilborn, R. 495
Hill, C.E. 453
Hill, G. 836
Hill, J. 595
Hill, W.G. 79, 142, 495, 500, 525–528, 594, 812
Hiller, J. 314
Hillier, L.W. 947
Hillis, D.M. 215, 391
Hilmer, J.K. 549
Hinch, A.G. 77, 269

- Hinchliffe, A. 530
Hindorff, L.A. 812
Hindson, C.M. 760
Hingorani, A.D. 674
Hinton, G. 343, 759
Hinzmann, B. 946
Hirai, K. 726
Hirbo, J.B. 84
Hirschhorn, J.N. 76, 270–271, 625–627, 648, 650, 729, 793
Hirschhorn, K. 648
Hirsh, A.E. 144
Hirst, M. 796
Hirzel, A.H. 493
Hitomi, Y. 793
Hizon, R. 549
Hjalgrim, H. 834
Hjartarson, E. 79
Hjerling-Leffler, J. 759
Hladun, S.L. 729
Ho, G. 792
Ho, S.Y.W. 1014, 1016
Ho, V. 943
Hoang, A. 453
Hoban, S. 495
Hofer, B. 316
Höber, B. 291
Hobolth, A. 79, 85, 144, 362
Hoch, J.A. 1020
Hochberg, Y. 623, 725
Hochc, J.A. 346
Hockenberry, A.J. 417
Hodgson, J.A. 270
Hodoglugil, U. 318
Hoede, C. 84
Hoefsloot, H.C. 973
Hoekstra, H.E. 453
Hoekstra, J.M. 453
Hoen, P.A.C. 732
Hof, P.R. 726
Hofacker, G.L. 366
Hoffman, G.E. 726
Hoffman, K. 993
Hoffmann, A.A. 495
Hoffmann, C. 992–994, 996
Höffner, B. 291
Hoffs, M. 946
Hofker, M.H. 992
Hofman, A. 270, 732, 943
Hofmanová, Z. 48, 314, 317
Hofreiter, M. 270, 315, 317, 323
Hoggart, C.J. 974
Hohenlohe, P.A. 454
Hohmann, J.G. 726
Höhna, A.D.S. 216
Höhna, S. 215
Hoischen, A. 788, 791
Holden, M.T.G. 1017–1018
Holder, D. 731
Holder, M.T. 215
Holland, J.B. 530
Holle, R. 943
Hollenbeck, C.M. 495
Holler, E. 995
Hollfelder, N. 324
Holloway, A. 728
Holmans, P. 627
Holmes, C.C. 788, 872
Holmes, E.C. 391, 1016–1018
Holmes, I. 364
Holmes, S.P. 992
Holmlund, G. 270, 319
Holmström, L. 528
Holsinger, K.E. 417
Holt, R.A. 729
Homburger, J.R. 273
Homfray, T. 792
Homuth, G. 732
Hong, H. 625
Honkela, A. 871
Hood, R.L. 791
Hooiveld, G. 975
Hooning, M.J. 674
Hoover, J. 730
Hopf, T.A. 343
Hopfe, C. 292
Hopper, J.L. 674
Hoppin, J.A. 947
Hore, V. 727
Horejs, B. 48
Horikoshi, M. 625, 676
Horlbeck, M.A. 753
Hormozdiari, F. 675, 727
Horsnell, R. 522
Horton, R. 79, 943
Horvath, S. 727, 729, 731, 943–945
Höschele, I. 727
Hoskinson, D.C. 791

- Hospital, F. 525
Hothorn, T. 416
Hottenga, J.J. 728, 732
Hou, L. 943, 945, 947
Hou, R. 756
Hou, Y.-C.C. 795
Houck, J.T. 729
Houseman, E.A. 945
Houstis, N. 729
Houston, R.D. 812
Hovhannisyan, N.A. 270, 318
Howden, P.J. 946
Howe, D.G. 791
Howe, K.L. 946
Howell, G.R. 946
Howell, P. 522, 524
Howie, B. 81, 110–111, 627, 649, 873
Howland, T. 730
Hsaker, R.E. 790
Hsiao, C.J. 759
Hsie, L. 595
Hsieh, P.H. 291
Hsu, H.C. 726
Hsu, L. 836
Hsu, L.Y. 1017–1018
Hu, F. 834
Hu, H. 733, 791, 813
Hu, J.J. 270
Hu, M. 946
Hu, Y. 111, 291, 630, 755, 759, 875
Hu, Y., Wang, T. 629
Hu, Z. 527
Hua, J. 530
Hua, X. 1014
Huan, T. 728
Huang, B.E. 111, 525, 529
Huang, C.C. 972
Huang, E. 111
Huang, J. 110, 797, 943
Huang, K. 525
Huang, L. 628
Huang, M. 755, 759
Huang, P.-S. 344, 1019
Huang, Q.Q. 727
Huang, S. 732
Huang, X. 525
Huang, Y. 293, 755, 759
Huang, Y.-F. 364
Huang, Z.J. 754
Hubbard, T. 947
Hubbell, E. 595
Huber, P.J. 675
Huber, W. 727, 753
Hubisz, M.J. 83, 269, 316, 318, 418–419, 495
Hublin, J.-J. 290, 292, 321–323
Hubner, N. 730, 874
Huchard, E. 499
Huckins, L.M. 726
Huckle, E.J. 946
Hudgens, C.W. 993
Hudjashov, G. 271
Hudson, J. 973
Hudson, M.E. 992
Hudson, R.R. 79–80, 142, 172–174, 268, 270, 416–417
Hudson, T.J. 595, 944
Huelsenbeck, E.T. 270
Huelsenbeck, J.P. 213, 215–216, 270, 364, 392, 1017–1019
Huerta-Sánchez, E. 47, 292, 315, 317, 418, 420
Huet, S. 791
Huff, C.D. 791
Hufford, M.B. 526
Hughes, A.L. 392, 417
Hughes, J.D. 975
Hughes, T.K. 755
Hughes, T.R. 728
Hugot, J.P. 625
Huh, D. 757
Huh, I. 944
Hulsegrave, I. 813
Hulselmans, G. 753
Hultman, C.M. 112
Humbert, J.E. 80
Hume, J. 946
Humphreville, V.R. 80
Humphrey, M.B. 730
Humphries, S.E. 625
Hung, S.P. 973
Hunt, A.R. 946
Hunt, K.A. 793
Hunt, P.J. 946
Hunt, S. 82, 143, 624
Hunter, D. 674, 836
Hurgobin, B. 526
Hurles, M.E. 789, 792, 797–798, 946
Hurwitz, B.L. 525
Hüsing, A. 836

- Huson, D.H. 729
 Hussain, M. 796
 Hussin, J.G. 77
 Hutcheson, H.B. 419
 Hutchinson, D.S. 993
 Hutnik, K. 78, 81, 271
 Huttenhower, C. 995
 Hutzenthaler, M. 141
 Huynh, C. 754
 Huynh-Thu, V.A. 753
 Hwa, T. 344, 346, 1020
 Hwang, D.G. 364
 Hwu, P. 993
 Hwu, W.J. 993
- i**
- Ibegwam, C. 730
 Ibrahim, J.G. 871–872
 Ideker, T. 725
 Iizuka, M. 173
 Ikram, M.A. 677
 Ilicic, T. 755–756
 Im, H.K. 725–726, 732
 Imai, Y. 76
 Imbens, G. 671
 Imhann, F. 992
 Imholte, G.C. 875
 Imrichova, H. 753
 Imura, M. 727
 Ina, Y. 392
 Inagaki, Y. 364
 Indap, A.R. 416
 Indyk, P. 1017
 Ingason, A. 80, 111
 Ingelsson, E. 672, 835
 Ingles, S.A. 270
 Ingraham, J. 345–346
 Ingraham, J.B. 343
 Ingvarsson, P.K. 526
 Innan, H. 174
 Inouye, M. 112, 727, 833
 Ioannidis, J.P.A. 793
 Ioannidis, N.M. 792
 Iqbal, Z. 77
 Iriarte, E. 316
 Iribarren, K. 994
 Irigoyen, J. 522
 Irincheeva, I. 875
 Irish, J.D. 49
- Iritani, B.M. 730
 Irizarry, R.A. 757, 942, 944–946
 Isaacs, W. 270
 Isherwood, J. 946
 Ishwaran, H. 872
 Islam, S. 755, 759
 Isomura, M. 84
 Issaq, H.J. 872
 Ito, Y. 792
 Iuga, A. 730
 Iversen, R. 318
 Ives, A.R. 1017–1018
 Iwata, H. 530
 Iyengar, R. 725
 Iyer, R. 84
 Izaabel, H. 76
 Izar, B. 758
 Izzi, B. 945
- j**
- Jackson, A.P. 792
 Jackson, A.R. 794
 Jackson, D.K. 943
 Jackson, E. 569
 Jackson, G. 548
 Jackson, R. 524
 Jacob, D. 729
 Jacob, L. 946
 Jacob, S. 794, 947
 Jacobson, R.M. 628
 Jacquelin, B. 293
 Jacquelot, N. 994
 Jaeger, M. 992
 Jaenicke, V. 317
 Jaenisch, R. 944–945
 Jafari, N. 943
 Jaffe, A.E. 727, 942
 Jäger, C. 994
 Jager, P.L.D. 730
 Jain, P. 792
 Jain, S. 215, 756, 797
 Jaitin, D.A. 755
 Jakeli, N. 270, 317
 Jakobsson, M. 269–270, 273, 292, 316–324,
 1017–1018
 Jakubowska, A. 674
 Jalali, M. 730
 Jallow, M. 77, 625
 James, P. 840

- James, R.A. 794
Jamin, P. 526
Jana, B. 344
Janes, H. 836, 838
Janjic, N. 677
Jankauskas, R. 314
Janke, A. 213, 362
Jankipersadsing, S.A. 992
Jankovic, I. 628
Jannink, J.-L. 525
Jansen, J. 523
Jansen, R. 728, 732
Jansen, R.C. 725, 728, 732
Janss, L. 810
Janssens, A. 836
Jansson, J.K. 993
Janzen, F.J. 453
Jaroszewski, L. 343
Järvelin, M.R.,
Jarvik, G.P. 293, 788
Jarysz, R. 314
Jasinska, A.J. 293
Jasmine, F. 674
Jasra, A. 870, 872, 974
Jasson, S. 526
Jay, F. 273, 290, 292, 319–322, 324
Jayaraman, J. 80
Jaynes, E.T. 48
Jeddeloh, J.A. 944
Jeffery, P.L. 795
Jeffreys, A.J. 80
Jeffries, N.O. 648
Jégou, H. 755
Jehl, P. 343
Jellema, R.H. 973
Jenkins, M. 674
Jenq, R.R. 993
Jensen, H. 453
Jensen, J.D. 47, 291, 316, 418, 494
Jensen, J.L. 79, 364, 366
Jensen, V.B. 993
Jensen J.L. 362
Jeon, J. 836
Jeon, S. 315
Jeon, Y. 315
Jeong, H. 725
Jerby-Arnon, L. 754
Jewett, E.M. 628
Jewett, M.C. 417
Jha, A.R. 318, 321
Ji, C. 872
Ji, H. 755, 759, 944
Ji, R.R. 729
Ji, S. 759
Ji, Y. 872
Ji, Z. 755
Jia, C. 755
Jia, M. 529
Jia, Z. 872
Jian, M. 994
Jiang, B. 872
Jiang, C. 728
Jiang, H. 420, 873, 877, 993
Jiang, L. 675, 755
Jiang, N. 732
Jiang, Q. 346
Jiang, R. 797
Jiang, T. 677
Jiang, W. 80, 363
Jiang, X. 625, 797
Jiang, Y. 526, 726
Jiggins, C.D. 291
Jin, L. 85, 291, 293
Jin, V.X. 946
Jin, X. 317, 420
Jin, Z.-B. 792
Jinks, J.L. 526–527
Jobling, M.A. 498
Jochens, A. 548
Joe, L. 494
Joehanes, R. 728, 945
Johannsen, N.N. 318
John, E.M. 270, 674
John, J.A. 526, 530
Johns, D. 730
Johnson, A.D. 676
Johnson, A.J.A. 992
Johnson, B.P. 549
Johnson, C. 80
Johnson, D. 946
Johnson, D.L. 810
Johnson, D.S. 727
Johnson, E.A. 497
Johnson, G.C. 80
Johnson, J.E. 729
Johnson, J.M. 728, 731
Johnson, J.R. 273
Johnson, J.S. 726

- Johnson, M.R. 676, 839
 Johnson, M.S. 365, 367
 Johnson, N. 674
 Johnson, P.L.F. 290–292, 314, 316, 320–321
 Johnson, T. 142, 272, 418
 Johnson, W.E. 756, 945
 Johnston, A. 1019
 Johnston, J.J. 796
 Johnston, P. 731
 Jolivot, D. 526
 Jolley, K.A. 1019
 Jolly, R.A. 732
 Jombart, T. 321, 495, 1014–1018
 Jona, G. 993
 Jonasdóttir, A. 80
 Jonasson, K. 110, 593
 Jones, A.L. 84
 Jones, A.R. 726, 728
 Jones, B.N. 757
 Jones, D.R. 345, 366, 676
 Jones, D.T. 216, 342–345, 363–364, 366–367
 Jones, E.R. 270, 315, 317–319
 Jones, M.C. 946
 Jones, O.R. 495
 Jones, P.A. 947
 Jones, P.J. 548
 Jones, S. 568, 946
 Jones, W. 78
 Jonkers, D. 992
 Jonsson, F. 111
 Jónsson, H. 317, 322
 Joost, S. 755
 Joosten, L. 992
 Jordan, A.T. 996
 Jordan, D.M. 361, 788
 Jordan, M.I. 1014, 1020
 Jorde, P.E. 495
 Jores, P. 314
 Jorgensen, A.L. 629
 Jorgensen, M.E. 627
 Jorsboe, E. 317
 Joseph, S.S. 946
 Joseph, T.A. 317
 Joshi, A. 837, 839
 Jost, L. 495
 Jost, M. 753
 Jostins, L. 836
 Joubert, B. 943
 Jousson, O. 995
 Juevas, J. 526
 Jukes, T.H. 142, 215, 364, 392
 Jun, G. 317
 Jung, M. 1019
 Junge, O. 270
 Junge, W. 344
 Juric, I. 291
 Jurynec, M.J. 80
 Just, A.C. 943
 Just, R.S. 549
 Juttukonda, L.J. 996

k

- Kaakinen, M. 626
 Kabisch, M. 674
 Kacem, S. 945
 Kadasi, L. 272
 Kaderbhai, C. 994
 Kadie, C.M. 626
 Kadin, J.A. 796
 Kaepller, S.M. 526
 Kahali, B. 626, 648
 Kähler, A.K. 790
 Kahn, J.P. 571
 Kahneman, D. 526
 Kahraman, A. 946
 Kaiser, E. 324
 Kaj, I. 174
 Kaján, L. 343
 Kakaradov, B. 758
 Kalari, K.R. 759
 Kalbfleisch, J. 838
 Kalender Atak, Z. 755
 Kalinowski, S.T. 496
 Kalisky, T. 756
 Kalogeropoulou, A. 974
 Kamat, M.A. 676–677
 Kamath, G.M. 757
 Kamath, R.S. 728
 Kamberov, Y.G. 80
 Kamburov, A. 972
 Kamisetty, H. 342–344, 1019
 Kamitaki, N. 756
 Kamm, J.A. 49, 84, 293, 323
 Kamoun, S. 497
 Kam-Thong, T. 625
 Kan, Z. 728, 731
 Kanehisa, M. 728
 Kang, E.Y. 675, 727

- Kang, H.M. 317, 594, 625
Kang, L. 291
Kangas, A.J. 975
Kanter, I. 756
Kao, C.H. 732
Kao, W.H. 270
Kaplan, J.M. 728
Kaplan, L.M. 727
Kaplan, N.L. 80, 142, 172–173, 416–417
Kaplan, R.C. 629
Karachanak-Yankova, S. 76, 318
Karbalaï, N. 625
Karchin, R. 788, 793
Karczewski, K.J. 81, 792, 795
Kardia, S.L.R. 50
Karim, L. 811
Karjalainen, J. 732
Karlin, S. 80, 142
Karmin, M. 273, 320–321
Karp, C.L. 728
Karp, R.L. 757
Karpinets, T.V. 993
Karr, J.R. 725
Karssen, L.C. 625
Kasarskis, A. 730–731
Kasela, S. 732
Kash, S.F. 731
Kasper, M. 755
Kass, R.E. 48, 215
Kasukawa, T. 792
Kathagen, G. 995
Kathail, P. 759
Kathireshan, S. 798
Katsanis, N. 726
Katsuya, T. 84
Kattenberg, M. 732
Katus, H. 317
Katz, Y. 872
Katzfuss, M. 872
Kauffman, S.A. 758
Kaufman, B. 835
Kaufmann, B.B. 753
Kauppi, L. 80
Kaur, G. 792
Kaushal, M. 797
Kavanagh, D.H. 726
Kawabata, T. 364
Kawachi, I. 677
Kawai, Y. 874
Kawaji, H. 792
Kawashima, M. 728
Kaye, J. 569
Kayser, M. 270, 324
Kazan, J. 524
Kazazian, H. 77
Ke, Z. 730
Keane, J.A. 1015
Keane-Moore, M. 728
Keates, S.G. 290
Keating, B.J. 626
Kechris, K.J. 946
Keebler, J. 791
Keenan, S. 82, 946
Keener, R.W. 48
Keeney, S. 83
Keerl, V. 319
Kehl, S.C. 549
Keightley, P.D. 416
Keigwin, M. 497
Keinan, A. 270–271
Keizer, P. 525
Kejariwal, A. 365
Kelleher, J. 80, 82, 141–142, 317
Keller, A. 317
Keller, L.F. 497
Kelley, D.R. 875
Kelley, J.L. 495
Kelley, R. 872
Kelley, S. 994
Kellis, M. 730
Kelly, D.J. 1019
Kelly, K.A. 872
Kelly, M.A. 792
Kelly, S. 946
Kelly, T.K. 320
Kelsey, K.T. 945
Kelso, J. 290–293, 314–323
Kemp, B.M. 273
Kemp, D.M. 727
Kemp, D.W. 757
Kemper, K.E. 812
Kempton, R. 526
Kendal, D.G. 1017–1018
Ken-Dror, G. 625
Kendziorski, C. 753, 756, 872–874
Kendziorski, C.M. 728
Kennedy, S. 728
Kennett, D.J. 317

- Kenny, E.E. 271, 627
 Kent, M. 812
 Kent, W.J. 625
 Kerber, R. 836
 Kere, J. 945
 Kerem, B. 626
 Keren-Shaul, H. 755
 Kerkhoven, R.M. 731
 Kern, A.D. 419
 Kerr, K. 836
 Kerr, M.K. 526, 974
 Kerr, R.J. 526
 Kerr, S.M. 270, 318
 Kershaw, J.K. 946
 Kessler, M. 796
 Kessner, D.E. 50
 Kester, L. 755
 Kettunen, J. 732
 Keun, H.C. 972
 Kevrekidis, I.G. 758
 Key, F.M. 292
 Khairat, R. 317
 Khan, H. 673
 Khan, M.T. 996
 Khan, S. 344
 Khan, Z. 946
 Kharchenko, P.V. 756–757, 759
 Khatry, D.B. 726
 Khetani, R. 992
 Khokhlov, A. 314, 316, 319
 Khoury, M. 835, 839
 Khusainova, R. 318
 Khusnutdinova, E. 273, 318, 321
 Kibbe, W.A. 943
 Kibriya, M. 674
 Kichaev, G. 419
 Kidd, B.A. 730
 Kidd, K.K. 273, 549
 Kidd, M.J. 728, 731
 Kiel, D.P. 943, 945
 Kiesewetter, H. 320
 Kijas, J. 493
 Kijas, J.W. 496
 Kilburn, D. 595
 Kilinç, G.M. 316–317
 Kim, D. 343–344, 793, 872, 875
 Kim, J.B. 729
 Kim, J.J. 811
 Kim, J.K. 728, 754–756
 Kim, P. 345
 Kim, R. 793
 Kim, S. 872
 Kim, S.K. 324
 Kim, Y. 417–418, 420
 Kimmel, G. 273
 Kimmel, M. 272
 Kimura, M. 80, 82, 142, 215, 392, 417, 494
 Kimura, R. 84
 Kinch, L.N. 343
 Kind, C.-J. 314
 Kindlon, N. 81
 King, A.M. 728
 King, C.E. 392
 King, J.L. 142, 550
 King, K.S. 272, 418
 King, M.-C. 836
 King, N. 244
 Kinghorn, B.P. 110
 Kingman, J.F.C. 48, 80, 110, 142, 173, 243, 594,
 1017–1018
 Kingsbury, N. 526
 Kingsford, C. 757
 Kingsolver, J.G. 453
 Kioschis, P. 946
 Kircher, M. 291–292, 314, 316–322, 792
 Kirchgessner, T.G. 727
 Kirkpatrick, M. 142, 453
 Kirsanow, K. 48, 314, 318, 324
 Kirschner, K. 756
 Kiryu, H. 757
 Kirzhner, V.M. 142
 Kiselev, V.Y. 756
 Kisfali, P. 272
 Kishino, H. 214–217, 362, 366–367, 392
 Kiss, V. 314
 Kistler, L. 317
 Kitchener, A.C. 493
 Kittles, R.A. 270
 Kitzman, J.O. 292, 320, 796
 Kivisild, T. 271–273, 318, 321
 Kizilkaya, K. 811, 836
 Kjoglum, S. 812
 Klages, S. 946
 Klareskog, L. 945
 Klei, L.L. 726
 Klein, C. 994
 Klein, H.-U. 944
 Klein, R.J. 626

- Kleinjung, J. 344
Kleinman, C.L. 364
Klein, W.M.P. 796
Kline, L. 730
Klinkenberg, D. 1018
Klisovic, R. 947
Klitz, W. 318
Kloesges, T. 290
Klopfstein, S. 216
Klotzle, B. 943
Klug, A. 342
Kluger, Y. 728
Klughammer, J. 754
Klugman, K.P. 1015
Knapp, L.A. 499
Knight, J.A. 674
Knight, J.C. 732
Knight, R. 993–995
Knights, A.J. 946
Knoblich, J.A. 792
Knolwes, L.L. 243
Knoppers, B.M. 569, 795–796
Knowles, D.A. 759
Knowles, J.A. 726
Knowles, L.L. 498
Knudsen, B. 393
Ko, A. 292–293, 322–323
Ko, K.S. 1015
Ko, M.S.H. 757
Ko, S.B.H. 757
Koch, C. 726
Koch, G. 1020
Kochar, J. 732
Kocher, J.-P.A. 759
Kodira, C.D. 729
Koehrsen, M. 992
Koelle, K. 1020
Koenen, K.C. 677
Koenig, C. 943
Koenig, W. 676
Koeth, R. 993
Kohl, J. 728
Kohlbacher, O. 343
Köhler, S. 792
Kohn, R. 872, 874
Kojima, K. 143, 874
Koki, G. 293
Koks, S. 726
Kolaczyk, E.D. 974
Kolář, J. 314
Koller, D. 48, 728
Kolmert, J. 972
Kolodner, R. 394
Kolodziejczyk, A.A. 754–756
Colonel, L.N. 270
Kondrashov, A.S. 142
Kong, A. 80, 110–111, 593, 730
Kong, S.Y. 594
Kong, X. 529
König, I.R. 629
Köninger, J. 324
Konings, P. 796
Kono, H. 80
Koo, M.S. 496
Koornneef, M. 525
Kooy, K. 731
Kopelman, N.M. 1018
Koppert, L.B. 674
Koptekin, D. 317
Korbel, J.O. 994
Korem, T. 993
Koren, A. 790
Koren, S. 1019
Korlach, J. 944
Kormaksson, M. 942
Korneliussen, T.S. 76, 273, 293, 318, 321–323, 416, 420, 492, 496, 498, 593
Korol, A.B. 142
Korolev, A. 530
Korte, A. 836
Korthauer, K.D. 756
Korzun, V. 522
Kosakovsky Pond, S.L. 363, 393, 417
Kosambi, D.D. 594
Koscielny, G. 792
Kosciolek, T. 343
Köser, C.U. 1017–1018
Koshi, J.M. 364, 393
Kosintsev, P.A. 290
Kosiol, C. 364
Kosiura, A. 946
Kosma, V.M. 674
Kosmicki, J.A. 795
Kostem, E. 675, 727
Kota, K. 992
Kote-Jarai, Z. 623
Kotsakis, K. 48
Kottmann, R. 994

- Kousathanas, A. 48, 314, 317–318
Kouvatsi, A. 270
Kovacic, J.C.,
Kovacs, P. 270, 318
Kovári, I. 315
Kovar-Smith, C. 946
Kover, P.X. 526
Kowalczyk, A. 833
Koyano-Nakagawa, N. 756
Koyuk, A. 498
Kraft, C.L. 729
Kraft, P. 674, 677, 836, 839
Krainitzki, H. 318
Krakauer, J. 418
Kramer, R. 726
Kramer, R.S. 726
Kranenburg, T. 523
Krarup, N. 836
Krause, J. 243, 270, 292, 314, 316–324
Krausgruber, T. 754
Kravitz, K. 595
Kravitz, S.A. 729
Kravitz, S.N. 788
Krawczak, M. 270, 548–549
Kreitman, M. 393, 417–418
Kremer, A. 324
Kren, V. 730
Kresovich, S. 527
Kreth, K. 343
Kretschmer, A. 943
Kretzschmar, W. 111–112, 627, 649
Kreutzer, S. 48, 314, 317
Kriiska, A. 314
Krimbas, C.B. 496
Krings, M. 318
Krishna, A. 872
Krishnan, T. 49
Krishnaraj, R. 792
Krishnaswamy, S. 759
Kristal, B.S. 973
Kristensen, T.N. 495
Kristensen, V. 674
Kristiansen, K. 81, 314, 318, 321, 994, 1015
Kriwacki, R. 345
Kroemer, G. 994
Krogan, N.J. 728
Krogh, A. 322, 873
Kroksmark, A.-K. 791
Krolewski, M. 731
Krone, S.M. 143, 173–174
Kronenberg, F. 943
Kroonen, G. 318
Kruger, M.J. 731
Kruglyak, L. 77, 594–595, 623, 629, 725, 728, 733
Kruijver, M. 549
Kruuk, L.E.B. 453–455
Kryazhimskiy, S. 315, 416
Kryshtafovych, A. 343, 365–366
Krzewinska, M. 317
Kubatko, L. 243–245
Kubzansky, L.D. 677
Kucan, Z. 316
Kuchel, H. 527
Kuchenbaecker, K. 836
Kuchler, A. 754
Kucukelbir, A. 869
Kudaravalli, S. 84, 320, 420
Kuhlwilrn, M. 292, 318, 320
Kuhn, J.M.M. 318
Kuhner, M. 215
Kuhner, M.K. 80, 85, 596
Kullback, S. 594
Kullo, I.J. 293
Kulp, D. 795
Kumar, A. 496, 758
Kumar, H. 758
Kumar, L. 530
Kumar, S. 393, 395
Kumar, T. 993
Kumar, V. 992
Kumaran, M.K. 730
Kunde, J. 943
Kundu, R.K. 729
Kundu, S. 836
Kunst, M. 316
Kunze, S. 943
Kuo, L. 217, 872
Kuperman, D. 728
Kural, D. 82
Kurilshikov, A. 992
Kurita, K. 992
Kurths, J. 974
Küry, S. 794
Kushniarevich, A. 318
Kusmec, A. 526
Kusterer, B. 529
Kusuma, P. 76

- Kutalik, Z. 272, 418, 813
 Kutt, L. 994
 Kutuev, I. 268
 Kuzmin, Y.V. 290
 Kuznetsov, P. 316, 319
 Kvalnes, T. 453
 Kvasz, A. 811
 Kwak, I.-Y. 756
 Kwan, J. 839
 Kwart, D. 794
 Kwiatkowski, D.P. 112
 Kwint, M.P. 788
 Kwon, D. 872
 Kyparissi-Apostolika, N. 48
 Kyrides, N. 344
 Kyrides, N.C. 1019
- I**
- Laakkonen, L. 316
 Labalette, C. 758
 Labbe, A. 875, 944–945
 Lacaze, X. 526
 Lach, R.P. 729
 Lachance, J. 291
 Lachenbruch, P.A. 1017–1018
 Lachmann, M. 290–292, 314, 316, 320
 Lackman-Ancrenaz, I. 495
 LaCroix, A.Z. 943
 Ladd-Acosta, C. 942, 944
 Ladouceur, M. 626, 630
 Laetsch, T.W. 793
 Lafferty, J. 291
 Lagace, R. 549
 Lagane, F. 324
 Lahad, A. 835–836
 Lahr, M.M. 321–322
 Lai, C.-Q. 630
 Lai, J. 526
 Lai, Z. 729
 Laidig, F. 526
 Laird, G.K. 946
 Laird, N. 47, 648
 Laird, P.W. 945, 947–948
 Lakhani, S. 837
 Lale, M. 789
 Lall, K. 648
 Laluezza-Fox, C. 291, 316, 318–322
 Lam, T.-W. 81
 La Manno, G. 755
 Lamason, R.L. 80
 Lamb, J. 731
 Lamb, J.R. 731–732
 Lambert, C.G. 792
 Lambert, D. 293, 322
 Lambert, J.A. 548
 Lambotte, R. 754
 Lamnisos, D. 872
 Lan, H. 728, 872
 Lan, Y. 342, 1016
 Lancaster, M.A. 792
 Lancet, D. 795
 Lande, R. 142–143, 417, 453
 Landell, M.G.A. 524
 Lander, E.S. 291, 316, 419, 549, 594–595, 626, 628, 729, 798, 945
 Landi, M.T. 872
 Landis, M.J. 215
 Landray, M. 84, 677
 Lang, B. 345
 Lang, G.I. 417
 Lange, C. 83
 Lange, J. 83
 Lange, K. 48, 111–112, 314, 492, 593–596
 Langefeld, C.D. 627
 Langenegger, F. 324
 Langfelder, P. 729
 Langford, C. 946
 Langit, R. 549
 Langley, C.H. 173, 416–417
 Langley, S.R. 869, 874
 Langmead, B. 80, 944–947
 Langmead, C. 342
 Lango Allen, H. 594
 Langston, M.A. 726
 Lanner, F. 757
 Lao, K. 758
 Lao, O. 270, 319
 Lapedes, A.S. 343
 Laport, B. 728
 Lara-Astiaso, D. 755
 Larbi, A. 732
 Laredj, L. 318
 Larget, B. 215–216
 Lari, M. 315
 Lariagon, C. 526
 Larkin, E. 270
 Laros, J.F.J. 790
 La Rosa, P. 993

- Larribe, F. 496
Larson, G. 365
Larson, W.A. 499
Lartillot, N. 215–216, 362, 364–366
Lasak, I. 314
Lascoux, M. 174, 323
Lasken, R. 993
Lasseigne, B.N. 789
Lathrop, M. 524
Lau, C. 726
Lau, J. 676
Laud, P. 876
Lauder, A. 994
Lauffenburger, D.A. 729
Laukens, K. 975
Laurie, C. 549, 725
Laurie, G. 569, 571
Laurila, E. 729
Lauritzen, S.J. 594
Lauritzen, S.L. 974
Laval, G. 496
Lavryashina, M. 273
Law, C.N. 526
Law, C.W. 875, 946
Lawi, S. 216
Lawlor, D. 672, 675
Lawlor, S. 946
Lawlor, T.J. 810
Lawrence, M.J. 526
Lawrence, N.D. 871
Lawrenz, F. 571
Lawson, D.J. 78, 81, 270–271, 318, 1017–1018
Lay, F.D. 947
Layer, R.M. 81, 788
Lazar, A.J. 993
Lazaridis, I. 270, 315–319, 417–418
Le, J.M. 943
Le, Q.S. 77, 362
Le, S.Q. 365, 1017–1018
Le, T.H. 626
Le, V.S. 362
Lea, A.J. 945
Leach, L. 732
Leaché, A.D. 243, 496
Lebofsky, R. 757
Leboreiro, I. 273
Lebrasseur, O. 319
Lecarpentier, J. 836
Le Chatelier, E. 993–994
Lecoeur, C. 323
LeCun, Y. 343
Leday, G.G. 876
Lee, A. 836
Lee, B. 732
Lee, C. 317, 754, 758, 789
Lee, D. 111, 993
Lee, E. 674, 730
Lee, I. 728
Lee, J.C. 873
Lee, J.D. 729
Lee, J.E. 993
Lee, J.H. 944
Lee, J.M. 792
Lee, K. 872
Lee, M. 836
Lee, P.H. 676
Lee, S.H. 626, 812–813, 874
Lee, S.I. 728
Lee, S.Y. 84
Lee, W. 626
Lee, Y.S. 675
Leek, J.T. 728, 945
Lees, J.A. 1017–1019
Lefebvre, E. 754
Leffler, E.M. 76
Legarra, A. 810, 812
Legler, M.M. 549
Legood, R. 837
Lehar, J. 729
Lehermeier, C. 522
Lehmann, E.L. 48
Lehrach, H. 947
Lei, Y. 729
Leibler, R.A. 594
Leidinger, P. 317
Leimar, O. 498
Lein, E.S. 292, 320, 726
Leistritz-Kessler, D. 794
Lek, M. 792–793
Lema, G. 291
Le Marchand, L. 270
Lemire, M. 944
Lencz, T. 497, 595
Lenffer, J. 792
Leng, N. 753, 873
Leng, S.N. 797
Lengyel, G. 270, 318
Lennon, A. 797

- Lennon, N.J. 759
Lenski, E.E. 453
Lenski, R.E. 417
Lenstra, J.A. 496
Lenz, M.W. 420
Leo, M.C. 788
Leonardson, A. 729, 731
Leonhardt, H. 760
Leon-Novelo, L.G. 873
Leonte, D. 874
Lepage, P. 994
Lepage, T. 216
Le Paslier, D. 994
Leppert, T. 549
Lerceteau, E. 875
Lercher, M.J. 994
Lerner, I.M. 526
Leroy, F. 324
Leroy, T. 324
Lesk, A.M. 362
Lesley, S.A. 995
Lesnar, P. 726
Lessard, S. 84
Lettre, G. 594, 836
Leuenberger, C. 47–50
Leutenegger, A. 594
Leutenegger, A.L. 110
Levenez, F. 994
Leversha, M. 946
Levin, A. 271
Levina, E. 973
Levine, A.J. 754
Levine, H.Z.P. 419
Levine, M.E. 943, 945
Levison, B. 993
Levitsky, A.A. 729
Levy, D. 728, 943, 945
Levy-Lahad, E. 836
Lewin, A. 728, 870, 873, 876
Lewin, J. 943
Lewis, D.A. 726
Lewis, J.D. 993–994
Lewis, K.L. 796
Lewis, L. 946
Lewis, M. 972
Lewis, P.O. 217
Lewis, S.J. 624
Lewontin, R. 81, 393, 418
Lewontin, R.C. 143
Ley, E. 873
Ley, R. 995
Leyrat, A.A. 757
Leyser, O. 526
Li, A. 529
Li, B. 630, 731, 754, 756, 946
Li, C. 756, 793, 834, 837–838, 875
Li, D. 994, 1015
Li, G. 725, 731
Li, H. 48, 81–82, 112, 173, 271, 290–292, 316,
318–321, 418, 527, 529, 727, 992–996
Li, J. 321, 674, 729, 756, 873, 994
Li, J.Z. 271
Li, K. 367, 729, 993
Li, L. 526–527, 946
Li, M. 755, 759, 793
Li, N. 77, 81, 111, 269, 271, 418, 757
Li, P.W. 729
Li, R. 81, 293, 994
Li, S. 80, 216, 273, 291, 322, 675, 759, 942, 994
Li, T. 791
Li, W.-H. 78, 172–173, 393, 416, 418
Li, W.V. 756
Li, X. 527, 876, 944
Li, Y. 81, 111, 293, 321, 630, 650, 732, 756, 945,
992, 994
Li, Z. 346, 729, 876
Liang, F. 873
Liang, G. 945, 947
Liang, H. 994
Liang, K.-Y. 394
Liang, L. 112, 677, 943
Liang, M. 271, 291, 416
Liang, S. 872
Liang, Y. 317, 420
Liao, Y. 344
Libertini, E. 945
Lichtenstein, P. 112
Liddle, J. 944
Liebeke, M. 973
Lienhard, M. 945
Liland, K.H. 974
Lilje, B. 320
Lim, B. 758
Lim, C.Y. 756
Lim, M. 758
Lim, P. 529
Lima, C. 324
Lin, A. 994

- Lin, C. 756
Lin, D.Y. 111, 648, 676
Lin, K. 418
Lin, M.F. 78, 82
Lin, S. 81, 111, 943–944, 946
Lin, W. 992–993
Lin, X. 626, 630, 943
Lin, Y. 873
Lin, Z. 527
Lindahl, T. 318
Lindberg, J. 972
Lindblom, A. 674
Lindgreen, S. 320–322
Lindgren, C.M. 649, 729
Lindgren, D. 527
Lindo, J. 273
Lindon, J.C. 973–974
Lindquist, M.A. 943
Lindsay, B.G. 48
Lindsay, H. 946
Lindsay, J. 836
Lindsay, S. 789
Lindstrom, S. 625, 649
Lines, M.A. 791
Linington, R.G. 992–993
Link, V. 48, 314, 317–318
Linnarsson, S. 755, 757, 759
Linneberg, A. 994
Linseisen, J. 943
Linsley, P.S. 731, 755
Linz, B. 1016
Liò, P. 365, 395
Lipman, D.J. 342
Lippert, C. 626
Lipsitch, M. 676, 1015, 1020
Lipska, B.K. 726
Lipson, M. 271–272, 318–319
Lisec, J. 528
Lissens, W. 593
Lister, R. 945
Listgarten, J. 626
Little, J. 793
Littlejohn, M.D. 812
Litvinov, S. 273, 318, 321
Liu, A.Y.H. 494
Liu, D. 624
Liu, F. 873
Liu, H. 944
Liu, H.Q. 293
Liu, J. 732, 834, 995
Liu, K. 111
Liu, L. 216, 243–244, 343, 759, 796
Liu, P. 756, 758, 873
Liu, S. 111
Liu, W. 759, 946
Liu, X. 82, 84, 629, 729, 793
Liu, Y. 344, 626–627, 732, 945
Liu, Z. 812
Livak, K.J. 759
Livesay, D. 343
Lizardi, P. 756
Llamas, B. 316, 322
Llewellyn, S.R. 730
Lloyd, C. 946
Lloyd, D.M. 946
Lloyd-Jones, L.R. 728, 813
Lobón, I. 320
Locke, A.E. 626, 648
Lockhart, P.J. 367
Lockless, S.W. 343
Loe, L. 322
Loerch, P.M. 728, 731
Logsdon, B.A. 726
Loh, P.R. 269, 271–272, 626, 672, 813, 837
Lohman, B.K. 788
Lohmueller, J. 83
Lohmueller, K.E. 321, 324, 626
Loman, N.J. 994
Lomedica, P. 394
Long, A.D. 869
Long, C.L. 244
Longhi, C. 314
Lönnberg, T. 755
Lönnerberg, P. 755, 759
Lonnstedt, I. 755
Loo, J.A. 731
Loo, R.-L. 974
Loog, L. 319
Loos, R.J. 649–650
Lopdell, T. 812
Lopes, F.M. 754
Lopes, M.C. 273, 321
Lopez, J.R. 729
Lopez, P. 365
López, S. 48, 314, 317
López-Bigas, N. 791
Lopiano, K.K. 757
Lorbeer, R. 732

- Lord, R.V. 946
 Lordkipanidze, D. 270, 317
 Lorgeou, J. 524, 526
 Loriot, Y. 994
 Losic, B. 726, 730
 Lotta, L.A. 676
 Lotterhos, K.E. 495
 Lou, D. 759
 Lou, H. 291
 Loukidis, T. 318
 Lousegued, H. 946
 Lourbakos, A. 791
 Lovci, M.T. 758
 Love, J.C. 755
 Love, M.I. 757
 Loveland, J.E. 946
 Lovell, F.L. 946
 Lovell, J.D. 946
 Lovell, S. 345
 Lökvist, C. 342, 1016
 Lowe, J.K. 79, 593, 731
 Lowe, W.H. 496
 Lowry, D.B. 495
 Loytynoja, A. 344
 Lozado, R. 946
 Lozano, M. 319
 Lozupone, C. 994
 Lrum, M.J. 792
 Lry, J.J.M. 792
 Lu, A.T. 943, 945
 Lu, C. 726
 Lu, D. 291
 Lu, F. 729
 Lu, H. 730
 Lu, J. 112, 946, 994
 Lu, K. 527
 Lu, L. 726
 Lu, T.T. 270
 Lu, Y. 291, 837
 Luan, J. 676, 813
 Luben, R. 674
 Lubieniecki, K.P. 812
 Lubinski, J. 674
 Lucas, J. 873
 Lucas, J.E. 870
 Lucas, L. 522
 Ludwig, T. 217
 Luedtke, A. 111
 Lueth, F. 314
 Lugon-Moulin, N. 493
 Lui, J.H. 757
 Lui, T.W.H. 367
 Luijk, R. 947
 Luijt, R.B.V.D. 674
 Luikart, G. 493–496, 498–500
 Luiselli, D. 272–273
 Lum, P.Y. 725, 727–732
 Lumey, L.H. 947
 Lumley, T. 732
 Lun, A.T.L. 755–757
 Lun, D.S. 549
 Lund, D. 569
 Lund, M.S. 810
 Lund, O. 321
 Lunetta, K.L. 943, 945
 Lunn, D. 496
 Lunt, B. 344
 Lunter, G. 77, 81
 Luo, C.-C. 393, 418
 Luo, H. 756
 Luo, L. 523
 Luo, Y. 272, 292, 320, 497
 Luo, Z. 732
 Lupski, J.R. 79, 798
 Lush, J.L. 453, 527
 Lusis, A. 725–727, 729, 731, 791
 Lüthy, R. 365
 Lutz, B. 342
 Lu X. 522
 Luz, M.F. 273
 Lv, J.-N. 792
 Ly, L.-H. 758
 Lycett, S. 1019
 Lykou, A. 873
 Lynch, M. 144, 453, 455, 527
 Lyne, R. 946
 Lynnerup, N. 314
 Lyons, A. 322
- m**
- Ma, C. 626
 Ma, D.S. 730
 Ma, J. 946
 Maas, P. 837
 Maatz, H. 730
 Mable, B.K. 524
 MacArthur, D. 527
 MacArthur, D.G. 793, 795

- MacArthur, J. 626, 793
Macaulay, I.C. 758
Macbeth, H. 569
MacCallum, M. 321
MacCawley, S. 730
MacDonald, M.E. 626, 732
Macek, M. 270
MacFadyen, J.G. 676
MacFie, T.S. 82
Mackay, I. 523, 525
Mackay, T.F.C. 453
MacLeod, I.M. 812
Macosko, E.Z. 756
MacPhee, R. 320
Macpherson, E. 493, 973
Macpherson, J.M. 269
Madar, V. 732
Madden, P.A. 596, 813
Madden, T.L. 342
Maddison, W.P. 173, 216
Madigan, D. 729
Madrigal, P. 754
Madsen, B.E. 626
Madsen, T. 316
Madur, D. 526
Maeda, O. 522
Maegi, R. 320
Maenhout, S. 527
Mafessoni, F. 292
Magdelenat, H. 870
Magi, R. 271, 648
Magnúsdóttir, S. 994
Magnusson, G. 79
Magnusson, P. 112
Magoon, A. 730
Magwene, P.M. 756
Mahajan, A. 624, 626, 676
Mahajan, M. 732
Mahajan, N. 674
Mahajan, S. 273
Maher, A.D. 973
Maheshwari, M. 946
Mahley, R.W. 318
Mahoney, M.W. 628
Maibaum, E. 974
Maiden, M.C.J. 1019
Maier, R.M. 812
Mailman, M.D. 792
Mailund, T. 1019
Maixner, F. 317
Majithia, A.R. 793
Majmudar, D. 729
Majoros, W.H. 729, 791
Mak, T. 837
Makino, S. 732
Mäki-Tanila, A. 527
Maksimiak, K. 345
Maksimovic, J. 945
Malaspinas, A.-S. 49, 291, 293, 314, 316, 319,
321–322, 418
Malaspinas, O. 319, 418
Malécot, G. 143
Malerba, G. 110
Malhi, R.S. 49, 273, 321
Maliepaard, C. 523
Maller, J. 83, 498, 595
Mallick, B.K. 869
Mallick, C.B. 271
Mallick, S. 270, 272, 291–293, 315–322, 418,
497
Mallon, A.-M. 788
Mallonee, A.B. 1019
Mallory, J. 315
Malmström, H. 273, 316, 319–323
Malone, K.E. 674
Malosetti, M. 523, 527, 529–530
Malvar, R.A. 528
Malyarchuk, B. 273
Mamanova, L. 789
Manchanda, R. 837
Mancini, M.C. 524
Manco, L. 319
Mancuso, N. 837
Mandel, J. 840
Mane, S.M. 873
Mangano, V.D. 77
Mangin, B. 522, 526
Mangino, M. 945
Mangion, J. 730, 840
Mangione, R. 498
Mangravite, L.M. 726
Mani, D.N. 496
Manica, A. 270, 273, 293, 315, 317, 319,
321–322
Manichanh, C. 994
Manly, K.F. 726
Mann, A. 293
Mann, F. 730

- Mannermaa, A. 674
Manneroas-Holm, L. 996
Manning, W.C. 726
Mano, H. 796
Manolescu, A. 649
Manolio, T.A. 793, 812
Mansilla, J. 273
Manson, J.E. 943
Mantel, N. 217
Mäntyniemi, S. 1015
Manzo, T. 993
Mao, K. 346
Mao, L. 529, 944
Mao, M. 731
Mao, Q. 757–758
Mao, X. 80
Maples, B.K. 271
Maranville, J.C. 677
Marc, S. 524
Marchand, L.L. 674
Marchini, J. 77, 79, 81–82, 110–112, 269, 624, 627, 629, 649, 727, 873
Marciniak, S. 291, 319
Marco, E. 757
Marconi, T.G. 524
Marcos, M.E.P. 322
Marcotte, E.M. 728
Marcucci, G. 943–944, 947
Mardia, K.V. 363
Mareddy, L. 873
Margara, L. 342
Margarint, M. 731
Margaryan, A. 293, 314, 322
Margolin, S. 674
Margulies, E.H. 791
Maricic, T. 291–292, 316, 321
Marin, J.M. 47, 494
Marioni, J.C. 730, 754–756, 759, 873
Marioni, R.E. 943, 945
Marjanovic, N.D. 754
Marjoram, P. 49, 78, 81–82, 143, 172–173, 417–418, 593, 627
Markiewicz, D. 626
Markov, A.A. 49
Markowitz, E. 214, 363
Marks, D. 343–346
Marks, J. 570
Marks, S.P. 570
Marme, F. 674
Marmor, R. 797
Marnetto, D. 292
Marques-Bonet, T. 291–292, 316, 321
Márquez-Luna, C. 837
Marshall, C. 873
Marshall, J.L. 789
Marsili, S. 345
Martersteck, E.M. 756
Marth, G.T. 77, 271
Martin, A. 837
Martin, C.-A. 792
Martin, J. 992–993
Martin, M.D. 319
Martin, N.G. 596, 813
Martin, O.C. 529
Martin, P.A. 570
Martin, S.H. 291
Martin, W. 214, 361, 530
Martinelli, N. 324
Martínez, I. 319
Martínez, P. 320
Martiniano, R. 48, 77, 270, 314–315, 319
Martins, E.R.F. 524
Martins Dos Santos, V.A.P. 995
Martinussen, T. 676–677
Marton, M.J. 728, 731
Marttinen, P. 1015, 1017–1018
Maru, Y. 796
Maruyama, Y. 873
Mascline, A.A.M. 992
Masel, J. 291
Mashkour, M. 314
Masignani, V. 84
Masinde, G.L. 876
Masip, J.R. 994
Mason, B.A. 811–812
Mason, C. 942
Mason, C.E. 873
Mason, D. 793
Mason-Suares, H. 791
Masquelier, D.A. 760
Massa, H. 84
Massam, H. 870
Mas-Sandoval, A. 76
Massingham, T. 365, 393
Masson, G. 80, 111
Masuda, M. 272
Matarin, M. 270
Mather, K. 527

- Mathias, R. 49, 83, 272
 Mathieson, I. 319, 418, 496
 Mathotaarachchi, S.S. 675
 Matot, E. 993
 Matreyek, K.A. 793
 Matsen, F. 994
 Matskevich, Z. 270, 317
 Matsui, Y. 78
 Matsumoto, H. 757
 Matsumoto, T. 393
 Matsumura, S. 314
 Matsunami, N. 549
 Matthews, L.H. 947
 Matthews, L.R. 974
 Matthews-Palmer, T.R. 345
 Mattiangeli, V. 315
 Matukumalli, L. 835
 Matukumalli, L.K. 811
 Matz, M.V. 496
 Matzas, M. 317
 Mau, B. 216
 Maude, S.L. 793
 Mavaddat, N. 836–837
 Mavrommatis, K. 992
 May, C.A. 80
 May, D. 730
 Mayer, J. 317
 Mayer, L.F. 730
 Maynard Smith, J. 81, 143
 Mayrose, I. 393, 1017–1018
 Mayr W.R. 549
 Mayzel, J. 1017–1018
 Mazet, O. 271
 Mazmanian, S.K. 995
 Mazutis, L. 759
 Mazzola, E. 837
 McAllister, F. 993
 McAndrew, B.J. 812
 McAuliffe, J.D. 869
 McCallum, A. 291
 McCallum, C.M. 527
 McCarroll, S.A. 83, 730, 756
 McCarthy, D.J. 755, 757–758
 McCarthy, M.I. 172, 416, 623, 627, 650, 727,
 793, 812
 McCarthy, S. 81, 111, 627, 649
 McCarty, C.A. 570
 McCaskill-Stevens, W. 835
 McClelland, R. 836
 McCoy, C. 994
 McCoy, R.C. 291, 293, 323
 McCray, A.T. 793
 McCue, K. 757, 873
 McCue, M. 322
 McCulloch, R. 870–871
 McCulloch, R.E. 871
 McDavid, A. 755
 McDermott, G.P. 759
 McDermott-Roe, C. 874
 McDonagh, P.D. 731
 McDonald, G.J. 83
 McDonald, J.H. 418
 McDonald, L.E. 944
 McDonald, M.J. 417
 McDonald, M.T. 793
 McElrath, M.J. 755
 McElwee, J. 732
 McEvoy, B. 840
 McEvoy, B.P. 271, 596, 630, 813
 McGee, L. 1015
 McGettigan, D. 78, 270, 318
 McGlynn, G. 314
 McGovern, D. 835
 McGregor, J. 80
 McGregor, K. 945
 McGuire, A. 837
 McGuire, J.A. 243, 496
 McHale, J.V. 570
 McIntosh, T.C. 730
 McIntyre, L.M. 757, 873
 McKusick, V.A. 789
 McLachlan, A.D. 365
 McLachlan, G. 49
 McLaren, P.J. 112
 McLaren, S. 947
 McLaughlin, R.L. 77, 270, 315, 317
 McLaughlin, S. 317
 McLay, K. 946
 McLean, C. 674
 McLean, K. 525
 McLeod, A. 367
 McManus, K.F. 418
 McMullan, G. 834
 McMullen, M.D. 525, 527, 530
 McMurdie, P.J. 992
 McMurray, A. 947
 McMurray, B. 726
 McNeill, L.H. 270

- McPartlan, H.C. 495
McPherson, A. 754
McRae, A. 728
McRae, A.F. 813, 943, 945
McRae, J.F. 795
McRae, T.A. 526
McRae K.B. 522
McSweeney, K.M. 790
McVean, G. 76–77, 80–82, 111, 142, 271, 317, 319, 496, 594, 627, 873
Mead, A. 527
Mead, R. 527
Medard, M. 549
Medema, M.H. 992, 994
Meder, B. 317
Medini, D. 84
Medland, S. 813, 840
Meehan, T.F. 788
Meese, E. 317
Meester, R. 549
Mehdi, S.Q. 272
Mehrabian, M. 731
Mehta, D. 840
Mehta, S. 730
Mehta, T. 876
Mei, L. 791
Meidl, P. 947
Meier, H. 217
Meijer, E. 290
Meijers-Heijboer, H. 674
Meiklejohn, C. 270, 315, 318, 418
Meindl, A. 674, 797, 947
Meirmans, P.G. 496
Meisinger, C. 943
Meisler, M.H. 793, 797
Meissner, A. 945, 948
Meitinger, T. 732, 797, 947
Mel, J.-L. 794
Melchinger, A.E. 528–529
Melchiotti, R. 732
Melegh, B. 272, 318
Melham, K. 569
Meller, H. 316, 319
Melmed, S. 729
Melnick, A. 942
Melquist, S. 732
Melsted, P. 754, 1019
Meltzer, D.J. 49, 273, 321
Memari, Y. 110
Mende, D. 994–995
Mende, D.R. 994
Mendel G. 594
Mendes, F.K. 419
Mendez, F.L. 291
Mendoza-Revilla, J. 76
Mendrick, D. 625
Menelaou, A. 111
Menéndez Hurtado, D. 344
Meng, C. 244
Meng, Q. 727
Menon, V. 757
Menozzi, P. 271
Mercader, J.M. 996
Mercer, J.M. 732
Mercier, B. 593
Mercier, P. 975
Meric, G. 1019
Merikangas, K. 528, 628
Merino, C. 419
Merker, J.D. 793
Merkevicius, A. 314
Merkulov, G.V. 729
Merkyte, I. 314
Merrett, D.C. 270, 315, 318
Merrigan, M. 78, 270, 318
Merriwether, D.A. 270, 293, 318
Meshveliani, T. 270, 317
Mesirov, J.P. 729, 796
Messaoudene, M. 994
Messer, P.W. 78, 417
Messier, W. 393
Messina, C. 523
Mest, J.R. 80
Methe, B. 993
Metropolis, N. 49, 216
Metspalu, A. 648, 728, 732
Metspalu, E. 268, 271–273, 318, 320–321
Metspalu, M. 268, 271–273, 314, 318, 320–321
Metzeler, K. 947
Metzger, J.M. 731
Metzker, M.L. 946
Metzler, D. 143, 525
Meudt, H.M. 245
Meuwissen, T. 143, 837
Mewada, A. 674
Mewes, H.W. 727
Meyer, A.G. 366

- Meyer, C.P. 495
 Meyer, K. 527
 Meyer, M. 290–292, 314–316, 318–321
 Meyer, M.R. 728, 731
 Meyer, R.S. 527
 Meyer, U. 526
 Meyers, D.A. 50
 Mézard, M. 529
 Mezey, J.G. 268
 Mezquita, L. 994
 Mezzavilla, M. 496
 Mi, H. 365
 Miano, M.G. 797
 Miao, Z. 754
 Michailidis, G. 973
 Michailidou, K. 649, 674, 837
 Michel, M. 319, 344–345
 Michels, K.B. 945
 Michener, C.D. 216
 Michoel, T. 726
 Middeldorp, C.M. 732
 Miga, K. 82
 Migliano, A.B. 321
 Mihola, O. 78
 Mikacenic, C. 290
 Mikkelsen, T.S. 419, 759
 Milani, L. 732
 Mill, J. 731
 Millar, D.S. 944
 Millar, W. 836
 Millasseau, P. 524
 Miller, B. 342, 993
 Miller, C.L. 729
 Miller, D.G. 627
 Miller, H.W. 755
 Miller, J.R. 729
 Miller, K. 110
 Miller, W. 320, 342
 Milligan, C.J. 793
 Milligan, S.B. 731
 Millikan, R.C. 270
 Millstein, J. 729, 731
 Milne, I. 525
 Milne, R.L. 674
 Milne, S. 947
 Milosavljevic, A. 945
 Milshina, N.V. 729
 Mimori, T. 874
 Min, J.L. 110
 Minard-Colin, V. 994
 Mindell, D.P. 363–364, 393
 Minelli, C. 648, 672, 674
 Miner, G. 947
 Mineur, F. 493
 Ming, C. 293
 Ming, G. 758
 Minikel, E. 797
 Minikel, E.V. 792–793
 Minnoye, L. 755
 Minx, P.J. 947
 Miotto, R. 730
 Miragaia, R.J. 758
 Mirarab, S. 244
 Mirazon Lahr, M. 293
 Miretti, M. 79
 Mirowska, K.K. 674
 Mishra, P. 496
 Mistry, S.L. 947
 Misztal, I. 810, 812
 Mitchell, J. 273
 Mitchell, J.A. 793
 Mitchell, R. 837
 Mitchell, R.J. 454
 Mitchell, S. 529
 Mitchell, T. 873
 Mitrea, C. 947
 Mitreva, M. 992–993
 Mittelmann, H.D. 362
 Mittmann, M. 595
 Mittnik, A. 315–316, 318
 Miyamoto, M.M. 393
 Miyamoto, R. 393
 Miyamoto, S. 393
 Miyata, T. 393
 Miyazawa, S. 365, 393
 Mizrahi, A.-S. 270, 318
 Mkrtchyan, R. 314
 Mni, M. 811
 Mobarry, C.M. 729
 Mochalov, O. 316, 319
 Moen, E.L. 946
 Moen, T. 812
 Moerman, T. 753
 Moffatt, M.F. 80, 943
 Moghaddar, N. 810
 Mohammadi, M. 530
 Mohan, S. 876
 Mohideen, M.-A.P. 80

- Möhle, M. 143, 173
Möhring, J. 524, 528–529
Moiseyev, V. 293, 314, 316, 319, 322
Moldovan, O.T. 315
Molina, G. 873
Molina, J. 318
Molitor, J. 49, 81, 627
Moll, R.H. 523
Mollet, I. 270
Molinari, M. 524, 529
Molloy, A.M. 78
Molloy, P.L. 944, 946
Molnar, P. 319
Molony, C. 727, 731–732
Moltke, I. 76, 82, 272–273, 292–293, 318,
 320–322, 415, 492, 496–497, 594, 627
Momber, G. 324
Monasson, R. 342
Monastyrskyy, B. 343, 366
Monda, K. 839
Mondol, S. 497
Mondragón, L. 994
Monge, J.M. 270, 318
Mongodin, E.F. 83
Mongru, D.A. 730
Monks, S.A. 729, 731
Monni, S. 873
Monod, H. 522, 524
Montana, G. 974
Montano, C. 944
Montesclaros, L. 759
Montesinos-López, O.A. 526
Montgomery, G.W. 596, 732–733, 813
Montgomery, M. 346
Montgomery, R.R. 870
Montgomery, S.B. 729–730
Montinaro, F. 76
Montooth, K.L. 418
Moon, K.R. 759
Moore, A.Z. 943
Moore, B.R. 215
Moore, G. 530
Moore, J.H. 623, 628
Moore, J.L. 996
Moore, J.M. 78
Moore, T. 677
Moorjani, P. 271–272, 290, 292, 320, 497, 793
Mootha, V.K. 729
Morales, A. 792
Morales, J. 793
Moralli, D. 77
Moran, J. 839
Moran, J.L. 112
Moran, P.A.P. 143
Morcos, F. 344
Mordelet, F. 869
Moreau, L. 522, 526
Moreau, Y. 794, 796
Morell, M. 523
Moreno, L.I. 549
Moreno-Estrada, A. 269, 273, 317
Moreno-Mayar, J.V. 49, 273, 321
Moreno-Moral, A. 873
Moretti, T.R. 549
Morfitt, D.C. 732
Morgan, K. 84
Morgan, M.D. 755
Morgan, M.T. 141, 171
Morgante, M. 527
Mori, S. 796
Moritz, C. 243, 495
Moriwaki, K. 80
Moriyama, E.N. 393
Morley, M. 726, 729, 945
Morling, N. 548–549
Morrell, N.W. 677
Morrey, J.D. 570
Morris, A. 648, 840
Morris, A.P. 111, 623–624, 626–627, 629–630,
 813
Morris, M.K. 729
Morris, M.S. 595
Morris, R. 839
Morris, S. 947
Morrison, J. 650
Morrison, M. 569
Morrissey, M.B. 454–455
Morrow, J. 836
Morse, A.M. 757
Morse, M. 759
Mort, M. 796
Mortazavi, A. 754, 757, 873
Mortha, A. 730
Mortier, F. 495
Mortimer, R. 322
Morton, J. 993
Moser, G. 812–813
Mosher, M.J. 273

- Moskvina, V. 627
 Mosley, T. 50
 Mossel, E. 244
 Mosteller, F. 676
 Mott, R. 526
 Motwani, R. 1017–1018
 Motyer, A. 77, 109
 Moult, J. 365
 Mount, C. 758
 Mountain, J. 318
 Mountain, J.L. 269, 498
 Mountier, E.I. 793
 Mountz, J.D. 726
 Moutsianas, L. 627
 Mouy, M. 111
 Moy, L. 730
 Moy, M.C. 730
 Mozaffari, S.V. 726
 Mrode, R.A. 527
 Mu, T. 527
 Mujagic, Z. 992
 Mukherjee, S. 269
 Mukhopadhyay, N.D. 975
 Mulder, H.A. 525
 Muller, C. 319, 321
 Müller, I. 947
 Muller, P. 871–872, 975
 Müller, W. 270, 317
 Muller, Z. 493
 Muller-Myhsok, B. 625
 Müller-Trutwin, M. 293
 Mulligan, C.J. 270
 Mullikin, J.C. 270, 291–292, 316, 320, 791, 947
 Mullin, T.J. 527
 Mumford, J.A. 729, 1017–1018
 Mundlos, S. 796
 Munoz, O. 314
 Munro, H.M. 81, 111
 Munson, P.J. 728
 Munsterkotter, M. 727
 Munters, A.R. 322
 Murabito, J.M. 943, 945
 Mural, R.J. 729
 Murphy, A. 270
 Murphy, B.J. 730
 Murphy, E.M. 315
 Murphy, K.M. 494
 Murphy, K.P. 49
 Murphy, M. 944, 947
 Murphy, S.D. 730
 Murray, D.L. 216
 Murray, J. 755, 759
 Murray, M.L. 794
 Murray, S.S. 629
 Murrell, B. 393
 Murtagh, M.J. 568
 Murvai, J. 271
 Musani, S. 270
 Muse, S.V. 216, 363–365, 393–394, 417–418
 Mustafaoglu, G. 317
 Musunuru, K. 840
 Mutayoba, B. 85, 500
 Muto, K. 796
 Mutshinda, C.M. 874
 Muýnoze, P.R. 522
 Muzny, D. 946
 Myers, A.J. 732
 Myers, C.T. 793–794
 Myers, E.W. 729
 Myers, G.S. 83
 Myers, R.M. 271
 Myers, S. 49, 78–83, 110–111, 269–272, 274, 318, 873, 1017–1018
 Myers, S.R. 77, 82, 143, 270
 Myres, N.M. 317
 Mysore, J. 732
- n**
- Nabika, T. 84
 Nachman, M.W. 173
 Nadeau, J. 729
 Nadel, D. 270, 318
 Nagasaki, M. 874
 Nagel, S. 292, 319
 Nagireddy, A. 496
 Nagylaki, T. 143, 173
 Naidoo, T. 316
 Naik, A.K. 729
 Nainys, J. 759
 Nakatsuka, N. 793
 Nakhleh, L. 245
 Näkkäläjärvi, K. 318
 Nalls, M.A. 732, 943
 Naltet, C. 994
 Nap, J.P. 728
 Narasimhan, V.M. 793
 Narayanan, M. 732

- Nariai, N. 874
Naseri, A. 82
Natarajan, K.N. 754, 756, 758
Nauck, M. 732
Nauen, D.W. 758
Navarro, A. 173
Navarro, P. 273
Nayfach, S. 994
Naylor, M. 838
Ndila, C.M. 77
Neale, B. 83, 498, 595, 628, 839
Neale, B.M. 111–112, 627, 672, 795, 798
Neapolitan, R.E. 625
Nee, S. 1019
Need, A.C. 793
Neftel, C. 758
Neher, E. 344
Neher, R.A. 1019
Nehlich, O. 321
Nei, M. 143, 174, 216–217, 244–245, 319, 392,
394–395, 418, 497–498, 1019
Neiman, E.M. 730
Nejentsev, S. 676, 813
Nejman, D. 947
Nekhrizov, G. 315
Nelis, M. 271, 732
Nelle, O. 324
Nelson, B.J. 292
Nelson, C.C. 795
Nelson, D.L. 947
Nelson, K.A. 730
Nelson, M.R. 269
Nelson, M.R. 77, 270, 272, 418
Nelson, P.S. 869
Nelson, S.F. 727
Nemesh, J. 756
Neophytou, C. 497
Nesheva, D. 318
Nessel, L. 993
Netea, M.G. 992
Nettelblad, C. 525
Nettleton, D. 526
Neudecker, A. 872
Neuditschko, M. 493
Neuenschwander, S. 50, 315
Neuhausen, S. 840
Neuhausen, S.L. 674
Neuhauser, C. 143, 173
Neumann, H. 732
Neumann, R. 80
Neumann, S. 975
Nevanlinna, H. 674
Neven, P. 674
Nevins, J. 873
Nevins, J.R. 870
Nevo, E. 142
Newcombe, P. 649
Newman, B. 836
Newman, M.H. 729
Newton, J. 975
Newton, M. 216, 753, 756, 874
Nezi, L. 993
Ng, A.Y. 1014
Ng, B.K. 548
Ng, L. 726
Ng, M. 755
Ng, M.C. 676
Ng, P.C. 365, 797
Ng, T.C. 493
Ngai, J. 754, 758
Nguyen, H. 78
Nguyen, N. 947
Nguyen, Q.T. 728
Nguyen, T. 729, 947
Nguyen, X. 877
Ni, J. 994
Ni, P. 317
Niblett, D. 944
Nica, A.C. 730
Nicholas, F.W. 792
Nicholas, S.L. 726
Nicholas, T.J. 788
Nichols, R.A. 76, 268, 548
Nicholson, G. 272
Nicholson, J.K. 973–974
Nicholson, M.R. 996
Nicholson, P. 530
Nickel, B. 315, 319
Nickerson, D.A. 77, 269, 623, 629
Nickerson, E.K. 1017–1018
Nicklisch, N. 316
Nicolae, D.L. 50, 111, 726, 732
Nicolau, J. 676
Nielsen, F.C. 76, 82, 314, 492, 593–594
Nielsen, H.B. 994
Nielsen, K. 321

- Nielsen, R. 49, 76, 79, 82, 142, 216, 243, 271–273, 291–293, 316, 318, 320–322, 365, 367, 394–395, 415–420, 492, 495–496, 593–594
- Nielsen, T. 994
- Nigst, P.R. 293, 322–323
- Nikaido, I. 757
- Nikiforova, S.V. 367
- Nikkel, S.M. 791
- Niknafs, N. 793
- Nikpay, M. 649
- Niroula, A. 793
- Nisbett, J. 729
- Nishikawa, K. 364
- Nishimura, S.Y. 760
- Nissim, I. 994
- Niu, B. 343
- Niu, N. 759
- Nivard, M.G. 728
- Nkadori, E. 730
- Noah, A. 174
- Noble, J. 80
- Nocedal, J. 49
- Noda, T. 796
- Nodell, M. 729
- Noggle, S. 794
- Nolan, G.P. 754
- Nolte, I.M. 813
- Noonan, J.P. 320
- Nordborg, M. 82, 171, 174, 293, 320
- Nordenfelt, S. 316, 318, 322
- Nordestgaard, B.G. 671
- Nordman, E. 758
- Nordsiek, G. 947
- Norman, M.F. 143
- Norman, P.J. 273
- Norman, T.M. 753–754
- Noronha, A. 994
- Norris, M. 757
- Northoff, B.H. 323
- Norton, H. 293, 323
- Notarangelo, L.D. 788
- Nothnagel, M. 270, 549
- Noto, M.J. 996
- Notohara, M. 174
- Nott, D.J. 874
- Noueiry, A. 874
- Novak, A.M. 78, 82
- Novak, M. 270, 318–319, 418
- Novak, S. 143
- Novelletto, A. 626
- Novembre, J. 76, 272, 314, 418, 492, 497
- Novod, N. 291, 316
- Novy, R.G. 529
- Nowak, M.A. 391
- Nowakowski, T.J. 757
- Nowicka, M. 946
- Ntranos, V. 757
- Ntzoufras, I. 871, 873
- Nugent, T. 344
- Nuida, K. 83
- Numminen, E. 1019
- Nunes, V. 593
- Nuri, Z. 730
- Nurnberg, P. 270
- Nusbaum, C. 291, 316, 595
- Nussbaum, R.L. 796
- Nussberger, B. 497
- Nussey, D.H. 455
- Nusskern, D.R. 729
- Nutile, T. 110
- Nuýnez, J.K. 753
- Nuzhdin, S.V. 757
- Nyakatura, G. 947
- Nyambo, T.B. 291
- Nyante, S. 270
- Nyholt, D. 837
- Nyholt, D.R. 595–596, 813
- Nyhus, J. 726
- Nykamp, K. 793
- Nyman, J. 758
- Nýomper, A. 569

O

- O'Brien, J. 994
- O'Connell, J. 77, 82, 109, 111–112, 628
- O'Connor, B.D. 317
- O'Dell, C.N. 947
- O'Donnell, C.J. 728
- O'Donnell-Luria, A.H. 795
- O'Donovan, M.C. 627
- O'Dushlaine, C. 839
- O'Farrell, A.M. 726
- O'Hara, R.B. 874
- Oakey, H. 527
- O'Maille, G. 974

- O'Neill, O. 570
 O'Reilly, K.M. 1016
 O'Reilly, S. 78, 270, 318
 O'Roak, B.J. 792
 Ober, C. 76
 Ober, E.S. 524
 Oberg, A.L. 757
 Oberhardt, M. 994
 Oddson, A. 79
 Odegård, J. 812
 Oefner, P. 995
 Oesterheld, M. 727
 Oettler, G. 528
 Ofengheim, D. 757
 Ogburn, E.L. 943
 Ogden R. 493
 Ogilvie, H.A. 244
 Ogilvy-Stuart, A.L. 1017–1018
 Ogunju, J. 838
 Ogunju, J.O. 528
 Oh, E.C. 726
 Ohkubo, T. 84
 Ohm, P. 570
 Ohno-Machado, L. 797
 Ohta, K. 80
 Ohta, T. 82, 394
 Okhuysen, P.C. 993
 Oksenberg, J.R. 269
 Okwuonu, G. 947
 Olalde, I. 320, 322
 Olason, P.I. 80, 111
 Oldfield, C. 345
 Oldham, M.C. 729
 Olek, A. 944
 Olivares, E.C. 944
 Oliver, K.L. 788
 Oliver-Williams, C. 677
 Olm, M.R. 992
 Olmstead, J.W. 522
 Olry, A. 795
 Olsen, K.M. 794
 Olson, J.E. 674
 Olsson, L.M. 996
 Oltvai, Z.N. 725, 730
 Omberg, L. 268
 Omrak, A. 316–317, 320, 323
 Onar, A. 730
 Ondov, B.D. 1019
 Ong, K.K. 673, 945
 Ongen, H. 730
 Ongyerth, M. 292, 320
 Onorato, A.J. 549
 Onozato, M.L. 759
 Onuchic, J. 344
 Onyango, P. 944
 Oosting, M. 992
 Opgen-Rhein, R. 974
 Opolon, P. 994
 Orcutt, B.C. 363
 Oriel, S. 756
 Orlando, L. 49, 83, 273, 293, 314–316,
 320–322, 324, 497
 Orozco-terWengel, P. 493
 Orr, H.A. 143
 Osada, N. 80
 Oshlack, A. 758, 875, 945
 Osipova, L. 273, 318
 Oskolkov, N. 790
 Osman, M. 84
 Ossó, A. 528
 Ostrander, E.A. 85, 500
 Ostrer, H. 268, 271
 Ota, T. 392–393
 Otecko, N.O. 293
 Otley, A.R. 993
 Otoo, E.J. 85
 Ott, J. 595
 Otto, T.D. 758
 Oualkacha, K. 675
 Oughtred, R. 789
 Ouwehand, W.H. 677, 945
 Ovchinnikov, S. 342–344, 1019
 Overington, J. 365
 Overington, J.P. 367
 Owen, M.J. 627
 Ozenne, B. 835
 Özer, F. 317
- p**
- Pääbo, S. 290–293, 314, 316–322, 362
 Paajanen, P. 322
 Pachter, L. 754, 757, 875
 Packer, J.S. 754
 Paderewski, J. 528
 Padhukasahasram, B. 82
 Padmanabhan, R. 798
 Padyukov, L. 945
 Paetkau, D. 497

- Pagani, L. 272, 274, 794
Page, A.J. 1015
Pagel, P. 727
Pagnani, A. 342–344, 1016
Pagnon, J. 727
Pai, A.A. 730, 869, 943
Paige, E. 677
Paigen, K. 82
Painter, I.S. 217
Paja, L. 314
Pak, R.A. 753
Pakendorf, B. 272
Pakstis, A.J. 549
Palamara, P.F. 81, 111, 497, 595
Pálfi, G. 314
Palin, K. 112
Palmedo, P. 345
Palmer, C.D. 175, 269
Palmer, L.J. 623
Palmer, S. 947
Palo, J. 270
Palomaki, G. 838
Palomo, A. 324
Palsson, B. 993–994
Palstra, F.P. 497
Palta, J.P. 529
Pamilo, P. 244, 394
Panagiotou, O. 839
Panayi, M. 80
Pandian, R. 947
Pandini, A. 344
Pankow, J.S. 943
Panousis, N.I. 730
Panula, S.P. 757
Pap, I. 315
Papadakis, J. S. 528
Papageorgopoulou, C. 48
Papalex, E. 757
Papanicolaou, G.J. 270
Papastamoulis, P. 874
Papenfuss, T.J. 496
Papili Gao, N. 757
Papin, J. 994
Papp, J.C. 111
Pappu, R. 345
Paquet, D. 794
Pardo, L. 366, 876
Parekh, S. 760
Parfitt, T. 268
Parham, P. 273
Pariani, M. 788
Parik, J. 268, 318, 323
Parisi, G. 363, 365
Park, C.M. 362
Park, H. 344, 1019
Park, J. 944, 946
Park, R.F. 530
Park, T. 874, 944
Park, Y. 730, 946
Parker, B.J. 320
Parker, D. 947
Parker, G.J. 549
Parker, K.A. 730
Parker, M. 569–570
Parkhill, J. 994, 1015, 1019
Parl, F.F. 628
Parmigiani, G. 837, 874, 975
Parnas, O. 753–754
Parrish, J. 947
Parry, B. 570
Parsawar, K. 549
Parson, W. 270, 548
Parts, L. 731, 875
Parvanov, E.D. 82
Parvizi, M. 726
Parvizi, P. 317
Pasanen, L. 528
Pasaniuc, B. 112, 268, 419, 630, 675, 727
Paschou, P. 628
Pascual, M. 493
Pasquale, L. 840
Pasquinelli, A. 728
Pasternak, S. 947
Pastina, M.M. 524
Pastinen, T. 945
Patch, A.-M. 790
Patel, A.A. 759
Patel, A.P. 758
Patel, D. 947
Patel, R.Y. 794
Patel, S. 727, 837
Patel, S.P. 993
Paten, B. 78, 82
Pathak, J. 293
Pati, A. 992
Pati, D. 869
Patrinos, G.P. 77
Patro, R. 757

- Patsopoulos, N.A. 726
Patterson, H.D. 528
Patterson, M. 112
Patterson, N. 49, 76, 83, 85, 111–113, 243, 269–272, 290–293, 315–322, 416–419, 497, 628, 672, 729
Paul, A. 754
Paul, C.L. 943–944
Paul, D.B. 570
Paul, D.S. 676–677
Paul, J.S. 112, 320
Pauling, L. 218, 394
Paulo, M.J. 525
Paulo, R. 873
Paunovic, M. 316
Pavlidis, P. 143, 418
Pavlopoulos, G.A. 1019
Pawitan, Y. 49, 626, 759
Pawluk, B.S. 791
Paxinos, P.D. 319
Payseur, B.A. 292
Pazos, F. 342, 344, 365
Pe'er, D. 728, 754, 757, 759
Pe'er, I. 77, 79, 317, 497, 593, 595, 628, 649–650, 794
Pearce, A.V. 947
Pearce, C.L. 626
Pearce, J.T. 972
Pearl, D.K. 243–244
Pearl, J. 676, 730
Pearse, D.E. 497
Pearson, D.M. 947
Pease, J.B. 419
Pedersen, A.-M.K. 364, 366–367, 395
Pedersen, B.S. 788, 946
Pedersen, C.-E.T. 497
Pedersen, C.N.S. 1019
Pedersen, J.S. 316, 320–321, 795
Pedersen, O. 994
Pedotti, P. 975
Pedro, M. 320
Pee, D. 834
Pei, J. 344
Peiffer, J.A. 525
Pejaver, V. 792
Pelan, S.E. 947
Pella, J. 272
Pelletier, E. 994
Pelletier, T.A. 243
Peltz, G. 728
Pemberton, T.J. 324
Pena, R.G. 316
Pencina, M. 834, 838
Peng, J. 273
Penney, K.L. 624
Penninger, J.M. 792
Pennings, P.S. 142, 417
Penninx, B.W. 732
Penny, D. 367
Penrose, L.S. 595
Pensar, J. 1019
Pepe, M. 838
Pepin, M.G. 794
Pereira, F.C.N. 291
Pereira, L. 268
Pereira, T. 319
Perez, J.I.A. 674
Perez, L. 947
Pérez, P. 526
Pérez-Elizalde, S. 523
Pericak-Vance, M.A. 624
Perisic, L. 729
Peristov, N.V. 290
Perkins, B. 674
Perkins, N. 595
Perkins, R. 625, 996
Perler, F. 394
Perlin, M.W. 549
Perola, M. 732
Perraudeau, F. 758
Perrigoue, J. 730
Perrin, N. 493
Perry, G.H. 317
Perry, J.R. 672–673, 676, 813, 945
Pers, T.H. 813
Pertea, G. 872, 875
Pertea, M. 995
Perumal, T.M. 726
Pesonen, M. 1019
Pessia, A. 1015
Petaccia de Macedo, M. 993
Peter, B.M. 292, 317, 320, 418
Peterlongo, P. 674
Peters, A. 943
Peters, E.C. 995
Peters, J.E. 677, 873

- Peters, L.A. 730
Peters, M.A. 726
Peters, M.J. 732, 943
Peters, T.J. 946
Peters, U. 838
Peterse, H.L. 731
Petersmann, A. 732
Peterson, C. 974
Peterson, J. 1019
Peterson, P. 732
Petkov, P. 82
Petkov, P.M. 82
Petkova, D. 77, 109, 497
Peto, J. 834
Peto, T.E. 1016
Peto, T.E.A. 1016
Petretto, E. 730, 869–870, 873–874
Petri, A.A. 290
Petrie, T. 47
Petropoulos, S. 757
Petrosino, J.F. 993
Petrou, S. 791
Petrov, D.A. 78, 417
Petrovski, S. 788–791, 793–794, 796, 798
Pettener, D. 273
Pettersson, F.H. 623–624
Pettett, R.M. 790, 944
Petukhov, V. 757
Petzelt, B. 273
Pevsner-Fischer, M. 993
Pfaffelhuber, P. 419
Pfannkoch, C.M.
Pfeifer, S.P. 418
Pfeiffer, R. 838, 840
Pharoah, P. 674, 834–835, 837–838
Philippe, H. 215–216, 363–366
Philippeau, G. 524
Phillippy, A.M. 1019
Phillips, A.D. 796
Phillips, J.W. 726, 729, 731
Phillips, K.A. 674
Phillips, M.J. 213
Phillips, M.S. 627
Phillips, P.C. 454
Phipson, B. 875
Piálek, J. 78
Piazza, A. 271
Picelli, S. 757
Pichler, I. 628
Pichler, S.L. 316
Pickrell, J.K. 271–272, 292, 320, 418–419, 628, 730, 943
Pidsley, R. 946
Piegorsch, W. 838
Piepho, H.-P. 522, 528, 838
Pierce, B. 676
Pierce, B.L. 628
Pierce, L.C.T. 796
Pierce-Hoffman, E. 795
Pierre, T.L. 273, 321
Pierson, E. 757, 759
Pietravalle, S. 524
Pietrusiak, P. 729
Pignone, M.L. 549
Piipari, M. 731
Pike, K.A. 273
Pike, M. 626
Pikkookana, P. 875
Pillai, N.S. 869
Pilling, L.C. 943
Pimentel, H. 754, 872, 875
Pinchin, R. 548
Pinelli, M. 788
Pinello, L. 755
Pinhasi, R. 270–271, 315, 317–320
Pinker, S. 570
Pinto, D. 726
Pinto, L.R. 524
Pinto, N. 548
Pique-Regi, R. 943
Piraux, F. 526
Pires, D.E.V. 794
Pirinen, M. 271, 628, 671, 795, 838
Piry, S. 494
Pitchappan, R. 271
Pitchford, W. 527
Piterman, N. 756
Piton, A. 794
Pjanic, M. 729
Plagnol, V. 49, 81, 292, 320, 324, 674
Plaisier, C. 727
Plaster, C. 272
Platko, J.V. 83, 419
Platonov, F. 318
Platt, A. 836
Platt, D. 320

- Platt, J.C.K. 788
Platt, R.W. 673
Platteel, M. 732
Plaza Reyes, A. 757
Plenge, R.M. 272, 628, 677
Pliner, H.A. 758
Plog, S. 317
Plomion, C. 324
Ploski, R. 270
Plotkin, J.B. 315, 416
Ploughman, L.M. 595
Plumb, R.S. 973
Pluskal, T. 974
Pluzhnikov, A. 82, 174, 630
Podlich, D.W. 523
Podtelezhnikov, A.A. 732
Poelwijk, F. 344
Poelwijk, F.J. 343
Poidinger, M. 732, 792
Poinar, H. 315, 320
Poirier, S. 530
Poirier-Colame, V. 994
Pokharel, P. 759
Pokutta, D. 314
Polak, P.P. 790
Poland, G.A. 628
Polanski, A. 272
Poldrack, R.A. 729
Pollak, E. 174, 497
Pollen, A.A. 757
Pollock, D.D. 344, 366, 392
Polson, N.G. 869–870, 874
Pommeret, D. 869
Pomp, D. 876
Pompan-Lotan, M. 993
Pompanon, F. 493, 497
Ponce de León, M.S. 273
Pond, S.L.K. 215
Ponder, B. 838
Pong-Wong, R. 811
Pons, J. 497
Pons, N. 993–994
Pontier, D. 1014
Pontil, M. 343
Pool, J.E. 272, 420
Poole, C. 673
Pooley, K.A. 648
Pooni, H.S. 526
Pop, M. 80
Popescu, E. 322
Poplin, R. 794
Popovic, D. 794, 796
Porsch, R. 837
Porteous, M. 837
Porter, K.M. 947
Portnoy, D.S. 495
Porto, M. 362
Porto Neto, L.R. 496
Pósa, A. 319
Posada, D. 216, 361, 1019
Posey, J.E. 794
Posma, J.M. 974
Pospieszny, L. 314
Poss, M.L. 495
Posth, C. 318
Postma, E. 454–455
Posukh, O.L. 273
Potash, J. 944
Potekhina, I.D. 324
Potter, B.A. 49
Potter, S.S. 755
Pounds, S.B. 730
Poustka, A. 947
Powell, J.E. 271, 729, 732–733, 813
Powell, J.R. 393
Powell, K. 84
Powell, W. 523, 525
Powers, M. 793
Poznik, G.D. 273, 321
Prabhakar, S. 758
Prabhu, S. 628
Prada, J.M. 1014
Pradel, R. 497
Prado, J.L. 322
Prado-Martinez, J. 318
Praestgaard, J.T. 732
Prainsack, B. 570
Prakhakaran, S. 757
Pramstaller, P.P. 498
Prangle, D. 47
Prato, J.D. 293
Pratto, F. 77
Pratts, E.C. 729
Pravenec, M. 730, 874
Preedy, K.F. 525
Preissner, R. 797
Prentice, R.L. 649–650, 838, 840
Prescott, N.J. 649

- Press, M.F. 270
 Press, W.H. 49
 Price, A.L. 49, 83, 111–112, 269–272, 292, 320, 418, 628, 672, 730
 Price, G.R. 454
 Price, N.D. 974
 Price, T.D. 314, 318
 Price, T.R. 943
 Prieto, D. 872
 Prieto, P.A. 993
 Prilusky, J. 795
 Pringle, T.H. 625
 Prins, B.P. 677
 Prins, J.F. 759
 Prinz, M. 548–549
 Pritchard, C.C. 869
 Pritchard, J.K. 83, 143, 172, 269, 272–273, 292, 320, 416, 418, 420, 493–498, 595, 730, 759, 943, 1016, 1019
 Prlic, M. 755
 Procaccini, A. 342
 Prokisch, H. 732
 Prokopenko, I. 626, 676
 Proserpio, V. 754
 Prout, T. 454
 Prudhomme, A.C. 730
 Prüfer, K. 290–292, 315–316, 318, 320
 Prügel-Bennett, A. 143
 Pruim, R.J. 628
 Prum, B. 594
 Pryce, J.E. 810, 812
 Przeworski, M. 76, 83, 272, 419–420
 Psaty, B.M. 270, 732
 Ptak, S.E. 316
 Pudovkin, A.I.
 Puechmaille, S.J. 498
 Puigserver, P. 729
 Puller, V. 1019
 Pulliam, H.R. 174
 Pupko, T. 363, 393
 Puranen, S. 1019
 Purcell, S. 83, 112, 498, 570, 595, 628, 676, 839
 Purdom, E. 754
 Puri, V.N. 730
 Purugganan, M.D. 526–527
 Pusch, C.M. 317
 Putz, B. 625
 Pybus, M. 322
 Pybus, O.G. 1016–1018
 Pyhäjärvi, T. 526
 Pyl, P.T. 792
 Pylkäs, K. 674
 Pyne, S. 756
- q**
- Qi, J. 320
 Qi, L. 649
 Qi, Q. 629
 Qi, R. 729
 Qi, X.G. 525
 Qian, F. 870
 Qiao, D. 83
 Qin, C. 344
 Qin, J. 994
 Qin, N. 994
 Qin, Y. 943
 Qin, Z.S. 81, 111
 Qiu, X. 754, 758
 Qu, A. 732
 Qu, B. 994
 Qu, Y. 726
 Quail, M.A. 758, 1017–1018
 Quang, D. 794
 Quertermous, T. 729
 Qui, W. 943
 Quince, C. 994
 Quinlan, A. 790
 Quinlan, A.R. 788
 Quinn, R.A. 993
 Quinodoz, M. 48
 Quinquis, B. 994
 Quirk, Y. 840
 Qureshi, H. 730
- r**
- Raaum, R.L. 270
 Rabbee, N. 112
 Rabiner, L.R. 272
 Rabinovic, A. 756
 Raby, B.A. 726
 Racimo, F. 291–293, 319–320, 419
 Raczky, P. 315
 Radke, D. 732
 Radovic, S. 527
 Radvanyi, F. 870
 Raes, J. 994–995
 Rafaels, N. 49–50, 83, 272
 Rafferty, I. 270

- Rafnar, T. 80, 111
Raftery, A.E. 48, 215, 871
Raghavachari, N. 728
Raghavan, M. 272, 320, 323
Raghavan, N.S. 794
Raghunathan, T.E. 870
Rahbari, R. 795
Rahimzadeh, V. 795
Rai, N. 793
Raimondi, D. 345
Rainville, P. 973
Raj, A. 273, 498, 755, 759
Raj, K. 945
Raj, T. 726
Rajan, D. 790
Rakocevic, G. 82
Rakyan, V.K. 943
Raley, J.C. 730
Ralph, P. 80, 83, 143
Ramachandran, S. 271, 273, 420
Ramakrishnan, U. 314
Ramalingam, N. 757
Raman, H. 530
Ramani, V. 754
Ramazzotti, D. 759
Rambaut, A. 213, 216, 1016–1018
Rambow, F. 753
Ramensky, V. 293
Ramirez, O. 320, 322
Ramos, C. 593
Ramos, E.M. 812
Rampp, M. 320
Ramsahoye, B. 945
Ramser, J. 946
Ramsey, Y. 947
Ramsköld, D. 754
Ramstein, M. 324
Ranciaro, A. 84
Rand, D.M. 394, 418
Rand, K. 837
Ranganathan, R. 343–345
Ranganathan, S. 525, 792
Rangel-Villalobos, H. 273
Rannala, B. 216–217, 244, 498, 500
Rantalainen, M. 759
Ranum, L.P.W. 797
Rao, A. 792
Rao, C.R. 454
Rao, J. 872
Rao, S. 835
Raoult, D. 994
Rasko, D.A. 83
Rasmussen, M. 273, 292, 314–315, 320–321
Rasmussen, S. 49, 272–273, 314, 318, 320–323
Rasnitsyn, A.P. 216
Rath, A. 795
Rathmann, W. 943
Ratmann, O. 1020
Rätsch, G. 83
Rattray, M. 871, 874
Ratz, S. 342
Rauber, C. 994
Rauch, A. 795
Rausch, T. 792, 796
Rausher, M.D. 454
Ravasz, E. 730
Ravcheev, D.A. 994
Rawlence, N.J. 291
Ray, N. 315
Raychowdhury, R. 754
Rayner, T. 530
Razhev, D.I. 290
Read, A.F. 1019
Readhead, B. 726
Réale, D. 455
Reardon, M.S. 730
Reaz, R. 244
Rebbeck, T.R. 291, 837
Rebhan, M. 795
Reboul, J. 974
Redd, A.J. 550
Reddy, A. 290
Redin, C. 794
Reding, M. 726
Redline, S. 270
Reed, F.A. 84
Reed, L.K. 495
Reents, R. 812
Rees, J. 673, 676
Reese, M.G. 791, 795
Reeser, J.C. 570
Reeve-Daly, M.P. 594
Regev, A. 753–754, 756, 758
Regier, M.A. 730
Regnaut, S. 494
Reher, D. 292
Rehm, H.L. 795
Reich, C.M. 811–812

- Reich, D. 49, 83, 85, 112–113, 175, 243, 269–272, 290–293, 315–316, 318–322, 416, 418, 624, 628
Reich, T. 595
Reichwald, K. 947
Reid, N.M. 243
Reid, S. 811
Reif, J.C. 526
Reijmers, T.H. 973
Reik, W. 756
Reilly, C., C., W. 874
Reiner, A.P. 270, 624, 943, 945
Reinert, K. 729
Reinhardt, F. 812
Reinius, B. 754, 757, 760
Reinius, L. 945
Reinmaa, E. 732
Reiter, E. 324
Reitman, M.L. 727
Relton, C.L. 673, 675
Remington, K.A. 729
Remm, M. 271
Remmert, M. 344
Ren, F. 366
Renaud, G. 292, 318, 320–323
Renault, P. 994
Rendeiro, A.F. 754
Renfrew, C. 321
Renner, M. 792
Renner, S.S. 528
Resende, M.F.R. 522
Reshef, Y.A. 81
Rest, J.S. 363
Reuben, A. 993
Reuther, J.D. 49
Rey, M.D. 530
Reynolds, D. 530
Reynolds, K.A. 345
Reynolds, P. 873
Reyondls, A. 268
Rezvani, K. 993
Rhodes, J.A. 242
Rhodes, M. 76
Rhodes, R. 570
Rhodes, S. 947
Riaz, A. 530
Ribaut, J.M. 527
Ribeiro, A. 758
Ricart, W. 996
Ricaut, F.-X. 76, 321
Rice, C.M. 947
Rice, D.P. 417
Rice, J. 595
Rice, K. 874
Rich, S.S. 270
Richard, C. 994
Richards, A.J. 528
Richards, B. 675
Richards, J.B. 110, 626, 630
Richards, M.B. 270, 318
Richards, M.P. 290
Richards, S. 795, 946
Richards, S.M. 790
Richardson, D. 673
Richardson, J.E. 796
Richardson, S. 214, 728, 759, 791, 869–870, 873–874, 876, 972
Richmond, C. 874
Richter, D.J. 83, 419
Ricklefs, R.E. 528
Ridderstrale, M. 729
Ridker, P.M. 111, 676
Ridler, K.A. 947
Riedelsheimer, C. 528
Rieder, M.J. 77, 269, 623
Rieseberg, L.H. 292, 500
Riess, O. 272
Riesselman, A. 346
Rietveld, C.A. 677
Riggi, N. 759
Riley, B. 839
Riley, D. 84
Riley, Z.L. 726
Rimm, E. 677
Ringbauer, H. 143
Ringler, E. 498
Ringler, M. 498
Rinn, J.L. 759, 875
Rio, J. 323
Rios, S. 366
RiouxB, J. 595
Ripatti, S. 109, 671, 732
Ripke, S. 672, 812–813, 833, 837, 839
Risch, N. 77, 273, 528, 548, 624, 628, 1019
Risch, R. 316
Rissman, A.I. 873
Risso, D. 754, 758
Ritchie, A.M. 1014

- Ritchie, M.D. 293, 628
Ritchie, M.E. 875
Ritchie, S.C. 727
Rivadeneira, F. 270, 594, 732
Rivas, G. 994
Rivas, M.A. 795
Rizvi, H. 994
Robbins, H. 49
Robert, C.P. 47, 494
Roberti, M.P. 994
Roberts, A. 875
Roberts, C. 731–732
Roberts, D. 367, 947
Roberts, G.O. 875
Roberts, T. 875
Robertson, A. 79, 142–143, 454, 495, 498
Robins, J. 674–676
Robinson, D.M. 216, 362, 366
Robinson, E. 595, 672
Robinson, J.D. 362
Robinson, J.T. 796
Robinson, M.D. 758, 875, 946
Robinson, M.R. 110, 733, 812–813
Robson, H.K. 314
Robson, P. 757
Roch, S. 244
Rocheleau, G. 172, 416
Rocke, D.M. 549
Rodan, A.Y. 626
Rödelsperger, C. 795
Roden, D.M. 293
Roden, M. 732
Rodman, C. 758
Rodrigo, A.G. 321, 392
Rodrigue, N. 364, 366
Rodrigues, J. 343
Rodrigues, P. 529
Rodriguez, B. 947
Rodriguez, P.C. 528
Rodriguez, W. 271
Rodriguez-Cintron, W. 268, 628
Rodriguez-Gil, J.L. 270
Rodriguez-Rivas, J. 345
Rodriguez-Santana, J. 268
Rodriguez-Varela, R. 316
Roeder, K. 548, 624, 726
Roetker, N.S. 943
Roewer, L. 548–549
Roff, D.A. 454
Roger, A.J. 364, 367
Rogers, A.R. 321
Rogers, C. 530
Rogers, J. 944, 947
Rogers, R. 292
Roggiani, M. 994
Rogowski, W. 839
Rohl, C.A. 731
Rohland, N. 112, 269–272, 291–292, 315–322,
 418, 497, 837
Rohou, A. 346
Rokas, A. 244, 362
Roland, C. 49
Rollefson, G. 270, 318
Romagosa, I. 529
Roman, H.E. 362
Romano, J.P. 48
Romano, V. 318
Romay, M.C. 529
Romblad, D.L. 730
Romero, I.G. 271
Romero Navarro, J.A. 529
Romey, M. 593
Romm, J.M. 317
Rommens, J.M. 626
Ronan, M.T. 314, 316
Ronchetti, E.M. 675
Ronen, R. 419
Rong, S. 420, 834
Rong, Y. 834
Rongione, M. 944
Ronin, Y.I. 142
Rönnegård, L. 529
Ronninger, M. 945
Ronquist, F. 215–216, 1019
Roodenberg, J. 319
Roodenberg, S.A. 270, 318–319, 418
Roodi, N. 628
Rootsi, S. 268, 317, 320
Rosa, P. 675
Rosas, A. 291, 316
Rose, G.A. 522
Rose, L.E. 290
Rosen, M.J. 992
Rosenberg, J. 1019
Rosenberg, N.A. 242–244, 269–270, 273, 316,
 324, 494, 595, 628, 1017–1018
Rosenbluth, A.W. 49, 216
Rosenbluth, M.N. 49, 216

- Rosenfeld, J.A. 794
 Rosenthal, A. 947
 Rosenthal, J.S. 874–875
 Rosenzweig, B.K. 419
 Roses, A.D. 624
 Roskin, K.M. 625
 Rosovitz, M. 83
 Ross, A. 973
 Ross, J. 972
 Ross, L.F. 570
 Ross, M.T. 946
 Rosset, S. 268, 835
 Ross-Ibarra, J. 525
 Rost, B. 343
 Rotella, J. 494
 Roth, C. 316
 Roth, F.P. 727, 796
 Rothberg, J.M. 316
 Rothhammer, F. 318
 Rothman, K.J. 628
 Rothstein, J.H. 792
 Rotter, J.I. 270
 Rotzschke, O. 732
 Rouquier, S. 84
 Rousset, F. 174
 Roussos, P. 726, 730
 Routy, B. 994
 Rowland, C.M. 628
 Roy, J. 78, 419
 Royal, C.D.M. 570
 Royall, J.J. 726
 RoyChoudhury, A. 242
 Roystvik, E.C. 81, 271
 Rozek, L.S. 946
 Rozen, S. 595
 Rozenblatt-Rosen, O. 758
 Rual, J.F. 728
 Ruan, D. 974
 Ruano-Rubio, V. 794
 Rubba, P. 789
 Rubin, D. 47, 671, 676
 Rubin, E. 995
 Rubin, E.M. 320
 Rubin, J.P. 364
 Rubinstein, J. 346
 Rubtsov, D.V. 974
 Rücker, G. 676
 Ruczinski, I. 49–50, 83, 272
 Rudan, I. 76, 318
 Rudan, P. 291–292, 316
 Ruderfer, D.M. 726, 812
 Rudolph, A. 674
 Rue, H. 876
 Ruepp, A. 727
 Ruff, T.G. 731
 Ruffieux, H. 875
 Ruiz-Linares, A. 318
 Ruiz-López, F.J. 811
 Ruiz-Romero, J. 593
 Runarsson, A. 945
 Runz, H. 677
 Ruotti, V. 873, 945
 Ruppert, D. 870
 Rusch, D.B. 729
 Rushmore, T.H. 731
 Russ, C. 291, 316
 Russell, A. 731
 Russell, R.B. 366
 Russell, S. 795
 Ruth, K. 674
 Ruther, A. 270
 Rutherford, M. 874
 Ruusalepp, A. 726, 729
 Ruvkun, G. 728
 Ruzicka, F. 493
 Ruzzante, D.E. 497
 Ryan, M. 530, 793
 Rybicki, B.A. 270
 Ryman, N. 495–496, 498
 Ryvkin, P. 759

S

- Saadi, H. 873
 Saag, L. 273, 314, 321
 Sabanés Bové, D. 48
 Sabatine, M.S. 795
 Sabatti, C. 594, 625
 Sabeti, P.C. 76, 83, 112, 419
 Sabino, J. 995
 Sablin, M. 314
 Sabuncyan, S. 944
 Sachdeva, H. 144
 Sachs, A. 729, 731
 Sachs, A.B. 725, 732
 Sadowski, M.I. 345
 Saelens, W. 754
 Saether, B.-E. 453
 Saeys, Y. 754

- Saez, I. 975
Saez-Rodriguez, J. 729
Sagasser, S. 757
Sagitov, S. 143
Sagoo, G.S. 649
Sagulenko, P. 1019
Sahakyan, H. 318
Sahana, G. 810
Sainger, R. 795
Saitou, N. 1019
Sajantila, A. 270, 318
Sakai, R. 796
Sakrejda, K. 454
Salamat, M. 140
Salame, T.M. 755
Salanti, G. 649
Salas, A. 318
Salazar-Garcia, D.C. 290, 320
Salem, R.M. 111
Sá-Leýao, R. 1015
Salfati, E.L. 945
Sali, A. 365, 992
Salim, A. 759
Salles, G. 791
Salmon, N. 875
Salomaa, V. 109, 671, 732
Salter, L.A. 243
Salzano, F.M. 315
Salzberg, S.L. 729, 872, 875, 995
Samadelli, M. 317
Samann, P.G. 625
Samaras, K. 946
Sambo, F. 83
Samino, S. 975
Sammeth, M. 729
Samocha, K. 798
Samocha, K.E. 795
Sampson, J. 834, 839
Samson, D. 524
Samsonova, M.G. 757
Samuels, Y. 791
Sanchez, E.M.R. 993
Sanchez, J.J. 314
Sanchez-Cobos, A. 361
Sánchez-Quinto, F. 316, 322
Sanchis-Juan, A. 795
SanCristobal, M. 496
Sandberg, R. 757
Sander, C. 343–346
Sander, T. 629
Sanders, R. 729
Sanders, S.J. 795
Sandholt, C. 836
Sandhu, M.S. 273
Sandman, D. 726
Sandoval-Velasco, M. 320
Sands, C.J. 974
Sands, J.L. 494
Sanes, J.R. 756
Sanguinetti, G. 755
Sankararaman, S. 175, 268, 273, 292–293, 320,
 322
Sanna, S. 628, 756
Santella, R. 674
Santi, N. 812
Santiago, E. 419, 499
Santoni, S. 530
Santos, F. 494, 995
Santos, R. 728
Santos, R.D. 789
Santure, A.W. 499
Sanz, M. 320
Sapolsky, R. 595
Saqi, M.A.S. 366
Sargent, D.J. 870
Sarkar, A. 674, 730
Sarkar, D. 874
Sartor, M.A. 946
Sartori, A. 317
Sarwar, R. 874
Sasaki, M. 83
Sasani, T.A. 788
Satija, R. 754–758
Satler, J.D. 243
Sato, Y. 728
Satterstrom, F.K. 790
Saunders, A.M. 624
Saunders, I.W. 174
Saura, M. 498
Savage, L.J. 49
Savolainen, V. 322
Sawyer, E.J. 674
Sawyer, S. 292, 318, 320–323
Saxena, R. 113
Sayer, D. 322
Sayle, R.A. 366
Sayyari, E. 244
Scadden, D. 757

- Scaife, A.A. 529
 Scally, A. 83, 293
 Scarcelli, N. 526
 Scaturro, D. 78
 Schaafsma, G.C.P. 790
 Schaarschmidt, J. 366
 Schadt, E.E. 674, 725–733, 877
 Schaefer, J. 974
 Schaefer, R. 322
 Schaeffer, L. 757, 873
 Schäffer, A.A. 342, 728
 Schaffner, S.F. 76, 78, 83, 419, 798
 Schaid, D.J. 628, 943
 Scharfe, C. 343
 Scharpf, R.B. 945
 Schatz, M.C. 84
 Schaub, M.T. 756
 Scheet, P. 81, 83, 111–112, 270
 Scheffler, K. 363, 393–394
 Scheike, T.H. 835
 Scheiner, S.M. 454
 Scheinfeldt, L.B. 293
 Schelter, J.M. 731
 Schemper, M. 1017–1018
 Schemske, D.W. 453
 Schep, D. 346
 Scherb, H. 623
 Scherer, S. 731, 946
 Scherrer, M.P. 366
 Scheu, A. 48, 314, 318
 Scheynius, A. 945
 Schier, A.F. 758
 Schier, W. 324
 Schierup, M.H. 172, 292
 Schiettecatte, F. 788, 796
 Schifano, E.D. 947
 Schiffels, S. 83, 272–273, 322, 794
 Schillert, A. 629
 Schilling, J.W. 394
 Schimenti, J.C. 78
 Schimmack, G. 731
 Schirmer, M. 992
 Schisterman, E.F. 673
 Schlebusch, C.M. 273, 322
 Schlessinger, D. 947
 Schlotterbeck, G. 973
 Schlötterer, C. 418
 Schlüter, D. 454
 Schmeichel, D.E. 624
 Schmid, C.H. 676
 Schmidler, S.C. 362, 364, 872
 Schmidt, E.M. 630
 Schmidt, M.K. 674
 Schmidt, R.H. 526
 Schmitt, C.A. 293
 Schmitz, R.W. 291, 316, 318
 Schmitz Carley, C.A. 529
 Schmutzler, R.K. 674
 Schnabel, R.D. 813
 Schnable, P.S. 526
 Schnall-Levin, M. 760
 Schneider, A. 366–367
 Schneider, J.A. 78
 Schneider, M. 648
 Schneider, P.M. 548–549
 Schneider, R. 343
 Schnell, F.W. 529
 Schoch, K. 793
 Schoenherr, S. 81
 Schofield, E. 674
 Scholkopf, B. 625
 Scholz, S.W. 270
 Schön, C.C. 522
 Schönher, S. 110
 Schöniger, M. 216, 366
 Schönleben, M. 522
 Schooler, E. 1019
 Schork, A. 833
 Schork, N.J. 629
 Schrag, T.A. 529
 Schraiber, J.G. 290–291, 293, 318–319,
 322–323, 419
 Schramm, K. 732
 Schreiber, F. 342, 993
 Schreiber, G.J. 731
 Schreiber, S. 270
 Schrider, D.R. 419
 Schroeder, H. 314–315, 320, 322
 Schrooten, C. 812
 Schroth, G.P. 943
 Schubert, M. 316–317, 322, 324
 Schuelke, M. 795
 Schuenemann, V.J. 324
 Schuetz, E. 731
 Schug, A. 342
 Schulmeister, S. 216
 Schultz, P.G. 995

- Schultz, R. 291, 316
Schulze, C. 993
Schulzrinne, H. 1019
Schulz-Streeck, T. 528, 838
Schumann, E.L. 550
Schuppe-Koistinen, I. 972
Schurmann, C. 732
Schuster, L.C. 754
Schuster, S. 320
Schutz, J. 730
Schwartz, A.G. 270
Schwartz, D.A. 946
Schwartz, J. 943
Schwartz, M.K. 496, 498
Schwartz, R.M. 363
Schwarz, C. 320
Schwarz, D.F. 629
Schwarz, G. 758
Schwarz, J.M. 795
Schwarze, U. 794
Schweichel, R. 324
Schymick, J.C. 270
Scialdone, A. 754
Scolnick, E. 676
Scott, A.F. 788
Scott, A.J. 789
Scott, C.E. 946
Scott, G. 946
Scott, I. 839
Scott, J. 994
Scott, J.G. 870, 874–875
Scott, J.L. 730
Scott, L.J. 113, 626, 629, 649
Scott, R. 730
Scott, R.A. 629, 673
Scott, R.H. 797
Scott-Boyer, M.P. 875
Scutari, M. 522
Searle, S. 946
Sebaihia, M. 83
Sebat, J. 788
Seboun, E. 524
Seefried, F.R. 811
Seegobin, S. 839
Seelig, G. 796
Seelow, D. 795
Seemann, S. 944
Seemayer, S. 345
Segal, E. 993
Segata, N. 994–995
Seguin-Orlando, A. 273, 293, 316, 322–324
Segura, V. 836
Ségurel, L. 76
Sehra, H.K. 947
Seim, I. 795
Seldin, M. 791
Self, S.G. 394
Seligman, S.J. 788
Sell, C. 48, 314, 317–318
Sella, G. 144, 174
Seminara, D. 835
Semino, O. 268, 270, 317–318
Semler, M.W. 996
Semple, C. 293
Sendecki, J. 730
Sénécal, K. 797
Sengupta, D. 758
Sengupta, S. 630
Senn, H. 973
Senter, L. 839
Seoighe, C. 394
Serang, O.R. 524
Serino, M. 996
Serra, M. 319
Serre, D. 317
Service, P.M. 453
Service, S.K. 594, 625
Servin, B. 323, 529, 629
Sestan, N. 726
Sethurman, J. 875
Seung, H.S. 756
Severe, N. 757
Seymour, R. 493
Seynaeve, C. 674
Sforça, D.A. 524
Sgrò, C.M. 495
Sha, N. 872, 875
Shabalin, A. 732
Shadick, N.A. 272, 628
Shaffer, S. 755, 759
Shaffrey, L. 528
Shafi, A. 947
Shah, H. 730, 732
Shah, K.P. 725–726, 732
Shah, M. 317, 674
Shah, N. 794–795
Shah, R. 529
Shah, S. 943

- Shah, T. 649
 Shahin, A. 523
 Shahrezaei, V. 753
 Shalek, A.K. 755–756
 Shalizi, C.R. 495
 Sham, P. 498, 570, 595, 759, 837, 839
 Shameer, K. 730
 Shamoon-Pour, M. 270, 318
 Shan, G. 759
 Shanno, D.F. 49
 Shapiro, B. 320, 323, 1016
 Shapiro, J.L. 143
 Sharma, P. 993
 Sharma, R. 759
 Sharon, G. 995
 Sharopova, N.R. 728
 Sharp, K. 77, 82, 109–110, 112
 Sharp, P.M. 362
 Sharp, S. 677
 Sharp, S.J. 676
 Shashi, V. 790, 793
 Shattuck, M.R. 525
 Shaw, F.H. 454
 Shaw, R.G. 454
 Shaw-Smith, C. 947
 Shchetynsky, K. 945
 Shedlock, A.M. 85, 500
 Sheehan, N.A. 498, 672, 674
 Sheehan, S. 83, 144, 273, 419
 Shekhar, K. 756
 Shelburne, S. 993
 Sheldon, B.C. 452
 Shen, E. 726
 Shen, H. 947–948
 Shen, J. 943
 Shen, N. 595
 Shen, P. 972
 Shen, R. 943
 Shen, T.-C.D. 994
 Shendure, J. 292, 319–320, 754, 792, 796
 Shenfeld, D.K. 754
 Shennan, S.J. 48
 Sheppard, S.K. 1019
 Sheridan, E.M. 947
 Sheridan, R. 343–344
 Sherlock, R. 570
 Sherry, S.T. 77, 271, 548
 Shetty, P. 569
 Shi, F. 527
 Shi, H. 112, 419, 727
 Shi, J. 649, 834, 839
 Shi, L. 625
 Shi, P. 993–995
 Shi, S. 110
 Shi, W. 875
 Shianna, K. V. 793
 Shields, G.F. 497
 Shim, H. 494
 Shimizu, K. 83
 Shin, J. 758
 Shin, J.Y. 758
 Shin, M.J. 673, 675
 Shin, S.Y. 673, 675
 Shinobu, L.A. 726
 Shirasaki, D.I. 731
 Shiroishi, T. 80
 Shirsekar, G. 324
 Shishlina, N. 314
 Shizuya, H. 84
 Shmulevich, I. 974
 Shoemaker, D.D. 728, 731
 Shorrocks, B. 493
 Short, P.J. 795
 Shoshani, A. 85
 Shou, S. 726
 Shoush, O. 571
 Shownkeen, R. 947
 Shpak, M. 140
 Shpall, E.J. 993
 Shrine, N. 82, 112
 Shriner, D. 876
 Shringarpure, S.S. 273
 Shrubsole, M.J. 674
 Shu, X.O. 674
 Shue, B.C. 729
 Shuga, J. 757, 759
 Shukla, A.K. 496
 Shungin, D. 629
 Shunkov, M.V. 292, 319–323
 Siao, C.J. 595
 Sibbesen, J.A. 873
 Sibinga, N. 730
 Sicheritz-Ponten, T. 273, 314, 321, 994
 Sidorenko, J. 813
 Sidow, A. 362, 366, 789
 Siebauer, M. 292, 316, 319–320
 Sieberts, S.K. 726–727, 731
 Siegmund, M. 83, 316

- Siegert, S. 549
Siegismund, H.R. 497
Siegmund, K.D. 945, 947
Siepel, A. 83, 269, 316, 366, 419, 795
Sievers, F. 342–343, 345
Siewert, K.M. 419
Sifrim, A. 794, 796
Siggia, E.D. 754
Signer-Hasler, H. 493
Signorello, L.B. 270
Sigurdsson, A. 80
Sihag, S. 729
Sikic, M. 797
Sikora, M. 49, 293, 314–315, 321–323, 418
Silber, S. 593
Silberstein, L. 756
Silk, M. 796
Sillanpää, M.J. 493, 528, 874–875, 1015
Silva, C.B.C. 524
Silva, N.M. 319, 323
Silva, R.R. 524
Silverman, D. 840
Silverman, J.S. 84
Silvestri, V. 836
Silvey, V. 528
Sim, X. 629
Simard, J. 629, 674, 836
Simcha, D. 945
Simecek, N. 80
Simianer, H. 811
Simmonds, J. 530
Simon, I. 947
Simon, J. 728
Simon, P. 811
Simonds, E.F. 754
Simons, J.F. 316
Simons, Y.B. 174
Simon-Sanchez, J. 270
Simonsen, K.L. 174
Simonsen, M. 1019
Simonti, C.N. 293
Simpson, J.T. 947, 994
Sinelnikov, I.V. 975
Singarayer, J. 273
Singer, A. 758
Singh, L. 271, 292, 318
Singh, S. 797
Singh, T. 343
Singleton, A. 273, 322
Singleton, A.B. 270, 324, 732, 943
Sinha, D. 758
Sinha, R. 993
Sinha, S. 84
Sinsheimer, J.S. 111
Sirak, K. 270, 318–319, 418
Sirén, J. 78, 1015
Sirmans, E. 993
Sirota, M. 789
Sirota-Madi, A. 993–994
Siska, V. 270, 315, 317
Sisson, S.A. 49
Sitter, C.D. 730
Siuzdak, G. 974, 995
Sjoberg, M. 759
Sjödin, P. 273
Sjödin, P. 174, 322–323
Sjögren, K.-G. 314, 317, 323
Sjolander, A. 626
Skaar, E.P. 996
Skinner, M.K. 725
Sklar, P. 83, 112, 498, 595, 726
Skoglund, P. 273, 292, 315–323
Skoglund, T. 323
Skogsberg, J. 726
Skol, A.D. 649
Skolnick, M.H. 593, 595
Skorecki, K. 268
Skotte, L. 321, 323, 498
Skrzynia, C. 839
Skuce, C.D. 947
Skwark, M.J. 344–345, 1016, 1019
Skyrms, B. 47
Slade, D. 728
Slagboom, P.E. 947
Slatkin, M. 83, 110, 175, 243, 290–293,
 315–322, 324, 418–419, 500, 624
Slattery, M. 836
Slepchenko, S.M. 290
Slichter, C.K. 755
Slichter, D. 676
Slieker, R.C. 947
Sloan, R. 673
Sloane, R.S. 993
Slon, V. 321, 323
Sloofman, L.G. 726
Slooten, K. 549
Slupsky, C.M. 975
Small, D.S. 675

- Small, G.W. 624
Small, K.S. 112, 625
Smallwood, M. 730
Smart, A. 569–570
Smedby, K. 834
Smedley, D. 788
Smeekens, S.P. 992
Smeeth, L. 649
Smerick, J.B. 549
Smets, M. 760
Smilde, A.K. 973
Smit, J.H. 732
Smith, A.N. 80
Smith, A.V. 272
Smith, B.D. 524
Smith, C.A. 974
Smith, C.A.B. 594–595
Smith, C.I. 273
Smith, C.L. 796
Smith, D.G. 273
Smith, E.N. 733
Smith, G.D. 673
Smith, H.O. 729
Smith, J.L. 549
Smith, J.M. 175, 419
Smith, K.G. 873
Smith, K.P. 530
Smith, M. 758, 872
Smith, N.G. 78, 394
Smith, N.J. 112, 629
Smith, P. 834
Smith, R.C. 731, 839
Smith, S. 110
Smith, T.J. 729
Smits, B.M.G. 873
Smolich, B.D. 726
Smoller, J.W. 676
Smuga-Otto, M. 82
Smulders, M.J.M. 523
Smyth, D.J. 624
Smyth, G.K. 758, 875
Snape, J.W. 526
Sneath, P.H.A. 217
Sneddon, T. 793
Snell, R. 811
Snieder, H. 813
Sniegowski, P.D. 141
Sninsky, J.J. 416
Snyder, M. 728, 759
So, H.-C. 839
Sobel, E. 112, 594–595
Sobreira, N. 796
Sochard, M.R. 391
Sodergren, E. 417, 946
Söding, J. 344–345, 1019
Soenov, V.I. 314
Sofer, T. 947
Sohn, M. 995
Soininen, P. 975
Sokal, R.R. 216–217, 1015
Solberg, T. 812
Solignac, M. 494
Solovieff, N. 676
Somel, M. 317
Somera, A.L. 730
Soneson, C. 758
Song, G. 529
Song, H. 758
Song, J.-H. 1015
Song, M. 839
Song, P. 944
Song, W.M. 730
Song, Y.S. 49, 83–84, 112, 268, 273, 293, 320,
323, 419
Soodyall, H. 273, 322
Soraggi, S. 293
Soranzo, N. 110, 677
Sorek, R. 419, 993
Sorensen, S. 993
Sorger, P.K. 729
Soria, J.-C. 994
Sorrells, M.E. 525
Sorrels, M.R. 525
Sotheran, E.C. 947
Soules, G. 47
Sousa, V.C. 47, 269, 315, 319, 323
Southey, M.C. 674
Southwick, A.M. 271
Souza, A.P. 524
Souza, G.M. 524
Sova, P. 733
Spain, S.L. 629, 677
Spang, R. 995
Spangenber, G.C. 524
Spangler, J. 317
Sparks, R. 1019
Spector, T. 728
Speed, D. 144, 529, 676, 813, 839

- Speed, T. 873, 943
Speed, T.P. 112, 758–759, 944–945
Speed, W.C. 549
Speliotes, E. 839
Speller, C.F. 291
Spelman, R. 811
Spelman, R.J. 811
Spence, J.P. 293
Spencer, C.C. 81, 109, 628, 671
Spencer, C.C.A. 628–629, 1016
Spencer, C.E. 549
Spencer, C.L. 496
Spencer, C.N. 993
Spencer, J. 595
Spiegelhalter, D. 496
Spiegelman, B. 729
Spielman, R.S. 726, 729
Spielman, S.J. 363, 394
Spielmann, M. 796
Spiller, W. 676
Spitz, M. 270
Spitze, K. 419
Spong, G. 494
Sprague, A.C. 730
Springer, M. 343
Sørensen, K.D. 944
Sridhar, S. 273
Srinivasan, S. 526
Sripracha, R. 111
Stacey, D. 677
Stacey, S.N. 649
Stade, B. 317
Stadler, T.J. 215
Stafford, T.W. 273, 320–321
Stagegaard, J. 322
Stahl, E.A. 726, 790, 812
Staley, J.R. 676–677
Stam, L.F. 732
Stam, P. 144, 529
Stamatakis, A. 217, 1019
Stamatoyannopoulos, J.A. 623
Stamenova, E.K. 948
Stamler, J. 972, 974
Stammler, F. 995
Standen, V. 273
Stanton, D. 217
Staples, J. 629
Starikovskaya, E.B. 318
Starita, L.M. 793, 796
Stark, A.L. 946
Statham, A.L. 946
Stäuble, H. 324
Stec, A. 524
Stecher, G. 393
Steed, A. 530
Steel, M. 217, 245, 367, 394
Steel, M.F. 872–873
Steele, C. 549
Steele, M.A. 293
Steemers, F.J. 754
Steer, S. 839
Stefanescu, G. 270, 318
Stefanov, V.T. 595
Stefanski, L.A. 870
Stefansson, H. 111
Stefansson, K. 111, 730
Steffen, D. 946
Steffens, M. 625, 629
Stegeman, A. 1020
Stegle, O. 731, 754, 871, 875
Stégmár, B. 319
Stein, A.D. 947
Stein, C.M. 454
Stein, L. 595
Steinberg, S. 80, 111
Steinfeld, I. 947
Steingruber, H.E. 947
Steinrücken, M. 49, 272, 293, 323
Stenderup, J. 273, 314, 321
Stenson, P.D. 788, 793, 796
Stenzel, U. 291–292, 314, 316, 319, 321
Stepaniants, S.B. 728
Stephan, W. 143, 416–420
Stephany, J.J. 793
Stephens, D. 972
Stephens, J. 795
Stephens, M. 77, 81, 83, 85, 110–113, 143,
269, 271–273, 323, 418, 494–498, 500, 525,
629–630, 730, 813, 841, 870–877, 1016,
1019–1020
Stephens, Z.D. 84
Stephenson, L.D. 730
Stern, A. 973
Sternberg, M.J.E. 366
Sterne, J.A. 649, 676
Stern H.S. 453
Steuer, R. 974
Stevens, A. 726

- Stevens, C. 522
Stevens, J. 549
Stevison, L.S. 324
Steward, C.A. 947
Stewardson, K. 270, 316–319, 418
Stewart, C.-B. 393–394
Stewart, E. 730
Stewart, I.D. 676
Stewart, J. 593
Stewart, J.D. 945
Stewart, R. 756
Stewart, R.M. 753, 873
Steyerberg, E. 839
Steyn, M. 322
Stift, M. 524
Stigler, S.M. 49
Stiller, M. 497
Stinchcombe, J.R. 454
Stinchcombe, J.R.,
Stingo, F.C. 875, 974
Stirling, I. 497
Stirzaker, D.R. 392
Stitt, M. 528
Stoahzman, M. 996
Stoesser, N. 1016
Stoffel, M. 79, 593
Stojmirovic, A. 730
Stone, A. 318
Stone, D.J. 732
Stone, E.A. 362, 366, 789
Stoneking, M. 272, 292–293, 318, 321
Stong, N. 793
Stopfer, J. 835
Storá, J. 316–317, 319–321, 323
Storey, J.D. 623, 629, 728, 731, 871, 945
Storey, R. 947
Storfer, A. 495
Stormo, G.D. 343
Stoughton, R. 728, 731
Stoughton, R.B. 731
Strachan, D.P. 112
Stram, D.O. 650
Strandén, I. 813
Stranger, B.E. 730
Strauch, K. 732
Straussman, R. 947
Street, N.R. 526
Street, T. 76
Streets, A.M. 759
Strimmer, K. 974
Strittmatter, W.J. 624
Strizich, G. 629
Strobeck, C. 84, 175, 497
Strobel, M. 48
Strom, S.C. 731
Strom, S.S. 270
Strong, R.V. 730
Stubbington, M.J.T. 758
Studenski, S. 943
Stumpf, C.L. 973
Stumpf, M.P. 84
Stumpf, M.P.H. 754
Stumpflen, V. 727
Stumvoll, M. 270, 318
Stunnenberg, H.G. 796, 945
Sturcke, A. 728
Su, B. 291
Su, C. 875
Su, G. 810
Su, L.Y. 293
Su, M. 595
Su, S.-Y. 876
Su, Y.H. 293
Su, Z. 112, 625, 629
Suarez, B.K. 595
Subramanian, A. 729
Subramanian, G. 730
Suchard, M.A. 213, 216, 1016
Suchy-Dicey, A. 732
Sudbery, I. 758
Sudbrak, R. 946
Sudbury, A. 144
Sudlow, C. 84, 629, 677
Sudmant, P.H. 291–292, 318–320, 796
Sudoyo, H. 76
Süel, G.M. 345
Suez, J. 993
Sugden, L.A. 420
Sugnet, C.W. 625
Suh, E. 730
Suhre, K. 677
Sujobert, P. 791
Sukernik, R. 318
Sukhavasi, K. 726
Sukumaran, J. 498
Sul, J.H. 594, 625
Suleimanov, Y.V. 626
Sulem, P. 111, 649, 674

- Sullivan, P.F. 112, 676, 726, 732
Sulpice, R. 528
Sulston, J.E. 947
Sumlin, W.D. 497
Sun, B. 677
Sun, B.B. 676–677
Sun, C. 315
Sun, F. 85
Sun, G. 757
Sun, J. 730
Sun, L. 648, 650
Sun, M. 498
Sun, M.-M. 756
Sun, Q. 525, 527
Sun, S. 346
Sun, W. 732
Sun, Y.V. 50
Sunagawa, S. 994–995
Sundaresan, V. 496
Sunkin, S.M. 726
Sunyaev, S. 112, 791, 841
Sunyaev, S.R. 361, 416, 726, 788, 798
Surani, M.A. 758
Surendran, P. 676–677
Susko, E. 364, 367
Sutto, L. 345
Sutton, A.J. 649, 676
Sutton, G.G. 729
Sutton, R. 528
Suver, C. 727, 731–732
Suzuki, H. 80
Suzuki, Y. 394
Svardal, H. 293
Sved, J. 84
Sved, J.A. 454, 498
Sveinbjornsson, G. 629
Swendsen, M. 812
Svensson, E.M. 316, 320
Svensson, V. 758
Sverdlov, S. 595
Sverrisdóttir, O.O. 316, 323
Svetnik, V. 731
Swain, P.S. 754
Swallow, D.M. 84
Swaminathan, H. 549
Swan, A.A. 810
Swann, R.M. 947
Swanson, S.A. 677
Swanson, W.J. 395
Swarbreck, D. 947
Swarts, K. 529
Swennes, A.G. 993
Swenson, M.S. 244
Swerdlow, A.J. 674
Swerdlow, H.P. 758
Swertz, M.A. 992
Swofford, D.L. 215, 217
Syed, H. 629
Sykes, N. 319
Sylvester, K. 730
Syn, C.K.C. 548
Szafer, A. 726
Szczęśniak, M.W. 754
Szécsényi-Nagy, A. 316, 319
Szeverényi, V. 314
Szklarczyk, D. 796
Szpankowski, L. 757
Szpiech, Z.A. 270, 324, 628
Szurmant, H. 345, 1020
Szurmanc, H. 346
Szymczak, S. 629, 839
Szymura, J.M. 144
- t**
- Tabara, Y. 84
Taber, J.M. 796
Taberlet, P. 493, 497, 499
Tabor, H.K. 629
Tabor, P.E. 947
Tachmazidou, I. 650
Tadesse, M.G. 872–873, 875
Tadmor, M.D. 754
Taenzer, S. 243
Taggart, J.B. 812
Tagliabruni, A. 270
Tai, S. 321
Tajima, F. 84, 175, 420, 497
Takahata, N. 175, 245, 498
Takano, E. 992
Takashima, K. 796
Takebayashi, N. 500
Takeuchi, F. 84
Talamo, S. 292, 320–321
Talavera, D. 345
Talbot, K. 726
Talbot, M. 529
Tallmon, D.A. 496, 498
Talmor, Y. 391

- Talmud, P.J. 625
 Talukdar, H.A. 726
 Tam, H.-H. 947
 Tamayo, P. 729
 Tambets, K. 273, 314, 320–321, 323
 Tams, S.H. 528
 Tamura, K. 217, 393
 Tamura, M. 80
 Tamuri, A.U. 366, 394
 Tan, J. 80
 Tan, L. 80
 Tan, L.S. 974
 Tan, R. 839
 Tan, S.Y. 549
 Tanabe, M. 728
 Tanaka, H. 366
 Tanaka, M.M. 49
 Tanaka, R. 530
 Tanaka, T. 943, 945
 Tanay, A. 755
 Tandon, A. 49, 83, 269, 272, 291–292,
 319–320, 837
 Tang, F. 758, 871
 Tang, H. 271, 273, 974, 1019
 Tang, K. 80
 Tang, T. 527
 Tang, Y. 758
 Tang, Z. 995
 Tannier, E. 213
 Tanriverdi, K. 728
 Tao, R. 244
 Tap, J. 994
 Tapió, M. 493
 Taravella, A.M. 418
 Tarazona, S. 754
 Tarekegn, A. 272, 274, 318
 Tarrant, C. 569
 Tarrant, T.K. 730
 Taschner, P.E.M. 790
 Tasse, A.M. 569
 Tatt, I. 649
 Taub, M. 945
 Taudien, S. 243, 947
 Tautenhahn, R. 975
 Tavaré, S. 49, 81, 171–172, 174–175, 217, 245,
 366, 1019
 Tavtigian, S.V. 629, 796
 Tawbi, H. 993
 Tayeb, A. 875
 Taylor, D. 548, 550
 Taylor, F. 677
 Taylor, H.A. 270
 Taylor, H.M. 142
 Taylor, J. 838–839
 Taylor, K.D. 624
 Taylor, T. 947
 Taylor, W. 344
 Taylor, W.R. 342, 344–345, 364, 366
 Teague, B. 947
 Teare, H. 569
 Teasdale, M.D. 315
 Technow, F. 523, 527–528
 Tegel, W. 324
 Teh, Y.W. 1016
 Teichmann, S.A. 754–758
 Tekola-Ayele, F. 791
 Telenti, A. 795–796
 Telesca, D. 975
 Telesco, S.E. 730
 Telis, N. 172, 416
 Teller, A.H. 49, 216
 Teller, E. 49, 216
 Temple, R. 839
 Templeton, A. 323
 Tenaillon, O. 84
 Tenesa, A. 273, 498
 Teng, D.H. 629
 Teo, S. 794
 Teo, Y.Y. 112, 625
 Teplitsky, C. 452
 Terberger, T. 314
 Terhorst, J. 49, 84, 323
 Terpstra, P. 725
 Terradas, X. 324
 Terrin, N. 676
 Terry, J.M. 759
 Teschendorff, A.E. 947–948
 Teshima, K.M. 420
 Teslenko, M. 216
 Teslovich, T. 840
 Teslovich, T.M. 628
 Tetchner, S. 343
 Tetens, J. 493
 Tettelin, H. 84, 995
 Tetzlaff, M.T. 993
 Teumer, A. 732
 Teusink, B. 995
 Tewari, M. 728

- Thaiss, C. 993
Thakar, J. 725
Than, C. 245
Thangaraj, K. 271, 292, 318
Theis, F.J. 754
Theodoridis, G.A. 973
Theunert, C. 292, 320
Thiele, I. 994–995
Thieringer, R. 731–732
Thiery, A. 548
Thomas, A. 496
Thomas, D. 627, 649, 840
Thomas, E. 797
Thomas, K. 947
Thomas, L. 83, 498, 595, 839
Thomas, M.G. 48, 272, 274, 314, 318–319,
 323–324
Thomas, P.B. 795
Thomas, P.D. 365
Thomas, R. 730
Thompson, C. 944
Thompson, D. 674
Thompson, E.A. 85, 217, 492, 593–596
Thompson, J. 672, 732
Thompson, J.F. 322
Thompson, J.R. 672, 674
Thompson, R. 144, 527–529
Thompson, S. 672–673, 677
Thompson, W.C. 550
Thoms, J. 869
Thomson, C.E. 454
Thomson, J. 756
Thomson, J.A. 753, 873
Thomson, K.L. 797
Thomson, N.R. 83
Thong, Z.H. 548
Thorleifsson, G. 80, 111, 730
Thorne, J.L. 215–216, 362–363, 365–367
Thornberry, J. 527
Thornton, J. 394
Thornton, J.M. 364
Thornton, K. 80
Thornton, T.A. 624
Thorogood, A. 796
Thorpe, A. 947
Thorsteinsdóttir, U. 111, 796
Threadgill, D.W. 726, 731
Thu, M. 757
Thuillet, A.C. 530
Thun, M.J. 270
Thuren, T. 676
Tian, Y. 245, 731
Tiao, G. 792
Tibshirani, R. 811, 840, 871, 973, 992
Ticehurst, F. 569
Tiede, T. 530
Tiemeier, H. 677
Tier, B. 110, 526
Tierney, L. 217
Tigchelaar, E.F. 992
Tikhonov, A. 320
Tikunov, Y. 975
Tillier, E.R.M. 367
Timmermans, M.J. 493
Timms, K. 947
Timp, W. 947
Timpson, A. 323–324
Timpson, N.J. 110, 649
Tin, A. 972
Tinch, A.E. 812
Tines, D.E. 628
Tint, N.N. 730
Tittle, N. 111
Tiret, L. 869–870
Tirosh, I. 756, 758
Tirrell, L. 726
Tischfield, J. 732
Tischkowitz, M. 835–836
Tishkoff, S.A. 84, 269, 291, 318
Tito, R.Y. 995
Titsias, M.K. 875
Tiwari, R. 496
To, T.H. 1019
Tobi, E.W. 947
Tobin, D.J. 320
Tobin, M. 82, 112
Todd, J.A. 77, 109, 649, 674, 677
Todd, J.L. 794
Todd, L. 732
Todd-Brown, K. 83, 498, 595, 628, 839
Tofanelli, S. 76
Toffolo, G. 83
Togan, İ. 317
Tokiwa, G. 725
Toland, A.E. 674
Tolonen, A.C. 995
Tomlinson, I. 674
Tompa, P. 345

- Tompkins, D.E. 729
Tomsho, L. 320
Toncheva, D. 76
Tondelli, A. 522
Tong, L. 595
Tong, W. 625
Tönjes, A. 270, 318
Tonkin-Hill, G. 1020
Tooker, M.E. 813
Toombs, J. 323
Topaloglou, T. 595
Topham, C.M. 367
Topol, A. 726
Topol, E.J. 629
Toppani, D. 757
Torgersen, J. 812
Torgerson, D.G. 50, 268
Torgrip, R.J. 972
Toro, M.A. 498
Torralbo, A. 1019
Torre, E. 755, 759
Torrents, D. 996
Torres, D. 674
Torres, J.M. 725
Torres, R. 796
Torroni, A. 270, 318
Torstenson, E.S. 725
Táth, G. 314
Toth-Petroczy, A. 345
Totir, R. 523, 527
Touchon, M. 84
Toyoshiba, H. 726
Trabetti, E. 110
Tracey, A. 947
Trachulec, Z. 78
Traherne, J.A. 80
Traill, L.W. 499
Trakalo, J. 526
Tran, L. 732
Tranchevent, L.-C. 796
Trapnell, C. 80, 754, 758–759, 872, 875
Trask, B.J. 84
Travers, K.J. 944
Travis, J. 454
Traxler, P. 754
Traynelis, J. 796
Traynor, B.J. 270
Treangen, T.J. 1019
Tregouet, D. 869
Trehearne, A. 797
Tremaroli, V. 996
Trevanion, S. 947
Triantaphyllou, S. 48
Triche, T.J. 947
Tridico, S. 321
Trifanova, S.V. 314
Triggs, C.M. 548
Trimarchi, M. 947
Tringe, S. 995
Trink, A. 756
Triplett, M.A. 813
Trippa, L. 757
Tromans, A.C. 947
Trombetta, J.J. 756
Tromp, G. 293
Trouplin, V. 790
Trowsdale, J. 80
Truberg, B. 528
Trudeau, M.M. 797
Truong, A. 727
Truong, T. 674
Trygg, J. 973, 975
Trynka, G. 835, 839
Trzaskowsk, M. 812
Tsai, P.-C. 943, 945
Tsakas, S. 496
Tsan, C. 943
Tsang, I. 757
Tsang, J.C.H. 756
Tsang, J.S. 731
Tsao, P.S. 943, 945
Tse, D.N. 757
Tserel, L. 732
Tsinoremas, N.F. 731
Tsonis, C. 730
Tsuda, B. 793
Tsui, K. 874
Tsui, L.-C. 626
Tsujikawa, T. 993
Tsuruta, S. 810
Tu, Z. 733
Tucci, S. 290, 293, 314, 323
Tuch, B.B. 758
Tuck, A.C. 756
Tucker, G. 111, 626
Tucker, M.A. 270
Tuffley, C. 217, 367, 394
Tukey, J.W. 676

- Tukiainen, T. 789, 975
 Tulinius, M. 791
 Tulloch, B. 797
 Tully, R.E. 793
 Tumian, A. 82
 Tung, J. 945
 Tung, P.-Y. 759
 Tupper, P. 216
 Turchin, M.C. 175
 Turdikulova, S. 318
 Turelli, M. 140, 144, 454
 Turnbaugh, P. 995
 Turnbull, C. 650, 797
 Turner, C. 1015, 1019
 Turner, D.J. 789
 Turner, E.H. 790
 Turner, K. 994
 Turner, P. 1015, 1019
 Turner, S.W. 944
 Turner, T.R. 293
 Turro, E. 791, 876
 Tusher, V. 871
 Tutton, R. 570
 Tybjørg-Hansen, A. 671
 Tye-Din, J. 833
 Tyler-Smith, C. 80, 272–274, 322
 Tyrer, J. 837, 840
 Tzeng, J.Y. 732
- U**
 Uauy, C. 530
 Ud-Dean, S.M.M. 757
 Udupa, N. 419
 Ueda, H. 80
 Ueshima, H. 972
 Uhler, C. 316
 Uitterlinden, A.G. 270, 732, 943
 Uktveryte, I. 318
 Ullrich, S.E. 529
 Umbach, D. 840
 Underhill, P.A. 317
 Underwood, J.G. 759
 Unterlaender, M. 314
 Unterländer, M. 48, 324
 Urbero, B. 795
 Urem-Kotsou, D.D. 48
 Ureyña, I. 316
 Uribe, K.B. 789
 Ursem, R. 975
 Ursin, G. 674
 Usher-Smith, J. 840
 Usoskin, D. 759
 Utevska, O. 318
 Utro, F. 78
 Utz, H.F. 528
 Uversky, V. 346
 Uzumcu, M. 725
- V**
 Vaccino, P. 522
 Vacquier, V.D. 393
 Vaglio, P. 974
 Valçarcel, B. 975
 Valdar, W. 526, 529
 Valdes, P. 273
 Valdiosera, C. 272, 315–316, 320–323
 Valè, G. 522
 Valen, E. 320
 Valencia, A. 342–365
 Valencia, V. 345
 Välimäki, N. 1017–1018
 Vallabh, S.M. 793
 Valle, D. 796
 Vallejos, C.A. 759
 Valles-Colomer, M. 995
 van Ballegooijen, W.M. 1020
 Van Berloo, R. 975
 van Bers, N. 525
 van Binsbergen, R. 813
 van Bon, B.W.M. 789, 791
 van de Leemput, J. 270
 Van Den Berg, D. 947
 van den Berg, J. 217
 Van den Berg, L.H. 732
 Van DenEngh, G. 84
 van den Hauwe, M. 791
 van den Oord, J. 753
 Vandepitte, D. 995
 van der Ark, K.C.H. 995
 Van der Auwera, G.A. 797
 Vander Jagt, C.J. 812–813
 van der Kouwe, A. 726
 Van der Lee, R. 345
 van der Leeden, R. 290
 van der Linden, M. 1015
 van der Maaten, L. 759
 van derMark, P. 216
 Vanderploeg, T. 76

- Van Der Vaart, A.W. 876
 VanderWeele, T. 677, 995
 van derWerf, J.H. 110
 Van Deutekom, J.C. 791
 van de Vijver, M.J. 731
 van de Vorst, M. 788
 Van De Wiel, M.A. 876
 van Dijk, D. 759
 Vandiver, A.R. 943
 vanDorp, L. 48
 van Dorp, L. 270, 274, 314, 317
 van Driem, G. 318, 321
 van Duijn, C. 625, 840
 van Eeuwijk, F.A. 523, 525, 527–530, 813,
 973
 Vang, S. 320
 van Geest, G. 523
 van Grootheest, G. 732
 van Heck, R.G.A. 995
 Van Heerwaarden, J. 526
 Van Hoek, M. 840
 Van Iterson, M. 975
 van Kippersluis, H. 677
 Vanlier, J.M. 270
 van Meurs, J.B. 732
 van Nas, A. 731
 Van Ness, B. 571
 van Nimwegen, E. 342
 Vannucci, M. 870, 872, 875, 974
 van Ommen, G. 975
 van Ooyen, A. 217
 van Oudenaarden, A. 753, 755
 van Overseem Hansen, T. 492, 593
 van Sluys, M-A. 524
 Vansteelandt, S. 672, 676–677, 995
 Van Tassell, C.P. 811
 Van Vliet-Ostaptchouk, J. 813
 Van Vooren, S. 790
 Van Wieringen, W.N. 876
 van Zwet, E.W. 947
 Vapnik, V. 416
 Varadhan, R. 49
 Varghese, N. 344, 1019
 Varilly, P. 83, 112, 419
 Varin, C. 49
 Varjabedian, A. 726
 Varmus, H. 789
 Varshney, R.K. 525
 Varul, L. 314
 Vasan, R. 834, 838
 Vaser, R. 797
 Vasilevsky, N.A. 792
 Vasquez, R. 318
 Vassura, M. 342
 Vatanen, T. 992
 Vattathil, S. 293
 Vaudin, M. 947
 Vaughan, T. 837
 Vavoulis, D.V. 876
 Vaz-Drago, R. 797
 Vears, D.F. 797
 Véber, A. 140–141, 171
 Vedantam, S. 813
 Veeramah, K.R. 48, 50, 314, 797
 Veerkamp, R. 810
 Veerkamp, R.F. 813
 Vehkala, M. 1017–1018
 Vekemans, X. 175, 495, 523
 Velazquez, A.M.V. 320, 322
 Veldink, J.H. 732
 Velle, B. 812
 Veltman, J.A. 788
 Vence, L.M. 993
 Vencovsky, R. 524
 Vendruscolo, M. 362
 Venkatesh, S.S. 49
 Venn, O. 76
 Vens, M. 629
 Venteicher, A.S. 758
 Venter, E. 729
 Venter, J.C. 730, 795
 Vento-Tormo, R. 758
 Vera, J.L. 273
 Verbyla, A. P. 525
 Verbyla, K.L. 813
 Verdin, F. 324
 Verduzco, D. 947
 Vergilino, R. 524
 Vergnaud, A.-C. 972
 Verhoef, S. 674
 Verlingue, C. 593
 Verma, S.S. 293
 Vermesch, J.R. 796
 Vermeire, S. 995
 Verna, C. 291, 316
 Vernot, B. 290–293, 323
 Verny, C. 594
 Vert, J.-P. 758

- Vervoort, R. 797
 Veyrieras, J.B. 730
 Viboud, C. 1017–1018
 Vicente, D. 993
 Vicente, M. 322
 Vicentini, R. 524
 Vicze, M. 314
 Vidal, M. 727–728, 974
 Vieira, M.L. 524
 Vieira-Silva, S. 995
 Vieth, B. 760
 Vigl, E.E. 317
 Vignal, A. 524
 Vignieri, S.N. 453
 Vihinen, M. 793
 Vila, A.V. 992
 Vilhelmsen, L. 216
 Vilhjálmsson, B. 836, 840
 Villjáalmsson, B.J. 111, 813
 Villalta, J.E. 753
 Villanueva, B. 498, 811
 Villar-Briones, A. 974
 Villasana, D. 947
 Villeda, H.S. 527
 Villem, R. 268, 271, 273, 318, 321
 Villena, M. 318
 Vimaleswaran, K.S. 650
 Vinaixa, M. 975
 Vincent, S. 974
 Vincentz, M.A.G. 524
 Vinckenbosch, N. 317, 420
 Vingron, M. 727
 Vink, J.M. 732
 Vinkhuyzen, A. 840
 Vinner, L. 49, 314, 320
 Vinuela, A. 727
 Viola, B. 292, 315, 320–323
 Viola, T.B. 290
 Visschedijk, M. 732
 Visscher, P.M. 144, 271, 273, 495, 595, 626,
 648, 650, 728, 732–733, 812–813, 874, 943,
 945
 Visser, R.G.F. 523
 Vissers, L.E.L.M. 788
 Vitalis, R. 499
 Viterbi, A.J. 49
 Vitezica, Z.G. 812
 Vladar, H.P. 140
 Vlassis, N. 995
 Voevoda, M.I. 273, 321
 Vogler, A.P. 493, 497
 Vogogias, A. 525
 Vohr, S.H. 292, 320
 Voight, B.F. 84, 419–420
 Voineagu, I. 731
 Vokonas, P.S. 943
 Volante, A. 522
 Volker, U. 732
 Volz, E.M. 1020
 Von Haeseler, A. 243
 von Heydebreck, A. 727
 von Mering, C. 994
 Voorrips, R.E. 523, 530
 Vosman, B. 530
 Vu, H. 733
 Vu, T.N. 759, 975
 Vukcevic, D. 77, 109, 674, 727
 Vy, H.M.T. 420
- W**
- Waber, P. 77
 Wacholder, S. 838, 840
 Waddell, P.J. 361
 Wade, M.J. 452
 Wadsworth, M.H. 755
 Wager, T.D. 943
 Wagner, A. 754
 Wagner, S. 324
 Wahl, J. 316, 318
 Wahl, L.M. 342
 Wain, L. 82, 112
 Wainwright, M.J. 1020
 Waits, L. 499
 Wakefield, J. 291, 293
 Wakeley, J. 77, 79, 84, 141, 175, 324
 Wako, H. 367
 Waldenberger, M. 943
 Waldmann, P. 493, 875, 1015
 Waldron, L. 947, 995
 Walhout, A.J. 727
 Walker, A.S. 1016
 Walker, A.W. 994
 Walker, J. 346
 Walker, K. 648
 Walker, L. 834
 Walker, N.M. 624, 649
 Walker, S. 876
 Walker, T.M. 1016

- Wall, J.D. 81–82, 85, 143, 173, 269, 273, 291–293, 320, 322, 324, 418–419
Wall, M. 947
Wallace, C. 674, 677, 727
Wallace, S.E. 568, 571
Walling, C.A. 455
Wallinga, J. 1017–1018, 1020
Walsh, B. 453, 455, 527, 550
Walsh, E. 419
Walsh, J.B. 144
Walsh, M.D. 753
Walsh, R. 797
Walter, F. 840
Walter, K. 110, 797
Walter, S. 677
Walters, G.B. 80
Walters, R. 837
Walters, R.K. 627
Wambebe, C. 269
Wandeler, P. 497
Wang, A. 729
Wang, B. 321, 595, 759, 994
Wang, C. 324, 623, 873, 876
Wang, D.G. 595
Wang, E.T. 872
Wang, G. 730
Wang, H. 367, 650, 755–756, 759, 876
Wang, J. 81, 321, 346, 495, 498–499, 726, 729, 755, 759, 876, 994–995
Wang, K. 269–270, 835
Wang, L. 757–759
Wang, M. 293, 729–730, 732
Wang, N. 85
Wang, O. 729
Wang, P. 273, 759, 1019
Wang, Q. 674, 732, 790–792, 794, 796, 870, 873, 947
Wang, R. 728
Wang, R.Y.-R. 344
Wang, S. 80, 346, 725, 727, 729, 731, 797
Wang, T. 729
Wang, W. 269, 732
Wang, X. 83, 629, 729, 731, 757–758, 871, 876, 943
Wang, Y. 112, 291–292, 321, 627, 630, 758
Wang, Z. 270, 346, 732, 759, 836, 876, 993
Wang de, Y. 732
Wangkumhang, P. 274
Want, E.J. 974
Waples, R.K. 499
Waples, R.S. 494, 498–500
Ward, K. 677
Ward, N.L. 84
Ward, R. 419
Ward, T. 731
Ward-Caviness, C.K. 943
Ware, D.H. 525
Ware, J.S. 797
Ware, L.B. 996
Wareham, N.J. 945
Wargo, J.A. 993
Waring, J.F. 732
Warinner, C. 293
Wark, A. 80
Warmuth, V. 273
Warmuth, V.M. 273, 321
Warnow, T. 244
Warren, J. 947
Warren, W.C. 293
Warry, G.L. 947
Washington, M.K. 996
Washington, N.L. 788
Wason, J.M. 650
Wasser, S.K. 85, 497, 500
Waterman, M.S. 85
Waters, M.R. 273, 321
Waterston, R.H. 754, 947
Watkins, J.C. 291
Watson, A. 530
Watson, A.S. 317
Watt, F. 944
Watterson, G.A. 144, 174, 245
Waugh, R. 527
Weadick, C.J. 394
Weale, M.E. 630
Weatheritt, R. 345
Weber, D. 995
Weber, J.L. 273
Weber, L.M. 946
Weber, T.M. 726, 729
Webster, D.R. 944
Webster, M. 944
Webster, T. 272, 292, 320
Weckwerth, W. 974–975
Weedon, M.N. 594, 649
Weeks, D.E. 596
Weersma, R.K. 732, 992
Wegmann, D. 47–50, 314, 318, 839

- Wehenkel, L. 417
Wei, D. 346
Wei, S.C. 993
Wei, X. 947
Wei, Y. 346
Weight, M. 342
Weigt, M. 342–346, 1016, 1020
Weihmann, A. 291, 316
Weinberg, C. 838, 840
Weinberg, W. 50
Weinberger, A. 993
Weinblatt, M.E. 272, 628
Weiner, A. 755
Weinert, L.A. 321
Weinreb, C. 346
Weinreich, D.M. 394
Weinstock, G. 293, 947
Weir, B.S. 85, 417, 500, 530, 548–550, 594, 624
Weir, W.H. 794
Weisenberger, D.J. 945, 947
Weiss, G.H. 142
Weiss, K. 269
Weiss, N. 47
Weissenbach, J. 994
Weissman, J.S. 753–754
Weissmann, C. 217
Weitz, D.A. 756
Welch, J.D. 759
Welch, R.P. 628
Weljie, A.M. 975
Wellman, J.A. 795
Wells, S. 788
Wen, B. 944
Wen, D. 245
Wen, G. 946
Wen, X. 13, 84, 269, 420, 871, 876
Wendel, J. 524
Wendel, L. 569
Wenger, A.M. 793
Wenzel, D. 792
Werge, T. 270
Wernisch, L. 756
Wesolowska, A. 321
West, A. 947
West, J.A.A. 757
West, M. 870–873, 876, 975
West, T. 944
Westaway, M. 293, 322
Westaway, M.C. 319
Westerhuis, J.A. 973
Westoby, J. 759
Westra, H.J. 732
Westreich, D. 673
Whalen, S. 674
Wheeler, B. 834
Wheeler, D.A. 946
Wheeler, H.E. 725–726, 732
Wheeler, T. 342, 759, 993
Wheeler, W. 837, 839
Whelan, S. 345, 363, 367, 395
Whiffin, N. 797
Whitaker, D. 530
White, D. 217
White, G.C. 500
White, R.A. 346, 1020
Whitehead, S.L. 947
Whiteley, M.N. 947
Whiteman, D. 837
Whitley, E.A. 569
Whitlock, M.C. 495, 499–500
Whitman, S.P. 947
Whitsel, E.A. 945
Whitsett, J.A. 755
Whittaker, A. 946
Whittaker, J.C. 974
Whittemore, A.S. 674
Whittle, J. 623
Whong-Barr, M.T. 569
Wichmann, H.E. 270
Wides, R. 729
Widschwendter, M. 948
Wieczorek, D. 795
Wieczorek, E. 293
Wiehe, T.H. 420
Wieland, B. 993
Wiencke, J.K. 270
Wierzbicki, A.S. 789
Wiggans, G.R. 811
Wigginton, J.E. 630
Wijmenga, C. 732, 992
Wikoff, W.R. 995
Wilberg, M.J. 500
Wilcox, S.K. 677
Wilde, S. 324
Wilde P. 522
Wilfert, A.B. 797
Wilke, C.O. 363, 366, 394
Wilkinson, D.G. 790

- Wilkinson, G.N. 524
Wilkinson, G.S. 454
Wilkinson, J.E. 947
Wilkinson-Herbots, H.M. 175
Willard, B. 869
Willard, H.F. 947
Willems, S.M. 676
Willemsen, G. 732
Willemsen, M.H. 789
Willer, C.J. 111, 594, 628, 630, 650
Willerslev, E. 49, 83, 273, 292, 314–316,
 319–323
Willey, D.L. 947
Willham, R.L. 455
Williams, A.L. 85, 113
Williams, B.A. 757, 873
Williams, B.L. 244
Williams, E.R. 526, 528, 530
Williams, G. 947
Williams, J.B. 568
Williams, K.A. 674
Williams, L. 947
Williams, M.S. 730
Williams, R.P. 812
Williams, R.W. 726
Williams, S. 269
Williams, S.M. 730
Williamson, A. 947
Williamson, E.G. 492, 500
Williamson, H. 947
Williamson, S. 418, 494
Williamson, S.H. 48, 269, 416, 420
Willis, L.E. 549
Willmitzer, L. 528
Wills, Q.F. 757, 759
Wills-Karp, M. 728
Willuweit, S. 548
Wilmes, P. 994
Wilming, L. 947
Wilson, A.C. 394
Wilson, A.J. 454–455
Wilson, D. 1016
Wilson, D.J. 1016
Wilson, G.A. 500
Wilson, I.D. 973–974
Wilson, J.F. 79, 110, 112, 269–270, 318
Wilson, J.G. 270
Wilson, R.J. 791
Wilson, R.K. 293, 947
Wilton, P.R. 77, 85, 144
Wiltshire, S. 625
Winberg, G. 757
Winchester, E. 595
Windle, J. 874
Windmeijer, F. 677
Windsor, S.M. 730
Wineberg, Y. 756
Wing, M.K. 624
Wingreen, N. 342
Winkelbach, L. 48
Winkler, C.A. 318
Winkler, T.W. 629
Winn, J. 731, 875
Winney, B. 81, 271
Winqvist, R. 674
Wirth, T. 1016
Wishart, D.S. 975
Withoff, S. 732
Witonsky, D. 416
Witte, J.S. 270
Witteman, J. 840
Witten, D.M. 792
Witteveen, A.T. 731
Wittke-Thompson, J.K. 630
Wiuf, C. 85, 144, 172, 175, 293
Wiviott, S.D. 795
Wjst, M. 623
Woerner, A. 324
Wohnoutka, P. 726
Wojcik, G. 837
Wojcik, G.L. 627
Wolak, M.E. 452
Wold, B. 757, 873
Wold, S. 975
Wolf, A. 754
Wolf, A.B. 293, 323
Wolf, D. E. 500
Wolf, G. 759
Wolf, P. 834
Wolf, S.M. 571
Wolfe, K. 730
Wolfe, K.H. 362
Wolpoff, M. 324
Won, H.H. 649
Won, S. 673
Wong, A.C. 80
Wong, E.H.M. 789
Wong, K.M. 217

- Wong, L.M. 726
 Wong, M. 757
 Wong, W.H. 873
 Wong, W.S.W. 395, 420
 Wong, Y.X. 548
 Wong W. 875
 Wood, A. 676
 Wood, A.M. 677
 Wood, A.R. 630, 732, 813
 Wood, D.E. 995
 Wood, J.L. 500
 Wood, J.L.N. 1017–1018
 Wood, S.N. 455
 Woodford, K.J. 729
 Woodhouse, K.A. 728
 Woodhouse, S. 756
 Woodman, S.E. 993
 Woodmansey, R.L. 947
 Woodruff, D.S. 494
 Wooldridge, J. 677
 Wooley, J. 343
 Woolhouse, M. 1017–1018
 Woolliams, J.A. 494, 498, 530, 811
 Wooster, R. 840
 Wootton, S. 549
 Worby, C.J. 1020
 Workman, R.E. 947
 Worku, M. 524
 Worl, R. 273
 Worland, A.J. 526
 Worley, K.C. 946
 Worth, C.L. 797
 Wortman, J.R. 729
 Wouters, J. 753
 Wray, N. 837, 840
 Wray, N.R. 626, 648, 733, 812–813, 874, 943,
 945
 Wray, P.W. 947
 Wrensch, M. 270
 Wricke, G. 530
 Wright, A.F. 112, 797
 Wright, C.F. 797
 Wright, D. 523
 Wright, F.A. 648, 728, 732
 Wright, J. 571
 Wright, P. 346
 Wright, S. 50, 144, 175, 324, 420, 455, 500,
 550, 630
 Write, M. 726
 Wu, A.R. 759
 Wu, C. 793, 876
 Wu, D. 756, 875
 Wu, G. 992, 996
 Wu, H. 876, 944, 946, 996
 Wu, J. 367, 792, 797, 840
 Wu, K. 85
 Wu, L.F. 731
 Wu, M.C. 626, 630
 Wu, M.M. 730
 Wu, M.S. 732
 Wu, S. 343
 Wu, W. 877
 Wu, X. 270, 944
 Wu, Y. 83, 245, 813
 Wu, Z. 759, 876, 944
 Wulff, B.B.H. 530
 Wulfridge, P. 947
 Wurfel, M.M. 290
 Würtz, P. 975
 Wuster, A. 795
 Wyatt, P.W. 760
 Wyckoff, G.J. 391
 Wyder, S. 796
 Wylie, K.M. 992
 Wymant, C. 1020
- X**
- Xavier, R.J. 992, 994–995
 Xavier Oms, F. 320
 Xhu, X. 273
 Xia, A.C. 729
 Xia, F. 996
 Xia, J. 975
 Xia, K. 732
 Xiang, F. 1015
 Xiang, J. 346
 Xiao, C. 729
 Xiao, F. 321
 Xiao, H.-S. 756
 Xiao, J. 530, 726
 Xiao, Y. 346
 Xie, C. 529
 Xie, D. 1016
 Xie, L. 726
 Xie, P. 798
 Xie, R. 649
 Xie, T. 732
 Xie, W. 217, 530

- Xie, X. 83, 794
Xie, Y. 994
Xifara, D.-K. 85
Xifra, G. 996
Xing, Y. 530
Xu, C. 876
Xu, E.Y. 733
Xu, J. 346, 994
Xu, J.L. 758
Xu, K.M. 947
Xu, L.Z. 650
Xu, N. 758
Xu, P. 216
Xu, Q. 733
Xu, S. 291, 293, 530, 872, 876
Xu, X.Q. 812
Xu, Y. 530, 755
Xu, Z.Z. 993
Xuan, J. 871
Xue, A. 813
Xue, L. 945
Xue, Y. 273
- y**
Yadav, R. 273, 321
Yaghoontkar, H. 732
Yajima, M. 755
Yaka, R. 317
Yakhini, Z. 947
Yamada, T. 994
Yamaguchi, Y. 395
Yamamoto, K. 84
Yamamura, Y. 270
Yamato, J. 80
Yan, P. 943–944, 947
Yan, W. 530
Yan, X. 729
Yan, Y. 1015
Yandell, B.S. 869, 876
Yandell, M. 729
Yanek, L.R. 50
Yanes, O. 975
Yang, C. 648, 834, 837
Yang, F. 1015
Yang, H. 944, 994
Yang, H.P. 525
Yang, I.V. 946
Yang, J. 596, 630, 650, 672, 728, 733, 812–813,
 837, 840
Yang, L. 731
Yang, M. 244–245
Yang, Q. 630
Yang, R. 876
Yang, S. 732
Yang, W.P. 727
Yang, X. 291, 727, 731, 947, 1015
Yang, X.W. 731
Yang, Y. 80, 291
Yang, Z. 213–216, 244–245, 363, 365–367,
 391–395, 418, 420, 498, 500, 1014
Yano, T. 214, 392
Yao, A. 729
Yao, C. 728
Yao, L. 948
Yao, W. 530
Yao, Y.G. 293
Yap, C.X. 813
Yasunaga, T. 393
Yates, F. 530
Yates, M.C. 500
Yates, P.D. 975
Yatsunenko, T. 995
Yau, C. 757, 875
Ye, C. 759
Ye, J. 729
Yeager, M. 839
Yearsley, J. 493
Yearsley, J.M. 494
Yelensky, R. 628, 794
Yell, M. 790–791
Yen, J. 947
Yengo, L. 172, 270, 318, 416, 813
Yeo, G.S. 649
Yeo, G.W. 758
Yet, I. 945
Yeung, K.Y. 871
Yi, N. 869, 876
Yi, Q. 77, 623
Yi, S.V. 944
Yi, X. 317, 420
Yilmaz, P. 994
Yin, Z. 732
Ying, L. 731
Ying, S.X. 728
Yip, S.H. 759
Yip, W.-K. 83
Yiu, A. 756
Yiu, S.-M. 81

- Yizhak, K. 758
 Yoder, A.D. 217
 Yofe, I. 755
 Yokota, M. 84, 500
 Yonezawa, T. 367
 Yoo, W. 674
 Yooséph, S. 729, 993
 York, J. 729
 York, T.L. 314, 415, 493
 Yosef, N. 754
 Young, A. 974
 Young, J.H. 972
 Young, L.J. 757
 Young, P. 595
 Young, R. 676
 Ypma, R.J.F. 1020
 Yu, C. 81, 994
 Yu, D. 759
 Yu, F. 112
 Yu, G. 627
 Yu, H. 728
 Yu, J. 112, 367, 530
 Yu, L. 244
 Yu, Y. 245
 Yuan, G.-C. 755, 757
 Yuan, J. 757
 Yuan, K. 291
 Yuan, M. 728, 872
 Yudkoff, M. 994
 Yudkovsky, G. 268
 Yule, G.U. 455
 Yun, Y. 245
 Yüncü, E. 317
 Yunusbayev, B. 268, 271
 Yvert, G. 725
- Z**
- Zackular, J.P. 996
 Zagury, J.-F. 77, 82, 110, 112, 624
 Zahler, A.M. 625
 Zaitlen, N.A. 418, 594, 625, 628
 Zajac, P. 755
 Zakharia, F. 268
 Zalcman, G. 994
 Zalloua, P.A. 269
 Zamparo, M. 342–343
 Zanetti, K.A. 270
 Zanke, B.W. 944
 Zaraneck, A.W. 82
 Zaveri, J. 730
 Zawati, M.H. 569
 Zaykin, D.V. 498
 Zecchina, R. 342, 344
 Zeevi, D. 993
 Zeggeri, E. 113, 626–627
 Zeigle, J. 494
 Zeisel, A. 755
 Zeiss, C. 626
 Zekavat, S.M. 790
 Zeller, G. 994–995
 Zellner, A. 877
 Zemunik, T. 76, 318
 Zeng, B. 728
 Zeng, C. 346
 Zeng, D. 648
 Zeng, J. 813
 Zeng, L. 293
 Zeng, P. 841
 Zeng, Z.B. 728, 732
 Zeven, A.C. 144
 Zhai, C. 84
 Zhai, W. 291, 316, 321
 Zhan, X. 624, 630
 Zhan, Y. 272, 292, 320, 497
 Zhang, A. 675, 995
 Zhang, B. 726–727, 729–730, 732–733
 Zhang, C. 291, 530, 731–732
 Zhang, F. 79, 733, 798, 946
 Zhang, G. 77, 497
 Zhang, H. 946
 Zhang, J. 395, 420, 630, 947, 993
 Zhang, J.H. 342
 Zhang, J.M. 757
 Zhang, K. 85, 291
 Zhang, L. 759, 943
 Zhang, L.V. 727
 Zhang, N.R. 755, 759
 Zhang, P. 525
 Zhang, Q. 530, 729
 Zhang, R. 346
 Zhang, S. 82, 759
 Zhang, W. 47, 76, 415, 493, 676, 727, 789, 877, 946, 1014
 Zhang, X. 530, 728, 754, 756, 792, 943, 994
 Zhang, Y. 346, 627, 630, 993, 996, 1015
 Zhang, Y.-M. 876
 Zhang, Y.P. 293
 Zhang, Z. 342, 630

- Zhao, C. 273
Zhao, H. 648, 674, 834, 837, 870
Zhao, J.-R. 756
Zhao, J.H. 650
Zhao, L. 877, 993
Zhao, Q. 729
Zhao, S. 674, 944
Zhao, W. 834
Zhao, Y. 346, 526, 730
Zhao, W. 834
Zheng, C. 85, 593, 596
Zheng, G.X.Y. 759
Zheng, H. 630, 994
Zheng, J. 725
Zheng, L. 729
Zheng, S.C. 947
Zheng, W. 270, 674, 870
Zheng, X.H. 729
Zhenli Liu, J. 727
Zhernakova, A. 732, 992
Zhernakova, D.V. 732, 992
Zhi, D. 82
Zhitenev, V. 314
Zhivotovsky, L.A. 273
Zhong, B. 367
Zhong, F. 729
Zhong, H. 650
Zhong, W. 729
Zhong, Y.-Q. 756
Zhou, A. 346
Zhou, B.-B. 756
Zhou, G. 530
Zhou, H. 111
Zhou, J. 530, 947
Zhou, X. 630, 813, 837, 841, 877, 945–946,
 1020
Zhou, Y. 109, 290–291, 314, 366, 994
Zhu, H. 729, 731, 994
Zhu, J. 245, 725–733, 759, 877, 973
Zhu, L. 420
Zhu, R. 758
Zhu, S. 245
Zhu, T. 245
Zhu, X. 270, 798, 1019
Zhu, Y. 758
Zhu, Z. 733, 812–813
Zhuang, J. 948
Ziegenhain, C. 760
Ziegler, A. 629
Ziegler, R.G. 270
Zielke, H.R. 726
Zilberman-Schapira, G. 992
Zilh ao, J. 320
Ziller, M.J. 945, 948
Ziman, M. 731
Zimmerman, E.S. 677
Zimmermann, T. 244
Zimmern, R. 838
Zink, A. 317
Zink, F. 79, 629
Ziota, C. 48
Zirah, S. 973
Ziraldo, S.B. 759
Zitvogel, L. 993–994
Ziv, E. 628, 630
Zobel, J. 833
Zobel, R.W. 524
Zoghbi, H. 947
Zolezzi, F. 732
Zollei, L. 726
Zoller, M.W. 944
Zoller, S. 367
Zollikofer, C.P.E. 273, 321
Z llner, S. 797
Zondervan, K.T. 623–624, 630
Zorilla, S. 947
Zou, C. 757
Zou, D. 944
Zou, F. 648, 732, 871
Zouali, H. 625
Zuber, V. 673
Zucker, D.M. 676
Zuckerkandl, E. 218, 394
Zuk, O. 798, 841
Zusmanovich, P. 80, 111
Zuzarte, P.C. 947
Zwickl, D.J. 218
Zwyns, N. 290

Subject Index

a

- ABBA-BABA method 241. *See also D-statistic*
ABC. *See approximate Bayesian computation (ABC)*
absolute risk model 818–821
 software for 820–821
absolute value penalty 961
acceptance rate 371
activators 700
adaptive molecular evolution 369–371
 along lineages
 likelihood calculation under models of
 variable ω ratios 380–381
 in primate lysozyme 381–382
amino acid sites under positive selection
 likelihood ratio test under models of
 variable ω ratios 384–386
 methods that test one site at time 386
 positive selection in HIV-1 *vif* genes 386–388
computer software 391
likelihood calculation on phylogeny 379–380
limitations of current methods 390–391
Markov model of codon substitution 371–372
non-synonymous and synonymous rates 370
non-synonymous/synonymous rate ratio (ω ratio) 370
reconstructed ancestral sequences based methods, comparison with 382–384
synonymous and non-synonymous substitution rates estimation
 Bayesian estimation 377
heuristic estimation methods 372–373
maximum likelihood estimation 374–377
numerical example 377–379
testing positive selection
 branch-site test 388–389
 clade models and other variants 389–390
adaptive peaks 133
additive genetic covariances 423
additive genetic value 422
additive genetic variance 423, 426
ADMIXTURE 254, 998
admixture 247, 275
admixture events, identifying/dating 262–263
DNA segments inherited from different sources 263–265
measuring decay of linkage disequilibrium 265–267
admixture model 480
African Genome Variation Project 776
age-related macular degeneration (AMD) 598
Akaike information criterion (AIC) 37, 185, 710
ALDER 265–267
alignment benchmark method 333
allele frequencies 2–3, 8, 12–14, 20–21, 32, 34–35, 400, 426
 changes in 397, 426–427
 genetic models for 535–539
 temporal changes in 466–470
allele frequency spectrum (AFS) 260–261
allele specific expression 846
allelic dropouts 471
allelic heterogeneity 576
Allen Brain Atlas 715

- All of Us Research Program 816
 allogamous species 504
 allopolyploids 506
 allosteric effects 330
 alpha designs 519
 ALPHAPHASE software 94
 altruistic participation 557
 American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) 766
 amino acid replacements 350–351
 empirically derived models of 348–351
 AMMI (additive main effects, multiplicative interaction) 513
 amplicon sequence variant (ASV) 981
 analogous folds 334
 ancestral alleles 535
 ancestral lineage 161
 ancestral recombination graph (ARG) 160–163, 412, 581
 ancestral selection graph 164
 ancestral sequence reconstruction 380
 ancestry tract inference methods 263–265
 ancient DNA (aDNA) 295–296
 challenges in working with, and processing aDNA data 296
 contamination 297–299
 genetic studies and understanding of human past 310
 archaic genomes and admixture with modern humans 310–311
 demographic changes during late Neolithic and Bronze Age 312–313
 Neolithic transition 311–312
 handling sequence data from 300
 additional filtering 300–301
 downstream analysis, effects of limited data on 301–302
 mapping and identification of endogenous DNA 300
 preprocessing of NGS data 300
 opportunities of 302–303
 allele frequency trajectories 308–310
 continuity 306–307
 demographic inference based on ancient genomes 308
 migration and admixture over time 307–308
 population differentiation in time and space 303–305
 sequence degradation 296–297
 animal breeding 502
 animal model 430–432
 annotator software 682
 anomalous gene trees 225
 anomaly zone 225
 ANOVA model 856
 approximate Bayesian computation (ABC) 31–37, 413, 470, 1007
 ABC-GLM 35–36
 ABC-MCMC 33–34
 ABC-REG 35
 ABC-SMC 34
 improved ABC sampling techniques 33–35
 insufficient summary statistics 37
 post-sampling adjustments 35–36
 approximate techniques
 composite likelihood 18–19
 Monte Carlo sampling 16–18
 using summary statistics 16
 archaic hominin admixture 275–276
 methods for testing 277
 D-statistic 279–282
 F_4 -statistics 282–283
 genetic drift and allele frequency divergence 277
 three-population test 277–279
 area under the precision–recall curve (AUC) 919
 ARGweaver method 415
 asexually propagated species 504
 asexual reproduction 504
 association studies 597–598
 ASTRAL (Accurate Species TRee ALgorithm) 232, 233
 asymptotic properties, ML estimator 6–7
 ATAC-seq 702
 autogamous species 504, 505
 autoployploids 506
 average effects of alleles 422
 average genome size (AGS) 982
- b**
- background selection 136, 168
 backSPIN 749
 bacterial population genomics 997–998
 gene content analysis 1013–1014

- genetic population structure
 background 998
 distance-based methods 1000–1001
 linkage disequilibrium 1000
 model-based clustering 998–1000
 genome-wide association studies
 1008–1012
 genome-wide epistasis analysis 1012–1013
 phylogenetics and dating analysis
 1001–1004
 transmission modeling 1004
 challenges 1004–1008
 balanced batches 743. *See also* single-cell RNA-sequencing (scRNA-seq)
 balancing selection 165–166, 399, 403
 Balding–Nichols formulation 535
 Balding’s sampling formula 536, 538, 545
 BAPS 998, 999
 barcoding 489
 BATMAN 952
 Baum–Welch algorithm 44–46
 BayesAss 484
 Bayes empirical Bayes (BEB) 385
 Bayes factor (BF) 185
 Bayesian algorithm 699
 Bayesian full-data methods, for species tree inference 234–235
 Bayesian genomic selection models 808–809
 Bayesian GPLVM (BGPLVM) 891
 Bayesian hierarchical modelling 844
 Bayesian inference 4, 20–37, 180–184
 approximate Bayesian computation 31–37
 Bayesian point estimates and confidence intervals 22–23
 empirical Bayes for latent variable problems 30–31
 Markov chain Monte Carlo 23–29, 195–198
 mechanics of 192–198
 prior distributions, choice of 21–22
 Bayesian Inference for Single-Cell clUstering and ImpuTing (BISCUIT) 745
 Bayesian information criterion (BIC) 37, 710
 Bayesian lasso 802, 862
 Bayesian linear regression (DBN) 902–905
 Bayesian logistic regression 860
 Bayesian methods, for gene expression analysis 843–844
 differential expression analysis
 microarray data 851–856
 RNA-sequencing data 856–857
 modelling microarray data
 gene variability 845–846
 modelling intensities 845
 normalization 846
 modelling RNA-sequencing reads
 gene-level read counts 850–851
 read-level data 847–848
 RNA-sequencing data 846–847
 transcript-level read counts 848–850
 multivariate gene selection models
 Bayesian shrinkage 861–863
 variable selection approach 857–861
 quantitative trait loci
 multiple-response models 864–868
 single-response models 863–864
 Bayesian model choice
 Bayes factor 39
 model posterior probabilities 38
 Bayesian network models 703
 Bayesian network reconstruction
 algorithm 721
 process
 integrating genetic data 720–721
 network priors in 721–722
 Bayesian piecewise linear regression
 (NH-DBN) 905–908
 Bayesian piecewise linear regression with coupled regression coefficients (coupled NH-DBN) 908–914
 Bayesian point estimates 22–23
 Bayesian Regression for Isoform Estimation (BRIE) 746
 Bayesian shrinkage 861–863
 Bayesian statistics 1. *See also* statistical inference, model-based
 Bayesian variable selection (BVS) 858
 Bayes’ theorem 15, 180–181
 BayesTraits 1010
 BCRAT1 830
 BCurve 938
 Beagle 91, 101–102
 BEAST 1003
 Bell number 474
 benefit maximisation models 554
 Benjamini–Hochberg method 706
 Berk–Jones test 690

- Bernoulli distribution 866
 best linear unbiased estimator (BLUE) 503
 best linear unbiased prediction (BLUP) 504,
 802
 single step 804–805
 beta distribution 20, 182, 536
 beta function 21
 beta prior probability distribution 182–183
 big data approaches 551
 BIMBAM 611
 binary disease phenotypes 599
 binary model 543–545
 binomial probability distribution 179
 Bioconductor R package 752
 bioethics 552
 BioGRID 781
 biological processes, modeling approaches for
 702–704
 biosynthetic gene clusters (BGC) 990
 birth–death models 1003
 birth–death process of cladogenesis 195
 bisulfite-sequencing (BS-seq) data 936–939.
 See also DNA methylation
 beta-binomial-based methods 936–938
 direct detection 938–939
 bivariate trait vector 707
 BLASTClust 333
 BLIMP (Best Linear Imputation) 104
 blocking, defined 913
 blocks of genome 138–140
 BLSMM 802
 BLUEPRINT project 681
 Bohring–Opitz syndrome 770
 BOLT-LMM 616
 Bonferroni correction 601, 956, 981, 1012
 Boolean network modeling 703
 BoolTraineR (BTR) 751
 bootstrap method 192
 BOXSET method 334
 branch-site test, of positive selection 388–389
 BratNextGen 1000
 Bravo TOPMed 771
 Bray–Curtis distance 986
 Breast and Ovarian Analysis of Disease
 Incidence and Carrier Estimation
 Algorithm (BOADICEA) model 831
 breeder’s equation 425, 426
 multivariate 425, 444
 in plant breeding 502–503
 breeding success 433
 breeding systems, of plants 504–505, 506
 breeding values 422
 Brownian motion 120, 121, 208, 1009
 Broyden–Fletcher–Goldfarb–Shanno (BFGS)
 algorithm 10
 bulk RNA-seq methods 747
 Bulmer’s iteration 443
 burden tests 619
 burn-in 27
- C**
- CADD 682
 calibration node 206
 Cannings model 118–119
 canonical correlation analysis (CCA) 753
 capillary electrophoresis (CE) 543, 950
 capture–mark–recapture (CMR) approach
 464, 470
 genetic CMR 471
 capture–sequence data 939–940. *See also*
 DNA methylation
 CardioClassifier 776
 cardiovascular disease (CVD) 817
 case–control studies 558
 IBD-based 589–590
 casual inference test (CIT) 710
 Catalog of Somatic Mutations in Cancer
 (COSMIC) 770
 CAT approach 354
 categorical phenotypes 599
 CCMpred method 327
 CD-hit 333
 CellRanger 741
 central limit theorem 426
 CFTR2 781
 Challis–Schmidler OU model of structural
 evolution 360
 Chapman–Kolmogorov equations 23
 Chapman–Kolmogorov theorem 375, 376
 chemical master equation (CME) 889
 China Kadoorie Biobank 268
 chip-based microarray technology 600
 ChIP-seq 702
 ChromHMM 684, 685
 CHROMOPAINTER 73, 256, 266
 chromosome painting 41–42, 43, 46
 methods 73
 chromosomes 574, 998

- clade models 389–390
 ClinGen 779, 784
 clinical trait QTL (cQTL) 708
 ClinVar 781, 782
 ClonalFrame 1005
 CLUMPAK 999
 CLUSTAL-Omega program 333
 ClusterFinder 990
 clustering algorithms 251–252
 admixture LD model 254–255
 admixture model 253–254
 on allele frequency patterns 252–253
 fineSTRUCTURE model 255–258
 interpreting genetic clusters 258–259
 STRUCTURE model 252–253
 Clustering through Imputation and Dimensionality Reduction (CIDR) 746
 cluster of orthologous gene sequences (COG) 1010
 cluster score 602
 coalescence trees 150, 156
 coalescent 145–146, 171
 approximation 148–151
 topology and branch lengths 148–149
 and ‘classical’ population genetics 169
 data analysis 145–146
 diploidy and segregation 157–159
 hermaphrodites 157–159
 males and females 159
 fundamental insights 146–148
 generalizing 151–155
 population structure on different time-scales 153–155
 robustness and scaling 151–152
 variable population size 152–153
 geographic structure and 155–157
 strong-migration limit 156–157
 structured coalescent 155–156
 Kingman 117–118
 multispecies (*see* multispecies coalescent)
 neutral mutations 168–169
 and phylogenetics 169–171
 recombination 159–164
 ancestral recombination graph 160–163
 properties and effects of 163–164
 selection 164–168
 background selection 168
 balancing selection 165–166
 selective sweeps 166–167
 sequentially Markov 139
 structured 137
 coalescent-based likelihood methods 469
 coalescent history 221–224
 coalescent independent sites (CIS) 236
 coalescent modelling 62–67
 genealogical history in 63
 LD in populations with geographical subdivision/admixture 67
 LD in recombining regions 65–66
 LD patterns in absence of recombination 63–65
 sequence of correlated trees in 63
 coalescent theory 17, 1003
 coalescent unit 220, 239
 Cochran–Armitage trend test 607
 codon-based models 355–356
 codon model 203–204
 coefficient of coancestry 423
 coefficient of gene differentiation 476
 coherence 718
 collider bias 670
 COLOC 713
 combined probability of exclusion (CPE) 542
 combined probability of inclusion (CPI) 542
 common disease common variant (CDCV)
 hypothesis 598
 CommonMind 680
 community consultation 567
 complete-data likelihood 11
 complete-data log-likelihood 12, 15, 88
 complex diseases 598
 composite likelihood methods 18–19, 70–71, 412–413
 computational analysis 739
 computational annotations 685
 conditional probability 179
 conditional random fields (CRFs) 287–289
 conditional structured coalescent 164
 confidence intervals, construction of 192
 confidentiality, in research 564–565
 confounders 404
 confounding factors 606–607
 conjugate priors 21, 183
 consent, in research 559–563
 conservation genetics 457, 491–492
 aim of 457
 census size, estimation of 470–472

- conservation genetics (*Continued*)
- capture–mark–recapture (CMR) approach 470–471
 - methods based on pedigree reconstruction 474–475
 - multilocus genotype mismatch method 472
 - pairwise relatedness analysis 472–473
 - pairwise relationships approach 473–474
 - deintroduction strategies 487–489
 - effective population size, estimation of 458–470
 - from heterozygosity excess approach 459–460
 - from linkage disequilibrium-based methods 460–463
 - present-time 460–462
 - in recent past 462–463
 - from methods based on relatedness 463–466
 - temporal approach 466–470
- genetic species delimitation 489–491
- genetic structure, inferring 475
- genetic differentiation, measurement of 475–477
 - inferring levels of recent gene flow 483–486
 - landscape genetics 486–487
 - population assignment 477–479
 - population clustering and ancestry proportions 479–483
 - overview 457–458
- conservative migration 157
- ConTest 333
- contingency-table test 57
- continuous model 546
- continuous-time Markov models 199. *See also* substitution models
- convergent evolution 1001
- CONVERGE study 95
- convolutional neural networks 337–338
- Coronary Artery Disease Genomewide Replication and Meta-analysis plus The Coronary Artery Disease (CARDIoGRAMplusC4D) 655
- correlated-drift approaches 262
- cosexuality 504
- counts per million (CPM) 742
- covariance selection differential 442
- covariation model 205–206
- Cox–Ingersoll–Ross process 208
- Cox proportional hazards model 820
- CpG dinucleotide 933
- CpG islands (CGI) 933
- Cramér–von Mises test 751
- credible intervals 1003
- CRISPR 777–778
- CRISPR-Cas9 551
- Critical Assessment of techniques for protein Structure Prediction (CASP) 330–332
- crossovers 575
- cross-sectional study 433
- curse of dimensionality 23
- d**
- Darwin's theory of natural selection 115
- data compression 74–75
- data mining approaches 618
- Dayhoff and Eck model 348–349
- Decipher 781
- Deciphering Developmental Disorders Study 776
- deep coalescence 170
- deep learning method 336–338
- convolutional neural networks 337–338
- DeepSea 684
- deintroduction strategies 487–489
- delimitation 489
- DeltaSVM 684
- demographic histories 259, 262
- demographic inference 17–18, 19
- de novo* mutations 766
- DESeq 747, 751
- detailed balance equations 25, 372
- differential expression (DE) 742
- differentially methylated cytosines (DMC) 936
- differentially methylated regions (DMR) 936
- direct detection of 938–939
- diffusion approximation 61–62
- diffusion coefficient 120
- diffusion process 120
- direct coupling analysis (DCA) 326–327, 1012
- directed acyclic graph (DAG) 3
- directional selection 398–399, 441, 442
- directional negative selection 398–399
 - directional positive selection 398

- directional selection differential 424–425, 434, 435, 439
 directional selection gradient 439
 Dirichlet distribution 39
 Dirichlet–Laplace prior 862
 Dirichlet process mixtures (DPM) 853
 DiscovEHR 771
 discrete-time Markov chains (DTMCs) 23
 discriminant analysis of principal components (DAPC) 1001
 disease risk models
 absolute risk model 818–821
 background 815–818
 breast cancer 829–832
 challenges 832–833
 clinical utility 828–829
 future directions 832
 model validation 827–828
 polygenic risk score 821–826
 PRSS and epidemiologic factors 826–827
 disomic inheritance 506
 dispersion tests 619
 disruptive selection 441, 442
 distance-based method 178, 479
 DIVAN 685
 divergence time estimation 206–211
 DNA 573–574
 ancient DNA 275 (*see also* ancient DNA (aDNA))
 descent of 575–576 (*see also* genetic mapping)
 inheritance of 575
 library 296
 markers 531–532
 polymorphisms 801
 sequencing 547
 DNA conservation metrics 681
 DNA methylation 701, 933–934
 differential
 bisulfite-sequencing data 936–939
 capture-sequence data 939–940
 HumanMethylation array data 940–941
 measuring 934–936
 double reduction 506–507
 doublet model 202–203
 downstream analysis 301–302
 drift coefficient 120
 drift load 126
 dropClust 749
 dropEst 741
 Drop-seq 737
Drosophila 433, 617
 D-statistic 279–282, 284
 duoHMM approach 93
 dynamical Bayesian networks (DBN) 900
 dysbiosis 978
- e**
- EAGLE2 method 96
 EAGLE v1 method 94
 East London Genes and Health 776
 eCaviar 713
 ecovalence 512
 edgeR 747
 EEMS software 487
 effective migration rates 487
 effective population size 152, 458
 defined 458
 estimation of 458–470
 efficiency, defined 718
 EIGEN 684, 685
 EIGENSTRAT (PCA software) 250
 EigenTHREADER 334
 elastic net penalty 961
 EM algorithm. *See* expectation-maximization (EM) algorithm
 emission probabilities 90
 EMMAX 616
 empirical Bayes analysis 194
 for latent variable problems 30–31
 empirical research 552
 Encyclopedia of DNA Elements (ENCODE) 680
 entropy measures, for Y-STR markers 540
 epigenetic annotation data 681. *See also* genome-wide association studies (GWAS)
 epigenetics 933
 epigenome-wide association studies (EWAS) 941
 epistasis 617
 equilibrium additive genetic variance 429
An Essay toward Solving a Problem in the Doctrine of Chances (Thomas Bayes) 20
 ethics 552
 bioethics 552
 defined 552

- ethics (*Continued*)
- and governance, case studies on 554–555
 - confidentiality and security 564–565
 - consent 559–563
 - 100,000 Genomes Project 556
 - incentives to participate 562–563
 - leaving the study 562
 - recruitment of participants 558–559
 - returning individual genetic research results 563–564
 - scientific and clinical value of research 556–558
 - UK Biobank 555–556
 - voluntariness 560–561
 - of population genetic research 553
 - benefit maximisation models 554
 - risk control models 553–554
 - stewardship and social issues 565–566
 - benefit sharing 566–567
 - community involvement and public engagement 567
 - race, ethnicity and genetics 567–568
- EVcouplings 326
- EvolBoosting 414
- evolutionary quantitative genetics 421
- additive genetic variances and covariances 423
 - Fisher's genetic decomposition 422
 - fitness 432
 - episodes of selection 433–434
 - individual 432–433
 - fitness functions and characterization of selection 441
 - gradients and local geometry of fitness surfaces 442–443
 - individual and mean fitness functions 441–442
 - inference of parameters 430–432
 - inference of selection gradients 447
 - flexible inference of fitness functions 449–450
 - normality and selection gradients 450–451
 - ordinary least squares analysis 448–449
 - infinitesimal model 426
 - allele frequency changes under 426–427
 - changes in variances 427–429
 - equilibrium additive genetic variance 429
- linearity of parent–offspring regressions under 426
- multivariate selection 444
- effects of genetic correlations 444–446
 - multivariate breeder's equation 444
 - selection gradients and traits affecting fitness 446–447
- opportunity for selection 437–438
- parent–offspring regressions and response to selection 423–424
- genetic and phenotypic covariance matrices 425
 - multiple-trait parent–offspring regressions 425
 - multivariate breeder's equation 425
 - selection differentials and breeder's equation 424–425
 - single-trait parent–offspring regressions 424
- Robertson–Price identity 435
- selection coefficients 438–439
- measures of selection on mean 439
 - measures of selection on variance 439–441
- theorems of selection 434
- description of 435–436
 - empirical operationalization 436–437
- Ewens sampling formula 64–65
- Ewens–Watterson homozygosity test 65
- ExAC 771
- Exomiser 781
- expectation–maximization (EM) algorithm 11–12, 262
 - application of 12–14
 - with numerical optimization 14–16
- expected Fisher information 21
- expected lod score 587
- explicit assumptions, on probability distributions 3–4
- explicit incorporation of prior information, in Bayesian analysis 181
- expression matrix, analysis of 747–753
- clustering 748–749
 - combining data sets 752–753
 - differential expression and marker genes 750–751
- dimensionality reduction and visualization 747

- feature selection 747–748
 network inference 751–752
 pseudotime 749–750
 expression quantitative trait loci (eQTL) 681, 698, 844
 for gene expression traits 701
 modeling for 705
 clinical trait linkage mapping 708–714
 heritability of expression traits 705–706
 joint 706–708
 single-trait 706
 extended haplotype homozygosity (EHH) 73, 410
 extended selection gradient 447
 External RNA Control Consortium 739
- f**
- factor loadings 962
 false alleles 471
 false discovery rate (FDR) 601, 706, 773, 855, 956
 family-based association studies 611–612
 fastPHASE 91
 fastsimcoal2 17, 19
 FDR. *See* false discovery rate (FDR)
 Felsenstein equation 227
 Felsenstein's pruning algorithm 178, 348, 352–353, 357
 Felsenstein's substitution model 15
 fineSTRUCTURE 73, 257, 998, 1000
 Firth's correction 1011
 Fisher information 7
 matrix 8–9
 Fisher's genetic decomposition 422
 fitness
 components 433
 frequency-dependent 441
 function 441
 individual 432–433
 fitness landscapes 441
 fixed-effects likelihood (FEL) model 386
 fixed-sites models 384
 FluidigmC1 chip 736
 fluorescent activated cell sorting (FACS) 737
 fMRI data 715
 Fokker–Planck equation 123
 forensic genetics 531–532
 behavior of likelihood ratio 546–547
 mixtures 542
 combined probabilities of inclusion and exclusion 542
 likelihood ratios 542–546
 principles of interpretation 532–534
 profile probabilities 534–535
 genetic models for allele frequencies 535–539
 multi-locus dependencies 538–539
 population structure 535–537
 relatedness 537–538
 Y-STR profiles 539–542
 single nucleotide polymorphism, sequence and omic data 547–548
 forward-backward algorithm 42–43
 fossilized birth–death process 211
 four-gamete test 164
 four-population test. *See D*-statistic
 four-state general time-reversible (GTR) models 349
 fragmentation of DNA strands 296–297
 Framingham Risk Score 817
 FRAPPE 254
 free induction decay (FID) 951
 frequentist statistics 1. *See also* statistical inference, model-based
 F -statistics 856, 998
 F_4 -statistics 282–283
 functional gene 574
 fundamental theorem of selection 435
 FUN-LDA 685
 funnel plots 644
- g**
- Gail model 817
 gamma rate variation model 205
 Gaussian graphical model (GGM) 967
 Gaussian kernel 745
 Gaussian latent variable model 744
 Gaussian process (GP) 879
 bulk time series expression data 883–884
 differential equation models 888–889
 identifying differential expression 884–885
 replicates and clusters 887–888
 two time course experiments 885–887
 covariance function 880–882
 inference 882–883

- Gaussian process (GP) (*Continued*)
modelling single-cell data
 dimensionality reduction and pseudotime
 inference 891–892
 modelling branching dynamics
 892–893
 modelling single-cell trajectory data
 889–891
- Gaussian process latent variable model (GPLVM) 891
- Gaussian random field 487
- Gauss–Markov process 890
- GEMMA 616
- GenABEL 606
- genealogical distortion 167
- gene-based analyses 619
- GENEBPM 619, 620
- GENECAP 472
- GeneCards 782
- GENECLUSTER 619–620
- gene content analysis 1013–1014
- gene editing 509
- gene expression 850
- gene-level association analysis,
 imputation-based 688–690
- gene-level read counts 850–851
- GeneMatcher 784
- Gene Ontology (GO) 716
- general genotype model 608
- general linear model (GLM) 743
- general time-reversible (GTR) model
 200–201, 1002
- gene regulatory networks
 Bayesian network reconstruction process
 721–722
 building from the bottom up or top down
 714–715
 integrating genetic data 720–721
 predictive Bayesian networks 722
 predictive gene networks 714
 predictive network models 718–720
 reconstruct coexpression networks
 715–718
- genetics 551, 552. *See also* ethics
- genetic analysis 573
- genetic architecture of traits, in plants
 510–511
- genetic covariance and partitioning heritability
 685–688
- genetic differentiation among populations
 401. *See also* natural selection
- genetic diversity 51
- genetic drift 121, 148, 277, 398
 strength of 458
- genetic evidence, interpretation of 532–533
- genetic interference 575
- genetic linkage 576
- genetic map distance 575
- genetic mapping 510, 573–574
 association mapping 576
 associations between markers and traits,
 IBD-based detection of 583–589
 goal of 574
- IBD-based case–control studies 589–590
- IBD-based linkage likelihoods
 for major gene models 585–587
 for random-effects models 587–589
- IBD-based mapping 576–577
- inference of IBD from genetic marker data
 577
- IBD at a locus 577–579
- inferring local IBD from marker data
 581–583
- modeling probabilities of patterns of IBD
 580–581
- probabilities of patterns of IBD
 579–580
- meiosis and descent of DNA 574–576
- patterns of IBD in affected relatives
 590–592
- SNP marker data for 574
- trait data probabilities
 for major gene models 583–584
 for random-effects models 584–585
- genetic recombination 157
- genetic research 553
- genetics 573
- Genetics Home Reference 779
- genetic species delimitation 489–491
- genetic variation 458
 background 651–652
 causal inference with genetic data 665–667
 Mendelian randomization 652–655,
 667–670
- monogenic Mendelian randomization
 analyses 655–656
- polygenic Mendelian randomization analyses
 656–664

- gene tree 219
likelihood based on 229–230
probabilities 221–227
and species tree 170, 219–221
topology probabilities 221–225
gene tree topologies, likelihood based on 229
genic variance 426, 427
GenoCanyon 684, 685
genome 51
blocks of 138–140
transmission of 138
genome browsers and annotator software 682
genome-scale metabolic models (GEM) 990
1000 Genomes Project 74, 87, 95, 103, 104,
609, 613, 771
100,000 Genomes Project 556
genome-wide association studies (GWASs)
95, 97, 104, 597–599, 620–623, 631–632,
679–680, 764, 997, 1012
in bacteria
background 1008–1009
phylogenetic methods 1009–1011
regression-based methods 1011–1012
of binary disease outcomes 604, 606, 615
design concepts 599
genome-wide significance and correction
for multiple testing 601–602
GWAS genotyping technology design
600–601
phenotype definition 599
replication 602
sample size considerations 601
structure of common genetic variation
599–600
functional annotation data in
DNA conservation 681
epigenetic annotation data 681
transcriptomic annotation data 680–681
genetic structure in 611–612
identification of related individuals
612–613
identify ethnic outliers and account for
population stratification 613–614
mixed modelling approaches 614–615
software 615–616
meta-analysis 640–647
methods to integrate functional
annotations in
future directions 690
gene-level association analysis 688–690
partitioning heritability and genetic
covariance 685–688
multiple SNP association analysis 616
gene-based analyses 619
haplotype-based analyses 616–617
SNP-SNP interaction analyses 617–618
software 619–620
personalised medicine, delivery of 620
quality control 602–603
sample quality control procedures
604–605
SNP quality control procedures
603–604
software for 606
replication 632–635
single SNP association analysis 606
accounting for confounding factors
606–607
Bayesian methods 611
coding of SNP genotypes 607–609
generalised linear modelling framework
606
imputed genotypes 609
interactions with non-genetic risk factors
609–610
software for 611
visualisation of results of 609
synthesise annotation data
computational annotations 685
genome browsers and annotator software
682
supervised learning methods 682–684
unsupervised learning methods 684
use of genotype imputation in 98–99
winner's curse 635–640
genome-wide epistasis analysis 997,
1012–1013
genomic best linear unbiased prediction
(GBLUP) method 489, 516
genomic estimates of breeding values (GEBVs)
515, 517, 518
genomic medicine and variant interpretation
current challenges 761–765
effect of variant 765–771
functional assays of genetic variation
777–779
future challenges 783–786
holistic variant interpretation 782–783

- genomic medicine and variant interpretation
(Continued)
- human and model phenotype resources 779–782
- large human reference cohorts 771–776
- genomic rearrangements, in plants 509–510
- Genomics England 565, 566
- genomic sequencing 566
- GenoSkyline 685
- genotype 347
- genotype calling 5, 31
- from microarrays 94
 - from sequencing 94
- genotyped markers 87
- genotype–environment interaction, in crops 511–514
- genotype frequencies 5, 38, 39–40
- genotype imputation 97–98
- accuracy, factors affecting 104–105
 - ancestry 105–106
 - genotyping microarray 106
 - imputation methods 107
 - reference panel size and SNP allele frequency 105
- basic idea of 98
- future directions 109
- haploid imputation 99–100
- methods
- Beagle 101–102
 - imputation servers 103
 - IMPUTE 100–101
 - MaCH/minimac 101
 - positional Burrows–Wheeler transform 102
 - SNP tagging approaches 102–103
- quality control for imputed data 107–109
- summary statistic imputation 104
- and testing for association 103–104
 - use of, in GWASs 98–99
- genotype likelihood (GL) 94
- genotype likelihood models 4
- genotype–phenotype relationship 347
- Genotype-Tissue Expression (GTEx) project 680
- genotyping errors 471
- genotyping microarray 106
- GGE analysis 513
- Gibbs sampling 999
- Gibbs variable selection (GVS) 858
- Gillespie algorithm 889
- Glimmer 990
- Global Alliance for Genomics & Health (GA4GH) 784
- Global LipidsGenetics Consortium 655
- GLOBETROTTER 73, 263, 266–267
- globular protein, structure of 328
- GMMAT 616
- gnomAD 771
- GNOVA 688
- Godambe information matrix 19
- GPflow package 892
- GQT toolkit 74
- GrandPrix package 892
- Granger causality 752
- GREML algorithm 686
- GREMLIN method 327–328, 330
- GREMLIN server 334
- group penalty 961
- GTEx 781
- Gubbins 1005
- GWASs. *See* genome-wide association studies (GWASs)
- GWAVA 682
- h***
- Haldane genetic map 575
- HAPGEN program 96
- HapHedge 96
- HAPI-UR (Haplotype Inference for Unrelated Samples) 93
- haploid imputation 99–100
- haplotype 51, 87, 616
- analyses 616–617
 - clustering 617, 619
 - defined 87
 - homozygosity 54
 - use of 87
- haplotype-based methods, to detect selection 410–412
- haplotype estimation 87–88
- challenges in 87–88
- hidden Markov models for 89–93
- Beagle 91
 - fastPHASE 91
 - HAPI-UR 93
 - IMPUTE and MaCH 90–91

- PHASE and Li and Stephens model 89–90
SHAPEIT approach 91–92
measuring phasing performance 96–97
from reference panel 95–96
in related samples 93–94
simple haplotype frequency model 88–89
using sequencing data 94–95
- Haplotype Reference Consortium (HRC)
project 94
haplotype reference panels 98
haplotyping. *See* haplotype estimation
- HapMap Project 102, 105
HapMap SNPs 687
HapMix 73
hard sweep 403
Hardy–Weinberg equilibrium (HWE) 3, 6,
459, 464, 534, 603
deviation from 604
Hardy–Weinberg genotype frequency 459
Hastings ratio 25, 26, 28
heritability 424
in plant breeding 503–504
hermaphrodites 157–159
heterotachy 354
heterotic pool 518
heterozygosity 14–16
heterozygosity excess, effective population size
estimation from 459–460
heuristic algorithms 740
heuristic estimation methods 372–373
HGMD 781
hidden Markov models (HMMs) 40–46, 256,
287–289, 352–353, 581–582, 900
Baum–Welch algorithm 44–46
Bayesian inference of hidden states 42–43
for phasing 89–93
Viterbi algorithm 43–44
hierarchical Bayesian analysis 194
hierarchical models 2
hierBAPS 998, 1000, 1001
higher-level abstraction 336
highest posterior density (HPD) 22
high-throughput sequencing (HTS) 94, 978
hitchhiking effect 401–403
HKA test 406
HKY mutation model 371
HLI 771
HMMER 990
- HMM forward backward algorithm 101
HMMs. *See* hidden Markov models (HMMs)
homozygosity mapping 591
Hotelling's T^2 test 939
Hudson's composite likelihood approach 71
Human Gene Mutation Database (HGMD)
682
human genome 51
Human Genome Epidemiology Network
(HuGENet) 640
Human Metabolome Database 952
HumanMethylation array data 940–941.
See also DNA methylation
HumanMethylationEPIC array 935
Human Phenotype Ontology (HPO) 782
human *versus* experimental models 704–705
hybrid breeding 518
Hybrid-Coal (software) 225
hybridization and gene flow 240–241
hybridization parameter 240
hypergeometric model 137
hypermutator strains 1006
hypothesis-free approach 598
- i**
- identical by descent (IBD) 94, 574–575
detection of associations between markers
and traits 583–589
mapping based on 576–577 (*see also*
genetic mapping)
segments 466
single-locus 61
two-locus 61
identity by state (IBS) metric 612–613
identity matrix 707
iHS test 73
Illumina HiSeq 935
improper priors 21
imputation servers 103
IMPUTE 100–101
IMPUTE v1 100
IMPUTE v2 90–91, 100–101
IMPUTE v4 101
inbreeding 28–29
inbreeding effective size 459
incomplete lineage sorting (ILS) 279
incomplete selective sweeps, identification of
73–74
incomplete sweep 408

independence assumptions 2–3
 independent-rate models, for rate variation 209
 index of total selection 437
 individual fitness 432–433
 individual fitness function 441–442
InDrop 737
 inference methods and algorithms 4. *See also* statistical inference, model-based
 inference of genotype frequencies 5
 infinite-alleles model 169
 infinitely many alleles model 118
 infinitely many sites model 118
 infinitesimal model 137–138, 426
 allele frequency changes under 426–427
 changes in variances 427–429
 equilibrium additive genetic variance 429
 linearity of parent–offspring regressions under 426
 infinite-sites model 169
 Infinium Human-Methylation450K 935
 inflammatory bowel disease (IBD) 722
 information measures 108
 inheritance probability 240
 instrument strength independent of direct effect (*InSIDE*) 662
 integrated haplotype score (iHS) 410–411
 intermediate quantitative traits 599
 International Breast Cancer Intervention Study (IBIS) 830
 International HapMap Consortium 609, 613
 International HapMap Project 599–600
 International Maize and Wheat Improvement Center (CIMMYT) 503
 intractable likelihoods 16–18
 introgressed sequences identification 283–284
 advantages and disadvantages of 289
 hidden Markov and conditional random field models 287–289
 S^* -statistic 284–287
 invariance property, ML estimator 6
 IQ-TREE 1002

j

jackHMMER 335
 Jacobian matrix 10
 Jeffreys' prior 21–22
 Jensen's inequality 8

JLIM 713
jModelTest 1002
 Jones group 327
 Jones–Taylor–Thornton (JTT) model, of amino acid replacement 350

k

Kalman filter 890, 953
 KEGG 715
 kernel-based regression, in microbiome studies 986–987
 kernel function 880
 Kimura's stepping stone model 127–130
 kinetic models 703
 Kingman coalescent 117–118
 kinship matrix 707, 1011
 k-mers 1009
 k-nearest-neighbours batch effect test (*kBET*) 744
 Kolmogorov equations 123–124
 backward equation 123
 forward equation 123–124
 Kolmogorov–Smirnov test 751
 Kruskal–Wallis test 751
 Kullback–Leibler information 589

l

labeled histories 195
Lactase gene 54
 Lagrange multiplier 5, 6
 Lagrangian function 6
 Lander–Green algorithm 93
 landscape genetics 486–487
 Langevin equation 890
 Lassosum 825
 latent variable models 16
 LC-MS. *See* liquid chromatography – mass spectrometry (LC-MS)
 LD. *See* linkage disequilibrium (LD)
 LD score regression 686–687
 Leiden Open Variation Database 3.0 781
 Lewontin's paradox 136
 lifetime fitness 432–433
 lifetime reproductive success 432
 likelihood 374
 calculation of 374–375
 likelihood-based partition method 475
 likelihood computation, for multi-allelic markers 468

- likelihood function 4, 181
 likelihood methods, for estimating population recombination rate 70
 likelihood of hypothesis 179
 likelihood principle 4
 likelihood ratios 533–534, 542–543
 behavior of 546–547
 binary model 543–545
 continuous model 546
 semi-continuous model 545–546
 likelihood ratio statistic 37–38
 likelihood ratio test (LRT) 37, 58, 184, 374
 likelihood tuning parameter 257
 limma-voom 751
 lineage association 1009
 lineage sorting 170
 linear mixed model (LMM) 1011
 linear noise approximation (LNA) 890
 linear regression methods 959–960. *See also*
 metabolomics, statistical methods in
 linkage analysis 93
 linkage disequilibrium (LD) 51–53, 131, 163,
 248, 423, 535, 576, 600, 681, 800, 824,
 1000
 and additive genetic variance 427
 admixture LD 248, 254–255
 background LD 248–249, 255
 coalescent modelling 62–67
 complete LD 600
 data analysis 69–75
 decay of 265–267
 effective population size estimation from
 460–463
 extensions of two-locus LD measures 60
 haplotype patterns with different levels of
 52
 historical sketch of mathematical treatments
 of 60–62
 hitchhiking and 401
 LD coefficients 52
 matrix methods and diffusion
 approximations 61–62
 measuring 53–60
 relating genealogical history to 67–69
 score regression 666–667
 single-number summaries of 54–56
 spatial distribution of 56–60
 spatial structure of 59
 and two-locus identity by descent (IBD) 61
 linkage equilibrium 52, 130
 liquid chromatography – mass spectrometry
 (LC-MS) 952–954
 $L_1 + L_2$ penalty 961
 local clocks 209
 local geometry of fitness surfaces 442–443
 locations 574
 LocusZoo 609
 lod score 586
 LOFTEE 770
 logic models 703
 log-likelihood function 5, 6, 375
 longitudinal study 433
 long-range phasing (LRP) 94
 loss function 22
 Louvain algorithm 749
 lowest common ancestor (LCA) 981
 lymphoblastoid cell lines 704
- m**
- MACAU 937
 machine learning 335, 413–414, 804
 deep learning method 336–338
 MetaPSICOV method 336
 PconsC 335–336
 phylogeny constraints 339–340
 sequence pairing 338–339
 MaCH (Markov Chain Haplotyping) method
 90–91, 101
 Madsen–Browning scheme 619
 MAF (minor allele count) 604
 major histocompatibility complex (MHC)
 370
 Manhattan plot 609, 610
 map_align algorithm 334
 marginal likelihood 20, 181
 computing 185
 marker genes 751
 marker imputation, in plant breeding
 519
 Markov Affinity-based Graph Imputation of
 Cells (MAGIC) 745
 Markov chain law of large numbers 197
 Markov chain Monte Carlo (MCMC) 23–29,
 185, 844, 903, 999
 algorithm 89
 for approximating posterior probability of
 phylogenies 195–198
 convergence and mixing 26–28

- Markov chain Monte Carlo (MCMC)
(Continued)
- Markov chains 23–25
- Metropolis–Hastings algorithm 25–26,
28–29
- Markov chains 23–25, 1000
- aperiodic 24
 - discrete-time 23
 - irreducible 24
 - reducible 24
 - reversibility 24
- Markov clustering algorithm 999
- Markov equivalent 719
- Markovian dependence 928
- Markov model of codon substitution 371–372
- Markov process 371
- Markov property 23
- Mash algorithm 1001
- mass spectrometry (MS) 950
- match probability 533, 538
- mate selection 517–518
- mathematical models in population genetics
115–116
- multi-locus models 130–131
 - linkage disequilibrium 134–140
 - linkage equilibrium 131–134
 - single-locus models 116–117
 - diffusion approximations 120–126
 - of panmictic populations 116
 - random drift and Kingman coalescent
117–120
 - spatially structured populations 126–130
- mating systems 459
- of flowering plants 504
 - of hermaphroditic populations 504
- MaxEntScan 767
- maximum *a posteriori* (MAP) estimate 22
- maximum *a posteriori* (MAP) tree 234
- maximum clade credibility (MCC) tree
234
- maximum composite likelihood estimator
(MCLE) 19
- maximum likelihood (ML) 4, 179–180
- estimation 374–377
 - mechanics of 192–193
- maximum likelihood estimator (ML estimator)
4–6, 468
- asymptotically unbiased 7–8
- properties of 6–8
- quantifying confidence with 8–9
- maximum likelihood inference 4–19, 60
- maximum posterior probability (MAP)
estimate 181
- McDonald–Kreitman (MK) test 406
- MCMC. *See* Markov chain Monte Carlo
(MCMC)
- mean-field substitution model 359
- mean fitness 132
- mean fitness landscape 441
- medical ethics 552
- meiosis 573, 574–575
- Mendelian randomization (MR) 621,
652–655, 712
- monogenic 655–656
 - polygenic 656–664
 - interactions and subsetting 663–664
 - median estimation methods 660
 - modal estimation methods 660
 - MR-Egger method 661–663
 - multivariable methods 663
 - practical advice 664
 - regularization methods 660–661
 - robust methods 661
- Mendel's first law 574
- mental capacity 559
- 31-mers 1009
- messenger RNA (mRNA) 700, 843
- MetaboAnalyst 3.0 957
- metabolome-wide significance level (MWSL)
956
- metabolomics, statistical methods in 949–950
- metabolite identification and pathway
analysis
 - pathway and metabolite set analysis
971–972
 - statistical correlation spectroscopy
969–970
- multivariate methods and chemometrics
- techniques 958–959
 - linear regression methods 959–960
 - shrinkage methods 960–961
- network analysis 966–969
- orthogonal projection methods 961
- onto latent structures 965–966
 - partial least squares 964–965
 - principal components analysis 962–964

- preprocessing and deconvolution
 liquid chromatography – mass spectrometry 952–954
 nuclear magnetic resonance spectroscopy 950–952
- univariate methods 954–956
 metabolome-wide significance levels 956–957
 sample size and power 957–958
- MetaNeighbour 752
- MetaPSICOV method 327
- methicillin-resistant *Staphylococcus aureus* (MRSA) 1004
- method of Lagrange multipliers 5
- MethylKit 937
- Metropolis–Hastings algorithm 25–26, 195–196
 in Bayesian inference 28–29
- Metropolis–Hastings step 94, 859, 904
- microarray data 851–856
 multi-class data 855–856
- microbial diversity index 983
- microbiome
 covariate 985–987
 future prospects of 991
 in human health and disease 977–978
 integrative analysis of 989–991
 mediator 987–989
 methods for analysis of 983–985
 16S rRNA and shotgun metagenomic sequencing data 980–983
- microsatellites 477, 597
- midparent–offspring regression 424
- migration
 conservative 157
 rate estimation 485
 strong 156–157
- minimum mean squared error (MMSE)
 estimate 22
- minimum spanning tree (MST) 750
- minor allele frequency (MAF) 803
- missense tolerance ratio (MTR) 771
- mitochondrial contamination estimation approach 298
- mixed stock analysis 480
- mixture models (MIX) 900
- mixture over marker (MOM) 866
- mnnCorrect 752
- mode of selection on trait 441
- molecular clock hypothesis 206
- molecular phylogenetics 170
- moment estimators 466–467
- moment methods, for estimating population recombination rate 69–70
- Monocle 750, 751
- Monte Carlo sampling, for intractable likelihoods 16–18
- Monte Carlo simulations 1010
- morality 552. *See also* ethics
- Moran model 120
- morgan 575
- most recent common ancestor (MRCA) 146, 577
- Mouse Genome Informatics (MGI) 781
- MP-EST (Maximum Pseudo-likelihood for Estimating Species Trees) 232, 233
- MR. *See* Mendelian randomization (MR)
- MrBayes program 202
- MULSEL method 333
- MultiBLUP 802
- multi-dimensional diffusion 124
- multi-dimensional scaling (MDS) 251, 613
- multilocus genotype mismatch method 472
- multi-marker multi-trait analysis 865–868
- multinomial distribution 6
- multiple hits 373
- multiple loci, population genetics of 130–131
 linkage disequilibrium 134–140
 applications 136–137
 approximations 137–138
 blocks of genome 138–140
 genotype frequencies, representing 134–136
 linkage equilibrium 131–134
 random drift 133–134
 selection gradients 131–133
- multiplexed assays for variant effect (MAVE) 778
- multispecies coalescent 219–221
 estimation of parameters at population and species levels 239
 hybridization and gene flow 240–241
 speciation times and population sizes 239–240
 species delimitation 241
 future prospects 242

- multispecies coalescent (*Continued*)
 gene trees and species tree 219–221
 probability distributions under 221
 gene tree probabilities 221–227
 model assumptions and violations 230
 site pattern probabilities 227–229
 species tree likelihoods 229–230
 species tree inference under 231
 Bayesian full-data methods 234–235
 empirical examples 237–238
 multilocus *versus* SNP data 236–237
 site pattern-based methods 235–236
 summary statistics methods 231–234
 multi-tissue HESS (MT-HESS) model 866
 multi-trait parent–offspring regression 425
 multivariate beta function 39
 multivariate breeder's equation 425
 multivariate selection 444–446
 mutation, models of 118
 mutual information (MI) 1013
 mutual nearest neighbours (MNN) 752
 M3VCF (data storage file format) 101
 MVNCALL 104
- n**
- naive empirical Bayes approach (NEB) 385
 National Research Council 536
 natural selection 369, 397–398
 methods to detect selection 405, 414–415
 approximate Bayesian computation 413
 composite likelihood methods 412–413
 haplotype-based methods 410–412
 intractable full likelihood methods 412
 machine learning methods 413–414
 methods based on genetic differentiation 408–410
 methods comparing substitutions and diversity 406–407
 methods using frequency spectrum 407–408
 substitution-based methods 405–406
 positive directional selection, signature of 400
 frequencies of selected alleles 400–401
 hitchhiking 401–403
 rates of substitution 400
 signature of selection in genome 399–400
 balancing selection 403
 confounders 404–405
 polygenic selection 403–404
 positive directional selection 400–403
 types of selection 398
 balancing selection 399
 directional selection 398–399
 polygenic selection 399
 NCBI RefSeq genomes 1001
 nearest-neighbor interchange (NNI) 192
 Needleman–Wunsch-like procedure 334
 neighbor joining (NJ) 1001
 Neolithic transition 311–312
 network inference models 752
 neutrality index 406
 neutral mutations 168–169
 neutral theory 369
 neutral Wright–Fisher model 116, 117, 146
 NEWHYBRIDS 485
 Newton's algorithm 9, 10
 Newton's method 9–11
 next generation sequencing (NGS) 296, 325, 997
NGS. See next generation sequencing (NGS)
 NH-DBN. See non-homogeneous dynamic Bayesian networks (NH-DBN)
 NHLBI Exome Sequencing Project 771
 NMR Suite 951
 NNSPLICE 767
 non-admixture model 480
 noncoding RNA (ncRNA) 701
 non-homogeneous dynamic Bayesian networks (NH-DBN) 900–901
 application
 morphogenesis in *Drosophila* 921–922
 synthetic biology in yeast 922–927
 methodology
 Bayesian linear regression 902–905
 Bayesian piecewise linear regression 905–908
 computational complexity 920
 coupled regression coefficients 908–914
 dynamic Bayesian network modelling 918–920
 dynamic Bayesian networks 901–902
 flexible allocation schemes 915–916
 time-varying network structures 916–918
 nonparametric differential expression for single cells (NODES) 742
 nonparametric statistical approach 3

- nonparametric techniques 55
 non-random mating 460
 non-ribosomal peptides (NRPS) 990
 normal distribution 9, 26, 27
 normality
 of phenotype 451
 of residuals 450
 and selection gradients 450–451
N
NormExpression 743
 nuclear magnetic resonance (NMR)
 spectroscopy 328, 950–952
 nucleotide substitution model 227
 null hypothesis 717
- O**
 odds ratio 608
 ODE. *See* ordinary differential equation (ODE)
 OD-seq analysis tool 333
 OLS. *See* ordinary least squares (OLS)
 omic data 547
 OMIM 779
 omni-genic model 679
 one-step transition matrix 23
 Online Mendelian Inheritance in Animals (OMIA) 781
On the Origin of Species (Darwin) 177
 operational taxonomic units (OTU) 980
 opportunity for selection 437–438
 optimal contribution theory 518
 oracle penalty 961
 ordinary differential equation (ODE) 703, 752, 888
 ordinary least squares (OLS) 959, 960
 Ornstein–Uhlenbeck (OU) process 126, 208, 360, 880
 Orphanet 779
 orthogonal signal correction (OSC) 965
 outcrossing species 505
 output feature map 337
 overlapping generations, models of 120
 overlapping mixture of Gaussian processes (OMGP) 892
 Oxford Phasing Server 96
- P**
 pain in the torus 127, 128
 pairwise relatedness analysis 472–473
 pairwise sequentially Markovian coalescent (PSMC) model 261–262
 paleobiochemistry 383
 pan-genome 998
 PANINI 1014
 parametric modeling 3
 parental haplotypes 94
 parent-independent mutation 133
 parent–offspring phenotypic covariance 423
 parent–offspring regressions 423–424
 breeder’s equation 425
 directional selection differential 424–425
 genetic and phenotypic covariance matrices 425
 multiple trait 425
 response to selection 424–425
 single-trait 424
 Parisi–Echave technique 357
 parsimony method 178
 partial least squares (PLS) 804, 961, 964–965
 regression 703
 PCA. *See* principal components analysis (PCA)
 peak of polymorphism 166
 Pearson correlation 707
 Pearson correlation coefficient 57
 pedigree-based linkage analysis 585–587
 pedigree-derived relationship matrix 430
 peer review 557
 penetrance probabilities 583
 percent splicing index (PSI) 746
 permutation procedures 601–602
 person of interest (POI) 532, 533, 547
 PHASE software 90, 619
 phasing. *See* haplotype estimation
 PhastCons 681
 Phen-Gen 685
 phenotype 347
 definition 599
 phenotype prediction and DNA variants 799–800
 accuracy of prediction 806–808
 additive genetic values 802–806
 Bayesian genomic selection models 808–809
 DNA polymorphisms 801
 genetic variation affecting phenotype 800–801
 genomic prediction 809–810
 phenotypic covariance matrix 425
 phenotypic value of offspring 422
 phenotypic variance 503

- Phevor 685
 philosophical ethics 552
 phylogenetic inference 170
 phylogenetic networks 240
Phylogenetic Systematics (Hennig) 177
 phylogenetic trees 1001
 calibrating 209–211
 phylogeny
 constraints 339–340
 likelihood calculation on 379–380
 phylogeny estimation, by likelihood-based methods 177–179, 212
 applications of likelihood-based methods 199–211
 divergence time estimation 206–211
 expanding the model around groups of sites 202–204
 rate variation across sites 204–206
 substitution models 199–202
 Bayesian inference 180–184
 mechanics of 192–198
 calculating likelihood for phylogenetic model 186
 calculating probability of character history 187–188
 character matrices and alignments 186
 continuous-time Markov model 188–189
 marginalizing over character histories 189–191
 phylogenetic model 186–187
 choosing among models 184–185
 cladists *vs.* pheneticists 177–178
 maximum likelihood method 179–180
 mechanics of 192–193
 statistical phylogenetics 178
 PhyloP 681
 phyloscanner 1006
 piecewise constant population sizes 261
 PINES 685
 plant breeding 501–502
 breeding systems 504–505, 506
 experimental design and analysis 519–521
 genetic architecture of traits 510–511
 genomic rearrangements 509–510
 genomic selection
 and cross prediction 517
 genomic prediction of hybrid performance and heterosis 518
 genotype–environment interaction 514–515
 marker imputation 519
 mate selection 517–518
 and phenotyping cost 517
 quantitative trait loci and major genes 516–517
 sequential selection 518
 heritability and breeder’s equation in 502–504
 polyploidy in plants and genetic consequences 505–509
 response to environment and plasticity 511–514
 plasmids 998
 pleiotropy 423
 pleiotropy effects test (PET) 716
 PLINK 606, 619, 824
 PLINKv2 606, 611, 615
 plmDCA method 326, 327
 PLS. *See* partial least squares (PLS)
 PMEN2 lineage 1014
 Poisson-beta distribution 747
 Poisson distribution 739
 Poissonization 168
 Poisson model 850
 Pólya–gamma latent variables 860
 polygenic effect 585
 polygenic risk score (PRS) 803, 817
 polygenic selection 399, 403
 polyketides (PKS) 990
 polymerase chain reaction (PCR) 737, 765, 934
 polyploidy in plants 505–509
 Pool-seq 5, 10–11
 population assignment analysis 477–479, 484
 population-based association studies 597–598
 population branch statistic (PBS) 409
 population, definition of 251–252
 population genetics 53, 116
 multi-locus models 130–131
 linkage disequilibrium 134–140
 linkage equilibrium 131–134
 single-locus models 116–117
 diffusion approximations 120–126
 of panmictic populations 116

- random drift and Kingman coalescent 117–120
- spatially structured populations 126–130
- population mean fitness 441–442
- population recombination rate 55. *See also* recombination rate, estimation of
- population-scale sequencing projects 74
- population size changes and split times, inferring 259–260
- allele frequency spectrum approach 260–261
- whole-genome sequencing approaches 261–262
- population splits 247
- population-structure parameter 535
- positional Burrows–Wheeler transform (pBWT) 74, 92, 102
- posterior distribution 20
- posterior mean estimate 22
- posterior probability 184
of tree 195
- post-mortem degradation (PMD) 298–299
- potential scale reduction factors (PSRF) 919
- Potts models 1012
- Prdm9* 75
- precision medicine 556
- PredictABEL 820, 821
- PrediXcan 689, 690
- PRIMUS 615
- principal components analysis (PCA) 249–251, 613, 703, 747, 892, 961–964
- Principles of Numerical Taxonomy* (Sokal and Sneath) 177
- prior distribution 4, 20
choice of 21–22
- probabilistic model of protein evolution 347–348, 360–361
codon-based models 355–356
dependence among positions 357–359
heterogeneity of replacement rates among sites 351
- models of amino acid replacement
Dayhoff and Eck model 348–349
descendants of Dayhoff model 350–351
models with physicochemical basis 355
protein structural environments 351–353
stochastic models of structural evolution 359–360
- variation of preferred residues among sites 353–355
- probabilistic quotient normalization (PQN) 954
- probability distributions 3–4
emission probabilities 41–42
initial state distribution 41
under multispecies coalescent 221
gene tree probabilities 221–227
model assumptions and violations 230
site pattern probabilities 227–229
species tree likelihoods 229–230
- transition probabilities 41
- Procrustean design 519
- product of approximate conditionals (PAC) scheme 71
- professional ethics 552
- profile drift 332
- proportion of variance explained (PVE) 863
- protein
disordered 329
globular 328
protein–protein interactions 329–330
- protein evolution, probabilistic model of 347–348, 360–361
codon-based models 355–356
dependence among positions 357–359
heterogeneity of replacement rates among sites 351
- models of amino acid replacement
Dayhoff and Eck model 348–349
descendants of Dayhoff model 350–351
models with physicochemical basis 355
protein structural environments 351–353
stochastic models of structural evolution 359–360
- variation of preferred residues among sites 353–355
- protein–protein interaction (PPI) 721
- protein structural environments 351–353
- protein-truncating variants (PTV) 766
- pseudocounted quantiles (pQ) 742
- pseudo-haploid model 95
- pseudo-likelihood method (PLM) 233, 326
- pseudo likelihoods 18–19
- pseudotime 749–750
- PSI-BLAST 335
- PSICOV method 327

- public engagement 567
p-value 547, 716
- q**
 QCTOOL 606
 Q-function 11–13
 QIIME 981
 quadratic convergence 10
 quadratic function 22
 quadratic penalty 960–961
 quadratic selection differential 442
 quadratic selection gradient 441
 quality control (QC) 953–954
 GWAS 602–603
 automated procedures 603
 sample quality control procedures 604–605
 SNP quality control procedures 603–604
 software for 606
 quantile–quantile plot, GWAS 612
 quantitative genetics 422. *See also*
 evolutionary quantitative genetics
 quantitative trait loci (QTLs) 502, 509, 511, 514, 516–518, 844
 multiple-response models 864–868
 single-response models 863–864
 quantitative traits 136
 quasi likelihoods 18–19
 quasi-Newton methods 10
- r**
 RADMeth 937
 random drift 133–134
 random regression model 804
 random-sites models 384
 RapidNJ 1001
 RaptorX-contact 337
 rare alleles 467
 rates of substitution 400
 rate variation over protein sites 351
 RAxML 1002
 real-time polymerase chain reaction (RT-PCR) 922
 recent common ancestry, use of 72–73
 recombination 159–160, 575
 ancestral recombination graph 160–163
 breakpoints 161–163
 clocks 73
 properties and effects of 163–164
 recombination rate, estimation of 69
 approximating coalescent 71–72
 approximating likelihood 70–71
 likelihood methods 70
 moment methods 69–70
 RECON project 1007
 recruitment of participants 558–559
 reduced representation bisulfite sequencing (RRBS) 935
 reference genome 846
 reference panels 105
 reference transcriptome 846
 rejection algorithm 31–32
 relaxed-clock models 207–208
 autocorrelated models 208–209
 grouping branch rates 209
 independent branch-rate models 209
 relaxed molecular clocks 207–209
 remove unwanted variation (RUV) 744
 replicates and clusters, hierarchical models of 887–888
 replication 632–635
 forms of 632–634
 heterogeneity 635
 motivation 632
 significance thresholds 634–635
 two-stage GWAS design 634
 repressor proteins 700
 reproduction ‘events’ 130
 residual blocks 337
 RettBASE 781
 reverse graph embedding 750
 reversible jump Markov chain Monte Carlo 71, 202, 490
 rhodopsin parsed contact maps 335
 ribosomally encoded and posttranslationally modified peptides (RiPP) 990
 ridge regression 961
 risk control models, of genetics 553–554
 risk function 22
 R language 1000
 RNA-sequencing (RNA-seq) 697, 735
 data 846–847
 RNA structure prediction 328–329
 Roadmap Epigenomics Project 680, 681
 Robertson–Price identity 435, 437
 Robinson model, of protein-coding DNA sequence evolution 358
 ROLLOFF 265–267

S

Saccharomyces cerevisiae 617
 sample-centric approach (genotypes storage) 74
 Sanger Imputation Server 102
 SCODE 752
 scone package scores 742
 score function 5
 SCOTTI 1007
 SCUBA 750
 secondary theorem of selection 435
 seemingly unrelated regression (SUR) model 868
 SEER software 1011
 segment parsing 334–335
 segregation load 126
 Segway 684, 685
 selection coefficients 398, 438–441
 selection differentials 424–425, 434
 selection, fundamental and secondary theorems of 435–436
 selection gradients 443, 446–447
 correlational 448
 inference of 447–451
 normality and 450–451
 quadratic 448
 selection intensity 437, 439
 selection on mean 439
 selection on trait, detecting 439
 selection on variance 439–441
 selective sweeps 166–167, 401
 semi-continuous model 545–546
 sense codons 371
 separation of time-scales 134, 157
 sequence alignment 332
 alignment benchmarking and improvement 333
 family membership validation 332–333
 sequence covariation analysis, in biological polymers 325–326, 340–341
 applications
 allostery and dynamics 330
 CASP 330–332
 globular protein fold prediction 328
 protein disordered regions 329
 protein–protein interactions 329–330
 RNA structure prediction 328–329
 transmembrane protein prediction 328
 CCMpred method 327

comparison to known structures 333–334
 DCA method 326–327
 GREMLIN method 327–328
 machine learning 335
 deep learning method 336–338
 MetaPSICOV method 336
 PconsC 335–336
 phylogeny constraints 339–340
 sequence pairing 338–339
 plmDCA method 327
 PSICOV method 327
 segment parsing 334–335
 sequence alignment 332
 alignment benchmarking and improvement 333
 family membership validation 332–333
 sequence data 138
 sequence pairing 338–339
 direct contact iteration 339
 phylogenetic similarity 339
 sequence path 358
 sequencing reads 95
 sequential kernel association test (SKAT) 619
 sequentially Markov coalescent (SMC) 65, 139, 415
 sequential Monte Carlo (SMC) 34
 sequential selection 518
 SeqWell 737
 Seurat 751
 Seurat algorithm 749
 Seurat R package 753
 sex systems 504
 sexual selection 433
 SHAPEIT approach 91–92, 96, 619
 SHAPEIT v1 92, 93
 SHAPEIT v2 92, 93, 95, 96
 SHAPEIT v3 92
 SHAPEIT v4 92
 Sherlock 713
 SHiC 414
 shifting balance model of evolution 133
 shortcut methods. *See* summary statistics methods
 shrinkage methods 960–961. *See also* metabolomics, statistical methods in shuttle breeding 503
 sibship frequency approach 464–465
 SINCERA 749

- Single-cell Analysis Via Expression Recovery (SAVER) 745
- single-cell consensus clustering (SC3) 749
- single-cell Interpretation via Multikernel LeaRning (SIMLR) 749
- single-cell latent variable models (scLVM) method 744
- SINgle CELl Regularized Inference using TIme-stamped Expression profileS (SINCERITIES) 752
- Single Cell rEgulatory Network Inference and Clustering (SCENIC) 749
- single-cell RNA-sequencing (scRNA-seq) 735–736
- experimental platforms and low-level analysis
 - computational analysis 739
 - high-throughput methods 737–739
 - low-throughput methods 736–737
 - novel statistical challenges
 - estimating transcript levels 739–747
 - expression matrix analysis 747–753
 - single-locus models, of panmictic populations 116–117
 - diffusion approximations 120–126
 - adding selection and mutation 121–122
 - diffusion process 120–121
 - Gaussian fluctuations and drift load 125–126
 - Kolmogorov equations 123–124
 - multiple alleles 124–125
 - Wright–Fisher model and 121
 - random drift and Kingman coalescent 117–120
 - adding mutation 118
 - Cannings model 118–119
 - Kingman coalescent 117–118
 - limitations 119–120
 - Moran model 120
 - neutral Wright–Fisher model 117
 - spatially structured populations 126–130
 - continuous space 126–127
 - Kimura’s stepping stone model 127–130
 - spatial Lambda–Fleming–Viot process 130
 - single-marker multi-trait analysis 865
 - single nucleotide polymorphisms (SNPs) 53, 87, 98, 104, 301, 547–548, 574, 597, 598, 600, 633, 679, 706, 800, 817, 942
 - coding of SNP genotypes 607–609 (*see also* genome-wide association studies (GWASs))
 - low-quality SNPs 603, 604
 - quality control procedures 603–604
 - tag SNPs 600
 - single nucleotide variants (SNV) 765
 - SingleSplice 746
 - singleton density score (SDS) 411
 - singular value decomposition (SVD) 744, 962
 - site frequency spectrum (SFS) 400. *See also* allele frequency spectrum (AFS)
 - site-likelihoods 348
 - site pattern probabilities 227–229
 - site-wise likelihood ratio (SLR) test 386
 - SKAT-O 619
 - SLICER 750
 - SMARTPCA 616
 - Smartseq2 737
 - snapclust 1000, 1001
 - SNAPP 237
 - SNPs. *See* single nucleotide polymorphisms (SNPs)
 - SNP tagging approaches 102–103
 - SNPTEST 611
 - SNPtools approach 94
 - social issues, by genetics research 567–568
 - soft sweep 403, 404
 - software
 - genetic structure in GWAS 615–616
 - gene tree topology probabilities 225
 - for GWAS quality control 606
 - for haplotype reconstruction in GWASs 619–620
 - for single-variant analysis 611
 - STRUCTURE 136
 - software library 74
 - solidary participation 557
 - South Asian Genome project 776
 - spatial Lambda–Fleming–Viot process 130
 - Spearman correlation network 752
 - speciation times and population sizes, estimation of 239–240
 - species delimitation 239, 241
 - species tree 170, 219–221. *See also*
 - multispecies coalescent
 - species tree inference, under multispecies coalescent 231
 - Bayesian full-data methods 234–235

- empirical examples 237–238
 multilocus *versus* SNP data 236–237
 site pattern-based methods 235–236
 summary statistics methods 231–234
 species tree likelihoods, under multispecies
 coalescent 229–230
 gene trees and 229–230
 gene tree topologies and 229
 multilocus data and 230
 speed breeding platform 503
 SpeedGene data format 74
 splines 449
 SpydrPick 1013
 squared correlation 105
 SQUAREM method 11–12
 S^* -statistic 284–287
 stabilizing selection 439, 441
 differential 440
 gradients 448
 STARNET 680
 stationary distribution 24
 stationary frequencies 201
 statistical coupling 330
 statistical genetics 1
 statistical inference, model-based 1
 Bayesian inference 20–37
 hidden Markov models 40–46
 maximum likelihood inference 4–19
 model selection 37–40
 statistical models and inference 1–4
 statistical models 1, 2. *See also* statistical
 inference, model-based
 independence assumptions 2–3
 probability distributions 3–4
 statistical phylogenetics 178
 statistical total correlation spectroscopy
 (STOCSY) 969, 970
 stepping stone model 127–130
 stewardship 565–567
 STITCH (Sequencing To Imputation Through
 Constructing Haplotypes) method 95
 stochastic differential equation 120
 stochastic models 115
 of structural evolution 359–360
 stochastic search variable selection (SSVS)
 858
 stop codons 371
 Storey's method 706
 strict clock 1003
 strong migration 156–157
 structural variants (SV) 765
 STRUCTURE 136, 480–483, 998
 clustering algorithms 252–253
 structured coalescent 137, 155–156
 structured Wright–Fisher model 154
 Structure (v2) model 1000
 subjective probabilities, in Bayesian analysis
 181
 subset quantile normalization (SQN) 940
 substitution-based methods, to detect selection
 405–406
 substitution models 184, 199–202
 GTR model 200–201
 Jukes and Cantor (1969) model 199–200
 Kimura (1980) model 200
 reversible-jump MCMC 202
 time-reversible 201–202
 subtree pruning and regrafting (SPR) 192
 succinct tree sequence 75
 summary statistics methods 231–234
 SuperDCA 1013
 supervised learning methods 682–684
 support vector machine (SVM) 414
 SVDQuartets method 236
 sweep
 hard 137, 403, 411
 incomplete 408
 selective 166–167, 401
 soft 137, 403, 404, 411
 sweep from standing variation (SSV) 403
 Swendsen–Wang algorithm 861
 switching model 390
 symmetrical models 137
 symmetric matrix 425
 Systematic Long Range Phasing (SLRP) 94
 systematic review 557

t

- Targeting Induced Local Lesions in Genomes
 (TILLING) 509
 TASC 751
 technical replicates 845
 temporal method 466–470
 TensorFlow package 892
 TESLA 921
 theory of speed and scale 123
 therapeutic misconception 557, 559
 therapeutic privilege 563

- three-population test 277–279
 threshold number 472
 Time reconstruction in Single-Cell RNAseq ANalysis (TSCAN) 750
 time-reversible model 201
 time to the most recent common ancestor (TMRCA) 261
 tip dating approach 211
 TopMed study 105
 topology 148–149
 total evidence approach 211
 total inbreeding coefficient 535
 trait SDS (tSDS) 411
trans-acting elements 701
 transcription 700–702
 transcription factor binding-associated proteins (TAF) 700
 transcription factor binding site (TFBS) 721
 transcriptomic annotation data 680–681.
See also genome-wide association studies (GWASs)
 transcripts per million (TPM) 742
trans eQTL 712, 713
 trans-ethnic fine-mapping 622
 transgressive segregation 502
 transition probabilities 89, 90, 348
 transition–transversion rate difference 373,
 378–379
 TransPhylo 1006
 transversions 200
 TraP 767
 tree
 neighbors 192
 topology 192
 tree bisection and reconnection (TBR) 192
 TreeTime 1004
 treeWAS 1011
 t-SNE 1014
 TWAS 689, 690
- U**
- UK Biobank 268, 555–556, 566, 622
 UK10K 771
 UK National Health Service (NHS) 555
 Ultimate MRCA 161
 unbiased estimator 7
- UniFrac 986
 UniProt 781
 unique molecular identifiers (UMI) 739–740
 UNPHASED 619
 unsupervised learning methods 684
 UTMOST 690
- V**
- variance effective size 459
 variance selection differential 442
 variance-standardized selection differential 437
 Variant Call Format 74
 variant-centric approach (data compression) 74
 Variational Bayes (VB) 864
 VEGAS 690
 Viterbi algorithm 43–44
- W**
- WAG model 350
 Wald confidence intervals 9
 Wald statistic 8
 Wald test 606
 walk through tree space 160, 163
 Wanderlust 750
 Waterfall 750
 Watterson estimator 90
 Wellcome Trust Case Control Consortium (WTCCC) 598, 633, 637
 WhatsHap method 95
 whole-exome sequencing (WES) 765
 whole-genome bisulfite sequencing (WGBS) 935
 whole genome sequencing (WGS) 105, 275,
 547, 601, 765, 999. *See also* genome-wide association studies (GWASs)
 Wilcoxon test 955
 Wilcox rank-sum test 751
 winner's curse 635–640
 Wishart distribution 868
 Wright–Fisher diffusion 121
 Wright–Fisher model 25, 116, 117, 121
 Wright–Fisher transition matrix 468
 Wright–Malécot formula 127, 128
 Wright's fixation index 475
 Wright's *F*-statistics 535
 Wright's inbreeding coefficient 605

x

X chromosome, genotype data from 605

y

Y-STR database 539

Y-STR profiles 539–542

Y-typing 539

z

Zebrafish Information Network (ZFIN) 781

Zellner–Siow Cauchy 863

zero-inflated factor analysis (ZIFA) 747

zero-inflated negative binomial-based wanted

variation extraction (ZINB-WaVE) 747

Z-scores 104, 687