



# Population Genetics

**Matthew B Hamilton**



WILEY-BLACKWELL

# Population Genetics

Dedication

For my wife and best friend, I-Ling.

# **Population Genetics**

**Matthew B. Hamilton**

 **WILEY-BLACKWELL**

A John Wiley & Sons, Ltd., Publication

This edition first published 2009, © 2009 by Matthew B. Hamilton

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical and Medical business to form Wiley-Blackwell.

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

*Editorial offices:* 9600 Garsington Road, Oxford, OX4 2DQ, UK  
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK  
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at [www.wiley.com/wiley-blackwell](http://www.wiley.com/wiley-blackwell)

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

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 the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

*Library of Congress Cataloguing-in-Publication Data*

Hamilton, Matthew B.

Population genetics / Matthew B. Hamilton.

p. ; cm.

Includes bibliographical references and index.

ISBN 978-1-4051-3277-0 (hbk. : alk. paper) 1. Population genetics. I. Title.

[DNLM: 1. Genetics, Population. QU 450 H219p 2009]

QH455.H35 2009

576.5'8—dc22

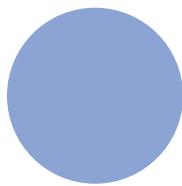
2008042546

A catalogue record for this book is available from the British Library.

Set in 10/12.5pt Photina by Graphicraft Limited, Hong Kong

Printed and bound in Malaysia

1 2009



# Contents

Preface and acknowledgments, xi

## 1 Thinking like a population geneticist, 1

- 1.1 Expectations, 1
  - Parameters and parameter estimates*, 2
  - Inductive and deductive reasoning*, 3
- 1.2 Theory and assumptions, 4
- 1.3 Simulation, 6
  - Interact box 1.1 The textbook website, 7

Chapter 1 review, 8

Further reading, 8

## 2 Genotype frequencies, 9

- 2.1 Mendel's model of particulate genetics, 9
- 2.2 Hardy–Weinberg expected genotype frequencies, 13
  - Interact box 2.1 Genotype frequencies, 14
- 2.3 Why does Hardy–Weinberg work?, 17
- 2.4 Applications of Hardy–Weinberg, 19
  - Forensic DNA profiling*, 19
  - Problem box 2.1 The expected genotype frequency for a DNA profile, 22
  - Testing for Hardy–Weinberg*, 22
  - Box 2.1 DNA profiling, 22
  - Interact box 2.2  $\chi^2$  test, 26
  - Assuming Hardy–Weinberg to test alternative models of inheritance*, 26
  - Problem box 2.2 Proving allele frequencies are obtained from expected genotype frequencies, 27
  - Problem box 2.3 Inheritance for corn kernel phenotypes, 28
- 2.5 The fixation index and heterozygosity, 28
  - Interact box 2.3 Assortative mating and genotype frequencies, 29
  - Box 2.2 Protein locus or allozyme genotyping, 32
- 2.6 Mating among relatives, 33
  - Impacts of inbreeding on genotype and allele frequencies*, 33
  - Inbreeding coefficient and autozygosity in a pedigree*, 34
  - Phenotypic consequences of inbreeding*, 37
  - The many meanings of inbreeding*, 40
- 2.7 Gametic disequilibrium, 41
  - Interact box 2.4 Decay of gametic disequilibrium and a  $\chi^2$  test, 44
  - Physical linkage*, 45
  - Natural selection*, 46
  - Interact box 2.5 Gametic disequilibrium under both recombination and natural selection, 46

- Mutation*, 47
- Mixing of diverged populations*, 47
- Mating system*, 48
- Chance*, 48
  - Interact box 2.6 Estimating genotypic disequilibrium, 49
- Chapter 2 review, 50
- Further reading, 50
- Problem box answers, 51
  
- 3 Genetic drift and effective population size, 53**
- 3.1 The effects of sampling lead to genetic drift, 53
  - Interact box 3.1 Genetic drift, 58
- 3.2 Models of genetic drift, 58
  - The binomial probability distribution*, 58
  - Problem box 3.1 Applying the binomial formula, 60
  - Math box 3.1 Variance of a binomial variable, 62
  - Markov chains*, 62
  - Interact box 3.2 Genetic drift simulated with a Markov chain model, 65
  - Problem box 3.2 Constructing a transition probability matrix, 66
  - The diffusion approximation of genetic drift*, 67
- 3.3 Effective population size, 73
  - Problem box 3.3 Estimating  $N_e$  from information about N, 77
- 3.4 Parallelism between drift and inbreeding, 78
- 3.5 Estimating effective population size, 80
  - Interact box 3.3 Heterozygosity, and inbreeding over time in finite populations, 81
  - Different types of effective population size*, 82
  - Problem box 3.4 Estimating  $N_e$  from observed heterozygosity over time, 85
  - Breeding effective population size*, 85
  - Effective population sizes of different genomes*, 87
- 3.6 Gene genealogies and the coalescent model, 87
  - Math box 3.2 Approximating the probability of a coalescent event with the exponential distribution, 93
  - Interact box 3.4 Build your own coalescent genealogies, 94
- 3.7 Effective population size in the coalescent model, 96
  - Interact box 3.5 Simulating gene genealogies in populations with different effective sizes, 97
  - Coalescent genealogies and population bottlenecks*, 98
  - Coalescent genealogies in growing and shrinking populations*, 99
  - Interact box 3.6 Coalescent genealogies in populations with changing size, 101
- Chapter 3 review, 101
- Further reading, 102
- Problem box answers, 103
  
- 4 Population structure and gene flow, 105**
- 4.1 Genetic populations, 105
  - Method box 4.1 Are allele frequencies random or clumped in two dimensions?, 110
- 4.2 Direct measures of gene flow, 111
  - Problem box 4.1 Calculate the probability of a random haplotype match and the exclusion probability, 117
  - Interact box 4.1 Average exclusion probability for a locus, 117
- 4.3 Fixation indices to measure the pattern of population subdivision, 118
  - Problem box 4.2 Compute  $F_{IS}$ ,  $F_{ST}$ , and  $F_{IT}$ , 122
  - Method box 4.2 Estimating fixation indices, 124

- 4.4 Population subdivision and the Wahlund effect, 124
    - Interact box 4.2 Simulating the Wahlund effect, 127
    - Problem box 4.3 Account for population structure in a DNA-profile match probability, 130
  - 4.5 Models of population structure, 131
    - Continent-island model, 131*
    - Interact box 4.3 Continent-island model of gene flow, 134
    - Two-island model, 134*
    - Infinite island model, 135*
    - Interact box 4.4 Two-island model of gene flow, 136
    - Math box 4.1 The expected value of  $F_{ST}$  in the infinite island model, 138
    - Problem box 4.4 Expected levels of  $F_{ST}$  for Y-chromosome and organelle loci, 139
    - Interact box 4.5 Finite island model of gene flow, 139
    - Stepping-stone and metapopulation models, 141*
  - 4.6 The impact of population structure on genealogical branching, 142
    - Combining coalescent and migration events, 143*
    - The average length of a genealogy with migration, 144*
    - Interact box 4.6 Coalescent events in two demes, 145
    - Math box 4.2 Solving two equations with two unknowns for average coalescence times, 148
- Chapter 4 review, 149  
 Further reading, 150  
 Problem box answers, 151

## 5 Mutation, 154

- 5.1 The source of all genetic variation, 154
  - 5.2 The fate of a new mutation, 160
    - Chance a mutation is lost due to Mendelian segregation, 160*
    - Fate of a new mutation in a finite population, 162*
    - Interact box 5.1 Frequency of neutral mutations in a finite population, 163
    - Geometric model of mutations fixed by natural selection, 164*
    - Muller's Ratchet and the fixation of deleterious mutations, 166*
    - Interact box 5.2 Muller's Ratchet, 168
  - 5.3 Mutation models, 168
    - Mutation models for discrete alleles, 169*
    - Interact box 5.3  $R_{ST}$  and  $F_{ST}$  as examples of the consequences of different mutation models, 172
    - Mutation models for DNA sequences, 172*
  - 5.4 The influence of mutation on allele frequency and autozygosity, 173
    - Math box 5.1 Equilibrium allele frequency with two-way mutation, 176
    - Interact box 5.4 Simulating irreversible and bi-directional mutation, 177
  - 5.5 The coalescent model with mutation, 178
    - Interact box 5.5 Build your own coalescent genealogies with mutation, 181
- Chapter 5 review, 183  
 Further reading, 183

## 6 Fundamentals of natural selection, 185

- 6.1 Natural selection, 185
  - Natural selection with clonal reproduction, 185*
  - Problem box 6.1 Relative fitness of HIV genotypes, 189
  - Natural selection with sexual reproduction, 189*
- 6.2 General results for natural selection on a diallelic locus, 193
  - Math box 6.1 The change in allele frequency each generation under natural selection, 194
  - Selection against a recessive phenotype, 195*
  - Selection against a dominant phenotype, 196*

- General dominance*, 197  
*Heterozygote disadvantage*, 198  
*Heterozygote advantage*, 198  
*The strength of natural selection*, 199  
Math box 6.2 Equilibrium allele frequency with overdominance, 200  
6.3 How natural selection works to increase average fitness, 200  
*Average fitness and rate of change in allele frequency*, 201  
Problem box 6.2 Mean fitness and change in allele frequency, 203  
*The fundamental theorem of natural selection*, 203  
Interact box 6.1 Natural selection on one locus with two alleles, 203  
Chapter 6 review, 206  
Further reading, 206  
Problem box answers, 206

## 7 Further models of natural selection, 208

- 7.1 Viability selection with three alleles or two loci, 208  
*Natural selection on one locus with three alleles*, 209  
Problem box 7.1 Marginal fitness and  $\Delta p$  for the *Hb C* allele, 211  
Interact box 7.1 Natural selection on one locus with three or more alleles, 211  
*Natural selection on two diallelic loci*, 212  
7.2 Alternative models of natural selection, 216  
*Natural selection via different levels of fecundity*, 216  
*Natural selection with frequency-dependent fitness*, 218  
*Natural selection with density-dependent fitness*, 219  
Math box 7.1 The change in allele frequency with frequency-dependent selection, 219  
Interact box 7.2 Frequency-dependent natural selection, 220  
Interact box 7.3 Density-dependent natural selection, 222  
7.3 Combining natural selection with other processes, 222  
*Natural selection and genetic drift acting simultaneously*, 222  
Interact box 7.4 The balance of natural selection and genetic drift at a diallelic locus, 224  
*The balance between natural selection and mutation*, 225  
Interact box 7.5 Natural selection and mutation, 226  
7.4 Natural selection in genealogical branching models, 226  
*Directional selection and the ancestral selection graph*, 227  
Problem box 7.2 Resolving possible selection events on an ancestral selection graph, 230  
*Genealogies and balancing selection*, 230  
Interact box 7.6 Coalescent genealogies with directional selection, 231  
Chapter 7 review, 232  
Further reading, 233  
Problem box answers, 234

## 8 Molecular evolution, 235

- 8.1 The neutral theory, 235  
*Polymorphism*, 236  
*Divergence*, 237  
*Nearly neutral theory*, 240  
Interact box 8.1 The relative strengths of genetic drift and natural selection, 241  
8.2 Measures of divergence and polymorphism, 241  
Box 8.1 DNA sequencing, 242  
*DNA divergence between species*, 242  
*DNA sequence divergence and saturation*, 243  
*DNA polymorphism*, 248

- 8.3 DNA sequence divergence and the molecular clock, 250  
*Interact box 8.2 Estimating  $\pi$  and  $S$  from DNA sequence data, 251*  
*Dating events with the molecular clock, 252*  
*Problem box 8.1 Estimating divergence times with the molecular clock, 254*
- 8.4 Testing the molecular clock hypothesis and explanations for rate variation in molecular evolution, 255  
*The molecular clock and rate variation, 255*  
*Ancestral polymorphism and Poisson process molecular clock, 257*  
*Math box 8.1 The dispersion index with ancestral polymorphism and divergence, 259*  
*Relative rate tests of the molecular clock, 260*  
*Patterns and causes of rate heterogeneity, 261*
- 8.5 Testing the neutral theory null model of DNA sequence evolution, 265  
*HKA test of neutral theory expectations for DNA sequence evolution, 265*  
*MK test, 267*  
*Tajima's D, 269*  
*Problem box 8.2 Computing Tajima's D from DNA sequence data, 271*  
*Mismatch distributions, 272*  
*Interact box 8.3 Mismatch distributions for neutral genealogies in stable, growing, or shrinking populations, 274*
- 8.6 Molecular evolution of loci that are not independent, 274  
*Genetic hitch-hiking due to background or balancing selection, 278*  
*Gametic disequilibrium and rates of divergence, 278*
- Chapter 8 review, 279  
Further reading, 280  
Problem box answers, 281

## 9 Quantitative trait variation and evolution, 283

- 9.1 Quantitative traits, 283  
*Problem box 9.1 Phenotypic distribution produced by Mendelian inheritance of three diallelic loci, 285*  
*Components of phenotypic variation, 286*  
*Components of genotypic variation ( $V_G$ ), 288*  
*Inheritance of additive ( $V_A$ ), dominance ( $V_D$ ), and epistasis ( $V_I$ ) genotypic variation, 291*  
*Genotype-by-environment interaction ( $V_{G\times E}$ ), 292*  
*Additional sources of phenotypic variance, 295*  
*Math box 9.1 Summing two variances, 296*
- 9.2 Evolutionary change in quantitative traits, 297  
*Heritability, 297*  
*Changes in quantitative trait mean and variance due to natural selection, 299*  
*Estimating heritability by parent–offspring regression, 302*  
*Interact box 9.1 Estimating heritability with parent–offspring regression, 303*  
*Response to selection on correlated traits, 304*  
*Interact box 9.2 Response to natural selection on two correlated traits, 306*  
*Long-term response to selection, 307*  
*Interact box 9.3 Response to selection and the number of loci that cause quantitative trait variation, 309*  
*Neutral evolution of quantitative traits, 313*  
*Interact box 9.4 Effective population size and genotypic variation in a neutral quantitative trait, 314*
- 9.3 Quantitative trait loci (QTL), 315  
*QTL mapping with single marker loci, 316*  
*Problem box 9.2 Compute the effect and dominance coefficient of a QTL, 321*  
*QTL mapping with multiple marker loci, 322*

Problem box 9.3 Derive the expected marker-class means for a backcross mating design, 324

*Limitations of QTL mapping studies, 325*

*Biological significance of QTL mapping, 326*

Interact box 9.5 Effect sizes and response to selection at QTLs, 328

Chapter 9 review, 330

Further reading, 330

Problem box answers, 331

## 10 The Mendelian basis of quantitative trait variation, 334

10.1 The connection between particulate inheritance and quantitative trait variation, 334

*Scale of genotypic values, 334*

Problem box 10.1 Compute values on the genotypic scale of measurement for *IGF1* in dogs, 335

10.2 Mean genotypic value in a population, 336

10.3 Average effect of an allele, 337

Math box 10.1 The average effect of the  $A_1$  allele, 339

Problem box 10.2 Compute the allele average effect of the *IGF1*  $A_2$  allele in dogs, 341

10.4 Breeding value and dominance deviation, 341

Interact box 10.1 Average effects, breeding values, and dominance deviations, 345

*Dominance deviation, 345*

10.5 Components of total genotypic variance, 348

Interact box 10.2 Components of total genotypic variance,  $V_G$ , 350

Math box 10.2 Deriving the total genotypic variance,  $V_G$ , 350

10.6 Genotypic resemblance between relatives, 351

Chapter 10 review, 354

Further reading, 354

Problem box answers, 355

## 11 Historical and synthetic topics, 356

11.1 Historical controversies in population genetics, 356

*The classical and balance hypotheses, 356*

*How to explain levels of allozyme polymorphism, 358*

*Genetic load, 359*

Math box 11.1 Mean fitness in a population at equilibrium for balancing selection, 362

*The selectionist/neutralist debates, 363*

11.2 Shifting balance theory, 366

*Allele combinations and the fitness surface, 366*

*Wright's view of allele-frequency distributions, 368*

*Evolutionary scenarios imagined by Wright, 369*

*Critique and controversy over shifting balance, 372*

Chapter 11 review, 374

Further reading, 374

## Appendix, 376

Statistical uncertainty, 376

Problem box A.1 Estimating the variance, 378

Interact box A.1 The central limit theorem, 379

Covariance and correlation, 380

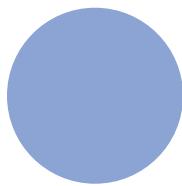
Further reading, 382

Problem box answers, 382

References, 383

Index, 396

*Color plates appear in between pages 114–115*



## Preface and acknowledgments

This book was born of two desires, one simple and the other more ambitious, both of which were motivated by my experiences learning and teaching population genetics. My first desire was to create a more up-to-date survey text of the field of population genetics. Several of the widely employed and respected standard texts were originally conceived in the mid-1980s. Although these texts have been revised over time, aspects of their organization and content are inherently dated. At the same time, I set out with the more ambitious goal of offering an alternative body of materials to enrich the manner in which population genetics is taught and learned.

Much of population genetics during the twentieth century was hypothesis-rich but data-poor. The theory developed between about 1920 and 1980 spawned manifold predictions about basic evolutionary processes. However, most of these predictions could not be tested or tested with only very limited power for lack of appropriate or sufficient genetic data. In the last two decades, population genetics has become a field that is no longer data-limited. With the collection and open sharing of massive amounts of genomic data and the technical ability to collect large amounts of genetic information rapidly from almost any organism, population genetics has now become data-rich but relatively hypothesis-poor. Why? Because mainstream population genetics has struggled to develop and employ alternative testable hypotheses in addition to those offered by traditional null models. Innovation in developing context-specific and testable alternative population genetic models is as much a requirement for hypothesis testing as empirical data. Such innovation, of course, first requires a sound understanding of the traditional and well-accepted models and hypotheses.

It is often repeated that the major advance in population genetics over the last decade or two is the availability of huge amounts of genetic data generated by the ability to collect genetic data and to sequence entire genomes. It is certainly true that advances in molecular biology, DNA sequencing

technology, and bioinformatics have provided a wealth of genetic data, some of it in the form of divergence or polymorphism data that is grist for the mill of population genetics hypothesis testing.

An equally fundamental advance in population genetics has been the emergence of new models and expectations to match the genetic data that are now readily available. Coalescent or genealogical branching theory is primary among these conceptual advances. During the past two decades, coalescent theory has moved from an esoteric problem pursued for purely mathematical reasons to an important conceptual tool used to make testable predictions. Nonetheless, teaching of coalescent theory in undergraduate and graduate population genetics courses has not kept pace with the growing influence of coalescent theory in hypothesis testing. A major impediment has been the lack of teaching materials that make coalescent theory truly accessible to students learning population genetics for the first time. One of my goals was to construct a text that met this need with a systematic and thorough introduction to the concepts of coalescent theory and its applications in hypothesis testing. The chapter sections on coalescent theory are presented along with traditional theory of identity by descent on the same topics to help students see the commonality of the two approaches. However, the coalescence chapter sections could easily be assigned as a group.

Another of my primary goals for this text was to offer material to engage the various learning styles possessed by individuals. Learning conceptual population genetics in the language of mathematics is often relatively easy for abstract and mathematical learners. However, my aim was to cater to a wide range of learning styles by building a range of features into the text. A key pedagogical feature in the book is formed by boxes set off from the main text that are designed to engage the various learning styles. These include Interact boxes that guide students through structured exercises in computer simulation utilizing software in the public domain. The simulation

problems are active rather than reflective and should appeal to trial-and-error or visual learners. Additionally, simulations uniquely demonstrate the outcome of stochastic processes where the evaluation of numerous replicates is required before a pattern or generalization can be seen. Because understanding the biological impact of stochastic processes is a major hurdle for many students, the Interact boxes should improve learning and retention. Problem boxes placed in the text rather than at the end of chapters are designed to provide practice and to reinforce concepts as they are encountered, appealing to experiential learners. Math boxes that fully explain mathematical derivations appeal to mathematical and logical learners and also provide a great deal of insight for all readers into the many mathematical approximations employed in population genetics. Finally, the large number of two-color illustrations in the text were designed to appeal to and help cultivate visual learning.

The teaching strategy employed in this text to cope with mathematics proficiency deserves further explanation. The undergraduate biology curricula employed at most US institutions has students take calculus in their first year and usually does not require the application of much if any mathematics within biology courses. This leads to students who have difficulty in or who avoid courses in biological disciplines that require explicit mathematical reasoning. Population genetics is built on basic mathematics and, in my experience, students obtain a much richer and nuanced understanding of the subject with some comprehension of these mathematical foundations. Therefore, I have attempted to deconstruct and offer step-by-step explanations the basic mathematics (mostly probability) required for a sound understanding of population genetics. For those readers with more interest or facility in mathematics, such as graduate students, the book also presents more difficult and detailed mathematical derivations in boxes that are separated from the main narrative of the text as well as chapter sections containing more mathematically rigorous content. These sections can be assigned or skipped depending on the level and scope of a course using this text. The Appendix further provides some very basic background in statistical concepts that are useful throughout the book and especially in Chapter 3 on genetic drift and Chapters 9 and 10 on quantitative genetics. This approach will hopefully provide students with the tools to develop their abilities in basic mathematics through application, and at the same time learn population genetics more fully.

Members of my laboratory and the students who have taken my population genetics course provided a range of feedback on chapter drafts, figures, and effective means to explain the concepts herein. This feedback was absolutely invaluable and helped me shape the text into a more useful and usable resource for students. James Crow graciously reviewed each chapter and offered many insightful comments on points both nuanced and technical. Rachel Adams, Genevieve Croft, and Paulo Nuin provided many useful comments on each of the chapters as I wrote them. A.W.F. Edwards reviewed the material on the fundamental theorem in Chapter 6 and also provided the photograph of R.A. Fisher. Sivan Rottenstreich and Judy Miller patiently helped me with numerous mathematical points and derivations, including material included in the Math boxes. John Braverman supplied me with insights and thought-provoking discussions that contributed to this book. Ronda Rolfs and Martha Weiss also provided comments and suggestions. I also thank Paulo Nuin for his collaboration and hard work on the creation of PopGene.S<sup>2</sup>. I also thank the anonymous reviewers from Aberdeen University, Arkansas State University, Cambridge University, Michigan State University, University of North Carolina, and University of Nottingham who provided feedback on some or all of the draft chapters.

John Epifanio provided the allozyme gel picture in Chapter 2. Eric Delwart provided the original data used to draw a figure in Chapter 6. Michel Veuille shared information on *Drosophila simulans* DNA sequences used in an Interact box in Chapter 8. Peter Armbruster shared unpublished mosquito pupal mass data used in Chapter 9. John Dudley and Stephen Moose generously shared the Illinois Long-Term Selection experiment data used in Chapter 9. Robert J. Robbins kindly provided high-resolution scans from Sewall Wright's Chapter in an original copy of the *Proceedings of the Sixth International Congress of Genetics* (see [www.esp.org](http://www.esp.org)).

I am grateful to Nancy Wilton for pushing me at the right times and for getting this project off the ground initially. Elizabeth Frank, Haze Humbert, and Karen Chambers of Wiley-Blackwell helped bring this book to fruition. I thank Nik Prowse for his expertise as a copy editor. I owe everyone at the Mathworks an enormous debt of gratitude since all of the simulations and many of the figures for this text were produced using Matlab.

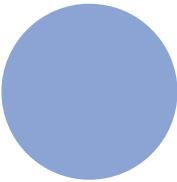
Matthew B. Hamilton  
September 2008

**Interact boxes**

Throughout this book you will encounter Interact boxes. These boxes contain opportunities for you to interact directly with the material in the text using computer simulations designed to demonstrate fundamental concepts of population genetics. Each box will contain step-by-step instructions for you to follow in order to carry out a simulation. For instructions on how to get started with the simulations, see Interact box 1.1 on page 7.

A companion website is available with interactive computer simulations for each chapter at:  
[www.wiley.com/go/hamiltongenetics](http://www.wiley.com/go/hamiltongenetics)





## CHAPTER 1

# Thinking like a population geneticist

All scientific fields possess a body of concepts that define their domain as well as a specialized vocabulary used to express these concepts precisely. Population genetics is no different and the entirety of this book is designed to introduce, explain, and demonstrate these concepts and their vocabulary. What may be unique about population genetics among the natural sciences is the way that its practitioners approach questions about the biological world. Population genetics is a dialog between predictions based on principles of Mendelian inheritance and results obtained from empirical measurement of genotype and allele frequencies that form the basis of hypothesis tests. Idealized predictions stemming from general principles form the basis of hypotheses that can be tested. At the same time, empirical patterns observed within and among populations require explanation through the comparison of various processes that might have caused a pattern. This first chapter will explore some of the ways in which population genetics approaches and defines problems that are relevant to the topics in all chapters. The chapter is also intended to give some insight into how to approach the study of population genetics.

### 1.1 Expectations

- What do we expect to happen?
- Expectations are the basis of understanding cause and effect.

In our everyday lives there are many things that we expect to occur or not to occur based on knowledge of our surroundings and past experience. For example, you probably do not expect to get hit by a meteorite walking to your next population genetics class. Why not? Meteorites *do* impact the surface of the Earth and on occasion strike something noticeable to people nearby. A few times in the distant past, in fact, large meteors have hit the Earth and left evidence like the Barringer Meteor crater in Arizona, USA.

What influences your lack of concern? It is probably a combination of basic knowledge of the principles of physics that apply to meteors as well as your empirical observations of the frequency and location of meteor strikes. Basic physics tells us that a small meteor on a collision course with Earth is unlikely to hit the surface since most objects burn up from the friction they experience traveling through the Earth's atmosphere. You might also reason that even if the object is big enough to pass through the atmosphere intact, and there are many fewer of these, then the Earth is a large place and just by chance the impact is unlikely to be even remotely near you. Finally, you have most probably never witnessed a large meteorite impact or even heard of one occurring during your lifetime. You have combined your knowledge of the physical world and your experience to arrive (perhaps unconsciously) at a prediction or an expectation: meteorite strikes are possible but are so infrequent that the risk of being struck on the way to class is minuscule. In this very same way, you have constructed models of many events and processes in your physical and social world and used the resulting predictions to make comparisons and decisions.

**Expectation** The expected value of a random variable, especially the average; a prediction or forecast.

The study of population genetics similarly revolves around constructing and testing expectations for genetic variation in populations of individual organisms. Expectations attempt to predict things like how much genetic variation is present in a population, how genetic variation in a population changes over time, and the pattern of genetic variation that might be left behind by a given biological process

that acts over time or through space. Building these expectations involves the use of first principles or the set of very basic rules and assumptions that define how natural systems work at their lowest, most basic levels. A first principle in physics is the force of gravity. In population genetics, first principles are the very basic mechanisms of Mendelian particulate inheritance and processes such as mutation, mating patterns, gene flow, and natural selection that increase, decrease, and shape genetic variation. These foundational rules and processes are used and combined in population genetics with the ultimate goal of building a comprehensive set of predictions that can be applied to any species and any genetic system.

Empirical study in population genetics also plays a central role in constructing and evaluating predictions. In population genetics as in all sciences, empirical evidence is not just from informal experiences, but is drawn from intentional observations, cleverly constructed comparisons, and experiments. Genetic patterns observed in actual populations are compared with expected patterns to test models constructed using general principles and assumptions. For example, we could construct a mathematical or computer simulation model of random genetic drift (change in allele frequency due to sampling from finite populations) based on abstract principles of sampling from a finite population and biological reproduction. We could then compare the predictions of such a model to the observed change in allele frequency through time in a laboratory population of *Drosophila melanogaster* (fruit flies). If the change in allele frequency in the fruit fly population matched the change in allele frequency predicted using the model of genetic drift, then we could conclude that the model effectively summarizes the biological sampling processes that take place in fruit fly populations.

It is also possible to use well-tested and accepted model expectations as a basis to hypothesize what processes caused an observed pattern in a biological population. Again to use a *Drosophila* population as an example, we might ask whether an observed change in allele frequency over some generations in a wild population could be explained by genetic drift. If the observed allele frequency change is within the range of the predicted change in allele frequencies based on a model of genetic drift, then we have identified a possible *cause* of the observed pattern. Comparing expected and observed genetic patterns in populations often requires modifications to

existing models or the construction of novel models in order to develop appropriate expectations. For example, a model of genetic drift constructed for *Drosophila* might naturally assume that all individuals in the population are diploid (individuals possess paired sets of homologous chromosomes). If we wanted to use that same model to predict genetic drift in a population of honey bees, we would have to account for the fact that in honey bee males are haploid (individuals possess single copies of each chromosome) while females are diploid. This change in reproductive biology could be taken into account by altering the assumptions of the model of genetic drift to make the prediction appropriate for honey bee populations. Note that without some modification, a single model of genetic drift would not accurately predict allele frequencies over time in both fruit flies and honey bees since their patterns of reproduction and chromosomal inheritance are different.

### **Parameters and parameter estimates**

While developing the expectations of population genetics in this book, we will most often be working with idealized quantities. For example, allele frequency in a population is a fundamental quantity. For a genetic locus with two alleles, A and a, it is common to say that  $p$  equals the frequency of the A allele and  $q$  equals the frequency of the a allele. In mathematics, **parameter** is another term for an idealized quantity like an allele frequency. It is assumed that parameters have an exact value. Put another way, parameters are idealized quantities where the messy, real-life details of how to measure the quantities they represent are completely ignored.

Empirical population genetics measures quantities such as allele frequencies to give **parameter estimates** by sampling and then measuring the alleles and genotypes present in actual populations. All experiments, observations, and even simulations in population genetics produce parameter estimates of some sort. There is a subtle notational convention used to indicate an estimate, the hat or  $\hat{}$  character above a variable. Estimates wear hats whereas parameters do not. Using allele frequency as an example, we would say  $\hat{p}$  (pronounced “ $p$  hat”) equals the number of A alleles sampled divided by the total number of alleles sampled. Intuitively, we can see from the denominator in the expression for  $\hat{p}$  that the allele frequency estimate will depend on the sample we gather to make the estimate.

In all populations a parameter has one true value. For the allele frequency  $p$ , knowing this true value would require examining the genotype of every individual and counting *all A* and *a* alleles to determine their frequency in the population. This task is impractical or impossible in most cases. Instead, we rely on an estimate of allele frequency,  $\hat{p}$ , obtained from a sample of individuals from the population. Sampling leads to some uncertainty in parameter estimates because repeating the sampling and parameter estimate process would likely lead to a somewhat different parameter estimate each time. Quantifying this uncertainty is important to determine whether repeated sampling might change a parameter estimate by just a little or change it by a lot. When dealing with parameters, we might expect that  $p + q = 1$  exactly if there are only two alleles with allele frequencies  $p$  and  $q$ . However, if we are dealing with estimates we might say the two allele frequency estimates should sum to approximately one ( $\hat{p} + \hat{q} \approx 1$ ) since each allele frequency is estimated with some error. The more uncertain the estimates of  $\hat{p}$  and  $\hat{q}$ , the less we should be surprised to find that their sum does not equal the expected value of one.

**Parameter** A variable or constant appearing in a mathematical expression; a value (usually unknown) used to represent a certain population characteristic; any factor that defines a system and determines or limits its performance.

**Estimate** An indication of the value of an unknown quantity based on observed data; an approximation of a true score, parameter, or value; a statistical estimate of the value of a parameter.

It could be said that statistics sits at the intersection of theoretical and empirical population genetics. Parameters and parameter estimates are fundamentally different things. Estimation requires effort to understand sampling variation and quantify sources of error and bias in samples and estimates. The distinction between parameters and estimates is critical when comparing actual populations with expectations to test hypotheses. When large, random samples can be taken, estimates are likely to have minimal error. However, there are many cases

where estimates have a great deal of uncertainty, which limits the ability to evaluate expectations. There are also instances where very different processes may produce very similar expected results. In such cases it may be difficult or impossible to distinguish the different potential causes of a pattern due to the approximate nature of estimates. While this book focuses mostly on parameters, it is useful to bear in mind that testing or comparing expectations requires the use of parameter estimates and statistics that quantify sampling error. The Appendix provides a review of some basic statistics that are used in the text.

### **Inductive and deductive reasoning**

Population genetics employs both **inductive** and **deductive reasoning** in an effort to understand the biological processes operating in actual populations as well as to elucidate the general processes that cause population genetic phenomena. The inductive approach to population genetics involves assembling measures of genetic variation (parameter estimates) from various populations to build up evidence that can be used to identify the underlying processes that produced the observed patterns. This approach is logically identical to that used by Isaac Newton, who used knowledge of how objects fall to the surface of the Earth as well as knowledge of the movement of planets to arrive at the general principles of gravity. Application of inductive reasoning requires detailed familiarity with the various empirical data types in population genetics, such as DNA sequences, along with the results of studies that report observed patterns of genetic variation. From this accumulated empirical information it is then possible to draw more general conclusions about the qualities and quantities of genetic variation in populations. Model organisms like *D. melanogaster* and *Arabidopsis thaliana* play a large role in population genetic conclusions reached by inductive reasoning. Because model organisms receive a large amount of scientific effort, to completely sequence their genomes for example, a great deal of available genetic data are accumulated for these species. Based on this evidence, many firm conclusions have been made about the population genetics of particular model species. Although model organisms provide very rich sources of empirical information, the number of species is limited by definition so that any generalizations may not apply universally to all species.

**Deductive reasoning** Using general principles to reach conclusions about specific instances.

**Inductive reasoning** Utilizing the knowledge of specific instances or cases to arrive at general principles.

The study of population genetics can also be approached using deductive reasoning. The actions of general processes such as genetic drift, mutation, and natural selection are represented by parameters in the mathematical equations that make up population genetic models. These models can then be used to make predictions about the quantity of genetic variation and patterns of genetic variation in space and time. Such population genetic models make general predictions about things like rates of change in allele frequency, the eventual equilibrium of allele or genotype frequencies, and the net outcome of several processes operating at the same time. These predictions are very general in that they apply to any population of any species since the predictions arose from general principles in the first place. At the same time, such general predictions may not be directly applicable to a specific population because the general principles and assumptions used to make the prediction are not specific enough to match an actual population.

Historically, the field of population genetics has developed from an interplay between arguments and evidence developed using both inductive and deductive reasoning approaches. Nonetheless, most of the major ideas in population genetics can be first approached with deductive reasoning by learning and understanding the expectations that arise from the principles of Mendelian heredity. This book stresses the process of deductive reasoning to arrive at these fundamental predictions. Empirical evidence related to expectations is included to illustrate predictions and also to demonstrate hypothesis tests that result from expectations. Because the body of empirical results in population genetics is very large, readers should resist the temptation to generalize too much from the limited number of empirical studies that are presented. Detailed reviews of particular areas of population genetics, many of which are cited in the Further reading sections at the end of each chapter, are a better source for comprehensive summaries of empirical studies.

In the next chapter we will start by building expectations for the frequencies of diploid genotypes based on the foundation of particulate inheritance: that alleles are passed unaltered from parents to offspring. There is ample support for particulate inheritance both from molecular biology, which identifies DNA as the hereditary molecule, and from allele and genotype frequencies that can be observed in actual populations. The general principle of particulate inheritance has been used to formulate a wide array of expectations about allele and genotype frequencies in populations.

## 1.2 Theory and assumptions

- What is a theory and what are assumptions?
- How can theories be useful with so many assumptions?

In colloquial usage, the word theory refers to something that is known with uncertainty, or a quantity that is approximate. On a day you are running late leaving work you might say, “In theory, I am supposed to be home at 6:00 pm.” In science, theory has a very different meaning. Theory is the accumulation of expectations and observations that have withstood tests and critical scrutiny and are accepted by at least some practitioners of a scientific field. Theory is the collection of all of the expectations developed for specific cases or individual biological processes that together form a more comprehensive set of general principles. The combination of Darwin’s hypothesis of natural selection with the laws of Mendelian particulate inheritance is often called the *modern synthesis* of evolutionary biology since it is a comprehensive theory to explain the causes of evolutionary change. The modern synthesis can offer causal explanations for biological phenomena ranging from antibiotic resistance in bacteria to the behavior of elephants to the rate of DNA sequence change as well as make predictions to guide animal and plant breeders. In all of the modern synthesis, population genetics plays a central role.

It is common for the uninitiated to ask the question “what good is theory if it is based on so many assumptions?” A body of theory is a useful tool to articulate assumptions and generate testable predictions. Theory that generates many testable predictions about the world also offers many opportunities to falsify its predictions and assumptions. Since hypotheses cannot be proven directly, but alternative

hypotheses can be disproven, the generation of plausible, testable alternative hypotheses is a requirement for scientific inquiry. Strong theories are able to make accurate predictions, offer causal explanations for diverse observations, and generate alternative hypotheses based on revised assumptions.

The words theory and assumption can seem abstract, but you should not be intimidated by them. Theories are just collections of expectations, each with a set of assumptions that place bounds on the prediction being made. If you understand what motivates an expectation, its predictions, and its assumptions, then you understand theory. Most expectations in population genetics will have at least a few, and often many, assumptions used to define and bound the situation. For example, we might assume something about the size of a population or the absence of mutation, or that all genotypes are diploid with two alleles. This is a way of limiting the prediction to appropriate circumstances and also a way of defining which quantities and conditions can vary and which are fixed. Each of these assumptions can influence the generality of an expectation. Each assumption can also be relaxed or altered to see how strongly it influences the expectation. To return to the example in the last section, if one day meteorites were falling around us with regularity we would be forced to call into question some of the basic assumptions originally used to formulate our expectation that meteorite strikes should be rare events. In this way, assumptions are useful tools to ask “what if . . . ?” as part of the process of developing a prediction. If our initial “what if . . . ?” conditions are badly off the mark, then the resulting prediction will probably also be poor.

In population genetics, as in much of science where theory and expectations are involved, empirical data and model expectations are routinely compared. Imagine observing a set of genotype frequencies in a biological population. It would then be natural to construct an idealized population using theory that approximates the biological population. This is an attempt to construct an idealized population that is *equivalent* to the actual population from the perspective of the processes influencing genotype frequencies. For example, a large population may behave exactly like a small, randomly mating ideal population in terms of genotype frequencies. This equivalence allows us to use expectations for ideal populations with one or a few variables specified in order to describe an actual population where there are many more, usually unknown, parameters.

What we strive to do is to focus on those variables that strongly influence genotype frequencies in the actual population. In this way it is often possible to reduce the complexity of a real population and determine the key variables that strongly influence a property like genotype frequencies. The ideal population is not meant to match the actual population in every detail.

**Theory** A scheme or system of ideas or statements held as an explanation or account of a group of facts or phenomena; the general laws, principles, or causes of something known or observed.

**Infer** To draw a conclusion or make a deduction based on facts or indications; to have as a logical consequence.

From the comparison of expectation and observation, we infer that the first principles used to construct the expectation are sound if they can be used to explain patterns observed in the biological world. For example, before we had detailed knowledge of the processes involved in plant and animal reproduction, organisms were thought to be produced by so-called spontaneous generation from non-living materials. Eventually, controlled experiments revealed that the expectations of spontaneous generation were not met. The emergence of flies from pieces of rotting meat was taken as proof of spontaneous generation (a phenomenon that by itself turned out to be consistent with expectations from two incompatible theories). However, when meat was placed in screened containers and allowed to rot, no flies emerged. This latter evidence was inconsistent with the expectations of spontaneous generation and indicated that some of the assumptions behind the theory were not accurate. It was Louis Pasteur in 1859 who used heat sterilization to show conclusively that the expectations of spontaneous generation were not met.

If observations match what is expected from theory in an ideal population, we infer that the observed pattern may be caused by the processes represented in the ideal population. However, there is a major distinction between considering an actual and an idealized population *equivalent* and considering them *identical*. This is seen in cases where the observed pattern in an actual population is consistent with the

expectations from several model populations built around distinct and incompatible assumptions. In such cases, it is not possible to infer the processes that cause a given pattern without additional information. A common example in population genetics are cases of genetic patterns that are potentially consistent with the random process of genetic drift and at the same time consistent with some form of the deterministic process of natural selection. In such cases unambiguous inference of the underlying cause of a pattern is not possible without additional empirical information or more precise expectations.

### 1.3 Simulation

- A method of practice, trial and error learning, and exploration.

Imagine learning to play the piano without ever touching a piano or practicing the hand movements required to play. What if you were expected to play a difficult concerto after extensive exposure (perhaps a semester) to only verbal and written descriptions of how other people play? Such a teaching style would make learning to play the piano very difficult because there would be no opportunity for practice, trial and error, or exploration. You would not have the opportunity for direct experience nor incremental improvement of your understanding. Unfortunately, this is exactly how science courses are taught to some degree. You are expected to learn and remember concepts with only limited opportunity for directly observing principles in action. In fairness, this is partly due to the difficulty of carrying out some of the experiments or observations that originally lead someone to discover and understand an important principle.

In the field of population genetics computer simulations can be used to effectively demonstrate many fundamental genetic processes. In fact, computer simulations are an important research tool in population genetics. Therefore, when you conduct simulations you are both learning by direct experience and learning using the same methods that are used by researchers. Simulations allow us to view how quantities like allele frequencies change over time, observe their dynamics, and determine whether a stable end point is reached: an equilibrium. With simulations we can view dynamics (change over time) and equilibria over very long periods of time and under a vast array of conditions in an effort to

reach general conclusions. Without simulations, it would be impossible for us to directly observe allele frequencies over such long periods of time and in such diverse biological situations.

Simulations are an effective means to understand some of the fundamental predictions of population genetics. Mathematical expressions are frequently used to express dynamics and equilibria in population genetics, but the equations alone can be opaque at first. Simulations provide a means to explore the relationships among variables that are summarized in the compact language of mathematics. Many people feel that a set of mathematical equations is much more meaningful after having the chance to explore what they describe with some actual numerical values. Simulation provides the means to explore what equations predict and can make learning population genetics an easier, more rewarding experience.

Carrying out simulations has the potential to make the expectations of population genetics much more accessible and understandable. Conducting simulations is not much extra work, especially once you get into the practice of using the text and simulation software in concert. You can approach simulations as if they are games, where each one shows a visual scene that helps to solve a puzzle. In addition, simulations can help you develop a more intuitive understanding of population genetic predictions so you do not have to approach the expectations of population genetics as disembodied or unanimated “facts.”

It is important to approach simulations in a systematic and organized fashion, not as just a collection of buttons to press and text entry boxes to be filled in on a whim. It is absolutely imperative that you understand the meaning behind each variable that you can control as well as the meaning of the results you obtain. To do so successfully you will need to be aware of both specific details and larger patterns, or both individual trees and the forest that they compose. For example, in a simulation that presents results as a graph, it is important that you understand the details of what variables are represented on each axis and the range of axis values. Sometimes these details are not always completely obvious in simulation software, requiring you to use both your intuition and knowledge of the population genetic processes being simulated.

Once you are comfortable with the details of a simulation, you will also want to keep track of

the “big picture” patterns that emerge as you view simulation results. Seeing these patterns will often require that you examine the results over a range of conditions. Try approaching simulations as experiments by changing only one variable at a time until you understand its effects on the outcome. Changing several things all at once can lead to confusion and an inability to see cause-and-effect relationships, unless you have fully understood the effects of individual variables. Finally, try writing down parameter values you have tried in a simulation and sketching or tabulating results on paper as you work with a simulation. Use all of your skills as a scientist and student when conducting simulations and they will become a powerful learning tool. Eventually, you may even use scripting and programming to carry out your own simulations specifically designed to explore your own genetic hypotheses.

### Interact box 1.1 The textbook website

Throughout this book you will encounter Interact boxes. These boxes contain opportunities for you to interact directly with the material in the text using computer simulations designed to demonstrate fundamental concepts of population genetics. Each box will contain step-by-step instructions for you to follow in order to carry out a simulation. By following the instructions you will get started with the simulation. However, always feel free to use your own imagination and intuition. After following the instructions in the Interact box and understanding the point at hand, enter different values, push more buttons, and even read the documentation. You can also return to Interact boxes at a later time, perhaps after you have read and understood more of the text, to reconsider a simulation or view it in a different light. You can also use the simulations to answer questions that may occur to you, or to test hypotheses that you may have. Questions in population genetics that start off “What would happen if . . . ?” are often begging to be answered with simulation.

Using Interact boxes will require that you are in front of a computer with a connection to the internet. You will also need user privileges to download and install programs in some cases. Some simulation programs have versions for multiple operating systems (e.g. Windows and Macintosh) whereas others can be used on only one operating system. It might be a good idea to think of locations now where you have access to computers with various operating systems. Finally, bear in mind that the simulation programs have all been donated to the scientific community by their respective authors and were often written in the author’s spare time. Don’t be surprised if the programs have a few rough edges or even bugs: focus on the population genetics concepts and remember that someone devoted their time to help you learn.

You will begin many of the Interact boxes by connecting to this textbook’s website, whereas for others you will use a program downloaded in an earlier Interact box. The worldwide web address (URL) for each of the simulation programs will be given on the textbook website rather than in the text itself. This prevents problems if web addresses change because the textbook website can be updated while your copy of the text cannot.

- Step 1 Open a web browser and enter <http://www.wiley.com/go/hamiltongenetics>.
- Step 2 If you are working on a computer you use regularly, bookmark the text website so you can reach it easily in the future.
- Step 3 Click on the link to Interact boxes.
- Step 4 Verify that the page gives links for each of the Interact boxes listed by their number. You could also bookmark this page so you can reach it directly in the future.

Congratulations! You have completed the first Interact box.

## Chapter 1 review

- General principles and direct measurements taken in actual populations combine to form comprehensive expectations about amounts, patterns, and cause-and-effect relationships in population genetics.
- The theory of population genetics is the collection of well-accepted expectations used to articulate a wide array of predictions about the biological processes that shape genetic variation.
- Parameters are idealized quantities that are exact while parameter estimates wear notational “hats” to remind us that they have statistical uncertainty.
- Population genetics uses both inductive reasoning to generalize from knowledge of specifics and deductive reasoning to build up predictions from general principles that can be applied to specific situations.
- Population genetics is not a spectator sport! Direct participation in computer simulation provides the opportunity to see population genetic processes in action. You can learn by trial and error and test your own understanding by making predictions and then comparing them with simulation results.

## Further reading

For a history of population genetics from Darwin to the 1930s see:

Provine WB. 1971. *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago, IL.

You don't have to be a computer programmer to create simulations. Check out this book for step-by-step instructions to construct simulations using the spreadsheet application Microsoft Excel:

Donovan TM and Welden CW. 2002. *Spreadsheet Exercises in Ecology and Evolution*. Sinauer Associates, Sunderland, MA.

For two personal and historical essays on the past, present, and assumptions of theoretical population genetics see:

Lewontin RC. 1985. Population genetics. In Greenwood PJ, Harvey PH, and Slatkin M (eds), *Evolution: Essays in Honour of John Maynard Smith*, pp. 3–18. Cambridge University Press, Cambridge.

Wakeley J. 2005. The limits of theoretical population genetics. *Genetics* 169: 1–7.

## CHAPTER 2

# Genotype frequencies

### 2.1 Mendel's model of particulate genetics

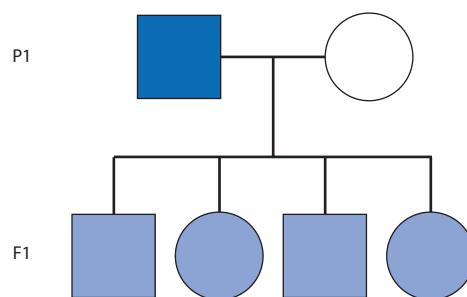
- Mendel's breeding experiments.
- Independent assortment of alleles.
- Independent segregation of loci.
- Some common genetic terminology.

In the nineteenth century there were several theories of heredity, including inheritance of acquired characteristics and blending inheritance. Jean-Baptiste Lamarck is most commonly associated with the discredited hypothesis of inheritance of acquired characteristics (although it is important to recognize his efforts in seeking general causal explanations of evolutionary change). He argued that individuals contain “nervous fluid” and that organs or features (phenotypes) employed or exercised more frequently attract more nervous fluid, causing the trait to become more developed in offspring. His widely known example is the long neck of the giraffe, which he said developed because individuals continually stretched to reach leaves at the tops of trees. Later, Charles Darwin and many of his contemporaries subscribed to the idea of blending inheritance. Under blending inheritance, offspring display phenotypes that are an intermediate combination of parental phenotypes (Fig. 2.1).

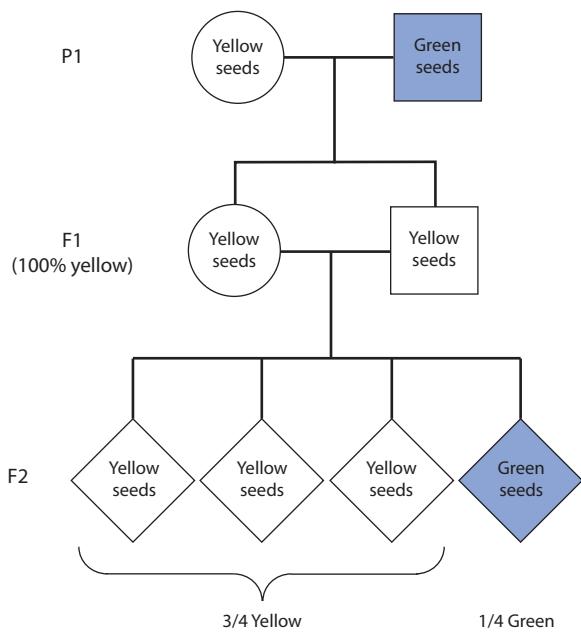
From 1856 to 1863, the Augustinian monk Gregor Mendel carried out experiments with pea plants that demonstrated the concept of particulate inheritance. Mendel showed that phenotypes are determined by discrete units that are inherited intact and unchanged through generations. His hypothesis was sufficient to explain three common observations: (i) phenotype is sometimes identical between parents and offspring; (ii) offspring phenotype can differ from that of the parents; and (iii) “pure” phenotypes of earlier generations could skip generations and reappear in later generations. Neither blending inheritance nor inheritance of acquired characteristics are satisfactory explanations for all of these observations. It

is hard for us to fully appreciate now, but Mendel's results were truly revolutionary and served as the very foundation of population genetics. The lack of an accurate mechanistic model of heredity severely constrained biological explanations of cause and effect up to the point that Mendel's results were “rediscovered” in the year 1900.

It is worthwhile to briefly review the experiments with pea plants that Mendel used to demonstrate independent assortment of both alleles within a locus and of multiple loci, sometimes dubbed Mendel's first and second laws. We need to remember that this was well before the Punnett square (named after Reginald C. Punnett), which originated in about 1905. Therefore, the conceptual tool we would use now to predict progeny genotypes from parental genotypes was a thing of the future. So in revisiting Mendel's experiments we will not use the Punnett square in an attempt to follow his logic. Mendel only observed the phenotypes of generations of pea plants that he had hand-pollinated. From these phenotypes and their patterns of inheritance he inferred the



**Figure 2.1** The model of blending inheritance predicts that progeny have phenotypes that are the intermediate of their parents. Here “pure” blue and white parents yield light blue progeny, but these intermediate progeny could never themselves be parents of progeny with pure blue or white phenotypes identical to those in the P1 generation. Crossing any shade of blue with a pure white or blue phenotype would always lead to some intermediate shade of blue. By convention, in pedigrees females are indicated by circles and males by squares, whereas P refers to parental and F to filial.



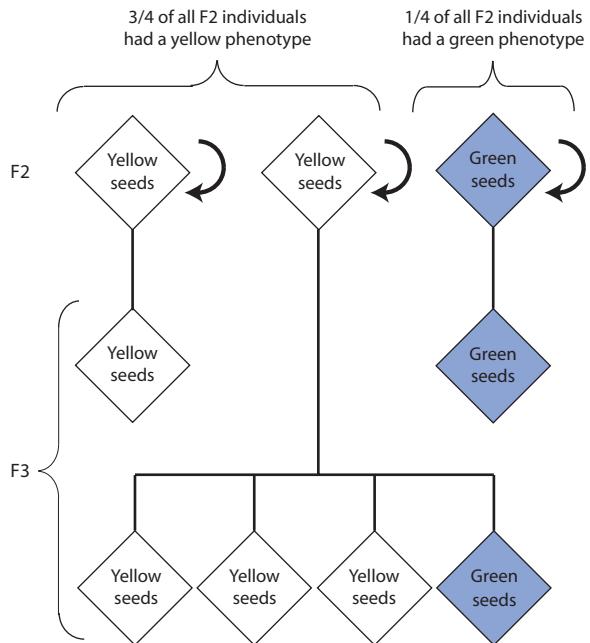
**Figure 2.2** Mendel's crosses to examine the segregation ratio in the seed coat color of pea plants. The parental plants (P1 generation) were pure breeding, meaning that if self-fertilized all resulting progeny had a phenotype identical to the parent. Some individuals are represented by diamonds since pea plants are hermaphrodites and can act as a mother, a father, or can self-fertilize.

existence of heritable factors. His experiments were actually both logical and clever, but are now taken for granted since the basic mechanism of particulate inheritance has long since ceased to be an open question. It was Mendel who established the first and most fundamental prediction of population genetics: expected genotype frequencies.

Mendel used pea seed coat color as a phenotype he could track across generations. His goal was to determine, if possible, the general rules governing inheritance of pea phenotypes. He established “pure”-breeding lines (meaning plants that always produced progeny with phenotypes like themselves) of peas with both yellow and green seeds. Using these pure-breeding lines as parents, he crossed a yellow- and a green-seeded plant. The parental cross and the next two generations of progeny are shown in Fig. 2.2. Mendel recognized that the F1 plants had an “impure” phenotype because of the F2 generation plants, of which three-quarters had yellow and one-quarter had green seed coats.

His insightful next step was to self-pollinate a sample of the plants from the F2 generation (Fig. 2.3). He considered the F2 individuals with yellow and green seed coats separately. All green-seeded F2 plants produced green progeny and thus were “pure”

green. However, the yellow-seeded F2 plants were of two kinds. Considering just the yellow F2 seeds, one-third were pure and produced only yellow-seeded progeny whereas two-thirds were “impure” yellow since they produced both yellow- and green-seeded progeny. Mendel combined the frequencies of the F2 yellow and green phenotypes along with the frequencies of the F3 progeny. He reasoned that three-quarters of all F2 plants had yellow seeds but these could be divided into plants that produced pure yellow F3 progeny (one-third) and plants that produced both yellow and green F3 progeny (two-thirds). So the ratio of pure yellow to impure yellow in the F2 was  $(1/3 \times 3/4) = 1/4$  pure yellow to  $(2/3 \times 3/4) = 1/2$  “impure” yellow. The green-seeded progeny comprised one-quarter of the F2 generation and all produced green-seeded progeny when self-fertilized, so that  $(1 \times 1/4 \text{ green}) = 1/4$  pure green. In total, the ratios of phenotypes in the F2 generation were 1 pure yellow : 2 impure yellow : 1 pure green or 1 : 2 : 1. Mendel reasoned that “the ratio of 3 : 1 in which the distribution of the dominating and recessive traits take place in the first generation therefore resolves itself into the ratio of 1 : 2 : 1 if one differentiates the



**Figure 2.3** Mendel self-pollinated (indicated by curved arrows) the F2 progeny produced by the cross shown in Figure 2.2. Of the F2 progeny that had a yellow phenotype (three-quarters of the total), one-third produced all progeny with a yellow phenotype and two-thirds produced progeny with a 3 : 1 ratio of yellow and green progeny (or three-quarters yellow progeny). Individuals are represented by diamonds since pea plants are hermaphrodites.

meaning of the dominating trait as a hybrid and as a parental trait" (quoted in Orel 1996). During his work, Mendel employed the terms "dominating" (which became dominant) and "recessive" to describe the manifestation of traits in impure or heterozygous individuals.

With the benefit of modern symbols of particulate heredity, we could diagram Mendel's monohybrid cross with pea color in the following way.

P1	Phenotype	Yellow × green
	Genotype	GG      gg
	Gametes produced	G      g
F1	Phenotype	All "impure" yellow
	Genotype	Gg
	Gametes produced	G, g

A Punnett square could be used to predict the phenotypic ratios of the F2 plants:

	G	g
G	GG	Gg
g	Gg	gg

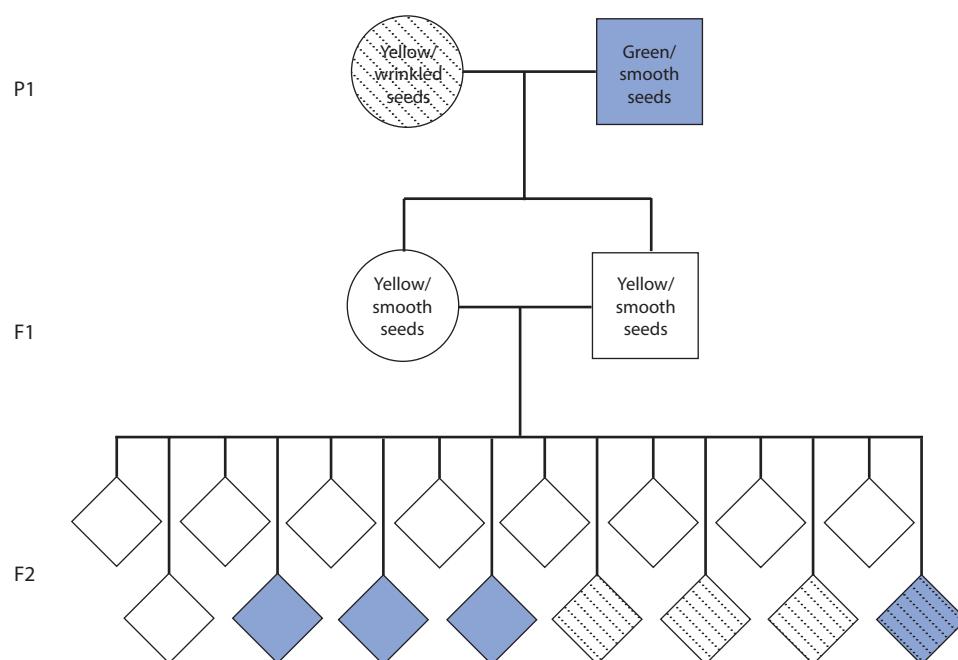
F2	Phenotype	3 Yellow : 1 green
	Genotype	GG      Gg      gg
	Gametes produced	G      G, g      g

and another Punnett square could be used to predict the genotypic ratios of the two-thirds of the yellow F2 plants:

	G	g
G	GG	Gg
g	Gg	gg

**Mendel's first "law"** Predicts independent segregation of alleles at a single locus: two members of a gene pair (alleles) segregate separately into gametes so that half of the gametes carry one allele and the other half carry the other allele.

Individual pea plants obviously have more than a single phenotype and Mendel followed the inheritance of other characters in addition to seed coat color. In one example of his crossing experiments, Mendel tracked the simultaneous inheritance of both seed coat color and seed surface condition (either wrinkled ("angular") or smooth). He constructed an initial cross among pure-breeding lines identical to what he had done when tracking seed color inheritance, except now there were two phenotypes (Fig. 2.4).



**Figure 2.4** Mendel's crosses to examine the segregation ratios of two phenotypes, seed coat color (yellow or green) and seed coat surface (smooth or wrinkled), in pea plants. The hatched pattern indicates wrinkled seeds while white indicates smooth seeds. The F2 individuals exhibited a phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow : 1 wrinkled/green.

The F<sub>2</sub> progeny appeared in the phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow: 1 wrinkled/green.

How did Mendel go from this F<sub>2</sub> phenotypic ratio to the second law? He ignored the wrinkled/smooth phenotype and just considered the yellow/green seed color phenotype in self-pollination crosses of F<sub>2</sub> plants just like those for the first law. In the F<sub>2</sub> progeny, 12/16 or three-quarters had a yellow seed coat and 4/16 or one-quarter had a green seed coat, or a 3 yellow : 1 green phenotypic ratio. Again using self-pollination of F<sub>2</sub> plants like those in Fig. 2.3, he showed that the yellow phenotypes were one-quarter pure and one-half impure yellow. Thus, the segregation ratio for seed color was 1 : 2 : 1 and the wrinkled/smooth phenotype did not alter this result. Mendel obtained an identical result when focusing instead on the wrinkled/smooth phenotype and ignoring the seed color phenotype.

Mendel concluded that a phenotypic segregation ratio of 9 : 3 : 3 : 1 is the same as combining two independent 3 : 1 segregation ratios of two phenotypes since  $(3 : 1) \times (3 : 1) = 9 : 3 : 3 : 1$ . Similarly, multiplication of two (1 : 2 : 1) phenotypic ratios will predict the two phenotype ratio  $(1 : 2 : 1) \times (1 : 2 : 1) = 1 : 2 : 1 : 2 : 4 : 2 : 1 : 2 : 1$ . We now recognize that dominance in the first two phenotype ratios masks the ability to distinguish some of the homozygous and heterozygous genotypes, whereas the ratio in the second case would result if there was no dominance. You can confirm these conclusions by working out a Punnett square for the F<sub>2</sub> progeny in the two-locus case.

**Mendel's second "law"** Predicts independent assortment of multiple loci: during gamete formation, the segregation of alleles of one gene is independent of the segregation of alleles of another gene.

Mendel performed similar breeding experiments with numerous other pea phenotypes and obtained similar results. Mendel described his work with peas and other plants in lectures and published it in 1866 in the *Proceedings of the Natural Science Society of Brünn* in German. It went unnoticed for nearly 35 years. However, Mendel's results were eventually recognized and his paper was translated into several languages. Mendel's rediscovered hypothesis of particulate

inheritance was also bolstered by evidence from microscopic observations of cell division that led Walter Sutton and Theodor Boveri to propose the chromosome theory of heredity in 1902.

Much of the currently used terminology was coined as the field of particulate genetics initially developed. Therefore, many of the critical terms in genetics have remained in use for long periods of time. However, the meanings and connotations of these terms have often changed as our understanding of genetics has also changed.

Unfortunately, this has lead to a situation where words can sometimes mislead. A common example is equating *gene* and *allele*. For example, it is commonplace for news media to report scientific breakthroughs where a "gene" has been identified as causing a particular phenotype, often a debilitating disease. Very often what is meant in these cases is that an *allele* with the phenotypic effect has been identified. Both unaffected and affected individuals all possess the gene, but they differ in their alleles and therefore in their genotype. If individuals of the same species really differed in their gene content (or loci they possessed), that would provide evidence of additions or deletions to genomes. For an interesting discussion of how terminology in genetics has changed – and some of the misunderstandings this can cause – see Judson (2001).

**Gene** Unit of particulate inheritance; in contemporary usage usually means an exon or series of exons, or a DNA sequence that codes for an RNA or protein.

**Locus** (plural *loci*, pronounced "low-sigh") Literally "place" or location in the genome; in contemporary usage is the most general reference to *any* sequence or genomic region, including non-coding regions.

**Allele** Variant or alternative form of the DNA sequence at a given locus.

**Genotype** The set of alleles possessed by an individual at one locus; the genetic composition of an individual at one or many loci.

**Phenotype** The morphological, biochemical, physiological, and behavioral attributes of an individual; synonymous with character, trait.

**Dominant** Where the expressed phenotype of one allele takes precedence over the expressed phenotype of another allele. The allele associated with the expressed phenotype is said to be dominant. Dominance is seen on a continuous scale that ranges between “complete” dominance (one allele completely masks the phenotype of another allele so that the phenotype of a heterozygote is identical to a homozygote for the dominant allele), “partial,” or “incomplete” dominance (masking effect is incomplete so that the phenotype of a heterozygote is intermediate to both homozygotes) and includes over- and under-dominance (phenotype is outside the range of phenotypes seen in the homozygous genotypes). The lack of dominance (heterozygote is exactly intermediate to phenotypes of both homozygotes) is sometimes termed “codominance” or “semi-dominance.”

**Recessive** The expressed phenotype of one allele is masked by the expressed phenotype of another allele. The allele associated with the concealed phenotype is said to be recessive.

(infrequent) alleles would disappear from populations over time. Godfrey H. Hardy (1908) and Wilhelm Weinberg (1908) worked independently to show that the laws of Mendelian heredity did not predict such a phenomenon (see Crow 1988). In 1908 they both formulated the relationship that can be used to predict allele frequencies given genotype frequencies or predict genotype frequencies given allele frequencies. This relationship is the well-known Hardy–Weinberg equation

$$p^2 + 2pq + q^2 = 1 \quad (2.1)$$

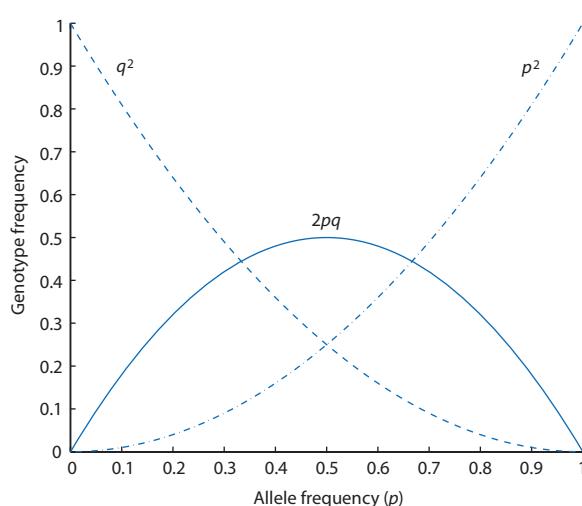
where  $p$  and  $q$  are allele frequencies for a genetic locus with two alleles.

Genotype frequencies predicted by the Hardy–Weinberg equation can be summarized graphically. Figure 2.5 shows Hardy–Weinberg expected genotype frequencies on the  $y$  axis for each genotype for any given value of the allele frequency on the  $x$  axis. Another graphical tool to depict genotype and allele frequencies simultaneously for a single locus with two alleles is the De Finetti diagram (Fig. 2.6). As we will see, De Finetti diagrams are helpful when examining how population genetic processes dictate allele and genotype frequencies. In both graphs it is apparent that heterozygotes are most frequent when the frequency of the two alleles is equal to 0.5. You can also see that when an allele is rare, the corresponding homozygote genotype is even rarer since the genotype frequency is the square of the allele frequency.

## 2.2 Hardy–Weinberg expected genotype frequencies

- Hardy–Weinberg and its assumptions.
- Each assumption is a population genetic process.
- Hardy–Weinberg is a null model.
- Hardy–Weinberg in haplo-diploid systems.

Mendel’s “laws” could be called the original expectations in population genetics. With the concept of particulate genetics established, it was possible to make a wide array of predictions about genotype and allele frequencies as well as the frequency of phenotypes with a one-locus basis. Still, progress and insight into particulate genetics was gradual. Until 1914 it was generally believed that rare



**Figure 2.5** Hardy–Weinberg expected genotype frequencies for AA, Aa, and aa genotypes ( $y$  axis) for any given value of the allele frequency ( $x$  axis). Note that the value of the allele frequency not graphed can be determined by  $q = 1 - p$ .

### Interact box 2.1 Genotype frequencies

PopGene.S<sup>2</sup> (short for Population genetics simulation software) is a population genetics simulation program that will be featured in several Interact boxes. Here we will use PopGene.S<sup>2</sup> to explore interactive versions of Figs 2.5 and 2.6. Using the program will require that you download it from a website and install it on a computer running Windows. Simulations that can be explored with PopGene.S<sup>2</sup> will be featured in Interact boxes throughout this book.

Find Interact box 2.1 on the text web page and click on the link for **PopGene.S<sup>2</sup>**. The PopGene.S<sup>2</sup> website has download and installation instructions (and lists computer operating system requirements). But don't worry: the program is small, runs on most Windows computers, and is simple to install. After you have PopGene.S<sup>2</sup> installed according to the instructions provided in the PopGene.S<sup>2</sup> website, move on to Step 1 to begin the simulation.

**Step 1** Open PopGene.S<sup>2</sup> and click once on the information box to make it disappear. Click on the **Allele and Genotype Frequencies** menu and select **Genotype frequencies**. A new window will open that contains a picture like Fig. 2.5 and some fields where you can enter genotype frequency values. Enter 0.25 into  $P(AA)$  to specify the frequency of the AA genotype. After entering each value the program will update the  $p$  and  $q$  values (the allele frequencies). Now enter 0.5 into  $P(Aa)$ . Once two genotype frequencies are entered the third genotype frequency is determined and the program will display the value. Click the **OK** button. A dot will appear on the graph corresponding to the value of  $P(Aa)$  and the frequency of the a allele. Changing the  $P(AA)$  and  $P(Aa)$  frequencies to different values and then clicking OK again will add a new dot. Try several values for the genotype frequencies at different allele frequencies, both in and out of Hardy–Weinberg expected genotype frequencies.

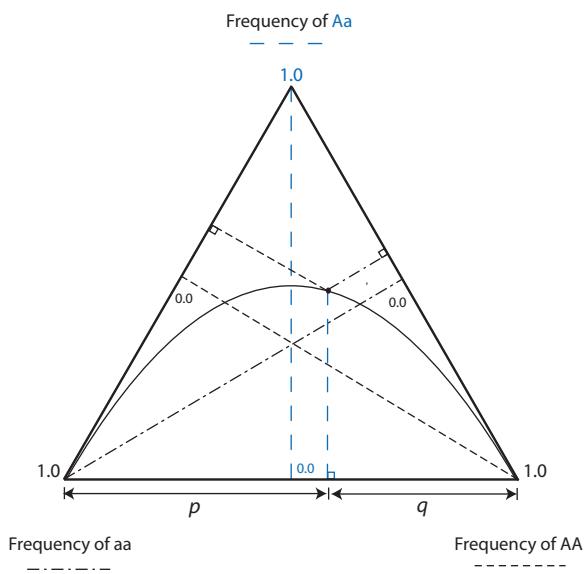
**Step 2** Leave the **Genotype frequencies** window open but move it to one side to make room for another window. Now click on the **Mating Models** menu and select **Autosomal locus**. A new window will open that contains a triangular graph like that in Fig. 2.6. To display a set of genotype frequencies, enter the frequencies for  $P(AA)$  and  $P(Aa)$  in the text boxes and click **OK**. (Frequency of the aa genotype or  $P(aa)$  is calculated since the three genotype frequencies must sum to one.) The point on the De Finetti diagram representing the user-entered genotype frequencies will be plotted as a red square and displayed under the heading of "Initial frequencies". Hardy–Weinberg expected genotype frequencies are plotted as a blue square based on the allele frequencies that correspond to the user-entered genotype frequencies. These Hardy–Weinberg expected genotype frequencies are displayed under the heading of "Panmixia frequencies". Try several sets of different genotype frequencies to see values for genotype frequencies at different allele frequencies, both in and out of Hardy–Weinberg expected genotype frequencies.

**Step 3** Compare the two ways of visualizing genotype frequencies by plotting identical genotype frequencies in each window.

A single generation of reproduction where a set of conditions, or assumptions, are met will result in a population that meets Hardy–Weinberg expected genotype frequencies, often called Hardy–Weinberg equilibrium. The list of assumptions associated with this prediction for genotype frequencies is long. The set of assumptions includes:

- the organism is diploid,
- reproduction is sexual (as opposed to clonal),
- generations are discrete and non-overlapping,

- the locus under consideration has two alleles,
- allele frequencies are identical among all mating types (i.e. sexes),
- mating is random (as opposed to assortative),
- there is random union of gametes,
- population size is very large, effectively infinite,
- migration is negligible (no population structure, no gene flow),
- mutation does not occur or its rate is very low,
- natural selection does not act (all individuals and gametes have equal fitness).



**Figure 2.6** A De Finetti diagram for one locus with two alleles. The triangular coordinate system results from the requirement that the frequencies of all three genotypes must sum to one. Any point inside or on the edge of the triangle represents all three genotype frequencies of a population. The parabola describes Hardy–Weinberg expected genotype frequencies. The dashed lines represent the frequencies of each of the three genotypes between zero and one. Genotype frequencies at any point can be determined by the length of lines that are perpendicular to each of the sides of the triangle. A practical way to estimate genotype frequencies on the diagram is to hold a ruler parallel to one of the sides of the triangle and mark off the distance on one of the frequency axes. The point on the parabola is a population in Hardy–Weinberg equilibrium where the frequency of AA is 0.36, the frequency of aa is 0.16, and the frequency of Aa is 0.48. The perpendicular line to the base of the triangle also divides the bases into regions corresponding in length to the allele frequencies. Any population with genotype frequencies not on the parabola has an excess (above the parabola) or deficit (below the parabola) of heterozygotes compared to Hardy–Weinberg expected genotype frequencies.

These assumptions make intuitive sense when each is examined in detail (although this will probably be more apparent after more reading and simulation). As we will see later, Hardy–Weinberg holds for any number of alleles, although equation 2.1 is valid for only two alleles. Many of the assumptions can be thought of as assuring random mating and production of all possible progeny genotypes. Hardy–Weinberg genotype frequencies in progeny would not be realized if the two sexes have different allele frequencies even if matings take place between random pairs of parents. It is also possible that just by chance not all genotypes would be produced if only a small number of parents mated, just like flipping

a fair coin only a few times may not produce an equal number of heads and tails. Natural selection is a process that causes some genotypes in either the parental or progeny generations to be more frequent than others. So it is logical that Hardy–Weinberg expectations would not be met if natural selection were acting. In a sense, these assumptions define the biological processes that make up the field of population genetics. Each assumption represents one of the conceptual areas where population genetics can make testable predictions via expectation in order to distinguish the biological processes operating in populations. This is quite a set of accomplishments for an equation with just three terms!

Despite all of this praise, you might ask: what good is a model with so many restrictive assumptions? Are all these assumptions likely to be met in actual populations? The Hardy–Weinberg model is not necessarily meant to be an exact description of any actual population, although actual populations often exhibit genotype frequencies predicted by Hardy–Weinberg. Hardy–Weinberg provides a **null model**, a prediction based on a simplified or idealized situation where no biological processes are acting and genotype frequencies are the result of random combination. Actual populations can be compared with this null model to test hypotheses about the evolutionary forces acting on allele and genotype frequencies. The important point and the original motivation for Hardy and Weinberg was to show that the process of particulate inheritance itself does not cause any changes in allele frequencies across generations. Thus, changes in allele frequency or departures from Hardy–Weinberg expected genotype frequencies must be caused by processes that alter the outcome of basic inheritance.

**Null model** A testable model of no effect. A prediction or expectation based on the simplest assumptions to predict outcomes. Often, null models make predictions based on purely random processes, random samples, or variables having no effect on an outcome.

In the final part of this section we will explore genotype frequency expectations adjusted to account for ploidy (number of homologous chromosomes)

differences between males and females as seen in chromosomal sex determination and haplo-diploid organisms. In chromosomal sex determination as seen in mammals, birds, and Lepidoptera (butterflies and moths), one sex is determined by possession of two identical chromosomes (the homogametic sex) and the other sex determined by possession of two different chromosomes (the heterogametic sex). In mammals females are homogametic (XX) and males heterogametic (XY), whereas in birds and Lepidoptera the opposite is true, with heterogametic females (ZW) and homogametic males (ZZ). In haplo-diploid species such as bees and wasps (Hymenoptera), males are haploid (hemizygous) for all chromosomes whereas females are diploid for all chromosomes.

Predicting genotype frequencies at one locus in these cases under random mating and the other assumptions of Hardy–Weinberg requires keeping track of allele or genotype frequencies in both sexes and loci on specific chromosomes. An effective method is to draw a Punnett square that distinguishes the sex of an individual as well as the gamete types

that can be generated at mating (Table 2.1). The Punnett square shows that genotype frequencies in the diploid sex are identical to Hardy–Weinberg expectations for autosomes, whereas genotype frequencies are equivalent to allele frequencies in the haploid sex. One consequence of different chromosome types between the sexes is that fully recessive phenotypes are more common in the heterogametic sex, where a single chromosome determines the phenotype and recessive phenotypes appear at the allele frequency. However, in the homogametic sex, fully recessive phenotypes appear at the frequency of the recessive genotype (e.g.  $q^2$ ) since they are masked in heterozygotes. Some types of color blindness in humans are examples of traits due to genes on the X chromosome (called “X-linked” traits) that are more common in men than in women due to haplo-diploid inheritance.

Later, in section 2.4, we will examine two categories of applications of Hardy–Weinberg expected genotype frequencies. The first set of applications arises when we assume (often with supporting evidence)

**Table 2.1** Punnett square to predict genotype frequencies for loci on sex chromosomes and for all loci in males and females of haplo-diploid species. Notation in this table is based on birds where the sex chromosomes are Z and W (ZZ males and ZW females) with a diallelic locus on the Z chromosome possessing alleles A and a at frequencies  $p$  and  $q$ , respectively. In general, genotype frequencies in the homogametic or diploid sex are identical to Hardy–Weinberg expectations for autosomes, whereas genotype frequencies are equal to allele frequencies in the heterogametic or haploid sex.

Hemizygous or haploid sex			Diploid sex		
Genotype	Gamete	Frequency	Genotype	Gamete	Frequency
ZW	Z-A Z-a W	$p$ $q$	ZZ	Z-A Z-a	$p$ $q$
Expected genotype frequencies under random mating					
Homogametic sex Z-A Z-A $p^2$ Z-A Z-a $2pq$ Z-a Z-a $q^2$ Heterogametic sex Z-A W $p$ Z-a W $q$					

that the assumptions of Hardy–Weinberg are true. We can then compare several expectations for genotype frequencies with actual genotype frequencies to distinguish between several alternative hypotheses. The second type of application is where we examine what results when assumptions of Hardy–Weinberg are not met. There are many cases where population genotype frequencies can be used to reveal the action of various population genetic processes. Before that, the next section builds a proof of the Hardy–Weinberg prediction that inheritance per se will not alter allele frequencies.

### 2.3 Why does Hardy–Weinberg work?

- A simple proof of Hardy–Weinberg.
- Hardy–Weinberg with more than two alleles.

The Hardy–Weinberg equation is one of the most basic expectations we have in population genetics. It is very likely that you were already familiar with the Hardy–Weinberg equation before you picked up this book. But where does Hardy–Weinberg actually come from? What is the logic behind it? Let's develop a simple proof that Hardy–Weinberg is actually true. This will also be our first real foray into the type of algebraic argument that much of population genetics is built on. Given that you start out knowing the conclusion of the Hardy–Weinberg tale, this gives you the opportunity to focus on the style in which it is told. Algebraic or quantitative arguments are a central part of the language and vocabulary of population genetics, so part of the task of learning population genetics is becoming accustomed to this mode of discourse.

We would like to prove that  $p^2 + 2pq + q^2 = 1$  accurately predicts genotype frequencies given the values of allele frequencies. Let's start off by making some explicit assumptions to bound the problem. The assumptions, in no particular order, are:

- 1 mating is random (parents meet and mate according to their frequencies);
- 2 all parents have the same number of offspring (equivalent to no natural selection on fecundity);
- 3 all progeny are equally fit (equivalent to no natural selection on viability);
- 4 there is no mutation that could act to change an A to a or an a to A;
- 5 it is a single population that is very large;
- 6 there are two and only two mating types.

Now let's define the variables we will need for a case with one locus that has two alleles (A and a).

$N$  = Population size of individuals ( $N$  diploid individuals have  $2N$  alleles)

Allele frequencies:

$$\begin{aligned} p &= \text{frequency(A allele)} \\ &= (\text{total number of A alleles})/2N \end{aligned}$$

$$\begin{aligned} q &= \text{frequency(a allele)} \\ &= (\text{total number of a alleles})/2N \end{aligned}$$

$$p + q = 1$$

Genotype frequencies:

$$\begin{aligned} X &= \text{frequency(AA genotype)} \\ &= (\text{total number of AA genotypes})/N \end{aligned}$$

$$\begin{aligned} Y &= \text{frequency(Aa genotype)} \\ &= (\text{total number of Aa genotypes})/N \end{aligned}$$

$$\begin{aligned} Z &= \text{frequency(aa genotype)} \\ &= (\text{total number of aa genotypes})/N \end{aligned}$$

$$X + Y + Z = 1$$

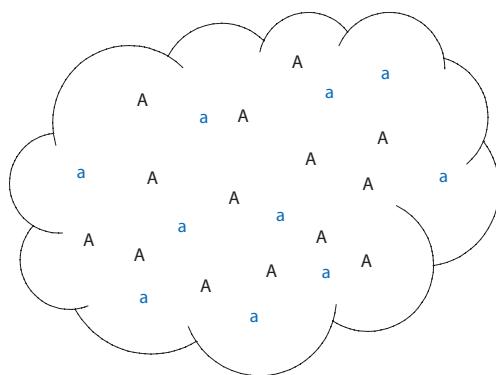
We do not distinguish between the heterozygotes Aa and aA and treat them as being equivalent genotypes. Therefore, we can express allele frequencies in terms of genotype frequencies by adding together the frequencies of A-containing and a-containing genotypes:

$$p = X + \frac{1}{2}Y \quad (2.2)$$

$$q = Z + \frac{1}{2}Y \quad (2.3)$$

Each homozygote contains two alleles of the same type while each heterozygote contains one allele of each type so the heterozygote genotypes are each weighted by half.

With the variables defined, we can then follow allele frequencies across one generation of reproduction. The first step is to calculate the probability that parents of any two particular genotypes will mate. Since mating is assumed to be random, the chance that two genotypes will mate is just the product of their individual frequencies. As shown in Fig. 2.7, random mating can be thought of as being like gas atoms in a balloon. As with gas atoms, each genotype or gamete bumps into others at random, with the probability of a collision (or mating or union) being the product of the frequencies of the two objects



**Figure 2.7** A schematic representation of random mating as a cloud of gas where the frequency of A alleles is 14/24 and the frequency of a alleles is 10/24. Any given A has a frequency of 14/24 and will encounter another A with probability of 14/24 or an a with the probability of 10/24. This makes the frequency of an A–A collision  $(14/24)^2$  and an A–a collision  $(14/24)(10/24)$ , just as the probability of two independent events is the product of their individual probabilities. The population of A and a alleles is assumed to be large enough so that taking one out of the cloud will make almost no change in the overall frequency of its type.

colliding. To calculate the probabilities of mating among the three different genotypes we can make a table to organize the resulting mating frequencies. This table will predict the mating frequencies among genotypes in the initial generation, which we will call generation  $t$ .

A parental mating frequency table (generation  $t$ ) is shown below.

Moms		Dads		
		AA	Aa	aa
	Frequency ...	X	Y	Z
AA	X	$X^2$	$XY$	$XZ$
Aa	Y	$XY$	$Y^2$	$YZ$
aa	Z	$ZX$	$ZY$	$Z^2$

The table expresses parental mating frequencies in the currency of genotype frequencies. For example, we expect matings between AA moms and Aa dads to occur with a frequency of  $XY$ .

Next we need to determine the frequency of each genotype in the offspring of any given parental mating pair. This will require that we predict the offspring genotypes resulting from each possible parental mating. We can do this easily with a Punnett square. We will use the frequencies of each parental

mating (above) together with the frequencies of the offspring genotypes. Summed for all possible parental matings, this gives the frequency of offspring genotypes one generation later, or in generation  $t + 1$ . A table will help organize all the frequencies, like the offspring frequency table (generation  $t + 1$ ) shown below.

		Offspring genotype frequencies		
Parental mating	Total frequency	AA	Aa	aa
AA × AA	$X^2$	$X^2$	0	0
AA × Aa	$2XY$	$XY$	$XY$	0
AA × aa	$2XZ$	0	$2XZ$	0
Aa × Aa	$Y^2/4$	$Y^2/4$	$(2Y^2)/4$	$Y^2/4$
Aa × aa	$2YZ$	0	$YZ$	$YZ$
aa × aa	$Z^2$	0	0	$Z^2$

In this table, the total frequency is just the frequency of each parental mating pair taken from the parental mating frequency table. We now need to partition this total frequency of each parental mating into the frequencies of the three progeny genotypes produced. Let's look at an example. Parents with AA and Aa genotypes will produce progeny with two genotypes: half AA and half Aa (you can use a Punnett square to show this is true). Therefore, the AA × Aa parental matings, which have a total frequency of  $2XY$  under random mating, are expected to produce  $(1/2)2XY = XY$  of each of AA and Aa progeny. The same logic applies to all of the other parental matings. Notice that each row in the offspring genotype frequency table sums to the total frequency of each parental mating.

The columns in the offspring genotype frequency table are the basis of the final step. The sum of each column gives the total frequencies of each progeny genotype expected in generation  $t + 1$ . Let's take the sum of each column, again expressed in the currency of genotype frequencies, and then simplify the algebra to see whether Hardy and Weinberg were correct.

$$\begin{aligned} AA &= X^2 + XY + Y^2/4 \\ &= (X + 1/2Y)^2 \text{ (recall that } p = X + 1/2Y) \\ &= p^2 \end{aligned}$$

$$\begin{aligned}
 aa &= Y^2/4 + YZ + Z^2 \\
 &= (Z + 1/2Y)^2 \text{ (recall that } q = Z + 1/2Y) \\
 &= q^2
 \end{aligned}$$
  

$$\begin{aligned}
 Aa &= XY + 2XZ + 2Y^2/4 + YZ \\
 &= 2(XY/2 + XZ + Y^2/4 + YZ/2) \\
 &= 2(X + Y/2)(Z + Y/2) \\
 &= 2pq
 \end{aligned} \tag{2.4}$$

So we have proved that progeny genotype and allele frequencies are identical to parental genotype and allele frequencies over one generation or that  $f(A)_t = f(A)_{t+1}$ . The major conclusion here is that *genotype frequencies remain constant over generations as long as the assumptions of Hardy–Weinberg are met*. In fact, we have just proved that under Mendelian heredity genotype and allele frequencies should not change over time unless one or more of our assumptions is not met. This simple model of expected genotype frequencies has profound conclusions! In fact, Hardy–Weinberg expected genotype frequencies serve as one of the most basic tools to test for the action of biological processes that alter genotype and allele frequencies.

You might wonder whether Hardy–Weinberg applies to loci with more than two alleles. For the last point in this section let's explore that question. With three alleles at one locus (allele frequencies symbolized by  $p$ ,  $q$ , and  $r$ ), Hardy–Weinberg expected genotype frequencies are  $p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$ . These genotype frequencies are obtained by expanding  $(p + q + r)^2$ , a method that can be applied to any number of alleles at one locus. In general, expanding the squared sum of the allele frequencies will show:

- the frequency of any homozygous genotype is the squared frequency of the single allele that composes the genotype ( $[\text{allele frequency}]^2$ );
- the frequency of any heterozygous genotype is twice the product of the two allele frequencies that comprise the genotype ( $2[\text{allele 1 frequency}] \times [\text{allele 2 frequency}]$ ); and
- there are as many homozygous genotypes as there are alleles and  $\frac{N(N - 1)}{2}$  heterozygous genotypes where  $N$  is the number of alleles.

Do you think it would be possible to prove Hardy–Weinberg for more than two alleles at one locus? The answer is absolutely yes. This would just require constructing larger versions of the parental genotype

mating table and expected offspring frequency table as we did for two alleles at one locus.

## 2.4 Applications of Hardy–Weinberg

- Apply Hardy–Weinberg to estimate the frequency of an observed genotype in a forensic DNA typing case.
- The  $\chi^2$  test gauges whether observed and expected differ more than expected by chance.
- Assume Hardy–Weinberg to compare two genetic models.

In the previous two sections we established the Hardy–Weinberg expectations for genotype frequencies. In this section we will examine three ways that expected genotype frequencies are employed in practice. The goal of this section is to become familiar with realistic applications as well as hypothesis tests that compare observed and Hardy–Weinberg expected genotype frequencies. In this process we will also look at a specific method to account for sampling error (see Appendix).

### Forensic DNA profiling

Our first application of Hardy–Weinberg can be found in newspapers on a regular basis and commonly dramatized on television. A terrible crime has been committed. Left at the crime scene was a biological sample that law-enforcement authorities use to obtain a multilocus genotype or DNA profile. A suspect in the crime has been identified and subpoenaed to provide a tissue sample for DNA profiling. The DNA profiles from the suspect and from the crime scene are identical. The DNA profile is shown in Table 2.2. Should we conclude that the suspect left the biological sample found at the crime scene?

**Table 2.2** Example DNA profile for three simple tandem repeat (STR) loci commonly used in human forensic cases. Locus names refer to the human chromosome (e.g. D3 means third chromosome) and chromosome region where the SRT locus is found.

Locus	D3S1358	D21S11	D18S51
Genotype	17, 18	29, 30	18, 18

**Table 2.3** Allele frequencies for nine STR loci commonly used in forensic cases estimated from 196 US Caucasians sampled randomly with respect to geographic location. The allele names are the numbers of repeats at that locus (see Box 2.1). Allele frequencies (Freq) are as reported in Budowle et al. (2001), Table 1, from FBI sample population.

	D3S1358	vWA	D21S11	D18S51	D13S317	FGA	D8S1179	D5S818	D7S820		
Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq
12	0.0000	13	0.0051	27	0.0459	<11	0.0128	8	0.0995	18	0.0306
13	0.0025	14	0.1020	28	0.1658	11	0.0128	9	0.0765	19	0.0561
14	0.1404	15	0.1122	29	0.1811	12	0.1276	10	0.0510	20	0.1454
15	0.2463	16	0.2015	30	0.2321	13	0.1224	11	0.3189	20.2	0.0026
16	0.2315	17	0.2628	30.2	0.0383	14	0.1735	12	0.3087	21	0.1735
17	0.2118	18	0.2219	31	0.0714	15	0.1276	13	0.1097	22	0.1888
18	0.1626	19	0.0842	31.2	0.0995	16	0.1071	14	0.0357	22.2	0.0102
19	0.0049	20	0.0102	32	0.0153	17	0.1556	23	0.1582	15	0.1097
				32.2	0.1122	18	0.0918	24	0.1378	16	0.0128
				33.2	0.0306	19	0.0357	25	0.0689	17	0.0026
				35.2	0.0026	20	0.0255	26	0.0179	27	0.0102
						21	0.0051	22	0.0026		

To answer this critical question we will employ Hardy–Weinberg to predict the expected frequency of the DNA profile or genotype. Just because two DNA profiles match, there is not necessarily strong evidence that the individual who left the evidence DNA and the suspect are the same person. It is possible that there are actually two or more people with identical DNA profiles. Hardy–Weinberg and Mendel’s second law will serve as the bases for us to estimate just how frequently a given DNA profile should be observed. Then we can determine whether two unrelated individuals sharing an identical DNA profile is a likely occurrence.

To determine the expected frequency of a one-locus genotype, we employ the Hardy–Weinberg equation (2.1). In doing so, we are implicitly accepting that all of the assumptions of Hardy–Weinberg are approximately met. If these assumptions were not met, then the Hardy–Weinberg equation would not provide an accurate expectation for the genotype frequencies! To determine the frequency of the three-locus genotype in Table 2.2 we need allele frequencies for those loci, which are found in Table 2.3. Starting with the locus D3S1358, we see in Table 2.3 that the 17-repeat allele has a frequency of 0.2118 and the 18-repeat allele a frequency of 0.1626. Then using Hardy–Weinberg, the 17, 18 genotype has an expected frequency of  $2(0.2118)(0.1626) = 0.0689$  or 6.89%. For the two other loci in the DNA profile of Table 2.2 we carry out the same steps.

$$\begin{aligned} \text{D21S11} & \quad 29\text{-Repeat allele frequency} = 0.1811 \\ & \quad 30\text{-Repeat allele frequency} = 0.2321 \\ & \quad \text{Genotype frequency} \\ & = 2(0.1811)(0.2321) = 0.0841 \\ & \quad \text{or } 8.41\% \end{aligned}$$

$$\begin{aligned} \text{D18S51} & \quad 18\text{-Repeat allele frequency} = 0.0918 \\ & \quad \text{Genotype frequency} = (0.0918)^2 \\ & = 0.0084 \text{ or } 0.84\% \end{aligned}$$

The genotype for each locus has a relatively large chance of being observed in a population. For example, a little less than 1% of white US citizens (or about 1 in 119) are expected to be homozygous for the 18-repeat allele at locus D18S51. Therefore, a match between evidence and suspect DNA profiles homozygous for the 18 repeat at that locus would not be strong evidence that the samples came from the same individual.

Fortunately, we can combine the information from all three loci. To do this we use the **product**

**rule**, which states that the probability of observing multiple independent events is just the product of each individual event. We already used the product rule in the last section to calculate the expected frequency of each genotype under Hardy–Weinberg by treating each allele as an independent probability. Now we just extend the product rule to cover multiple genotypes, *under the assumption that each of the loci is independent by Mendel’s second law* (the assumption is justified here since each of the loci is on a separate chromosome). The expected frequency of the three locus genotype (sometimes called the probability of identity) is then  $0.0689 \times 0.0841 \times 0.0084 = 0.000049$  or 0.0049%. Another way to express this probability is as an **odds ratio**, or the reciprocal of the probability (an approximation that holds when the probability is very small). Here the odds ratio is  $1/0.000049 = 20,408$ , meaning that we would expect to observe the three-locus DNA profile once in 20,408 white US citizens.

**Product rule** The probability of two (or more) independent events occurring simultaneously is the product of their individual probabilities.

**Odds ratio** The number of events divided by the number of non-events; one over the expected sample size required to observe a single instance or event.

Now we can return to the question of whether two unrelated individuals are likely to share an identical three-locus DNA profile by chance. One out of every 20,408 white US citizens is expected to have the genotype in Table 2.2. Although the three-locus DNA profile is considerably less frequent than a genotype for a single locus, it is still does not approach a unique, individual identifier. Therefore, there is a finite chance that a suspect will match an evidence DNA profile by chance alone. Such DNA profile matches, or “inclusions,” require additional evidence to ascertain guilt or innocence. In fact, the term prosecutor’s fallacy was coined to describe failure to recognize the difference between a DNA match and guilt (for example, a person can be present at a location and not involved in a crime). Only when DNA profiles do not match, called an “exclusion,” can a suspect be unambiguously and absolutely excluded as the source of a biological sample at a crime scene.

### Problem box 2.1 The expected genotype frequency for a DNA profile

Calculate the expected genotype frequency and odds ratio for the 10-locus DNA profile below. Allele frequencies are given in Table 2.3.

D3S1358	17, 18
vWA	17, 17
FGA	24, 25
Amelogenin	X, Y
D8S1179	13, 14
D21S11	29, 30
D18S51	18, 18
D5S818	12, 13
D13S317	9, 12
D7S820	11, 12

What does the amelogenin locus tell us and how did you assign an expected frequency to the observed genotype? Is it likely that two unrelated individuals would share this 10-locus genotype by chance? For this genotype, would a match between a crime scene sample and a suspect be convincing evidence that the person was present at the crime scene?

Current forensic DNA profiles use 10–13 loci to estimate expected genotype frequencies. Problem 2.1 gives a 10-locus genotype for the same individual in Table 2.2, allowing you to calculate the odds ratio for a realistic example. In Chapter 4 we will reconsider the expected frequency of a DNA profile with the added complication of allele-frequency differentiation among human racial groups.

#### Testing for Hardy–Weinberg

A common use of Hardy–Weinberg expectations is to test for deviations from its null model. Populations with genotype frequencies that do not fit Hardy–Weinberg expectations are evidence that one or more of the evolutionary processes embodied in the assumptions of Hardy–Weinberg are acting to determine genotype frequencies. Our null hypothesis is that genotype frequencies meet Hardy–Weinberg expectations within some degree of estimation error. Genotype frequencies that are not close to Hardy–Weinberg expectations allow us to reject this null hypothesis. The processes in the list of assumptions then become possible alternative hypotheses to explain observed genotype frequencies. In this section we will work through a hypothesis test for Hardy–Weinberg equilibrium.

The first example uses observed genotypes for the MN blood group, a single locus in humans that has

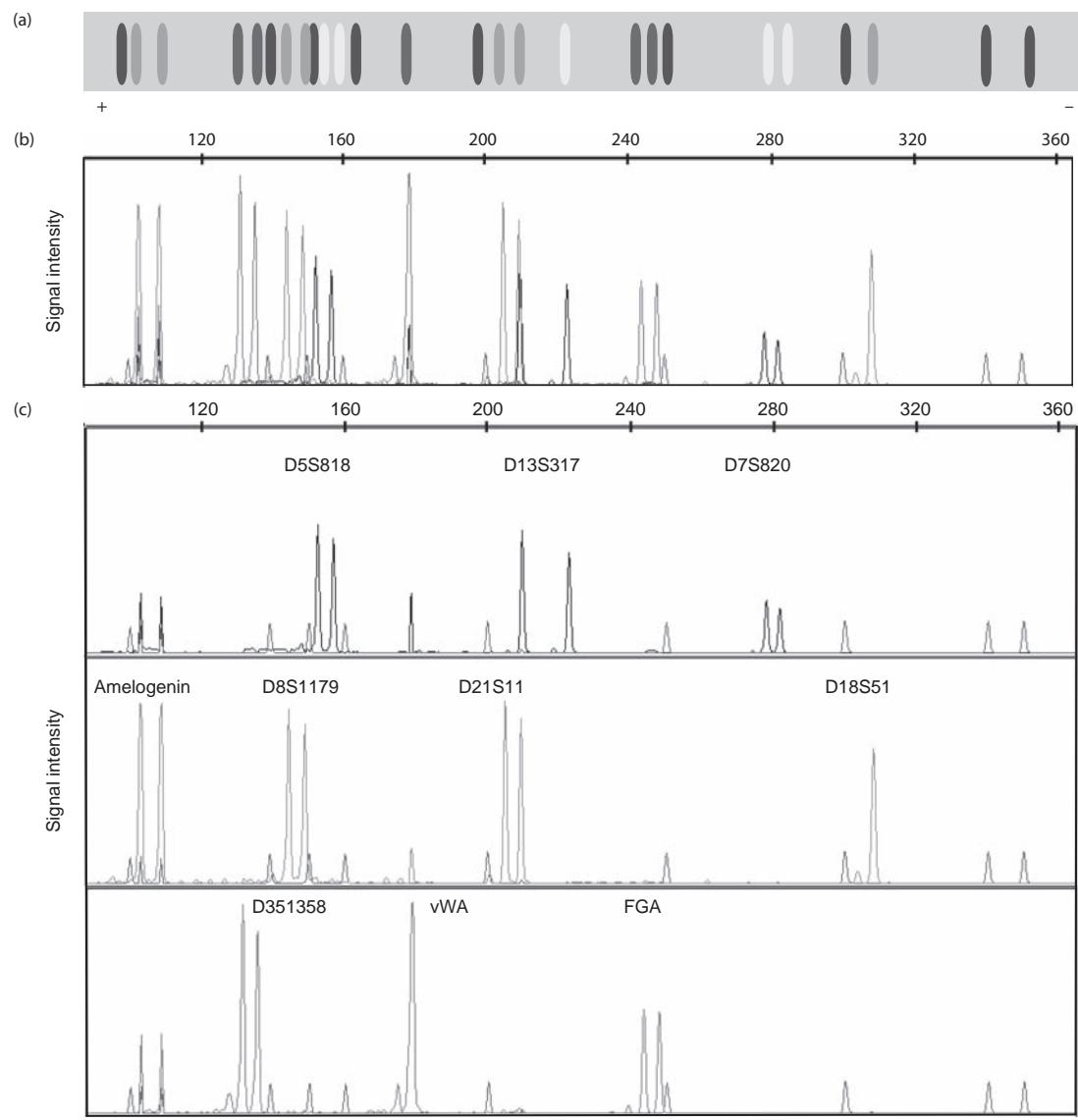
### Box 2.1 DNA profiling

The loci used for human DNA profiling are a general class of DNA sequence marker known as simple tandem repeat (STR), simple sequence repeat (SSR), or microsatellite loci. These loci feature tandemly repeated DNA sequences of one to six base pairs (bp) and often exhibit many alleles per locus and high levels of heterozygosity. Allelic states are simply the number of repeats present at the locus, which can be determined by electrophoresis of PCR amplified DNA fragments. STR loci used in human DNA profiling generally exhibit Hardy–Weinberg expected genotype frequencies, there is evidence that the genotypes are selectively “neutral” (i.e. not affected by natural selection), and the loci meet the other assumptions of Hardy–Weinberg. STR loci

are employed widely in population genetic studies and in genetic mapping (see reviews by Goldstein & Pollock 1997; McDonald & Potts 1997).

This is an example of the DNA sequence found at a microsatellite locus. This sequence is the 24.1 allele from the FGA locus (Genbank accession no. AY749636; see Fig. 2.8). The integral repeat is the 4 bp sequence CTTT and most alleles have sequences that differ by some number of full CTTT repeats. However, there are exceptions where alleles have sequences with partial repeats or stutters in the repeat pattern, for example the TTTCT and CTC sequences imbedded in the perfect CTTT repeats. In this case, the 24.1 allele is 1 bp longer than the 24 allele sequence.

(continued)

**Box 2.1 (continued)**

**Figure 2.8** The original data for the DNA profile given in Table 2.2 and Problem box 2.1 obtained by capillary electrophoresis. The PCR oligonucleotide primers used to amplify each locus are labeled with a molecule that emits blue, green, or yellow light when exposed to laser light. Thus, the DNA fragments for each locus are identified by their label color as well as their size range in base pairs. (a) A simulation of the DNA profile as it would appear on an electrophoretic gel (+ indicates the anode side). Blue, green, and yellow label the 10 DNA-profiling loci, shown here in grayscale. Other DNA fragments are size standards (originally in red) with a known molecular weight used to estimate the size in base pairs of the other DNA fragments in the profile. (b) The DNA profile for all loci and the size-standard DNA fragments as a graph of color signal intensity by size of DNA fragment in base pairs. (c) A simpler view of trace data for each label color independently with the individual loci labeled above the trace peaks. A few shorter peaks are visible in the yellow, green, and blue traces of (c) that are not labeled as loci. These artifacts, called pull-up peaks, are caused by intense signal from a locus labeled with another color (e.g. the yellow and blue peaks in the location of the green-labeled amelogenin locus). A full-color version of this figure is available on the textbook website.

```

GCCCCATAGGTTTGAACTCACAGAACTGTAAACCAAAATAAATTAGGCATTATTTACAAGCTA
GTTT CTTT CTTT
CTTT CTTT CTTT CTTT CTTT CTC CTTC CTTC CTTC CTTT CTTT TTTGCTGGCA
ATTACAGACAAATCAA

```

**Table 2.4** Expected numbers of each of the three MN blood group genotypes under the null hypotheses of Hardy–Weinberg. Genotype frequencies are based on a sample of 1066 Chukchi individuals, a native people of eastern Siberia (Roychoudhury and Nei 1988).

Frequency of M =  $\hat{p} = 0.4184$

Frequency of N =  $\hat{q} = 0.5816$

Genotype	Observed	Expected number of genotypes	Observed – expected
MM	165	$N \times \hat{p}^2 = 1066 \times (0.4184)^2 = 186.61$	-21.6
MN	562	$N \times 2\hat{p}\hat{q} = 1066 \times 2(0.4184)(0.5816) = 518.80$	43.2
NN	339	$N \times \hat{q}^2 = 1066 \times (0.5816)^2 = 360.58$	-21.6

two alleles (Table 2.4). First we need to estimate the frequency of the M allele, using the notation that the estimated frequency of M is  $\hat{p}$  and the frequency of N is  $\hat{q}$ . Note that the “hat” superscripts indicate that these are allele-frequency *estimates* (see Chapter 1). The total number of alleles is  $2N$  given a sample of  $N$  diploid individuals. We can then count up all of the alleles of one type to estimate the frequency of that allele.

$$\hat{p} = \frac{2 \times \text{Frequency(MM)} + \text{frequency(MN)}}{2N} \quad (2.5)$$

$$\hat{p} = \frac{2 \times 165 + 562}{2 \times 1066} = \frac{892}{2132} = 0.4184 \quad (2.6)$$

Since  $\hat{p} + \hat{q} \approx 1$ , we can estimate the frequency of the N allele by subtraction as  $\hat{q} = 1 - \hat{p} = 1 - 0.4184 = 0.5816$ .

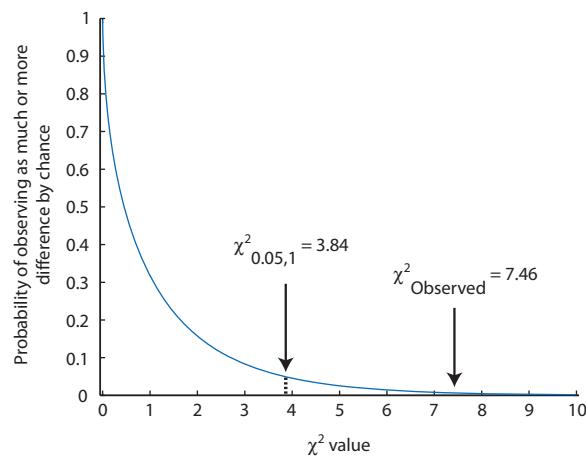
Using these allele frequencies allows calculation of the Hardy–Weinberg expected genotype frequency and number of individuals with each genotype, as shown in Table 2.4. In Table 2.4 we can see that the match between the observed and expected is not perfect, but we need some method to ask whether the difference is actually large enough to conclude that Hardy–Weinberg equilibrium does not hold in the sample of 1066 genotypes. Remember that any allele-frequency estimate ( $\hat{p}$ ) could differ slightly from the true parameter ( $p$ ) due to chance events as well as due to random sampling in the group of genotypes used to estimate the allele frequencies. Asking whether genotypes are in Hardy–Weinberg proportions is actually the same as asking whether a coin is “fair.” With a fair coin we expect one-half heads and one-half tails if we flip it a large number of times. But even with a fair coin we can get something other than exactly

50 : 50 even if the sample size is large. We would consider a coin fair if in 1000 flips it produced 510 heads and 490 tails. However, the hypothesis that a coin is fair would be in doubt if we observed 250 heads and 750 tails given that we expect 500 of each.

In more general terms, the expected frequency of an event,  $p$ , times the number of trials or samples,  $n$ , gives the expected number events or  $np$ . To test the hypothesis that  $p$  is the frequency of an event in an actual population, we compare  $np$  with  $n\hat{p}$ . Close agreement suggests that the parameter and the estimate are the same quantity. But a large disagreement instead suggests that  $p$  and  $\hat{p}$  are likely to be different probabilities. The Chi-squared ( $\chi^2$ ) distribution is a statistical test commonly used to compare  $np$  and  $n\hat{p}$ . The  $\chi^2$  test provides the probability of obtaining the difference (or more) between the observed ( $n\hat{p}$ ) and expected ( $np$ ) number of outcomes by chance alone if the null hypothesis is true. As the difference between the observed and expected grows larger it becomes less probable that the parameter and the parameter estimate are actually the same but differ in a given sample due to chance. The  $\chi^2$  statistic is:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (2.7)$$

The  $\chi^2$  formula makes intuitive sense. In the numerator there is the difference between the observed and Hardy–Weinberg expected number of individuals. This difference is squared, like a variance, since we do not care about the direction of the difference but only the magnitude of the difference. Then in the denominator we divide by the expected number of individuals to make the squared difference relative. For example, a squared difference of 4 is small if the expected number is 100 (it is 4%) but relatively



**Figure 2.9** A  $\chi^2$  distribution with one degree of freedom. The  $\chi^2$  value for the Hardy–Weinberg test with MN blood group genotypes as well as the critical value to reject the null hypothesis are shown (see text for details). The area under the curve to the right of the arrow indicates the probability of observing that much or more difference between the observed and expected outcomes.

larger if the expected number is 8 (it is 50%). Adding all of these relative squared differences gives the total relative squared deviation observed over all genotypes.

$$\chi^2 = \frac{(-21.6)^2}{181.61} + \frac{(43.2)^2}{518.80} + \frac{(-21.6)^2}{360.58} = 7.46 \quad (2.8)$$

We need to compare our statistic to values from the  $\chi^2$  distribution. But first we need to know how much information, or the degrees of freedom (commonly abbreviated as df), was used to estimate the  $\chi^2$  statistic. In general, degrees of freedom are based on the number of categories of data:  $df = \text{no. of classes compared} - \text{no. of parameters estimated} - 1$

for the  $\chi^2$  test itself. In this case  $df = 3 - 1 - 1 = 1$  for three genotypes and one estimated allele frequency (with two alleles: the other allele frequency is fixed once the first has been estimated).

Figure 2.9 shows a  $\chi^2$  distribution for one degree of freedom. Small deviations of the observed from the expected are more probable since they leave more area of the distribution to the right of the  $\chi^2$  value. As the  $\chi^2$  value gets larger, the probability that the difference between the observed and expected is just due to chance sampling decreases (the area under the curve to the right gets smaller). Another way of saying this is that as the observed and expected get increasingly different, it becomes more improbable that our null hypothesis of Hardy–Weinberg is actually the process that is determining genotype frequencies. Using Table 2.5 we see that a  $\chi^2$  value of 7.46 with 1 df has a probability between 0.01 and 0.001. The conclusion is that the observed genotype frequencies would be observed less than 1% of the time in a population that actually had Hardy–Weinberg expected genotype frequencies. Under the null hypothesis we do not expect this much difference or more from Hardy–Weinberg expectations to occur often. By convention, we would reject chance as the explanation for the differences if the  $\chi^2$  value had a probability of 0.05 or less. In other words, if chance explains the difference in five trials out of 100 or less then we reject the hypothesis that the observed and expected patterns are the same. The critical value above which we reject the null hypothesis for a  $\chi^2$  test is 3.84 with 1 df, or in notation  $\chi^2_{0.05,1} = 3.84$ . In this case, we can clearly see an excess of heterozygotes and deficits of homozygotes and employing the  $\chi^2$  test allows us to conclude that Hardy–Weinberg expected genotype frequencies are not present in the population.

**Table 2.5**  $\chi^2$  values and associated cumulative probabilities in the right-hand tail of the distribution for 1–5 df.

df	Probability					
	0.5	0.25	0.1	0.05	0.01	0.001
1	0.4549	1.3233	2.7055	3.8415	6.6349	10.8276
2	1.3863	2.7726	4.6052	5.9915	9.2103	13.8155
3	2.3660	4.1083	6.2514	7.8147	11.3449	16.2662
4	3.3567	5.3853	7.7794	9.4877	13.2767	18.4668
5	4.3515	6.6257	9.2364	11.0705	15.0863	20.5150

### Interact box 2.2 $\chi^2$ test

The program PopGene.S<sup>2</sup> (refer to the link for **Interact box 2.1** to obtain **PopGene.S<sup>2</sup>** if necessary) can be used to carry out a  $\chi^2$  test for one locus with two alleles under the null hypothesis of Hardy–Weinberg genotype frequencies. Launch PopGene.S<sup>2</sup> and select **Chi-square test** under the **Allele and Genotype Frequencies** menu.

Try these examples.

- Input the observed genotype frequencies in Table 2.4 to confirm the calculations and  $\chi^2$  value.
- Test the null hypothesis of Hardy–Weinberg genotype frequencies for these data from a sample of 459 Yugoslavians: MM, 144; MN, 201; NN, 114.

### *Assuming Hardy–Weinberg to test alternative models of inheritance*

Biologists are all probably familiar with the ABO blood groups and are aware that mixing blood of different types can cause blood cell lysis and possibly result in death. Although we take this for granted now, there was a time when blood types and their patterns of inheritance defined an active area of clinical research. It was in 1900 that Karl Landsteiner of the University of Vienna mixed the blood of the people in his laboratory to study the patterns of blood cell agglutination (clumping). Landsteiner was awarded the Nobel Prize

for Medicine in 1930 for his discovery of human ABO blood groups. Not until 1925, due to the research of Felix Bernstein, was the genetic basis of the ABO blood groups resolved (see Crow 1993a).

Landsteiner observed the presence of four blood phenotypes A, B, AB, and O. A logical question was, then, “what is the genetic basis of these four blood group phenotypes?” We will test two hypotheses (or models) to explain the inheritance of ABO blood groups that coexisted for 25 years. The approach will use the frequency of genotypes in a sample population to test the two hypotheses rather than an approach such as examining pedigrees. The hypotheses are that the four blood group phenotypes are explained by either two independent loci with two alleles each with one allele completely dominant at each locus (hypothesis 1) or a single locus with three alleles where two of the alleles show no dominance with each other but both are completely dominant over a third allele (hypothesis 2). Throughout, we will assume that Hardy–Weinberg expected genotype frequencies are met in order to determine which hypothesis best fits the available data.

Our first task is a straightforward application of Hardy–Weinberg in order to determine the expected frequencies of the blood group genotypes. The genotypes and the expected genotype frequencies are shown in Table 2.6. Look at the table but cover up the expected frequencies with a sheet of paper. The genotypes given for the two hypotheses would both explain the observed pattern of four blood groups. Hypothesis 1 requires complete dominance of the A and B alleles at their respective loci. Hypothesis 2 requires A and B to have no dominance with each other but complete dominance when paired with the O allele.

**Table 2.6** Hardy–Weinberg expected genotype frequencies for the ABO blood groups under the hypotheses of (1) two loci with two alleles each and (2) one locus with three alleles. Both hypotheses have the potential to explain the observation of four blood group phenotypes. The notation  $f_x$  is used to refer to the frequency of allele  $x$ . The underscore indicates any allele; for example,  $A\_$  means both AA and Aa genotypes. The observed blood type frequencies were determined for Japanese people living in Korea (from Berstein (1925) as reported in Crow (1993a)).

Blood type	Genotype		Expected genotype frequency		Observed (total = 502)
	Hypothesis 1	Hypothesis 2	Hypothesis 1	Hypothesis 2	
O	aa bb	OO	$fa^2fb^2$	$fo^2$	148
A	A_bb	AA, AO	$(1 - fa^2)fb^2$	$fA^2 + 2fAfO$	212
B	aa B_	BB, BO	$fa^2(1 - fb^2)$	$fb^2 + 2fBfO$	103
AB	A_B_	AB	$(1 - fa^2)(1 - fb^2)$	$2fAfB$	39

Now let's construct several of the expected genotype frequencies (before you lift that sheet of paper). The O blood group under hypothesis 1 is the frequency of a homozygous genotype at two loci (aa bb). The frequency of one homozygote is the square of the allele frequency:  $f_a^2$  and  $f_b^2$  if we use  $f_x$  to indicate the frequency of allele x. Using the product rule or Mendel's second law, the expected frequency of the two-locus genotype is the product of frequencies of the one-locus genotypes,  $f_a^2$  and  $f_b^2$ . For the next genotype under hypothesis 1 (A\_ bb), we use a little trick to simplify the amount of notation. The genotype A\_ means AA or Aa; in other words, any genotype but aa. Since the frequencies of the three genotypes at one locus must sum to one, we can write  $f_{A\_}$  as  $1 - f_{aa}$  or  $1 - f_a^2$ . Then the frequency of the A\_ bb genotype is  $(1 - f_a^2)f_b^2$ . You should now work out and write down the other six expected genotype frequency expressions: then lift the paper and compare your work to Table 2.6.

The next step is to compare the expected genotype frequencies for the two hypotheses with observed genotype frequencies. To do this we will need to estimate allele frequencies under each hypothesis and use these to compute the expected genotype frequencies. (Although these allele frequencies are parameter estimates, the "hat" notation is not used for the sake of readability.) For the hypothesis of two loci (hypothesis 1)  $f_b^2 = (148 + 212)/502 = 0.717$  so we can estimate the allele frequency as  $f_b = \sqrt{f_b^2} = \sqrt{0.717} = 0.847$ . The other allele frequency at that locus is then determined by subtraction:  $f_B = 1 - 0.847 = 0.153$ . Similarly for the second locus  $f_a^2 = (148 + 103)/502 = 0.50$  and  $f_a = \sqrt{f_a^2} = \sqrt{0.50} = 0.707$ , giving  $f_A = 1 - 0.707 = 0.293$  by subtraction.

### Problem box 2.2

#### Proving allele frequencies are obtained from expected genotype frequencies

Can you use algebra to prove that adding together expected genotype frequencies under hypotheses 1 and 2 in Table 2.7 gives the allele frequencies shown in the text? For the genotypes of hypothesis 1, show that  $f_{(aa\ bb)} + f_{(A\_ bb)} =$  the frequency of the bb genotype. For hypothesis 2 show the observed genotype frequencies that can be used to estimate the frequency of the B allele starting off with the relationship  $f_A + f_B + f_O = 1$  and then solving for  $f_B$  in terms of  $f_A$  and  $f_O$ .

For the hypothesis of one locus with three alleles (hypothesis 2) we estimate the frequency of any of the alleles by using the relationship that the three allele frequencies sum to one. This basic relationship can be reworked to obtain the expected genotype frequency expressions into expressions that allow us to estimate the allele frequencies (see Problem box 2.2). It turns out that adding together all expected genotype frequency terms for two of the alleles estimates the square of one minus the other allele. For example,  $(1 - f_B)^2 = f_O^2 + f_A^2 + 2f_Af_O$ , and checking in Table 2.7 this corresponds to  $(148 + 212)/502 = 0.717$ . Therefore,  $1 - f_B = 0.847$  and  $f_B = 0.153$ . Using similar steps,  $(1 - f_A)^2 = f_O^2 + f_B^2 + 2f_Bf_O =$

**Table 2.7** Expected numbers of each of the four blood group genotypes under the hypotheses of (1) two loci with two alleles each and (2) one locus with three alleles. Estimated allele frequencies are based on a sample of 502 individuals.

Blood	Observed	Expected number of genotypes	Observed – expected	$(\text{Observed} - \text{expected})^2 / \text{expected}$
Hypothesis 1 ( $f_A = 0.293$ , $f_a = 0.707$ , $f_B = 0.153$ , $f_b = 0.847$ )				
O	148	$502(0.707)^2(0.847)^2 = 180.02$	-40.02	8.90
A	212	$502(0.500)(0.847)^2 = 180.07$	31.93	5.66
B	103	$502(0.707)^2(0.282) = 70.76$	32.24	14.69
AB	39	$502(0.500)(0.282) = 70.78$	-31.78	14.27
Hypothesis 2 ( $f_A = 0.293$ , $f_B = 0.153$ , $f_O = 0.554$ )				
O	148	$502(0.554)^2 = 154.07$	-6.07	0.24
A	212	$502((0.293)^2 + 2(0.293)(0.554)) = 206.07$	5.93	0.17
B	103	$502((0.153)^2 + 2(0.153)(0.554)) = 96.85$	6.15	0.39
AB	39	$502(2(0.293)(0.153)) = 45.01$	-6.01	0.80

$(148 + 103)/502 = 0.50$ . Therefore,  $1 - fA = 0.707$  and  $fA = 0.293$ . Finally, by subtraction  $fO = 1 - fB - fA = 1 - 0.153 - 0.293 = 0.554$ .

The number of genotypes under each hypothesis can then be found by using the expected genotype frequencies in Table 2.6 and the estimated allele frequencies. Table 2.7 gives the calculation for the expected numbers of each genotype under both hypotheses. We can also calculate a Chi-squared value associated with each hypothesis based on the difference between the observed and expected genotype frequencies. For hypothesis 1  $\chi^2 = 43.52$ , whereas  $\chi^2 = 1.60$  for hypothesis 2. Both of these tests have one degree of freedom (4 genotypes – 2 for estimated allele frequencies – 1 for the test), giving a critical

value of  $\chi^2_{0.05,1} = 3.84$ . Clearly the hypothesis of three alleles at one locus is the better fit to the observed data. Thus, we have just used genotype frequency data sampled from a population with the assumptions of Hardy–Weinberg equilibrium as a means to distinguish between two hypotheses for the genetic basis of blood groups.

## 2.5 The fixation index and heterozygosity

- The fixation index ( $F$ ) measures deviation from Hardy–Weinberg expected heterozygote frequencies.
- Examples of mating systems and  $F$  in wild populations.
- Observed and expected heterozygosity.

The mating patterns of actual organisms frequently do not exhibit the random mating assumed by Hardy–Weinberg. In fact, many species exhibit mating systems that create predictable deviations from Hardy–Weinberg expected genotype frequencies. The term **assortative mating** is used to describe non-random mating. **Positive assortative mating** describes the case when individuals with like genotypes or phenotypes tend to mate. **Negative assortative mating** occurs when individuals with unlike genotypes or phenotypes tend to mate (also called disassortative mating). Both of these general types of non-random mating will impact expected genotype frequencies in a population. This section describes the impacts of non-random mating on genotype frequencies and introduces a commonly used measure of non-random mating that can be utilized to estimate mating patterns in natural populations.

Mating among related individuals, termed **consanguineous mating** or **biparental inbreeding**, increases the probability that the resulting progeny are homozygous compared to random mating. This occurs since relatives, by definition, are more likely than two random individuals to share one or two alleles that were inherited from ancestors they share in common (this makes mating among relatives a form of assortative mating). Therefore, when related individuals mate their progeny have a higher chance of receiving the same allele from both parents, giving them a greater chance of having a homozygous genotype. **Sexual autogamy** or **self-fertilization** is an extreme example of consanguineous mating where an individual can mate with itself by virtue of possessing reproductive organs of both sexes. Many plants and some animals, such as the nematode

### Problem box 2.3 Inheritance for corn kernel phenotypes

Corn kernels are individual seeds that display a wide diversity of phenotypes (see Fig. 2.10 and Plate 2.10). In a total of 3816 corn seeds, the following phenotypes were observed:

Purple, smooth	2058
Purple, wrinkled	728
Yellow, smooth	769
Yellow, wrinkled	261

Are these genotype frequencies consistent with inheritance due to one locus with three alleles or two loci each with two alleles?



**Figure 2.10** Corn cobs demonstrating yellow and purple seeds that are either wrinkled or smooth. For a color version of this image see Plate 2.10.

*Caenorhabditis elegans*, are hermaphrodites that can mate with themselves.

There are also cases of disassortative mating, where individuals with unlike genotypes have a higher probability of mating. A classic example in mammals is mating based on genotypes at major histocompatibility complex (MHC) loci, which produce proteins involved in self/non-self recognition in immune response. Mice are able to recognize individuals with similar MHC genotypes via odor, and based on these odors avoid mating with individuals possessing a similar MHC genotype. Experiments where young mice were raised in nests of either their true parents or foster parents (called cross-fostering) showed that mice learn to avoid mating with individuals possessing odor cues similar to their nest-mates' rather than avoiding MHC-similar individuals per se (Penn & Potts 1998). This suggests mice learn the odor of family members in the nest and avoid mating with individuals with similar odors, indirectly leading to disassortative mating at MHC loci as well as the avoidance of consanguineous mating. One hypothesis to explain the evolution of disassortative mating at MHC loci is that the behavior is adaptive since progeny with

higher heterozygosity at MHC loci may have more effective immune response. There is also evidence that humans prefer individuals with dissimilar MHC genotypes (Wedeckind & Füri 1997).

The effects of non-random mating on genotype frequencies can be measured by comparing Hardy–Weinberg expected frequency of heterozygotes, which assumes random mating, with observed heterozygote frequencies in a population. A quantity called the **fixation index**, symbolized by  $F$  (or sometimes  $f$ , although it never will be in this book, since  $f$  is reserved for the inbreeding coefficient as introduced later in section 2.6), is commonly used to compare how much heterozygosity is present in an actual population relative to expected levels of heterozygosity under random mating:

$$F = \frac{H_e - H_o}{H_e} \quad (2.9)$$

where  $H_e$  is the Hardy–Weinberg expected frequency of heterozygotes based on population allele frequencies and  $H_o$  is the observed frequency of heterozygotes. Dividing the difference between the expected and

### Interact box 2.3 Assortative mating and genotype frequencies

The impact of assortative mating on genotype frequencies can be simulated in PopGene.S<sup>2</sup>.

The program models several non-random mating scenarios that can be selected under the **Mating Models** menu. The results in each case are presented on a De Finetti diagram, where genotype and allele frequencies can be followed over multiple generations.

Start with the **Positive w/o dominance** model of mating. In this case only like genotypes are able to mate (e.g. AA mates only with AA, Aa mates only with Aa, and aa mates only with aa). Take the time to write out Punnett squares to predict progeny genotype frequencies for each of the matings that takes place. Enter initial genotype frequencies of  $P(AA) = 0.25$  and  $P(Aa) = 0.5$  ( $P(aa)$  is determined by subtraction) as a logical place to start. At first, run the simulation for 30 generations. With these values and mating patterns, what happens to the frequency of heterozygotes? What happens to the allele frequencies? Next try other initial genotype frequencies that vary the allele frequencies and that are both in and out of Hardy–Weinberg proportions.

Next run both the **Positive with dominance** and **Negative** (Dissassortative matings) models. In the **Positive with dominance** model, the AA and Aa genotypes have identical phenotypes. Mating can therefore take place among any pairing of genotypes with the dominant phenotype or between aa individuals with the recessive phenotype. In the **Negative** mating model, only unlike genotypes can mate. Take the time to write out Punnett squares to predict progeny genotype frequencies for each of the matings that takes place. In each case, use a set of the same genotype frequencies that you employed in the **Positive w/o dominance** mating model. How do all types of non-random mating affect genotype frequencies? How do they affect allele frequencies?

observed heterozygosity by the expected heterozygosity expresses the difference in the numerator as a percentage of the expected heterozygosity. Even if the difference in the numerator may seem small, it may be large relative to the expected heterozygosity. Dividing by the expected heterozygosity also puts  $F$  on a convenient scale of  $-1$  and  $+1$ . Negative values indicate heterozygote excess and positive values indicate homozygote excess relative to Hardy–Weinberg expectations. In fact, the fixation index can be interpreted as the correlation between the two alleles sampled to make a diploid genotype (see the Appendix for an introduction to correlation if necessary). Given that one allele has been sampled from the population, if the second allele tends to be identical there is a positive correlation (e.g. A and then A or a and then a), if the second allele tends to be different there is a negative correlation (e.g. A and then a or a and then A), and if the second allele is independent there is no correlation (e.g. equally likely to be A or a). With random mating, no correlation is expected between the first and second allele sampled to make a diploid genotype.

**Assortative mating** Mating patterns where individuals do not mate in proportion to their genotype frequencies in the population; mating that is more (positive assortative mating) or less (negative assortative mating) frequent with respect to genotype or genetically based phenotype frequency than expected by random combination.

**Consanguineous mating** Mating between related individuals that can take the form

of biparental inbreeding (mating between two related individuals) or sexual autogamy (self-fertilization).

**Fixation index ( $F$ )** The proportion by which heterozygosity is reduced or increased relative to the heterozygosity in a randomly mating population with the same allele frequencies.

Let's work through an example of genotype data for one locus with two alleles that can be used to estimate the fixation index. Table 2.8 gives observed counts and frequencies of the three genotypes in a sample of 200 individuals. To estimate the fixation index from these data requires an estimate of allele frequencies first. The allele frequencies can then be used to determine expected heterozygosity under the assumptions of Hardy–Weinberg. If  $p$  represents the frequency of the B allele,

$$\hat{p} = \frac{142 + 1/2(28)}{200} = 0.78 \quad (2.10)$$

using the genotype counting method to estimate allele frequency (Table 2.8 uses the allele counting method). The frequency of the b allele,  $q$ , can be estimated directly in a similar fashion or by subtraction ( $\hat{q} = 1 - \hat{p} = 1 - 0.78 = 0.22$ ) since there are only two alleles in this case. The Hardy–Weinberg expected frequency of heterozygotes is  $H_e = 2\hat{p}\hat{q} = 2(0.78)(0.22) = 0.343$ . It is then simple to estimate the fixation index using the observed and expected heterozygosities:

$$\hat{F} = \frac{0.343 - 0.14}{0.343} = 0.59 \quad (2.11)$$

**Table 2.8** Observed genotype counts and frequencies in a sample of  $N = 200$  individuals for a single locus with two alleles. Allele frequencies in the population can be estimated from the genotype frequencies by summing the total count of each allele and dividing it by the total number of alleles in the sample ( $2N$ ).

Genotype	Observed	Observed frequency	Allele count	Allele frequency
BB	142	$\frac{142}{200} = 0.71$	284 B	$\hat{p} = \frac{284 + 28}{400} = 0.78$
Bb	28	$\frac{28}{200} = 0.14$	28 B, 28 b	
bb	30	$\frac{30}{200} = 0.15$	60 b	$\hat{q} = \frac{60 + 28}{400} = 0.22$

**Table 2.9** Estimates of the fixation index ( $\hat{F}$ ) for various species and breeds based on pedigree or molecular genetic marker data.

Species	Mating system	$\hat{F}$	Method	Reference
Human <i>Homo sapiens</i>	Outcrossed	0.0001–0.046	Pedigree	Jorde 1997
Snail <i>Bulinus truncates</i>	Selfed and outcrossed	0.6–1.0	Microsatellites	Viard et al. 1997
Domestic dogs				
Breeds combined	Outcrossed	0.33	Allozyme	Christensen et al. 1985
German Shepherd	Outcrossed	0.10		
Mongrels	Outcrossed	0.06		
Plants				
<i>Arabidopsis thaliana</i>	Selfed	0.99	Allozyme	Abbott and Gomes 1989
<i>Pinus ponderosa</i>	Outcrossed	−0.37	Allozyme	Brown 1979

In this example there is a clear deficit of heterozygotes relative to Hardy–Weinberg expectations. The population contains 59% fewer heterozygotes than would be expected in a population with the same allele frequencies that was experiencing random mating and the other conditions set out in the assumptions of Hardy–Weinberg. Interpreted as a correlation between the allelic states of the two alleles in a genotype, this value of the fixation index tells us that the two alleles in a genotype are much more frequently of the same state than expected by chance.

In biological populations, a wide range of values has been observed for the fixation index (Table 2.9). Fixation indices have frequently been estimated with allozyme data (see Box 2.2). Estimates of  $\hat{F}$  are generally correlated with mating system. Even in species where individuals possess reproductive organs of one sex only (termed **dioecious** individuals), mating among relatives can be common and ranges from infrequent to almost invariant. In other cases, mating is essentially random or complex mating and social systems have evolved to prevent consanguineous mating. Pure-breed dogs are an example where mating among relatives has been enforced by humans to develop lineages with specific phenotypes and behaviors, resulting in high fixation indices in some breeds. Many plant species possess both male and female sexual functions (hermaphrodites) and exhibit an extreme form of consanguineous mating, self-fertilization, that causes rapid loss of heterozygosity. In the case of Ponderosa pines in Table 2.9, the excess of heterozygotes may be due to natural selection against homozygotes

at some loci (inbreeding depression). This makes the important point that departures from Hardy–Weinberg expected genotype frequencies estimated by the fixation index are potentially influenced by processes in addition to the mating system. Genetic loci free of the influence of other processes such as natural selection are often sought to estimate  $\hat{F}$ . In addition,  $\hat{F}$  can be estimated using the average of multiple loci, which will tend to reduce bias since loci will differ in the degree they are influenced by other processes and outliers will be apparent.

Extending the fixation index to loci with more than two alleles just requires a means to calculate the expected heterozygosity ( $H_e$ ) for an arbitrary number of alleles at one locus. This can be accomplished by adding up all of the expected frequencies of each possible homozygous genotype and subtracting this total from one or summing the expected frequencies of all heterozygous genotypes:

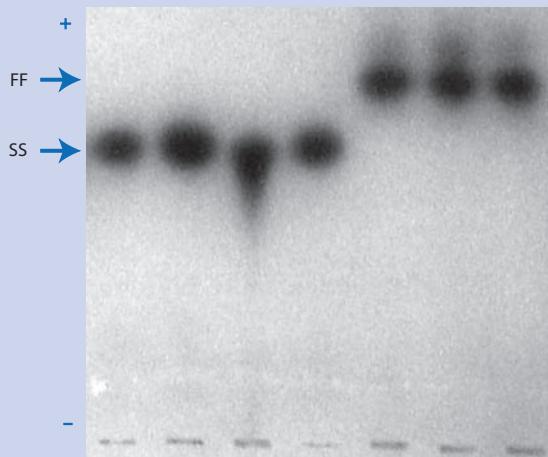
$$H_e = 1 - \sum_{i=1}^k p_i^2 = \sum_{i=1}^{k-1} \sum_{j=i+1}^k 2p_i p_j \quad (2.12)$$

where  $k$  is the number of alleles at the locus, the  $p_i^2$  and  $2p_i p_j$  terms represent the expected genotype frequencies based on allele frequencies, and the  $\sum_{i=1}^k$  (pronounced “sigma”) indicates summation of the frequencies of the  $k$  homozygous genotypes. This quantity is also called the **gene diversity** (Nei 1987). The expected heterozygosity can be adjusted

### Box 2.2 Protein locus or allozyme genotyping

Determining the genotypes of individuals at enzymatic protein loci is a rapid technique to estimate genotype frequencies in populations. Protein analysis was the primary molecular genotyping technique for several decades before DNA-based techniques became widely available. Alleles at loci that code for proteins with enzymatic function can be ascertained in a multi-step process. First, fresh tissue samples are ground up under conditions that preserve the function of proteins. Next, these protein extracts are loaded onto starch gels and exposed to an electric field. The electrical current results in electrophoresis where proteins are separated based on their ratio of molecular charge to molecular weight. Once electrophoresis is complete, the gel is then “stained” to visualize specific enzymes. The primary biochemical products of protein enzymes are not themselves visible. However, a series of biochemical reactions in a process called enzymatic coupling can be used to eventually produce a visible product (often nitro blue tetrazolium or NBT) at the site where the enzyme is active (see Fig. 2.11). If different DNA sequences at a protein enzyme locus result in different amino acid sequences that differ in net charge, then multiple alleles will appear in the gel after staining. The term allozyme (also known as isozyme) is used to describe the multiple allelic staining variants at a single protein locus. Allozyme electrophoresis and staining detects

only a subset of genetic variation at protein coding loci. Amino acid changes that are charge-neutral and nucleotide changes that are synonymous (do not alter the amino acid sequence) cannot be detected by allozyme electrophoresis methods. Refer to Manchenko (2003) for a technical introduction and detailed methods of allozyme detection.



**Figure 2.11** An allozyme gel stained to show alleles at the phosphoglucomutase or PGM locus in striped bass and white bass. The right-most three individuals are homozygous for the faster-migrating allele (FF genotype) while the left-most four individuals are homozygous for the slower-migrating allele (SS genotype). No double-banded heterozygotes (FS genotype) are visible on this gel. The + and – indicate the cathode and anode ends of the gel, respectively. Wells where the individuals samples were loaded into the gel can be seen at the bottom of the picture. Gel picture kindly provided by J. Epifanio.

for small samples by multiplying  $H_e$  by  $2N/(2N - 1)$  where  $N$  is the total number of genotypes (Nei & Roychoudhury 1974), a correction that makes little difference unless  $N$  is about 50 or fewer individuals. In a similar manner, the observed heterozygosity ( $H_o$ ) is the sum of the frequencies of all heterozygotes observed in a sample of genotypes:

$$\hat{H}_o = \sum_{i=1}^h H_i \quad (2.13)$$

where the observed frequency of each heterozygous genotype  $H_i$  is summed over the  $h = k(k - 1)/2$

heterozygous genotypes possible with  $k$  alleles. Finally, both  $H_e$  and  $H_o$  can be averaged over multiple loci to obtain mean heterozygosity estimates for two or more loci. Heterozygosity provides one of the basic measures of genetic variation in population genetics.

The fixation index as a measure of deviation from expected levels of heterozygosity is a critical concept that will appear in several places later in this text. The fixation index plays a conceptual role in understanding the effects of population size on heterozygosity (Chapter 3) and also serves as an estimator of the impact of population structure on the distribution of genetic variation (Chapter 4).

## 2.6 Mating among relatives

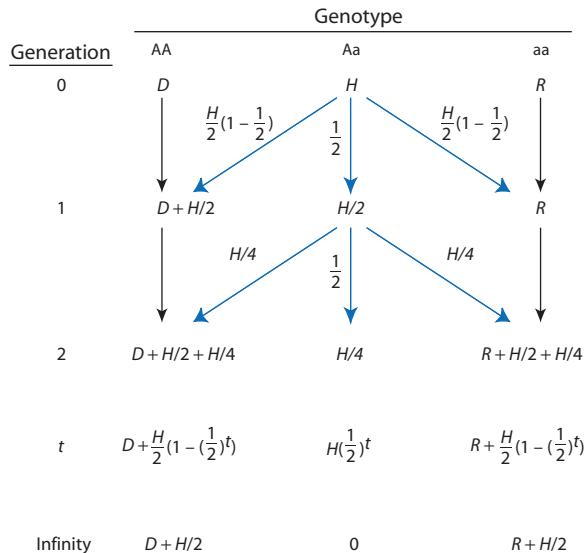
- Consanguineous mating alters genotype frequencies but not allele frequencies.
- Mating among relatives and the probability that two alleles are identical by descent.
- Inbreeding depression and its possible causes.
- The many meanings of inbreeding.

The previous section of this chapter showed how non-random mating can increase or decrease the frequency of heterozygote genotypes compared to the frequency that is expected with random mating. The last section also introduced the fixation index as well as ways to quantify heterozygosity in a population. This section will build on that foundation to show two concepts: (i) the consequences of non-random mating on allele and genotype frequencies in a population and (ii) the probability that two alleles are identical by descent. The focus will be on positive genotypic assortative mating (like genotypes mate) or inbreeding since this will eventually be helpful to understand genotype frequencies in small populations. The end of this section will consider some of the consequences of inbreeding and the evolution of autogamy.

### Impacts of inbreeding on genotype and allele frequencies

Let's develop an example to understand the impact of inbreeding on genotype and allele frequencies in a population. Under complete positive assortative mating or selfing, individuals mate with another individual possessing an identical genotype. Figure 2.12 diagrams the process of positive genotypic assortative mating for a diallelic locus, following the frequencies of each genotype through time. Initially, the frequency of the heterozygote is  $H$  but this frequency will be halved each generation. A Punnett square for two heterozygotes shows that half of the progeny are heterozygotes ( $H/2$ ). The other half of the progeny are homozygotes ( $H/2$ ), composed of one-quarter of the original heterozygote frequency of each homozygote genotype ( $H/2(1 - \frac{1}{2})$ ). It is obvious that matings among like homozygotes will produce only identical homozygotes, so the homozygote genotypes each yield a constant frequency of homozygous progeny each generation. In total, however, the frequency of the homozygous genotypes increases by a factor

of  $\frac{H}{2}\left(1 - \frac{1}{2}\right)$  each generation due to homozygous



**Figure 2.12** The impact of complete positive genotypic assortative mating (like genotypes mate) or self-fertilization on genotype frequencies. The initial genotype frequencies are represented by  $D$ ,  $H$ , and  $R$ . When either of the homozygotes mates with an individual with the same genotype, all progeny bear their parent's homozygous genotype. When two heterozygote individuals mate, the expected genotype frequencies among the progeny are one half heterozygous genotypes and one quarter of each homozygous genotype. Every generation the frequency of the heterozygotes declines by one-half while one-quarter of the heterozygote frequency is added to the frequencies of each homozygote (diagonal arrows). Eventually, the population will lose all heterozygosity although allele frequencies will remain constant. Therefore, assortative mating or self-fertilization changes the packing of alleles in genotypes but not the allele frequencies themselves.

progeny of the heterozygous genotypes. If the process of complete assortative mating continues, the population rapidly loses heterozygosity and approaches a state where the frequency of heterozygotes is 0.

As an example, imagine a population where  $p = q = 0.5$  that has Hardy–Weinberg genotype frequencies  $D = 0.25$ ,  $H = 0.5$ , and  $R = 0.25$ . Under complete positive assortative mating, what would be the frequency of heterozygotes after five generations? Using Fig. 2.12, at time  $t = 5$  heterozygosity would be  $H^{(1/2)^5} = H^{(1/32)} = 1/64$  or 0.016. This is a drastic reduction in only five generations.

Genotype frequencies change quite rapidly under complete assortative mating, but what about allele frequencies? Let's employ the same example population with  $p = q = 0.5$  and Hardy–Weinberg genotype frequencies  $D = 0.25$ ,  $H = 0.5$ , and  $R = 0.25$  to answer the question. For both of the homozygous genotypes, the initial frequencies would be  $D = R = (0.5)^2 = 0.25$ . In Fig. 2.12, the contribution of each

homozygote genotype frequency from mating among heterozygotes after five generations is  $H/2(1 - (1/2)^5) = H/2(1 - 1/32) = H/2(31/32)$ . With the initial frequency of  $H = 0.5$ ,  $H/2(31/32) = 0.242$ . Therefore, the frequencies of both homozygous genotypes are  $0.25 + 0.242 = 0.492$  after five generations. It is also apparent that the total increase in homozygotes ( $31/32$ ) is exactly the same as the total decrease in heterozygotes ( $31/32$ ), so that the allele frequencies in the population have remained constant. After five generations of assortative mating in this example, genotypes are much more likely to contain two identical alleles than they are to contain two unlike alleles. This conclusion is also reflected in the value of the fixation index for this example,  $\hat{F} = (0.5 - 0.016)/0.5 = 0.968$ . In general, *positive assortative mating or inbreeding changes the way in which alleles are “packaged” into genotypes*, increasing the frequencies of all homozygous genotypes by the same total amount that heterozygosity is decreased, but allele frequencies in a population do not change.

The fact that allele frequencies do not change over time can also be shown elegantly with some simple algebra. Using the notation in Fig. 2.12 and defining the frequency of the A allele as  $p$  and the a allele as  $q$  with subscripts to indicate generation, allele frequencies can be determined by the genotype counting method as  $p_0 = D_0 + 1/2H_0$  and  $q_0 = R_0 + 1/2H_0$ . Figure 2.12 also provides the expressions for genotype frequencies from one generation to the next:  $D_1 = D_0 + 1/4H_0$ ,  $H_1 = 1/2H_0$ , and  $R_1 = R_0 + 1/4H_0$ . We can then use these expressions to predict allele frequency in one generation:

$$p_1 = D_1 + 1/2H_1 \quad (2.14)$$

as a function of genotype frequencies in the previous generation using substitution for  $D_1$  and  $H_1$ :

$$p_1 = D_0 + 1/4H_0 + 1/2(1/2H_0) \quad (2.15)$$

which simplifies to:

$$p_1 = D_0 + 1/2H_0 \quad (2.16)$$

and then recognizing that the right hand side is equal to the frequency of A in generation 0:

$$p_1 = p_0 \quad (2.17)$$

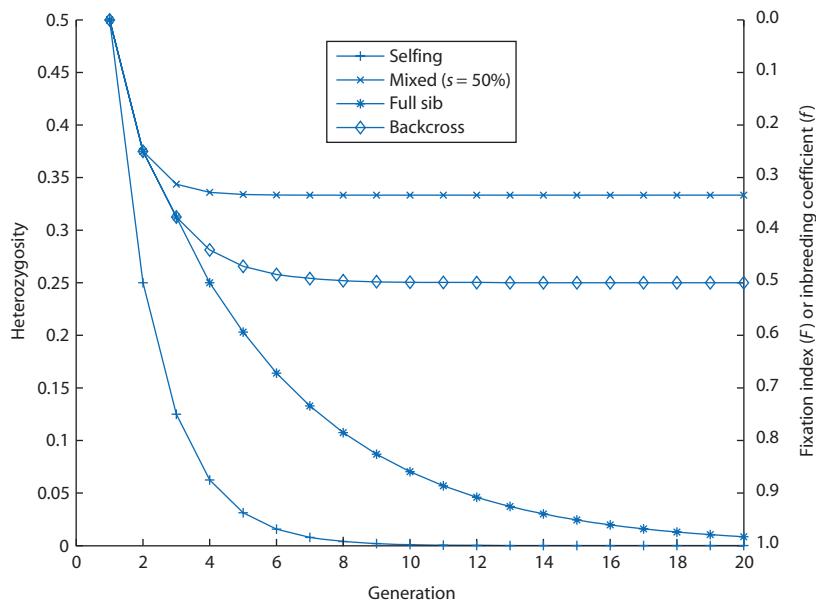
Thus, allele frequencies remain constant under complete assortative mating. As practice, you should carry out the algebra for the frequency of the a allele.

Under complete self-fertilization heterozygosity declines very rapidly. There can also be partial self-fertilization in a population (termed **mixed mating**), where some matings are self-fertilization and others are between two individuals (called **outcrossing**). In addition, many organisms are not capable of self-fertilization but instead engage in biparental inbreeding to some degree. In general, these forms of inbreeding will reduce heterozygosity compared to random mating, although they will not drive heterozygosity toward zero as in the case of complete selfing. The rate of decline in heterozygosity can be determined for many possible types of mating systems and a few examples are shown in Fig. 2.13. Regardless of the specifics of the form of consanguineous mating that occurs, it remains true that inbreeding causes alleles to be packaged more frequently as homozygotes (heterozygosity declines) and inbreeding does not alter allele frequencies in a population.

### *Inbreeding coefficient and autozygosity in a pedigree*

The effects of consanguineous mating can also be thought of as increasing the probability that two alleles at one locus in an individual are inherited from the same ancestor. Such a genotype would be homozygous and considered **autozygous** since the alleles were inherited from a common ancestor. If the two alleles are not inherited from the same ancestor in the recent past, we would call the genotype **allozygous** (*allo-* means other). You are probably already familiar with autozygosity, although you may not recognize it as such. Two times the probability of autozygosity (since diploid individuals have two alleles) is commonly expressed as the **degree of relatedness** among relatives. For example, full siblings (full brothers and sisters) are one-half related and first cousins are one-eighth related. Using a pedigree and tracing the probabilities of inheritance of an allele, the autozygosity and the basis of average relatedness can be seen.

Inbreeding in the autozygosity sense, often called the **coefficient of inbreeding** ( $f$ ), can best be seen in a pedigree such as that shown in Fig. 2.14. Fig. 2.14a gives a hypothetical pedigree for four generations. The pedigree can be used to determine the probability that the fourth-generation progeny, labeled G, have autozygous genotypes due to individual A being a common ancestor of both their maternal and paternal parents. To make the process simpler, Fig. 2.14b strips away all of the external

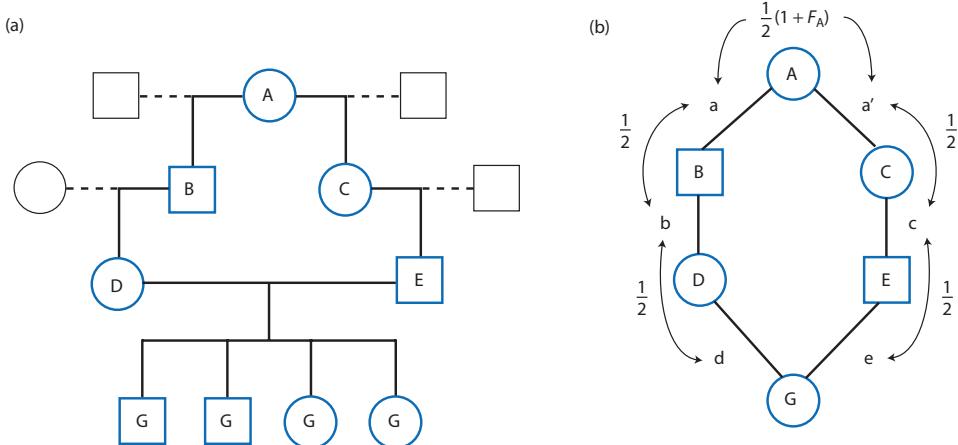


**Figure 2.13** The impact of various systems of consanguineous mating or inbreeding on heterozygosity, the fixation index ( $F$ ), and the inbreeding coefficient ( $f$ ) over time. Initially, the population has allele frequencies of  $p = q = 0.5$  and all individuals are assumed randomly mated. Since inbreeding does not change allele frequencies, expected heterozygosity ( $H_e$ ) remains 0.5 for all 20 generations. As inbreeding progresses, observed heterozygosity declines and the fixation index and inbreeding coefficient increase. Selfing is 100% self-fertilization whereas mixed mating is 50% of the population selfing and 50% randomly mating. Full sib is brother–sister or parent–offspring mating. Backcross is one individual mated to its progeny, then to its grand progeny, then to its great-grand progeny and so on, a mating scheme that is difficult to carry on for many generations. Change in the coefficient of inbreeding over time is based on the following recursion equations: selfing  $f_{t+1} = \frac{1}{2}(1 - F_t)$ ; mixed  $f_{t+1} = \frac{1}{2}(1 - F_t)(s)$  where  $s$  is the selfing rate; full sib  $f_{t+2} = \frac{1}{4}(1 + 2f_{t+1} + f_t)$ ; backcross  $f_{t+1} = \frac{1}{4}(1 + 2f_t)$  (see Falconer & MacKay 1996 for detailed derivations).

ancestors and shows only the paths where alleles could be inherited in the progeny from individual A.

To begin the process of determining the autozygosity for G, it is necessary to determine the probability that A transmitted the same allele to individuals B

and C, or in notation  $P(a = a')$ . With two alleles designated 1 and 2, there are only four possible patterns of allelic transmission from A to B and C, shown in Fig. 2.15. In only half of these cases do B and C inherit an identical allele from A, so  $P(a = a') = \frac{1}{2}$ .



**Figure 2.14** Average relatedness and autozygosity as the probability that two alleles at one locus are identical by descent. (a) A pedigree where individual A has progeny that are half-siblings (B and C). B and C then produce progeny D and E, which in turn produce offspring G. (b) Only the paths of relatedness where alleles could be inherited from A, with curved arrows to indicate the probability that gametes carry alleles identical by descent. Upper-case letters for individuals represent diploid genotypes and lower-case letters indicate allele copies within the gametes produced by the genotypes. The probability that A transmits a copy of the same allele to B and C depends on the degree of inbreeding for individual A, or  $F_A$ .

Path of inheritance	a	$a'$	a	$a'$	a	$a'$	a	$a'$	
Allele	1	1	2	2	1	2	2	1	
Alleles identical by descent (IBD)					Alleles not identical by descent				
by descent					by descent				

**Figure 2.15** The possible patterns of transmission from one parent to two progeny for a locus with two alleles. Half of the outcomes result in the two progeny inheriting an allele that is identical by descent. The a and  $a'$  refer to paths of inheritance in the pedigree in Fig. 2.14b.

This probability would still be  $\frac{1}{2}$  no matter how many alleles were present in the population, since the probability arises from the fact that diploid genotypes have only two alleles.

To have a complete account of the probability that B and C inherit an identical allele from A, we also need to take into account the past history of A's genotype since it is possible that A was itself the product of mating among relatives. If A was the product of some level of biparental inbreeding, then the chance that it transmits alleles identical by descent to B and C is greater than if A was from a randomly mating population. Another way to think of it is, with A being the product of some level of inbreeding instead of random mating, the chances that the alleles transmitted to B and C are not identical (see Fig. 2.14b) will be less than  $\frac{1}{2}$  by the amount that A is inbred. If the degree to which A is inbred (or the probability that A is autozygous) is  $F_A$ , then the total probability that B and C inherit the same allele is:

$$P(a = a') = \frac{1}{2} + \frac{1}{2}F_A = \frac{1}{2}(1 + F_A) \quad (2.18)$$

If  $F_A$  is 0 in equation 2.18, then the chance of transmitting the same allele to B and C reduces to the  $\frac{1}{2}$  expected in a randomly mating population.

For the other paths of inheritance in Fig. 2.14, the logic is similar to determine the probability that an allele is identical by descent. For example, what is the probability that the allele in gamete d is identical by descent to the allele in gamete b, or  $P(b = d)$ ? When D mated it passed on one of two alleles, with a probability of  $\frac{1}{2}$  for each allele. One allele was inherited from each parent, so there is a  $\frac{1}{2}$  chance of transmitting a maternal or paternal allele. This makes  $P(b = d) = \frac{1}{2}$ . (Just like with individual A,  $P(b = d)$  could also be increased to the extent that B was inbred, although random mating for all genotypes but A is assumed here for simplicity.) This same logic applies to all other paths in the pedigree that connect A and the

progeny G. The probability of a given allele being transmitted along a path is independent of the probability along any other path, so the probability of autozygosity (symbolized as  $f$  to distinguish it from the pre-existing autozygosity of individual A) over the entire pedigree for any of the G progeny is:

$$\begin{aligned} f_G = f_{DE} &= \frac{1/2}{P(b = d)} \times \frac{1/2}{P(a = b)} \times \frac{1/2(1 + f_A)}{P(a = a')} \times \frac{1/2}{P(a' = c)} \times \frac{1/2}{P(c = e)} \\ &= (\frac{1}{2})^5(1 + f_A) = \frac{1}{32}(1 + f_A) \end{aligned} \quad (2.19)$$

since independent probabilities can be multiplied to find the total probability of an event. This is equivalent to the average relatedness among half-cousins. In general for pedigrees,  $f = (\frac{1}{2})^i(1 + F_A)$  where A is the common ancestor and  $i$  is the number of paths or individuals over which alleles are transmitted. A trick is to write down the chain of individuals starting with the common ancestor and ending with the individuals of interest and count the individuals along paths of inheritance (not including the individuals of interest). That gives ~~GDBACEG~~ or five ancestors, yielding a result identical to equation 2.19.

Although it is useful to determine the inbreeding coefficient (autozygosity) for a specific pedigree, the more general point is to see mating among relatives as a process that increases autozygosity in a population. When individuals have common relatives, the chance that their genotype contains loci with alleles identical by descent is increased. Further, the inbreeding coefficient or autozygosity measured for a specific pedigree is identical in concept and interpretation to the departure from Hardy–Weinberg measured by the fixation index (assuming most of the deviation from Hardy–Weinberg expectations is caused by non-random mating and not other population genetic processes). Both express the probability that two alleles in a genotype are identical due to common ancestry.

The departure from Hardy–Weinberg expected genotype frequencies, the autozygosity or inbreeding coefficient, and the fixation index are all interrelated. Another way of stating the results that were developed in Fig. 2.12 is that  $f$  measures the degree to which Hardy–Weinberg genotype proportions are not met, due to inbreeding:

$$\begin{aligned} D &= p^2 + fpq \\ H &= 2pq - f2pq \\ R &= q^2 + fpq \end{aligned} \quad (2.20)$$

With consanguineous mating, the decline in heterozygosity is proportional to the increase in the inbreeding coefficient, shown by substituting  $H_e$  for  $2pq$  in equation 2.20 and rearranging to give

$$H = H_e(1 - f) \quad (2.21)$$

where  $H_e$  is the Hardy–Weinberg expected heterozygosity based on population allele frequencies. Rearranging equation 2.21 in terms of the inbreeding coefficient gives:

$$f = 1 - \frac{H}{H_e} \quad (2.22)$$

This is really exactly the same quantity as the fixation index (equation 2.9)

$$f = 1 - \frac{H}{H_e} = \frac{H_e}{H_e} - \frac{H}{H_e} = \frac{H_e - H}{H_e} \quad (2.23)$$

The inbreeding coefficient and the fixation index are measures of excess homozygosity and therefore are just different ways of expressing the heterozygosity. Returning to Fig. 2.13 helps show the equivalence of the inbreeding coefficient, the fixation index, and the decline in heterozygosity in several specific cases of regular consanguineous mating. Remember that in all cases in Fig. 2.13, the Hardy–Weinberg expected heterozygosity is  $1/2$ .

**Allozygous genotype** A homozygous or heterozygous genotype composed of two alleles not inherited from a recent common ancestor.

**Autozygosity or inbreeding coefficient ( $f$ )** The probability that two alleles in a homozygous genotype are identical by descent.

**Autozygous genotype** A homozygous genotype composed of two identical alleles that are inherited from a common ancestor.

**Identity by descent** Transmission from a common ancestor.

**Relatedness** The expected proportion of alleles between two individuals that are identical by descent; twice the autozygosity.

### Phenotypic consequences of inbreeding

The process of consanguineous mating or inbreeding is associated with changes in the mean phenotype within a population. These changes arise from two general causes: changes in genotype frequencies in a population per se and fitness effects associated with changes in genotype frequencies.

The mean phenotype of a population will be impacted by the changes in genotype frequency caused by inbreeding. To show this it is necessary to introduce terminology to express the phenotype associated with a given genotype, a topic covered in much greater detail and explained more fully in Chapters 9 and 10 in this volume. We will assign AA genotypes the phenotype  $+a$ , heterozygotes the phenotype  $d$ , and aa homozygotes the phenotype  $-a$ . Each genotype contributes to the overall phenotype based on how frequent it is in the population. The mean phenotype in a population is then the sum of each genotype-frequency-weighted phenotype (Table 2.10). When there is no dominance, the phenotype of the heterozygotes is exactly intermediate between the phenotypes of the two homozygotes and  $d = 0$ . In that case, it is easy to see that inbreeding will not change the mean phenotype in the population since both homozygous genotypes increase by the same amount and their effects on the mean phenotype cancel out (mean =  $ap^2 + aq^2 + d2pq - d^2pq - aq^2 - apq$ , where the heterozygote terms are crossed out since  $d = 0$ ). When there is some degree of dominance (positive  $d$  indicates the phenotype of Aa is like that of AA while negative  $d$  indicates the phenotype of Aa is like that of aa), then the mean phenotype of the population will change with consanguineous mating since heterozygotes will become less frequent. If dominance is in the direction of the  $+a$  phenotype ( $d > 0$ ), then inbreeding will reduce the population mean because the heterozygote frequency will drop. Similarly, if dominance is in the direction of  $-a$  ( $d < 0$ ) then inbreeding will increase the population mean again because the heterozygote frequency decreases. It is also true in the case of dominance that a return to random mating will restore the frequencies of heterozygotes and return the population mean to its original value before inbreeding. These changes in the population mean phenotype are simply a consequence of changing the genotype frequencies when there is no change in the allele frequencies.

There is a wealth of evidence that inbreeding has **deleterious** (harmful or damaging) consequences

**Table 2.10** The mean phenotype in a population that is experiencing consanguineous mating. The inbreeding coefficient is  $f$  and  $d = 0$  when there is no dominance.

Genotype	Phenotype	Frequency	Contribution to population mean
AA	$+a$	$p^2 + fpq$	$ap^2 + afpq$
Aa	$d$	$2pq - f2pq$	$d2pq - df2pq$
aa	$-a$	$q^2 + fpq$	$-aq^2 - afpq$

$$\text{Population mean: } ap^2 + d2pq - df2pq - aq^2 = a(p - q) + d2pq(1 - f)$$

and is associated with a decline in the average phenotype in a population, a phenomenon referred to as **inbreeding depression**. Since the early twentieth century, studies in animals and plants that have been intentionally inbred provide ample evidence that decreased performance, growth, reproduction, viability (all measures of fitness), and abnormal phenotypes are associated with consanguineous mating. A related phenomenon is **heterosis** or hybrid vigor, characterized by beneficial consequences of increased heterozygosity such as increased viability and reproduction, or the reverse of inbreeding depression. One example is the heterosis exhibited in corn, which has lead to the nearly universal use of F1 hybrid seed for agriculture in developed countries.

**Heterosis** The increase in performance, survival, and ability to reproduce of individuals possessing heterozygous loci (hybrid vigor); increase in the population average phenotype associated with increased heterozygosity.

**Inbreeding depression** The reduction in performance, survival, and ability to reproduce of individuals possessing homozygous loci; decrease in population average phenotype associated with consanguineous mating that increases homozygosity.

There is evidence that humans experience inbreeding depression, based on observed phenotypes in the offspring of couples with known consanguinity. For example, mortality among children of first-cousin marriages was 4.5% greater than for marriages between unrelated individuals measured in a range of human populations (see review by Jorde 1997). Human studies have utilized existing parental pairs with relatively low levels of inbreeding, such as uncle/niece, first cousins, or second cousins, in contrast

to animal and plant studies where both very high levels and a broad range of inbreeding coefficients are achieved intentionally. Drawing conclusions about the causes of variation in phenotypes from such observational studies requires extreme caution, since the prevalence of consanguineous mating in humans is also correlated with social and economic variables such as illiteracy, age at marriage, duration of marriage, and income. These latter variables are therefore not independent of consanguinity and can themselves contribute to variation in phenotypes such as fertility and infant mortality (see Bittles et al. 1991, 2002).

The Mendelian genetic causes of inbreeding depression have been a topic of population genetics research for more than a century. There are two classical hypotheses to explain inbreeding depression and changes in fitness as the inbreeding coefficient increases (Charlesworth & Charlesworth 1999; Carr & Dudash 2003). Both hypotheses predict that levels of inbreeding depression will increase along with consanguineous mating that increases homozygosity, although for different reasons (Table 2.11). The first hypothesis, often called the **dominance hypothesis**, is that increasing homozygosity increases the phenotypic expression of fully and partly recessive alleles with deleterious effects. The second hypothesis is that inbreeding depression is the result of the decrease in the frequency of heterozygotes that occurs with consanguineous mating. This explanation supposes that heterozygotes have higher fitness than homozygotes (heterosis) and is called the **overdominance hypothesis**. In addition, the fitness interactions of alleles at different loci (epistasis; see Chapter 9) may also cause inbreeding depression, a hypothesis that is particularly difficult to test (see Carr & Dudash 2003). These causes of inbreeding depression may all operate simultaneously.

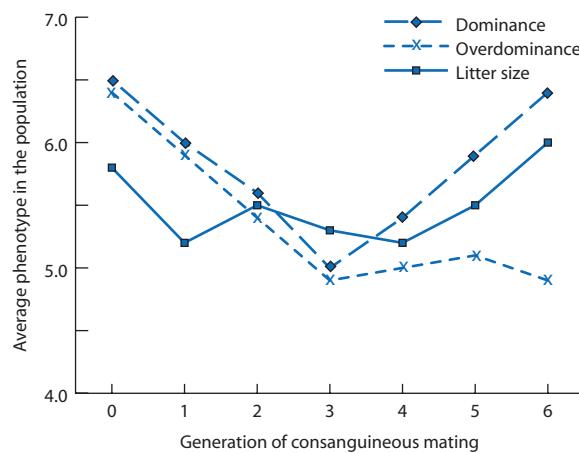
These dominance and overdominance hypotheses make different testable predictions about

**Table 2.11** A summary of the Mendelian basis of inbreeding depression under the dominance and overdominance hypotheses along with predicted patterns of inbreeding depression with continued consanguineous mating.

Hypothesis	Mendelian basis	Low-fitness genotypes	Changes in inbreeding depression with continued consanguineous mating
Dominance	Recessive and partly recessive deleterious alleles	Only homozygotes for deleterious recessive alleles	Purging of deleterious alleles that is increasingly effective as degree of recessiveness increases
Overdominance	Heterozygote advantage or heterosis	All homozygotes	No changes as long as consanguineous mating keeps heterozygosity low

how inbreeding depression (measured as the average phenotype of a population) will change over time with continued consanguineous mating. Under the dominance hypothesis, recessive alleles that cause lowered fitness are more frequently found in homozygous genotypes under consanguineous mating. This exposes the deleterious phenotype and the genotype will decrease in frequency in a population by natural selection (individuals homozygous for such alleles have lower survivorship and reproduction). This reduction in the frequency of deleterious alleles by natural selection is referred to as **purging of genetic load**. Purging increases the frequency of alleles that do not have deleterious effects when homozygous, so that the average phenotype in a population then returns to the initial average it had before the onset of consanguineous mating. In contrast, the overdominance hypothesis does not predict a purging effect with consanguineous mating. With consanguineous mating, the frequency of heterozygotes will decrease and not recover until mating patterns change (see Fig. 2.12). Even if heterozygotes are frequent and have a fitness advantage, each generation of mating and Mendelian segregation will reconstitute the two homozygous genotypes so purging cannot occur. These predictions highlight the major difference between the hypotheses. Inbreeding depression with overdominance arises from genotype frequencies in a population while inbreeding depression with dominance is caused by the frequency of deleterious recessive alleles in a population. Models of natural selection that are relevant to inbreeding depression on population genotype and allele frequencies receive detailed coverage in Chapter 6 in this volume.

Inbreeding depression in many animals and plants appears to be caused, at least in part, by deleterious recessive alleles consistent with the dominance hypothesis (Byers & Waller 1999; Charlesworth & Charlesworth 1999; Crnokrak & Barrett 2002). A classic example of inbreeding depression and recovery of the population mean for litter size in mice is shown in Fig. 2.16. Model research organisms



**Figure 2.16** A graphical depiction of the predictions of the dominance and overdominance hypotheses for the genetic basis of inbreeding depression. The line for dominance shows purging and recovery of the population mean under continued consanguineous mating expected if deleterious recessive alleles cause inbreeding depression. However, the line for overdominance as the basis of inbreeding depression shows no purging effect since heterozygotes continue to decrease in frequency. The results of an inbreeding depression experiment with mice show that litter size recovers under continued brother–sister mating as expected under the dominance hypothesis (Lynch 1977). Only two of the original 14 pairs of wild-caught mice were left at the sixth generation. Not all of the mouse phenotypes showed patterns consistent with the dominance hypothesis.

such as mice, rats, and *Drosophila*, intentionally inbred by schemes such as full-sib mating for 10s or 100s of generations to create highly homozygous, so-called pure-breeding lines, are also not immune to inbreeding depression. Such inbred lines are often founded from multiple families and many of these family lines go extinct from low viability or reproductive failure with habitual inbreeding. This is another type of purging effect due to natural selection that leaves only those lines that exhibit less inbreeding depression, which could be due to dominance, overdominance, or epistasis. Purging is not universally observed in all species and it is likely that inbreeding depression has several genetic causes within species as well as different predominant causes among different species.

The social and economic correlates of inbreeding depression in humans mentioned above are a specific example of environmental effects on phenotypes. Inbreeding depression can be more pronounced when environmental conditions are more severe or limiting. For example, in the plant rose pink (*Sabatia angularis*), progeny from self-fertilizations showed decreasing relative performance when grown in the greenhouse, a garden, and their native habitat, consistent with environmental contributions to the expression of inbreeding depression (Dudash 1990). In another study, the number of surviving progeny for inbred and random-bred male wild mice (*Mus domesticus*) were similar under laboratory conditions, but inbred males sired only 20% of the surviving progeny that random bred males did when under semi-natural conditions due to male–male competition (Meagher et al. 2000). However, not all studies show environmental differences in the expression of inbreeding depression. As an example, uniform levels of inbreeding depression were shown by mosquitoes grown in the laboratory and in natural tree holes where they develop as larvae and pupae in the wild (Armbruster et al. 2000).

The degree of inbreeding depression also depends on the phenotype being considered. In plants, traits early in the life cycle such as germination less often show inbreeding depression than traits later in the life cycle such as growth and reproduction (Husband & Schemske 1996). A similar pattern is apparent in animals, with inbreeding depression most often observed for traits related to survival and reproduction.

Inbreeding depression is a critical concept when thinking about the evolution of mating patterns in plants and animals. Suppose that a single locus

determines whether an individual will self or outcross and the only allele present in a population is the outcrossing allele. Then imagine that mutation produces an allele at that locus, which, when homozygous, causes an individual to self-fertilize. Such a selfing allele would have a transmission advantage over outcrossing alleles in the population. To see this, consider the number of allele copies at the mating locus transmitted from parents to progeny. Parents with outcrossing alleles mate with another individual and transmit one allele to their progeny. Self-fertilizing parents, however, are both mom and dad to their offspring and transmit two alleles to their progeny. In a population of constant size where each individual contributes an average of one progeny to the next generation, the selfing allele is reproduced twice as fast as an outcrossing allele and would rapidly become fixed in the population (Fisher 1999; see Lande & Schemske 1985). Based on this two-fold higher rate of increase of the selfing allele, complete self-fertilization would eventually evolve unless some disadvantage counteracted the increase of selfed progeny in the population. Inbreeding depression where the average fitness of outcrossed progeny exceeds the average fitness of selfed progeny by a factor of two could play this role. If outcrossed progeny are at a two-fold advantage due to inbreeding depression, then complete outcrossing would evolve. Explaining the existence of populations that engage in intermediate levels of selfing and outcrossing, a mating system common in plants, remains a challenge under these predictions (Byers & Waller 1999).

### ***The many meanings of inbreeding***

Unfortunately, the word inbreeding is used as a generic term to describe multiple distinct, although interrelated, concepts in population genetics (Jacquard 1975; Templeton & Read 1994). Inbreeding can apply to:

- consanguinity or kinship of two different individuals;
- autozygosity of two alleles either within an individual or sampled at random;
- the fixation index and Hardy–Weinberg expected and observed genotype frequencies, especially when there is an excess of homozygotes;
- inbreeding depression;
- the description of the mating system of a population or species (as in inbred);

- genetic subdivision of a species into populations that exchange limited levels of gene flow such that individual populations increase in autozygosity;
- the increase in homozygosity in a population due to its finite size.

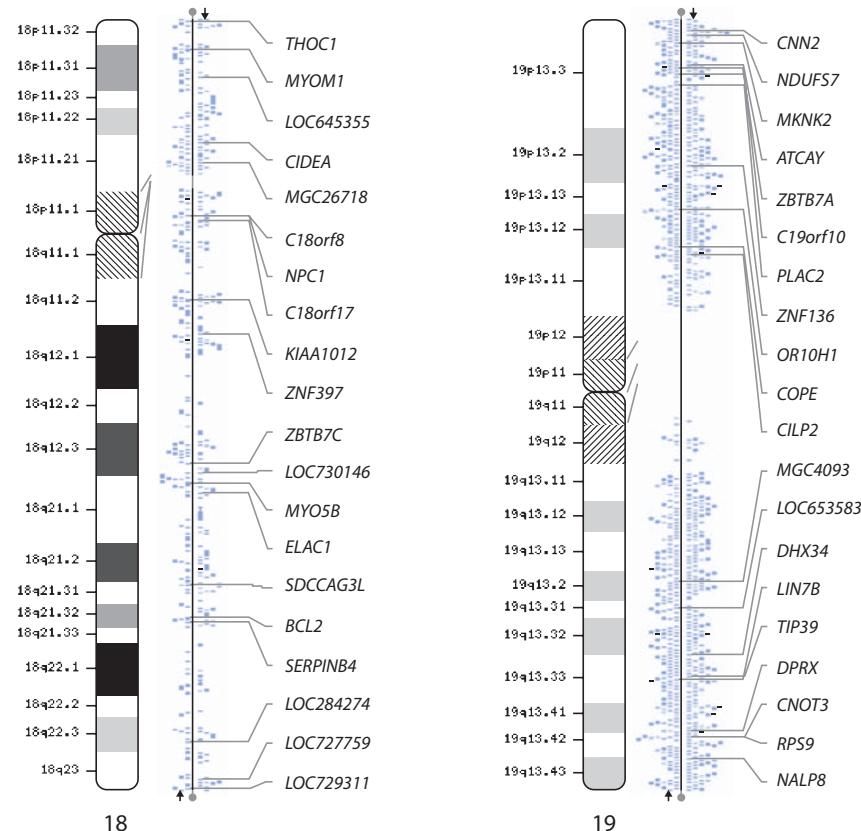
These different concepts all relate in some way to either the autozygosity or to genotype frequencies in a population, so the connection to inbreeding is clear. Awareness of the different ways the word inbreeding is used as well as an understanding of these different uses will prevent confusion, which can often be avoided simply by using more specific terminology. Remembering that the concepts are interrelated under the general umbrella of inbreeding can also help in realizing the equivalence of the population genetic processes in operation. The next chapter will show how finite population size is equivalent in its effects to inbreeding. Chapter 4 will take up the topic of population subdivision.

## 2.7 Gametic disequilibrium

- Estimating gametic disequilibrium with  $D$ .
- Approach to gametic equilibrium over time.
- Causes of gametic disequilibrium.

In 1902 Walter Sutton and Theodor Boveri advanced the chromosome theory of heredity. They observed cell division and hypothesized that the discrete bodies seen separating into sets at meiosis and mitosis contained hereditary material that was transmitted from parents to offspring. At the time the concept of chromosomal inheritance presented a paradox. Mendel's second law says that gamete **haplotypes** (**haploid genotype**) should appear in frequencies proportional to the product of allele frequencies. This prediction conflicted with the chromosome theory of heredity since there are not enough chromosomes to represent each hereditary trait.

To see the problem, take the example of *Homo sapiens* with a current estimate of around 30,000 genes in the nuclear genome. However, humans have only 23 pairs of chromosomes. There are a large number of loci but a small number of chromosomes. So if chromosomes are indeed hereditary molecules, many genes must be on the same chromosome (on average about 1300 genes per chromosome for humans if there are 30,000 genes). This means that some genes are physically **linked** by being located on the same chromosome (see Fig. 2.17). The solution to the paradox is the process of **recombination**. Sister chromatids touch at random points during

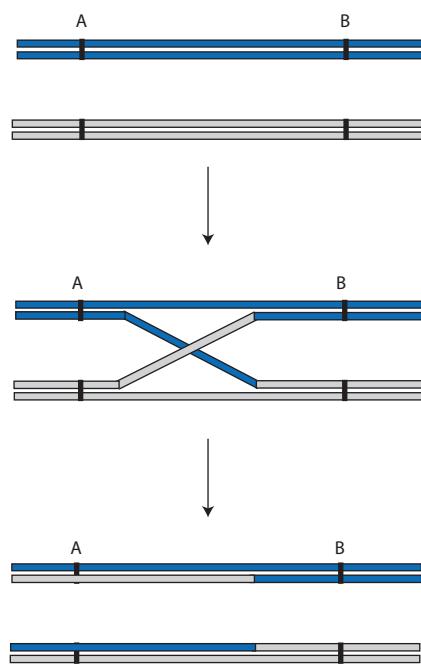


**Figure 2.17** Maps for human chromosomes 18 (left) and 19 (right) showing chromosome regions, the physical locations of identified genes and open reading frames (labeled orf) along the chromosomes, and the names and locations of a subset of genes. Chromosome 18 is about 85 million bp and chromosome 19 is about 67 million bp. Maps from NCBI Map Viewer based on data as of January 2008.

meiosis and exchange short segments, a process known as crossing-over (Fig. 2.18).

Linkage of loci has the potential to impact multi-locus genotype frequencies and violate Mendel's law of independent segregation, which assumes the absence of linkage. To generalize expectations for genotype frequencies for two (or more) loci requires a model that accounts explicitly for linkage by including the rate of recombination between loci. The effects of linkage and recombination are important determinants of whether or not expected genotype frequencies under independent segregation of two loci (Mendel's second law) are met. Autosomal linkage is the general case that will be used to develop expectations for genotype frequencies under linkage.

The frequency of a two-locus gamete haplotype will depend on two factors: (i) allele frequencies and (ii) the amount of recombination between the two loci. We can begin to construct a model based on the recombination rate by asking what gametes are generated by the genotype  $A_1A_2B_1B_2$ . Throughout



**Figure 2.18** A schematic diagram of the process of recombination between two loci, A and B. Two double-stranded chromosomes (drawn in different colors) exchange strands and form a Holliday structure. The crossover event can resolve into either of two recombinant chromosomes that generate new combinations of alleles at the two loci. The chance of a crossover event occurring generally increases as the distance between loci increases. Two loci are independent when the probability of recombination and non-recombination are both equal to  $1/2$ . Gene conversion, a double crossover event without exchange of flanking strands, is not shown.

this section loci are indicated by the letters, alleles at the loci by the numerical subscripts and allele frequencies indicated by  $p_1$  and  $p_2$  for locus A and  $q_1$  and  $q_2$  for locus B. The problem is easier to conceptualize if we draw the two locus genotype as being on two lines akin to chromosomal strands

$$\begin{array}{c} A_1 \quad B_1 \\ \hline A_2 \quad B_2 \end{array}$$

Given this physical arrangement of the two loci, what are the gametes produced during meiosis with and without recombination events?

$A_1B_1$  and  $A_2B_2$  “Coupling” gametes: alleles on the same chromosome remain together (term coined by Bateson and Punnett).

$A_1B_2$  and  $A_2B_1$  “Repulsion” gametes: alleles on the same chromosome seem repulsed by each other and pair with alleles on the opposite strand (term coined by Thomas Hunt Morgan).

The **recombination fraction**, symbolized as  $r$  (or sometimes  $c$ ), refers to the total frequency of gametes resulting from recombination events between two loci. Using  $r$  to express an arbitrary recombination fraction, let's build an expectation for the frequency of coupling and repulsion gametes. If  $r$  is the rate of recombination, then  $1 - r$  is the rate of non-recombination since the frequency of all gametes is one, or 100%. Within each of these two categories of gametes (coupling and repulsion), two types of gametes are produced so the frequency of each gamete type is half that of the total frequency for the gamete category. We can also determine the expected frequencies of each gamete under random association of the alleles at the two loci based on Mendel's law of independent segregation.

Gamete	Frequency		
	Expected	Observed	
$A_1B_1$	$p_1q_1$	$g_{11} = (1 - r)/2$	$1 - r$ is the frequency of all coupling gametes.
$A_2B_2$	$p_2q_2$	$g_{22} = (1 - r)/2$	
$A_1B_2$	$p_1q_2$	$g_{12} = r/2$	$r$ is the frequency of all recombinant gametes.
$A_2B_1$	$p_2q_1$	$g_{21} = r/2$	

The recombination fraction,  $r$ , can be thought of as the probability that a recombination event will occur between two loci. With independent assortment, the coupling and repulsion gametes are in equal frequencies and  $r$  equals  $1/2$  (like the chances of getting heads when flipping a coin). Values of  $r$  less than  $1/2$  indicate that recombination is less likely than non-recombination, so coupling gametes are more frequent. Values of  $r$  greater than  $1/2$  are possible and would indicate that recombinant gametes are more frequent than non-recombinant gametes (although such a pattern would likely be due to a process such as natural selection eliminating coupling gametes from the population rather than recombination exclusively).

The expected frequencies of gametes produced by all possible genotypes for two diallelic loci including the contributions of recombination are derived in Table 2.12. Summing the frequencies of each gamete produced by all genotypes gives the gamete frequencies that will found the next generation. This gives expected gamete frequencies for the more general case of a randomly mating population rather than for a single genotype. The table shows how only two of 10 possible genotypes contribute to the production of recombinant gametes. Most genotypes produce recombinant gametes that are identical to non-recombinant gametes (e.g. the  $A_1B_1/A_1B_2$  genotype produces  $A_1B_1$  and  $A_1B_2$  coupling gametes and  $A_1B_1$  and  $A_1B_2$  repulsion gametes).

We can utilize observed gamete frequencies to develop a measure of the degree to which alleles are

associated within gamete haplotypes. This quantity is called the **gametic disequilibrium** (or sometimes **linkage disequilibrium**) **parameter** and can be expressed by:

$$D = \frac{g_{11}g_{22}}{\text{(coupling term)}} - \frac{g_{12}g_{21}}{\text{(repulsion term)}} \quad (2.24)$$

where  $g_{xy}$  stands for a gamete frequency.  $D$  is the difference between the product of the coupling gamete frequencies and the product of the repulsion gamete frequencies. This makes intuitive sense: with independent assortment the frequencies of the coupling and repulsion gamete types are identical and cancel out to give  $D = 0$ , or **gametic equilibrium**. Another way to think of the gametic disequilibrium parameter is as a measure of the difference between observed and expected gamete frequencies:  $g_{11} = p_1q_1 + D$ ,  $g_{22} = p_2q_2 + D$ ,  $g_{12} = p_1q_2 - D$ , and  $g_{21} = p_2q_1 - D$  (note that observed and expected gamete frequencies cannot be negative). In this sense,  $D$  measures the deviation of gamete frequencies from what is expected under independent assortment. Since  $D$  can be both positive as well as negative, both coupling and repulsion gametes can be in excess or deficit relative to the expectations of independent assortment.

Gametic disequilibrium can be measured using several estimators, including the squared correlation coefficient ( $\rho^2$ , where  $\rho$  is pronounced “rho”), where  $\rho^2 = D^2/(p_1p_2q_1q_2)$ , which has a range of  $-1$  to  $+1$ . Different estimators of gametic disequilibrium

**Table 2.12** Expected frequencies of gametes for two diallelic loci in a randomly mating population with a recombination rate between the loci of  $r$ . The first eight genotypes have non-recombinant and recombinant gametes that are identical. The last two genotypes produce novel recombinant gametes, requiring inclusion of the recombination rate to predict gamete frequencies. Summing down each column of the table gives the total frequency of each gamete in the next generation.

Genotype	Expected frequency	Frequency of gametes in next generation			
		$A_1B_1$	$A_2B_2$	$A_1B_2$	$A_2B_1$
$A_1B_1/A_1B_1$	$(p_1q_1)^2$	$(p_1q_1)^2$			
$A_2B_2/A_2B_2$	$(p_2q_2)^2$		$(p_2q_2)^2$		
$A_1B_1/A_2B_2$	$2(p_1q_1)(p_1q_2)$	$(p_1q_1)(p_1q_2)$		$(p_1q_1)(p_1q_2)$	
$A_1B_1/A_2B_1$	$2(p_1q_1)(p_2q_1)$	$(p_1q_1)(p_2q_1)$			$(p_1q_1)(p_2q_1)$
$A_2B_2/A_1B_2$	$2(p_2q_2)(p_1q_2)$		$(p_2q_2)(p_1q_2)$	$(p_2q_2)(p_1q_2)$	
$A_2B_2/A_2B_1$	$2(p_2q_2)(p_2q_1)$		$(p_2q_2)(p_2q_1)$		$(p_2q_2)(p_2q_1)$
$A_1B_2/A_1B_2$	$(p_1q_2)^2$			$(p_1q_2)^2$	
$A_2B_1/A_2B_1$	$(p_2q_1)^2$				$(p_2q_1)^2$
$A_2B_2/A_1B_1$	$2(p_2q_2)(p_1q_1)$	$(1 - r)(p_2q_2)(p_1q_1)$	$(1 - r)(p_2q_2)(p_1q_1)$	$r(p_2q_2)(p_1q_1)$	$r(p_2q_2)(p_1q_1)$
$A_1B_2/A_2B_1$	$2(p_1q_2)(p_2q_1)$	$r(p_1q_2)(p_2q_1)$	$r(p_1q_2)(p_2q_1)$	$(1 - r)(p_1q_2)(p_2q_1)$	$(1 - r)(p_1q_2)(p_2q_1)$

have slightly different strengths and weaknesses (see Hedrick 1987; Flint-Garcia et al. 2003). The discussion here will focus on the classical estimator  $D$  to develop the conceptual basis of measuring gametic disequilibrium and to understand the genetic processes that cause it.

**Gametic disequilibrium** The non-random association or combination of alleles at multiple loci in a sample of gametes or haplotypes.

**Linkage** Co-inheritance of loci caused by physical location on the same chromosome.

**Recombination fraction** The proportion of “repulsion” or recombinant gametes produced by a double heterozygote genotype each generation.

Now that we have developed an estimator of gametic disequilibrium, it can be used to understand how allelic association at two loci changes over time or its dynamic behavior. If a population starts out with some level of gametic disequilibrium, what happens to  $D$  over time with recombination? Imagine a population with a given level of gametic disequilibrium at the present time ( $D_{t=n}$ ). How much gametic disequilibrium was there a single generation before the present at generation  $n - 1$ ? Recombination will produce  $r$  recombinant gametes each generation so that:

$$D_{tn} = (1 - r)D_{tn-1} \quad (2.25)$$

Since gametic disequilibrium decays by a factor of  $1 - r$  each generation,

$$\begin{aligned} D_{tn} &= (1 - r)D_{tn-1} = (1 - r)^2 D_{tn-2} \\ &= (1 - r)^3 D_{tn-3} \dots \end{aligned} \quad (2.26)$$

We can predict the amount of gametic disequilibrium over time by using the amount of disequilibrium initially present ( $D_{t0}$ ) and multiplying it by  $(1 - r)$  raised to the power of the number of generations that have elapsed:

$$D_{tn} = D_{t0}(1 - r)^n \quad (2.27)$$

Figure 2.19 shows the decay of gametic disequilibrium over time using equation 2.27. Initially there are only coupling gametes in the population and no repulsion gametes, giving a maximum amount of

gametic disequilibrium. As  $r$  increases, the approach to gametic equilibrium ( $D = 0$ ) is more rapid. Equation 2.27 and Fig. 2.19 both assume that there are no other processes acting to counter the mixing effect of recombination. Therefore, the steady-state will always be equal frequencies of all gametes ( $D = 0$ ), with the recombination rate determining how rapidly gametic equilibrium is attained.

### Interact box 2.4 Decay of gametic disequilibrium and a $\chi^2$ test

The program PopGene.S<sup>2</sup> has two simulation modules under the **Gametic Disequilibrium** menu that demonstrate two features of gametic disequilibrium covered so far.

The **Magnitude of  $D$**  simulation can be used to model and graph the decay in  $D$  over time given an initial value of  $D$  and a value for the recombination rate. This module can be used to produce graphs like that shown in Fig. 2.19.

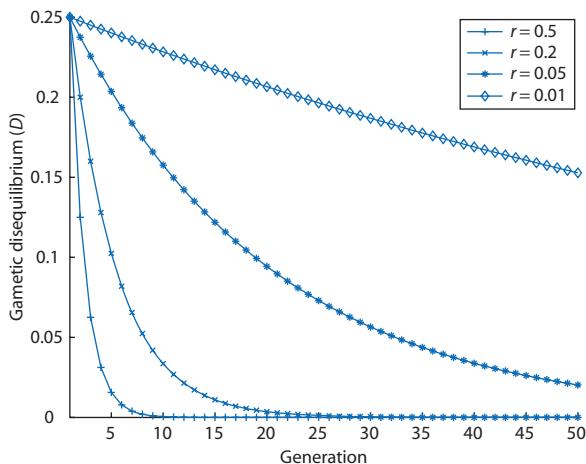
The **Gamete frequencies** simulation can be used to conduct a hypothesis test using a  $\chi^2$  test that the observed level of gametic disequilibrium is significantly different than expected under random segregation.

Another way to carry out this test without the aid of PopGene.S<sup>2</sup> is to use

$$\chi^2 = \frac{D^2 N}{p_1 p_2 q_1 q_2} \quad (2.28)$$

where  $N$  is the total sample size of gametes,  $D$  is the gametic disequilibrium parameter, and  $p$  and  $q$  are the allele frequencies at two diallelic loci. The  $\chi^2$  value has 1 df and can be compared with the critical value found in Table 2.5.

One potential drawback of  $D$  in equation 2.24 is that its maximum value depends on the allele frequencies in the population. This can make interpreting an estimate of  $D$  or comparing estimates of  $D$  from different populations problematic. For example, it is possible that two populations have very strong association among alleles within gametes (e.g. no repulsion gametes), but the two populations differ



**Figure 2.19** The decay of gametic disequilibrium ( $D$ ) over time for four recombination rates. Initially, there are only coupling ( $P_{11} = P_{22} = 1/2$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). Gametic disequilibrium decays as a function of time and the recombination rate ( $D_{tn} = D_{t0}(1 - r)^n$ ) assuming a single large population, random mating, and no counteracting genetic processes. If all gametes were initially repulsion, gametic disequilibrium would initially equal  $-0.25$  and decay to 0 in an identical fashion.

in allele frequency so that the maximum value of  $D$  in each population is also different. If all alleles are not at equal frequencies in a population, then the frequencies of the two coupling or the two repulsion gametes are also not equal. When  $D < 0$ ,  $D_{max}$  is the larger of  $-p_1q_1$  or  $-p_2q_2$ , whereas when  $D > 0$ ,  $D_{max}$  is the smaller of  $p_1q_2$  or  $p_2q_1$ .

A way to avoid these problems is to express  $D$  as the percentage of its largest value:

$$D' = D/D_{max} \quad (2.29)$$

This gives a measure of gametic disequilibrium that is normalized by the maximum or minimum value  $D$  can assume given population allele frequencies. Even though a given value of  $D$  may seem small in the absolute, it may be large relative to  $D_{max}$  given the population allele frequencies.

Measures of gametic disequilibrium can be used to test fundamental hypotheses regarding the processes that shape genotype frequencies in natural populations. In epidemiology, pathogens are often considered to reproduce predominantly clonally even if capable of sex and recombination. This assumption has implications for clinical treatment of infections, the emergence of new virulent strains, and vaccine development strategies. Clonal reproduction would accompany high levels of gametic disequilibrium since recombination would not occur, a hypothesis

that can be tested with genetic marker data. The protozoan parasite *Toxoplasma gondii* is one such example. It infects all mammals and birds and causes toxoplasmosis, an illness in humans, and was considered clonal despite sexual reproduction that occurs in cats. To test this hypothesis, Lehmann et al. (2004) sampled *T. gondii* from pigs, chickens, and cats, and then genotyped the protozoa at seven loci. The results revealed normalized gametic disequilibrium ( $|D'|$ ) between pairs of loci that ranged from 0.35 to 0.96. The two loci known to be physically linked showed the highest values of  $D'$  while all others have less gametic disequilibrium. This pattern is inconsistent with clonal reproduction, which would maintain gametic disequilibrium at all loci regardless of the physical distance between loci.

$D$  is frequently called the linkage disequilibrium parameter rather than the gametic disequilibrium parameter. This is a misnomer, since physical linkage only dictates the rate at which allelic combinations approach independent assortment. Recombination, determined by the degree of linkage, only causes a *reduction* in gametic disequilibrium over time, but it cannot cause an increase in gametic disequilibrium. Processes other than linkage are responsible for the production of deviations from independent assortment of alleles at multiple loci in gametes. Using the term gametic disequilibrium reminds us that the deviation from random association of alleles at two loci is a pattern seen in gametes or haplotypes. Although linkage can certainly contribute to this pattern, so can a number of other population genetic processes. It is even possible that several processes operating simultaneously produce a given pattern of gametic disequilibrium. Processes that maintain or increase gametic disequilibrium include those discussed in the following sections.

### Physical linkage

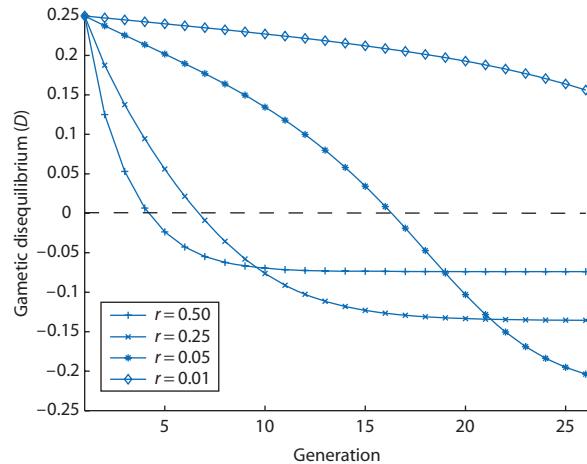
Linkage is the physical association of loci on a chromosome that causes alleles at the loci to be inherited in their original combinations. This association of alleles at loci on the same chromosome is broken down by crossing over and recombination. The probability that a recombination event occurs between two loci is a function of the distance along the chromosome between two loci. Loci that are very far apart (or on separate chromosomes) have recombination rates approaching 50% and are said to be unlinked. Loci located very near each other on the same chromosome might have recombination

rates of 5 or 1% and would be described as tightly linked. Therefore, the degree of physical linkage of loci dictates the recombination rate and thereby the decay of gametic disequilibrium.

Linkage-like effects can be seen in some chromosomes and genomes where gametic disequilibrium is expected to persist over longer time scales due to exceptional inheritance or recombination patterns. Organisms such as birds and mammals have chromosomal sex determination, as with the well-known X and Y sex chromosome system in humans. Loci located on X chromosomes experience recombination normally whereas those on Y chromosomes experience no recombination. This is caused by the Y chromosome lacking a homologous chromosome to pair with at meiosis since YY genotypes do not exist. In addition, we would expect that the rate of decay of gametic disequilibrium for X chromosomes is about half that of autosomes with comparable recombination rates, since X recombination takes place only in females (XX) at meiosis, and not at all in males (XY). Organelle genomes found in mitochondria and chloroplasts are a case where gametic disequilibrium persists indefinitely since these genomes are uniparentally inherited and do not experience observable levels of recombination.

### Natural selection

Natural selection is a process that can continuously counteract the randomizing effects of recombination. Imagine a case where genotypes have different rates of survival or different fitnesses. In such a case natural selection will reduce the frequency of lower



**Figure 2.20** The decay of gametic disequilibrium ( $D$ ) over time when both strong natural selection and recombination are acting. Initially, there are only coupling ( $P_{11} = P_{22} = 1/2$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). The relative fitness values of the  $AAbb$  and  $aaBB$  genotypes are 1 while all other genotypes have a fitness of  $1/2$ . Unlike in Fig. 2.19, gametic disequilibrium does not decay to zero over time due to the action of natural selection.

fitness genotypes, which will also reduce the number of gametes these genotypes contribute to forming the next generation. At the same time that natural selection is acting, recombination is also working to randomize the associations of alleles at the two loci. Figure 2.20 shows an example of this type of natural selection acting in concert with recombination to maintain gametic disequilibrium.

The action of natural selection acting on differences in gamete fitness can produce steady states other than  $D = 0$ , expected eventually under even free

### Interact box 2.5 Gametic disequilibrium under both recombination and natural selection

To simulate the combined action of recombination and natural selection on gametic disequilibrium, try the program **Populus**, which can be obtained by following the link on the text website.

In the Java version of Populus, use the **Natural Selection** menu to select the **Two-Locus Selection** simulation. Set  $pAB = pab = 0.5$  and  $pAb = paB = 0.0$  as a case where there is maximum gametic disequilibrium initially. Use fitness values of  $wAaBb = 1$ , all others = 0.5 and  $wAAbb = waaBB = 1$ , all others = 0.5 to generate strong natural selection (relative fitness values are explained in Chapter 6). Finally, try recombination values of  $r = 0.5$  and  $0.05$ . Focus your attention on the **D vs. t** plot. What do the two different fitness cases do to levels of gametic disequilibrium and how effective is recombination in opposing or accelerating this effect?

For those who would like to see the details behind the recombination and natural selection model in Populus, a spreadsheet version of this model is available in Microsoft Excel format. The spreadsheet model will allow you to see all the calculations represented by formulas along with a graph of gametic disequilibrium over time.

recombination. In such cases, the population reaches a balance where the action of natural selection to increase  $D$  and the action of recombination to decrease  $D$  cancel each other out. The point where the two processes are exactly equal in magnitude but opposite in their effects is where gametic disequilibrium will be maintained in a population. It is important to recognize that the amount of steady-state gametic disequilibrium depends on which genotypes have high fitness values, so there are also plenty of cases where natural selection and recombination act in concert to accelerate the decay of gametic disequilibrium more rapidly than just recombination alone.

### **Mutation**

Alleles change from one form to another by the random process of mutation, which can either increase or decrease gametic disequilibrium. First consider the case of mutation producing a novel allele not found previously in the population. Since a new allele is present in the population as only a single copy, it is found only in association with the other alleles on the chromosome strand where it originated. Thus, a novel allele produced by mutation would initially increase gametic disequilibrium. Should the novel allele persist in the population and increase in frequency, then recombination will work to randomize the other alleles found with the novel allele and eventually dissipate the gametic disequilibrium. Mutation can also produce alleles identical to those currently present in a population. In that case, mutation can contribute to randomizing the combinations of alleles

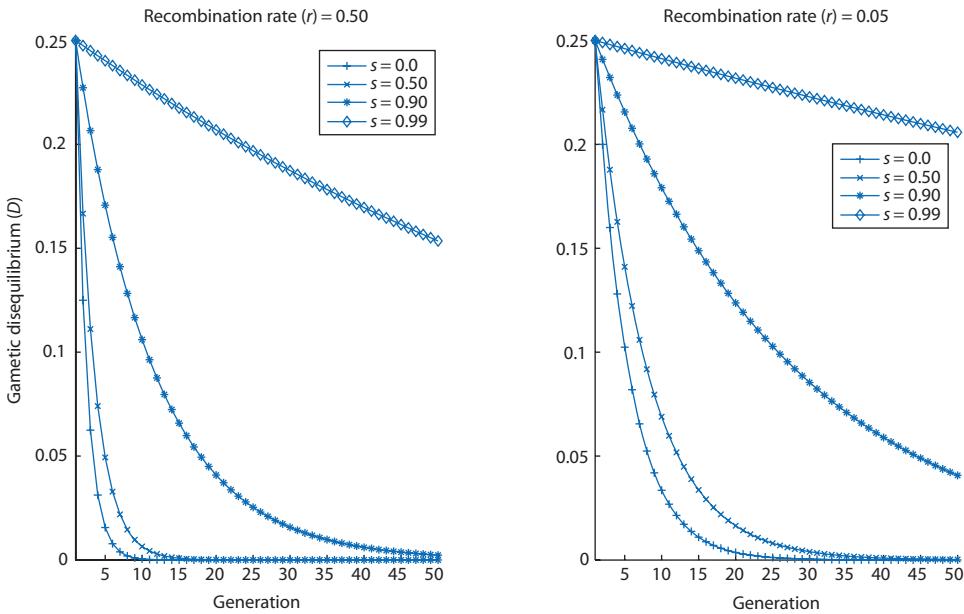
at different loci and thereby decrease levels of gametic disequilibrium. On the other hand, if the population is at gametic equilibrium mutation can create gametic disequilibrium by changing the frequencies of gamete haplotypes. However, it is important to recognize that mutation rates are often very low and the gamete frequency changes caused by mutation are inversely proportional to population size, so that mutation usually makes a modest contribution to overall levels of gametic disequilibrium.

### **Mixing of diverged populations**

The mixing of two genetically diverged populations, often termed admixture, can produce substantial levels of gametic disequilibrium. This is caused by different allele frequencies in the two source populations that result in different gamete frequencies at gametic equilibrium. Recombination acts to produce independent segregation but it does so only based on the allele frequencies within a group of mating individuals. Table 2.13 gives an example of gametic disequilibrium produced when two populations with diverged allele frequencies are mixed equally to form a third population. In the example, the allele-frequency divergence is large and admixture produces a new population where gametic disequilibrium is 64% of its maximum value. In general, gametic disequilibrium due to the admixture of two diverged populations increases as allele frequencies become more diverged between the source populations, and the initial composition of the mixture population approaches equal proportions of the source populations.

**Table 2.13** Example of the effect of population admixture on gametic disequilibrium. In this case the two populations are each at gametic equilibrium given their respective allele frequencies. When an equal number of gametes from each of these two genetically diverged populations are combined to form a new population, gametic disequilibrium results from the diverged gamete frequencies in the founding populations. The allele frequencies are: population 1  $p_1 = 0.1, p_2 = 0.9, q_1 = 0.1, q_2 = 0.9$ ; population 2  $p_1 = 0.9, p_2 = 0.1, q_1 = 0.9, q_2 = 0.1$ . In population 1 and population 2 gamete frequencies are the product of their respective allele frequencies as expected under independent segregation. In the mixture population, all allele frequencies become the average of the two source populations (0.5) with  $D_{max} = 0.25$ .

Gamete/ $D$	Gamete frequency	Population 1	Population 2	Mixture population
$A_1B_1$	$g_{11}$	0.01	0.81	0.41
$A_2B_2$	$g_{22}$	0.81	0.01	0.41
$A_1B_2$	$g_{12}$	0.09	0.09	0.09
$A_2B_1$	$g_{21}$	0.09	0.09	0.09
$D$		0.0	0.0	0.16
$D'$		0.0	0.0	$0.16/0.25 = 0.64$



**Figure 2.21**  
The decay of gametic disequilibrium ( $D$ ) over time with random mating ( $s = 0$ ) and three levels of self-fertilization. Initially, there are only coupling ( $P_{11} = P_{22} = \frac{1}{2}$ ) and no repulsion gametes ( $P_{12} = P_{21} = 0$ ). Self-fertilization slows the decay of linkage disequilibrium appreciably even when there is free recombination ( $r = 0.5$ ).

### Mating system

As covered earlier in this chapter, self-fertilization and mating between relatives increases homozygosity at the expense of heterozygosity. An increase in homozygosity causes a reduction in the effective rate of recombination because crossing over between two homozygous loci does not alter the gamete haplotypes produced by that genotype. The effective recombination fraction under self-fertilization is:

$$r_{\text{effective}} = r \left(1 - \frac{s}{2-s}\right) \quad (2.30)$$

where  $s$  is the proportion of progeny produced by self-fertilization each generation. This is based on the expected inbreeding coefficient at equilibrium  $F_{eq} = \frac{s}{2-s}$  (Haldane 1924). Figure 2.21 shows the decay in gametic disequilibrium predicted by equation 2.30 for four self-fertilization rates in the cases of free recombination ( $r = 0.5$ ) and tight linkage ( $r = 0.05$ ). Self-fertilization clearly increases the persistence of gametic disequilibrium, with marked effects at high selfing rates. In fact, the predominantly self-fertilizing plant *Arabidopsis thaliana* exhibits gametic disequilibrium over much longer regions of chromosome compared to outcrossing plants and animals (see review by Flint-Garcia et al. 2003).

### Chance

It is possible to observe gametic disequilibrium just by chance in small populations or small samples of gametes. Recombination itself is a random process in terms of where crossing over events occur in the genome. As shown in the Appendix, estimates are more likely to approach their true values as larger samples are taken. This applies to mating patterns and the number of gametes that contribute to surviving progeny in biological populations. If only a few individuals mate (even at random) or only a few gametes found the next generation, then this is a small “sample” of possible gametes that could deviate from independent segregation just by chance. When the chance effects due to population size and recombination are in equilibrium, the effects of population size can be summarized approximately by:

$$\rho^2 = \frac{1}{1 + 4N_e r} \quad (2.31)$$

where  $N_e$  is the genetic effective population size and  $r$  is the recombination fraction per generation (Hill & Robertson 1968; Ohta & Kimura 1969; the basis of this type of equation is derived in Chapter 4). As shown in Fig. 2.22, when the product of  $N_e$  and  $r$  is small, chance sampling contributes to maintaining some gametic disequilibrium since only a

few gametes contribute to the next generation when  $N_e$  is small or only a few recombinant gametes exist when  $r$  is small. The lesson is that  $D$  as we have used it in this section assumes a large population size

(similar to Hardy–Weinberg) so that actual gamete frequencies approach those expected based on allele frequencies, an assumption that is not always met in actual populations.

### Interact box 2.6 Estimating genotypic disequilibrium

In practice, the recombination fraction for two loci can be measured by crossing a double heterozygote with a double homozygote and then counting the recombinant gametes. However, this basic experiment cannot be carried out with species that cannot easily be mated in controlled crosses. An alternative but approximate means to test for gametic equilibrium is to examine the joint frequencies of genotypes at pairs of loci. If there is independent segregation at the two loci then the genotypes observed at one locus should be independent of the genotypes at the other locus. Such contingency table tests are commonly employed to determine whether genotypes at one locus are independent of genotypes at another locus.

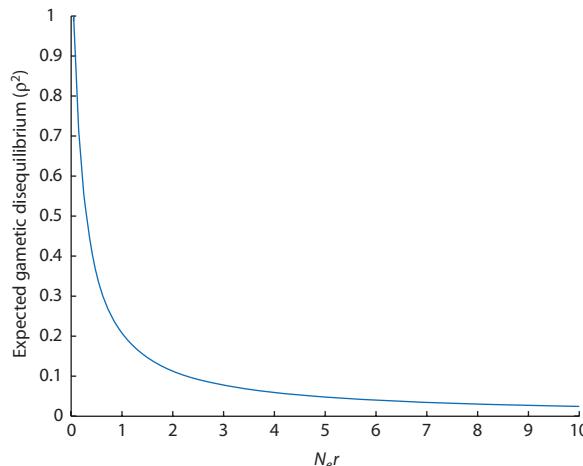
Contingency table tests involve tabulating counts of all genotypes for pairs of loci. In Table 2.14, genotypes observed at two microsatellite loci (AC25-6#10 and AT150-2#4) within a single population (the Choptank river) of the striped bass *Morone saxatilis* are given. The genotypes of 50 individuals are tabulated with alleles at each locus represented with numbers. For example, there were 15 fish that had a 22 homozygous genotype for locus AC25-6#10 and also had a 44 homozygous genotype for locus AT150-2#4. This joint frequency of homozygous genotypes is unlikely if genotypes at the two loci are independent, in which case the counts should be distributed randomly with respect to genotypes.

In the striped bass case shown here, null alleles (microsatellite alleles that are present in the genome but not reliably amplified by PCR) are probably the cause of fewer than expected heterozygotes that lead to a non-random joint distribution of genotypes (Brown et al. 2005). Thus, the perception of gametic disequilibrium can be due to technical limitations of genotyping techniques in addition to population genetic processes such as physical linkage, self-fertilization, consanguineous mating, and structured populations that cause actual gametic disequilibrium.

**Genepop on the Web** can be used to construct genotype count tables for pairs of loci and carry out statistical tests that compare observed to those expected by chance. Instructions on how to use Genepop and an example of striped bass microsatellite genotype data set in Genepop format are available on the text website along with a link to the Genepop site.

**Table 2.14** Joint counts of genotype frequencies observed at two microsatellite loci in the fish *Morone saxatilis*. Alleles at each locus are indicated by numbers (e.g. 12 is a heterozygote and 22 is a homozygote).

Genotype at locus AT150-2#4	Genotype at locus AC25-6#10					Row totals
	12	22	33	24	44	
22	0	0	1	0	0	1
24	1	4	0	4	1	10
44	2	15	0	0	0	17
25	0	3	0	0	0	3
45	0	8	0	1	0	9
55	1	1	0	0	0	2
26	0	1	0	2	0	3
46	1	3	0	0	0	4
56	0	0	0	1	0	1
Column totals	5	35	1	8	1	50



**Figure 2.22** Expected levels of gametic disequilibrium ( $p^2$ ) due to the combination of finite effective population size ( $N_e$ ) and the recombination rate ( $r$ ). Chance sampling will maintain gametic disequilibrium if very few recombinant gametes are produced (small  $r$ ), the population is small so that gamete frequencies fluctuate by chance (small  $N_e$ ), or if both factors in combination cause chance fluctuations in gamete frequency (small  $N_e r$ ).

## Chapter 2 review

- Mendel's experiments with peas lead him to hypothesize particulate inheritance with independent segregation of alleles within loci and independent assortment of multiple loci.
- Expected genotype frequencies predicted by the Hardy–Weinberg equation (for any number of alleles) show that Mendelian inheritance should lead to constant allele frequencies across generations. This prediction has a large set of assumptions about the absence of many population genetic processes. Hardy–Weinberg expected genotype frequencies therefore serve as a null model used as a standard of reference.
- The null model of Hardy–Weinberg expected genotype frequencies can be tested directly or assumed to be approximately true in order to test other hypotheses about Mendelian inheritance.
- The fixation index ( $F$ ) measures departures from Hardy–Weinberg expected genotype frequencies (excess or deficit of heterozygotes) that can be caused by patterns of mating.
- Mating among relatives or consanguineous mating causes changes in genotype frequencies (specifically a decrease in heterozygosity) but no changes in allele frequencies.
- Consanguineous mating can also be viewed as a process that increases the chances that alleles

descended from a common ancestor are found together in a diploid genotype (autozygosity).

- The fixation index, the autozygosity, and the coefficient of inbreeding are all interrelated measures of changes in genotype frequencies with consanguineous mating.
- Consanguineous mating may result in inbreeding depression, which ultimately is caused by overdominance (heterozygote advantage) or dominance (deleterious recessive alleles).
- The gametic disequilibrium parameter ( $D$ ) measures the degree of non-random association of alleles at two loci. Gametic disequilibrium is broken down by recombination.
- A wide variety of population genetic processes – natural selection, chance, admixture of populations, mating system, and mutation – can maintain and increase gametic disequilibrium even between loci without physical linkage to reduce recombination.

## Further reading

For a detailed history of Gregor Mendel's research in the context of early theories of heredity as well as the analysis of Mendel's results by subsequent generations of scientists see:

Orel V. 1996. *Gregor Mendel: the First Geneticist*. Oxford University Press, Oxford.

For recent perspectives on whether or not Gregor Mendel may have fudged his data, see a set of articles published together:

Myers JR. 2004 An alternative possibility for seed coat color determination in Mendel's experiment. *Genetics* 166: 1137.

Novitski E. 2004. Revision of Fisher's analysis of Mendel's garden pea experiments. *Genetics* 166: 1139–40.

Novitski E. 2004. On Fisher's criticism of Mendel's results with the garden pea. *Genetics* 166: 1133–6.

To learn more about the population genetics of DNA typing in criminal investigation consult:

Commission on DNA Forensic Science. 1997. *The evaluation of forensic DNA evidence. Proceedings of the National Academy of Sciences USA* 94: 5498–500 (an excerpt from the Executive Summary of the 1996 National Research Council Report).

Gill P. 2002. Role of short tandem repeat DNA in forensic casework in the UK – past, present, and future perspectives. *BioTechniques* 32: 366–85.

National Research Council, Commission on DNA Forensic Science. 1996. *The Evaluation of Forensic DNA Evidence: an Update*. National Academy of Sciences Press, Washington, DC.

To learn more about the Mendelian genetics of the ABO blood group, see the brief history:

Crow JF. 1993. Felix Bernstein and the first human marker locus. *Genetics* 133: 4–7.

Inbreeding depression has a large literature and probing its causes employs a wide array of methods. A good single source from which to learn more is:

Lynch M and Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

For more detail on ways to estimate gametic disequilibrium, consult:

Gaut BS and Long AD. 2003. The lowdown on linkage disequilibrium. *Plant Cell* 15: 1502–6.

A review of estimators of gametic disequilibrium, the genetic processes that influence its levels, and extensive references to past papers can be found in:

Flint-Garcia SA, Thornsberry JM, and Buckler ES. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54: 357–74.

## Problem box answers

### Problem box 2.1 answer

Using the allele frequencies in Table 2.3 we can calculate the expected genotype frequencies for each locus:

$$\begin{aligned} \text{D3S1358: } & 2(0.2118)(0.1626) = 0.0689; \\ \text{D21S11: } & 2(0.1811)(0.2321) = 0.0841; \\ \text{D18S51: } & (0.0918)^2 = 0.0084; \\ \text{vWA: } & (0.2628)^2 = 0.0691; \\ \text{FGA: } & 2(0.1378)(0.0689) = 0.0190; \\ \text{D8S1179: } & 2(0.3393)(0.2015) = 0.1367; \\ \text{D5S818: } & 2(0.3538)(0.1462) = 0.0992; \\ \text{D13S317: } & 2(0.0765)(0.3087) = 0.0472; \\ \text{D7S820: } & 2(0.2020)(0.1404) = 0.0567. \end{aligned}$$

As is evident from the allele designations, the amelogenin locus resides on the sex chromosomes and can be used to distinguish chromosomal males and females. It is a reasonable approximation to say that half of the population is male and assign a frequency of 0.5 to the amelogenin genotype. The expected frequency of the ten-locus genotype is therefore  $0.0689 \times 0.0841 \times 0.0084 \times 0.0691 \times 0.0190 \times 0.1367 \times 0.0992 \times 0.0472 \times 0.0567 \times 0.5 = 1.160 \times 10^{-12}$ . The odds ratio is one in 862,379,847,814. This 10-locus DNA profile is effectively a unique identifier since the current human population is approximately 6.5 billion and we would expect to observe this exact 10-locus genotype only once in a population 132 times larger than the current human population. In fact, it is likely that this 10-locus genotype has only occurred once in all of the humans who have ever lived.

### Problem box 2.2 answer

For hypothesis 1 the observed frequency of the bb genotype is given by:

$$f(aa\ bb) + f(A\_bb) = fa^2fb^2 + (1 - fa^2)fb^2$$

Expanding the second term gives:

$$= fa^2fb^2 + fb^2 - fa^2fb^2 = fb^2$$

The other allele frequencies can be obtained by similar steps. This must be true under Mendel's second law if the two loci are truly independent.

For hypothesis 2 the frequency of the B allele can be obtained from the fact that all the allele frequencies must sum to 1:

$$fA + fB + fO = 1$$

Then subtracting fB from each side gives:

$$fA + fO = 1 - fB$$

Squaring both sides gives:

$$(fA + fO)^2 = (1 - fB)^2$$

The left side of which can be expanded to:

$$\begin{aligned} (fA + fO)^2 &= (fA + fO)(fA + fO) \\ &= fO^2 + 2fAfO + fA^2 \end{aligned}$$

And then:

$$(1 - fB)^2 = fO^2 + 2fAfO + fA^2$$

The last expression on the right is identical to the sum of the first and second expected genotype frequencies for hypothesis 2 in Table 2.3. Expressions for the frequency of the A and O alleles can also be obtained in this fashion.

### Problem box 2.3 answer

The first step is to hypothesize genotypes under the two models of inheritance, as shown in Tables 2.6 and 2.7 for blood groups. Then these genotypes can be used to estimate allele frequencies (symbolized here with a *P* to indicate probability). For the hypothesis of two loci with two alleles each:

$$Pa^2 = \frac{769 + 261}{3816} = 0.270, Pa = 0.52$$

$$Pb^2 = \frac{728 + 261}{3816} = 0.260, Pb = 0.51$$

$$PA = 1 - Pa = 0.48, PB = 1 - Pb = 0.49$$

For the hypothesis of one locus with three alleles:

$$(1 - PB)^2 = (PA + PC)^2 = PA^2 + 2PAPC + PC^2 \\ = \frac{728 + 261}{3816} = 0.260$$

$$1 - PB = \sqrt{0.260} = 0.51 \\ PB = 0.49$$

$$(1 - PA)^2 = (PB + PC)^2 = PB^2 + 2PBPC + PC^2 \\ = \frac{769 + 261}{3816} = 0.270$$

$$1 - PA = \sqrt{0.270} = 0.52 \\ PA = 0.48 \\ PC = 1 - PA - PB = 0.03$$

The expected numbers of each genotype as well as the differences between the observed and expected genotype frequencies are worked out in the tables. For the hypothesis of two loci with two alleles each,  $\chi^2 = 0.266$ , whereas  $\chi^2 = 19,688$  for the hypothesis of one locus with three alleles. Both of these tests have one degree of freedom (4 genotypes – 2 for estimated allele frequencies – 1 for the test), giving a critical value of  $\chi^2_{0.05,1} = 3.84$  from Table 2.5. The deviations between observed and expected genotype frequencies could easily be due to chance under the hypothesis of two loci with two alleles each. However, the observed genotype frequencies are extremely unlikely under the hypothesis of three alleles at one locus since the deviations between observed and expected genotype frequencies are very large.

Phenotype	Genotype	Observed	Expected number of genotypes	Observed-expected	$(\text{Observed}-\text{expected})^2/\text{expected}$
Hypothesis 1: two loci with two alleles each					
Purple/smooth	A_B_	2058	$3816(1 - 0.52^2)(1 - 0.51^2) \\ = 2060.0$	2.0	0.002
Purple/wrinkled	A_bb	728	$3816(1 - 0.52^2)(0.51)^2 \\ = 724.2$	3.8	0.020
Yellow/smooth	aaB_	769	$3816(0.52)^2(1 - 0.51^2) \\ = 763.5$	5.5	0.040
Yellow/wrinkled	aabb	261	$3816(0.52)^2(0.51)^2 = 268.4$	-7.4	0.204
Hypothesis 2: one locus with three alleles					
Purple/smooth	AB	2058	$3816(2(0.48)(0.49)) = 1795$	63.0	38.5
Purple/wrinkled	AA, AC	728	$3816((0.48)^2 + 2(0.48)(0.03)) \\ = 989.1$	-261.1	68.9
Yellow/smooth	BB, BC	769	$3816((0.49)^2 + 2(0.49)(0.03)) \\ = 1028.4$	-259.4	63.7
Yellow/wrinkled	CC	261	$3816(0.03)^2 = 3.4$	257.6	19,517.0

# CHAPTER 3

## Genetic drift and effective population size

### 3.1 The effects of sampling lead to genetic drift

- Biological populations are finite.
- A simple sampling experiment with micro-centrifuge tube “populations.”
- The Wright–Fisher model of sampling.
- Sampling error and genetic drift in biological populations.

In Chapter 2, that population size is very large, effectively infinite, was among the assumptions listed for Hardy–Weinberg expected genotype frequencies to be realized. This entire chapter will be devoted to the changes in allele and genotype frequency that occur when this assumption is not met and populations are small or at least finite. Population size has profound effects on allele frequencies in biological populations and has a specific definition in the context of population genetics. A variable for population size in one form or another appears in many of the fundamental equations used to predict genotype or allele frequencies in populations. In those expectations where no explicit variable for population size appears there is an assumption instead, just as in the Hardy–Weinberg expectation for genotype frequencies. There is a strong biological motivation behind this attention to population size. All biological populations, without exception, are finite. Therefore, no actual population ever exactly meets the population size assumption of Hardy–Weinberg, although some may be large enough to show few genetic effects of finite size over relatively short periods of time. There is also a tremendous range of population sizes in the biological world. An understanding of the population genetic effects of population size will help to explain why some populations and species violate the assumptions to a greater degree than others, making sense of both the factors that cause differences in population size and the consequences of such differences. The causes and allele-frequency consequences of finite population size can

be understood and modeled in a variety of ways. Those models and concepts critical to understanding the population genetic impacts of finite population size will be the topics of this chapter.

A simple, hands-on demonstration can be used to show the role that population size plays in allele frequency in a population from one generation to the next. A plastic beaker filled with micro-centrifuge tubes can be used to represent gametes (Fig. 3.1). The micro-centrifuge tubes are of two different colors, say blue and clear, and there are 50 of each per beaker. Each beaker approximates a population with one diallelic locus where the allele frequencies are  $p = q = 0.5$ . Imagine sampling four tubes from a beaker and recording the resulting frequencies of the blue and clear tubes. Then imagine (after returning the four tubes and mixing the contents of the beaker) drawing out a sample of 20 tubes and recording the frequency of the blue and clear tubes. These handfuls of micro-centrifuge tubes represent the sampling process that occurs during reproduction and can be used to understand what happens to allele frequencies over time in a finite population.



**Figure 3.1** Beakers filled with micro-centrifuge tubes can be used to simulate the process of sampling and genetic drift. For a color version of this image see Plate 3.1.

Trial	$N = 4$			$N = 20$			Table 3.1 Typical results of sampling from beaker populations of micro-centrifuge tubes where the frequency of both blue ( $p$ ) and clear tubes is $1/2$ . After each draw, all tubes are replaced and the beaker is mixed to randomize the tubes for the next draw.
	Blue	Clear	$p$	Blue	Clear	$p$	
1	1	3	0.25	12	8	0.60	
2	2	2	0.50	10	10	0.50	
3	3	1	0.75	9	11	0.45	
4	0	4	0.0	7	13	0.35	
5	2	2	0.50	8	12	0.40	
6	1	3	0.25	11	9	0.55	
7	2	2	0.50	11	9	0.55	
8	3	1	0.75	12	8	0.60	
9	2	2	0.50	10	10	0.50	
10	1	3	0.25	9	11	0.45	

Some results typical for sampling from these micro-centrifuge tube populations are given in Table 3.1. The results have several striking patterns. First, micro-centrifuge tube “allele” frequencies fluctuate quite a bit in the samples. In some cases the results are 0.50/0.50 like in the ancestral population, but the results range from 0.35/0.65 in the sample of 20 to 0.0/1.0 in the sample of four. This latter result is called **fixation or loss**, since one allele composed the entire sample (its frequency went to 1) and the other allele was not sampled at all (its frequency went to zero). Second, the amount of fluctuation in the allele frequencies appears to be related to the size of the sample that was taken from the original population. The samples of four had greater fluctuations in allele frequency including a case of fixation and loss. The samples of 20 deviated somewhat less from the original allele frequencies of 0.50/0.50 and in those 10 trials no fixation/loss events were observed.

Compare these micro-centrifuge tube sampling results with what would be expected in an infinite population with  $p = q = 0.5$ . In the infinite population there would not be a sample of four or 20 drawn to found the next generation, the entire population would be used to found the next generation. Within the bounds of the micro-centrifuge tube population analogy, that would be like taking the entire beaker and just pouring it into another beaker to found the next generation: the allele frequencies would remain identical to the original frequency. This would also mean that if all other assumptions of Hardy–Weinberg were met, genotype frequencies would also remain constant ( $1/4$  AA,  $1/2$  Aa, and  $1/4$  aa with  $p = q = 0.5$ ).

Now return to the finite micro-centrifuge tube populations with a sample of two individuals, or

four gametes, drawn to found the population in the next generation. What are the chances that this next generation will consist of only AA genotypes? This is the same as asking what is the probability of sampling four blues or clear tubes in a handful of four. Since drawing one clear or blue tube has an independent probability of  $1/2$ , the probability of getting four is  $(1/2)^4 = 1/16$ . The same result can be seen from the perspective of genotypes by asking what are the chances of founding a population with two homozygous genotypes. If the source population is in Hardy–Weinberg equilibrium then  $1/4$  of all genotypes are one of the two homozygotes. The chance of drawing two identical homozygous genotypes is the product of their independent probabilities, or  $(1/4)^2 = 1/16$ .

The micro-centrifuge tube populations are a low-tech demonstration that genotype and allele frequencies fluctuate from one generation to the next due to small samples, or **sampling error**, in a process called **genetic drift**. The amount of genetic drift increases as the size of the sample used to found the next generation decreases. Another way to restate the population size assumption of Hardy–Weinberg is to say instead that Hardy–Weinberg assumes that there is very little or no genetic drift occurring.

**Sampling error** The difference between the value found in a finite sample from a population and the true value in the population.

**Genetic drift** Random changes in allele frequency from one generation to the next in

biological populations due to the finite samples of individuals, gametes, and ultimately alleles that contribute to the next generation. The amount of genetic drift increases as the size of the sample used to found the next generation decreases.

**Stochastic process** A process where individual outcomes are dictated by chance but the average of a large number of outcomes can be described as a probability distribution based on initial conditions.

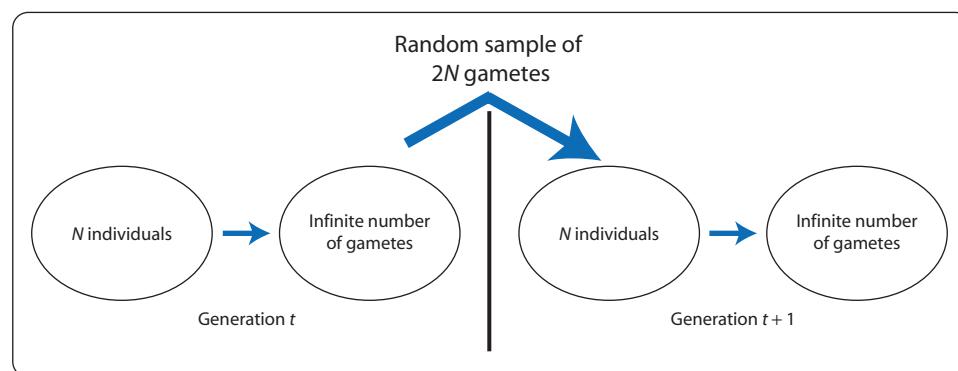
**Wright–Fisher model** A simplified version of the biological life cycle where all sampling to found the next generation occurs from an infinite pool of gametes built from equal contributions of all individuals. This approximation is commonly employed to model genetic drift.

To extend and generalize the model of genetic drift started with the micro-centrifuge tube populations, a model of the biological process of reproduction is helpful. To do this, let's consider the process of reproduction in populations. During reproduction, individual adult organisms produce gametes. These gametes are exchanged with mates and fuse to form zygotes, and these zygotes develop into a new generation of adult organisms (Fig. 3.2). This schematic of the biological life cycle is called the Wright–Fisher model of sampling (introduced by Sewall Wright (1931) and Ronald A. Fisher (1999; originally

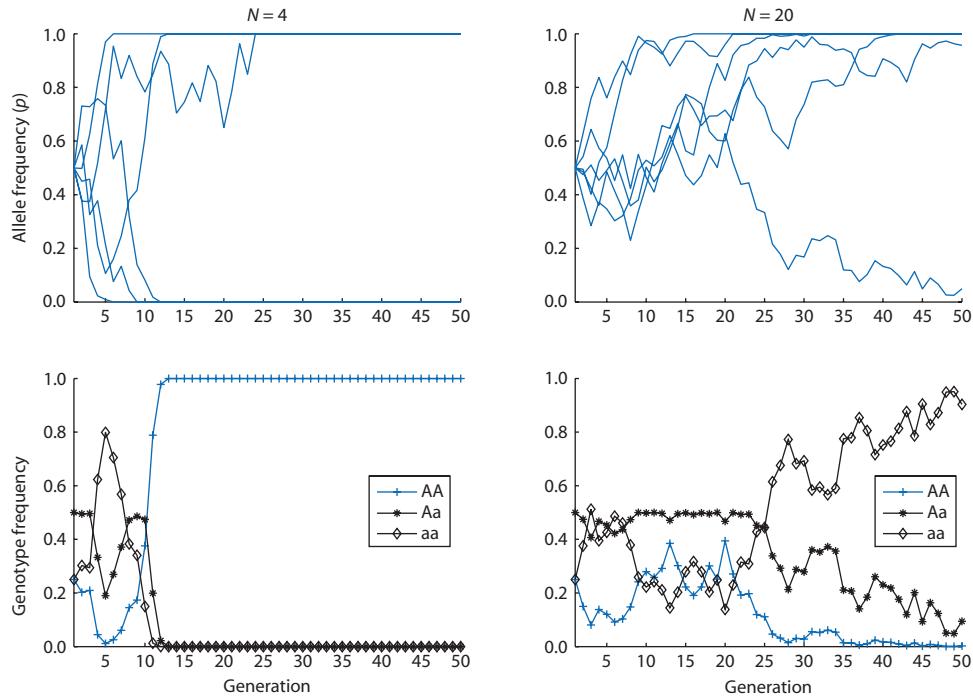
published in 1930)). It is not completely biologically realistic. There is obviously not an infinite number of gametes in any real population and sampling events can take place at many points during the life history of a population of organisms. But it allows the process of genetic drift to be reduced to a point that it can be modeled in a simple fashion. The Wright–Fisher model makes assumptions identical to those of Hardy–Weinberg (see section 2.2), with the exception that the population is finite rather than approaching infinite. Particularly critical assumptions include:

- generations are discrete and do not overlap, equivalent to adults that reproduce synchronously but only once during their lifetime;
- the numbers of females and males are equal;
- the size of the population ( $N$  individuals) remains constant through time; and
- all individuals are equal in their production of gametes and all gametes are equally viable, equivalent to no natural selection.

These assumptions reduce the complexity of sampling error in biological populations, concentrating all sampling into a single step as an approximation. This simplification approximates the genetic drift that occurs in biological populations. For example, sampling at several points in the life cycle can be equivalent in its effect on allele frequencies to the same total amount of sampling at a single point in the life cycle. As will be shown later in this chapter, sampling events may occur at many stages of the life



**Figure 3.2** The Wright–Fisher model of genetic drift uses a simplified view of biological reproduction where all sampling occurs at one point: sampling  $2N$  gametes from an infinite gamete pool. In this case  $N$  diploid individuals ( $N/2$  of each sex) generate an infinite pool of gametes where allele frequencies are perfectly represented, and a finite sample of  $2N$  alleles is drawn from this gamete pool to form  $N$  new diploid individuals in the next generation. Genetic drift takes place only in the random sample of  $2N$  gametes to form the next generation. Major assumptions include non-overlapping generations, equal fitness of all individuals, and constant population size through time. The model can easily be adjusted for haploid individuals or loci by assuming  $2N$  individuals or sampling  $N$  gametes to form the next generation.



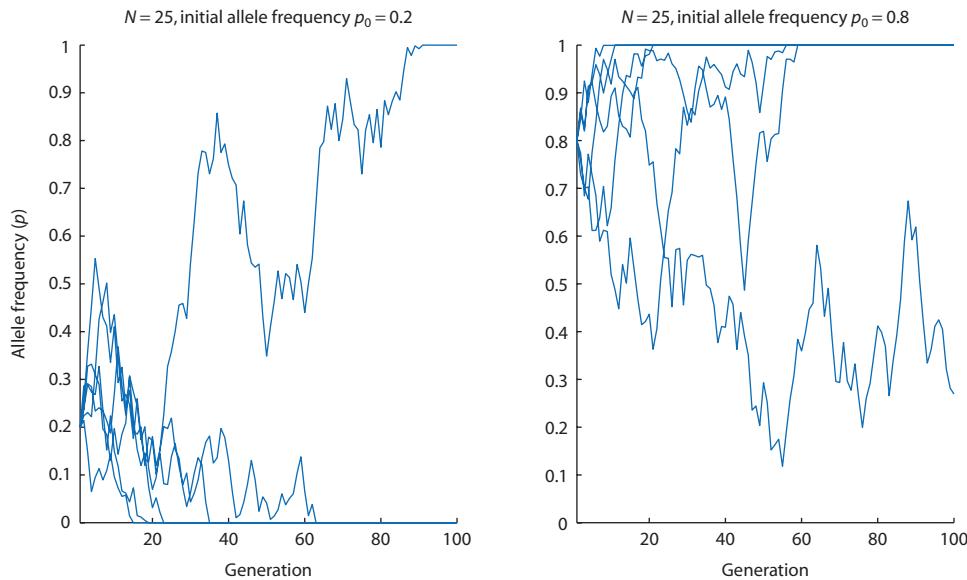
**Figure 3.3** The results of genetic drift continued every generation in populations of  $N = 4$  and  $N = 20$ . In the top panels, the six lines represent independent replicates or independent populations experiencing genetic drift starting at the same initial allele frequency ( $p = 0.5$ ). The random nature of genetic drift can be seen by the zig-zag changes in allele frequency that have no apparent direction. Allele frequencies that reach the upper or lower axes represent cases of fixation or loss. In the bottom panels, the genotype frequencies are shown for the allele frequencies represented by the black and blue lines under the assumption of random mating within each generation. The changes in genotype frequencies are a consequence of changes in allele frequencies due to genetic drift.

cycle in actual biological populations. The implications of relaxing other assumptions such as constant population size are topics of later chapters.

A major limitation of the micro-centrifuge tube demonstration is that it shows the effects of genetic drift over only one generation. A more general model of the effect of sampling error is needed to predict what may happen to allele frequencies over many generations. A general model would sample from one generation to found the next generation, then build a large pool of gametes (like the beakers of micro-centrifuge tubes) with those new allele frequencies. The sampling process would then be continued for many generations. Figure 3.3 shows computer-simulated allele frequencies based on this more general model for many generations under the assumptions of the Wright–Fisher model (note that the populations in Fig. 3.3 are twice as large as the samples of micro-centrifuge tubes). The effects of genetic drift are more obvious over longer periods of time. Allele frequencies over a few generations change at random, both increasing and decreasing, sometimes changing very little in one generation

and other times changing more substantially. There is a clear trend that over time in these genetic drift simulations that the frequency of one allele reaches either fixation ( $p = 1.0$ ) or loss ( $p = 0.0$ ), identical to loss ( $q = 0.0$ ) and fixation ( $q = 1.0$ ) for the alternate allele. There is also a trend that fixation or loss occurs in fewer generations with the smaller between-generation sample size and more slowly with the larger between-generation sample size. Since these simulations do not include any processes that could reintroduce genetic variation, once a population has reached fixation or loss there can be no further change in allele frequency.

The bottom panels of Fig. 3.3 show genotype frequencies based on random mating for one of the populations represented in the top panels. Genetic drift clearly causes genotype frequencies to change over time along with the changes in allele frequency. This is in contrast to the processes considered in Chapter 2, such as consanguineous mating, that result in changes in genotype frequency only but do not alter allele frequency. Genetic drift is most commonly modeled and demonstrated from the perspective of



**Figure 3.4** The results of genetic drift with different initial allele frequencies. The two panels have identical population sizes ( $N = 25$ ) but initial allele frequency is  $p = 0.2$  on the right and  $p = 0.8$  on the left. The chances of fixation are equal to the initial allele frequency, a generalization that can be seen by examining a large number of replicates of genetic drift. Consistent with this expectation, more replicates go to loss on the left and more reach fixation on the right. Even with the difference in initial allele frequencies, the random trajectory of allele frequencies is apparent.

allele frequencies since it is easier to summarize a diallelic locus in a population as two allele frequencies rather than three genotype frequencies. But remember that genotype frequencies are affected by genetic drift too, and genotype frequencies can always be obtained by multiplication given allele frequencies under the assumption of random mating.

Up to this point, the beakers of micro-centrifuge tubes and computer simulations only considered cases of alleles at equal initial frequencies ( $p = q = 0.5$ ). Figure 3.4 shows results of computer simulations for genetic drift where initial allele frequencies are  $p = 0.2$  and  $p = 0.8$  with identical population sizes of  $N = 25$ . Identical simulations under the assumptions of the Wright–Fisher model were used to produce both Figs 3.3 and 3.4, so the results in the two cases can be compared directly. Initial allele frequencies do influence the outcome of genetic drift in the simulations shown in Fig. 3.4. The lower initial allele frequency is associated with more frequent loss of the allele (five of six replicates) while the higher initial allele frequency fixed in five of six replicates. These results are consistent with what would occur in a much larger number of replicate simulations with the same initial allele frequencies and population size. A larger sample would show that the probability that an allele reaches fixation under genetic drift is the same as the initial allele frequency. This makes

intuitive sense, since with a lower initial frequency a population is closer in frequency to loss than to fixation. If the direction and magnitude of genetic drift in allele frequencies is random, there is a better chance of reaching zero on average than reaching one. This same pattern would be true for any initial values of the allele frequency closer to zero or to one, except in the special case of equal allele frequencies where the chances of fixation or loss for an allele would be equal.

Under the Wright–Fisher model the following are general conclusions about the action of genetic drift in finite populations:

- the direction of changes in allele frequency is random;
- the magnitude of random fluctuations in allele frequencies from generation to generation increases as the population size decreases;
- fixation or loss is the equilibrium state if there are no other processes acting to counteract genetic drift or reintroduce genetic variation;
- genetic drift changes allele frequencies and thereby genotype frequencies; and
- the probability of eventual fixation of an allele is equal to its initial frequency (or the probability of ultimate loss of an allele is equal to one minus its initial frequency).

### Interact box 3.1 Genetic drift

Genetic drift can be simulated with either Populus or PopGene.<sup>S<sup>2</sup></sup>. Try using Populus and following the instructions here.

Under the **Model** menu, select **Mendelian Genetics** and then **Genetic Drift**. Make sure the **Monte Carlo** tab is selected. The simulation dialog has entry fields for **Population Size (N)** and **Number of Loci** (or replicates). Number of generations shown in the graphs can be specified as **3N** or a fixed number of generations (**Other:**) with radio buttons (the program will automatically set the number of generations if all loci go to fixation or loss before the fixed number is reached). Set allele frequencies of all loci collectively. Press the **View** button to see the results graph. The model can be rerun with the **Iterate** button in the results window.

- Set the parameters for 10 loci and an initial allele frequency of 0.5, and view the results for 800 generations so that all fixation/loss events are visible. Examine drift for population sizes of 4, 20, 50, and 100. Record the generation of fixation or loss for 40 replicates of each population size. What is the average number of generations to fixation or loss? Are there equal numbers of fixation and loss events?
- Progressively reduce (or increase) the initial allele frequencies by intervals of 0.1 for a single population size. Record the generation of fixation or loss for 50 replicates of initial allele frequency. What is the observed relationship between initial allele frequency and probability of fixation or loss? Do these averages change if the population size changes?

Hint: using a spreadsheet program like Microsoft Excel can speed the calculation of averages for a list of values. Enter the values in columns and then use the average function ("=AVERAGE( )" with the range specified in the parentheses, such as "C1 : C10" to indicate the values of cells 1 through 10 in column C). A wide range of other useful functions is provided under **Functions . . .** in the **Insert** menu.

The next section of the chapter will develop two probability models of sampling error to provide more rigorous evidence for the conclusions reached in this section and to reach additional generalizations about the action of genetic drift.

### 3.2 Models of genetic drift

- An introduction to the binomial distribution and Markov chains.
- The diffusion approximation of genetic drift.

The last section demonstrated the phenomenon of genetic drift caused by sampling error and drew some general conclusions based on the results of computer simulations. Building on this foundation, this section will introduce three probability models that can be used to confirm these and other general properties of the process of genetic drift. The first model, the binomial distribution, will be used to show that the magnitude of genetic drift from one generation to

the next depends on allele frequencies in the population. The second model, the Markov chain, will be used to show the rate of change of allele frequencies under genetic drift. The third model, a continuous time approximation to the Markov chain, will be introduced to show how genetic drift can be modeled as the diffusion of particles.

#### ***The binomial probability distribution***

To develop the first model, let's return to the microcentrifuge tube populations from the last section. When sampling a tube from the beaker there are only two outcomes, a blue tube or a clear tube, which are used to represent the two alleles at one locus. The tubes are a specific case of a **Bernoulli random variable** (sometimes called a binomial random variable), or a variable representing a trial or sample that can have only two outcomes. Coin flips with either heads or tails outcomes are another example of a Bernoulli random variable. What we often

want to know is, what are the chances of obtaining a given set of Bernoulli outcomes? For example, what are the chances of obtaining four heads when flipping a coin four times? In our micro-centrifuge tube samples, what are the chances of one of the possible outcomes (20 blue, 19 blue, 18 blue, . . . , 0 blue) when sampling 20 tubes from the beaker? Answers to these types of questions require a means to estimate a probability distribution.

The binomial (literally, “two names”) formula defines the probability distribution for the sum of  $N$  independent samples of a Bernoulli variable:

$$P_{(i=A)} = \binom{2N}{i} p^i q^{2N-i} \text{ where } \binom{2N}{i} = \frac{(2N)!}{i!(2N-i)!} \quad (3.1)$$

The binomial formula gives the probability of sampling  $i$  A alleles in a sample of  $2N$  from a population where the A allele has a frequency of  $p$  and the alternate a allele is at a frequency of  $q$ . The  $p^i$  and  $q^{2N-i}$  terms estimate the probability of observing  $i$  and  $2N - i$  independent events each with probability

$p$  and  $q$ , respectively. The term  $\binom{2N}{i}$  (pronounced “two N draw  $i$ ”) serves as a way to enumerate the different ways (or permutations) of obtaining  $i$  As in a sample of  $2N$ .

Applying the binomial to the micro-centrifuge tube sampling results from the last section will illustrate how the binomial provides the probability for a specific sampling outcome. In the beaker, the blue and clear tubes were both at a frequency of  $p = q = 1/2$ . In the draws of  $N = 4$ , a result of two blue and two clear tubes occurred in four out of 10 draws or 40% of the time (Table 3.1). When drawing samples of tubes there are  $2^N$  possible combinations. So, for samples of  $N = 4$  there are  $2^4 = 16$  combinations (like the number of genotypes in a Punnett square for four alleles at a locus). Of these 16 combinations, there are exactly six (bbcc, bc<sub>b</sub>c, bcc<sub>b</sub>, cbcb, cbbc, ccbb) which yield two blue and two clear tubes. This same result can be obtained by using

$$\binom{4}{2} = \frac{4!}{2!2!} = 6 \quad (3.2)$$

to enumerate the number of possible permutations of outcomes of one type in a sample of  $2N$  objects.

The ! notation stands for factorial, and  $n!$  equals  $1 \times 2 \times 3 \times \dots \times n - 1 \times n$  and  $0! = 1$ . The other part of the binomial formula calculates the probability of obtaining a sample of  $i$  blue tubes and  $2N - i$  clear tubes. The blue tubes are at a frequency of  $p$  in the population, so the probability of sampling  $i$  of them is  $p^i$  since each is an independent event. The same logic applies to the clear tubes, whose frequency in the population is  $q$  and the number sampled is the remaining sample size not made up of blue tubes or  $2N - i$ , to give a probability of  $q^{2N-i}$ . Bringing both of these components of the binomial formula together,

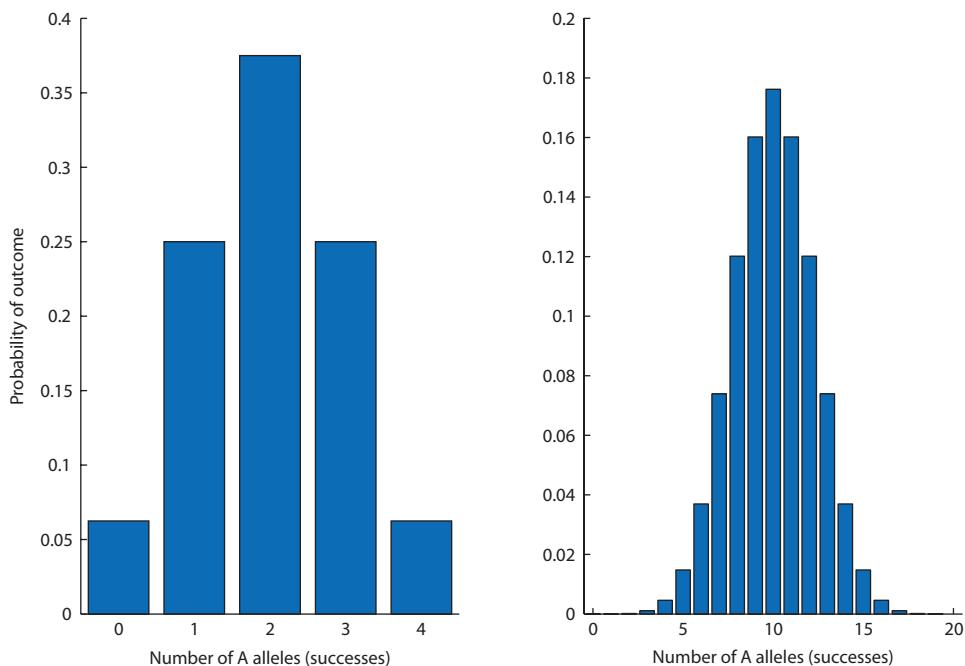
$$\begin{aligned} P_{b=2} &= \frac{4!}{2!(4-2)!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 \\ &= \frac{24}{4} \left(\frac{1}{4}\right) \left(\frac{1}{4}\right) = \frac{6}{16} \text{ or } 0.375 \end{aligned} \quad (3.3)$$

gives the expected frequency of draws of two blue and two clear tubes when sampling four tubes. This expected value is very close to what was observed in 10 draws of four tubes in Table 3.1.

The binomial formula can be used to calculate the expected probability of observing each of the possible outcomes when drawing samples of  $2N = 4$  and  $2N = 20$  micro-centrifuge tubes from beaker populations. These probability distributions (Fig. 3.5) summarize what we would expect to find if we drew many independent samples and then tabulated the results. The probability for each bar in the histograms of Fig. 3.5 was determined using the binomial formula. For example, the expected frequency of sampling 12 blue tubes in a total sample of 20 tubes is

$$P_{b=12} = \frac{20!}{(12!)(8!)} (0.5)^{12} (0.5)^8 = 0.1201 \quad (3.4)$$

These probability distributions explain why a fixation/loss event was observed when  $2N = 4$  but not for  $2N = 20$ , since the former outcome is expected in one out of 16 draws but the latter only once in 1,048,576 draws. With knowledge of the binomial probability distribution, the Wright–Fisher model of genetic drift (Fig. 3.2) makes a lot of sense. It was constructed, in fact, to articulate the assumptions that underlie the use of the binomial formula and binomial probability distributions to model genetic drift.



**Figure 3.5**  
Probability distributions for binomial random variables based on samples of  $N = 4$  (left) and  $N = 20$  (right) from populations where  $p = q = 0.5$ . These distributions describe the expected probability of each of the possible outcomes of the micro-centrifuge sampling experiment described in the text.

### Problem box 3.1 Applying the binomial formula

Two independent laboratory populations of the fruit fly *Drosophila melanogaster* were observed for two generations. The populations each had a size of  $N = 24$  individuals with an equal number of males and females. In the first generation, both populations were founded with  $f_A = p = 0.5$ . In the second generation, one population showed  $f_A = p = 0.458$  and the other  $f_A = p = 0.521$ . What are the chances of observing these allele frequencies after one generation of genetic drift?

**Bernoulli or binomial random variable**  
A variable representing a trial or sample that can have only two possible outcomes, such as zero or one.

Applying the binomial formula to determine expected probabilities associated with particular sampling outcomes is useful, and there is an even broader lesson that can be learned from the binomial. The examples up to this point have focused on the expected value.

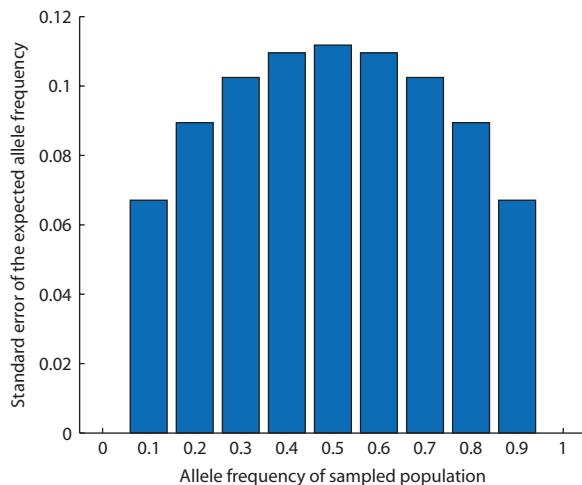
Shifting our perspective, we can use the binomial to explore how variable allele-frequency changes under genetic drift should be. The variance of a Bernoulli or binomial random variable is:

$$\sigma^2 = pq \quad (3.5)$$

This result is derived in Math Box 3.1 for those readers who would like to work through the details. The maximum variability will occur when  $p = q = 1/2$ . The standard deviation ( $\sigma = \sqrt{pq}$ ) and standard error of the allele frequency,

$$SE = \sqrt{\frac{pq}{2N}} \quad (3.6)$$

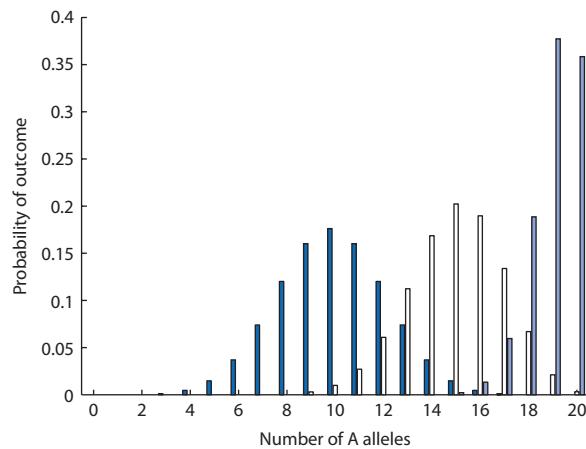
are also easy to obtain (see Appendix). The standard error is the standard deviation of a mean, and the mean in this case is the expected value or allele frequency  $p$  or  $q$ . For genetic drift under the Wright–Fisher model, equally frequent alleles will give the widest range of outcomes for a given sample size (Fig. 3.6). The variability in allele frequency caused by genetic drift decreases as a population approaches fixation or loss, causing  $pq$  to approach zero (Fig. 3.7). This result makes intuitive sense. When alleles are equally frequent, sampling error is equally likely to increase or decrease allele frequency and could produce an outcome anywhere along the spectrum of possible allele frequencies. At the other extreme,



**Figure 3.6** The standard error of the allele frequency ( $\sigma = \sqrt{\frac{pq}{2N}}$ ) for a binomial random variable for a sample size of  $N = 10$  for a range of allele frequencies. The standard error of the allele frequency decreases as the allele frequency approaches fixation or loss. In the same way, genetic drift is less effective at spreading out the distribution of allele frequencies as alleles approach fixation or loss. The standard deviation is 0 when the allele frequencies are 0 or 1 since there is no genetic variation and any size sample will faithfully reproduce the allele frequencies in the source population.

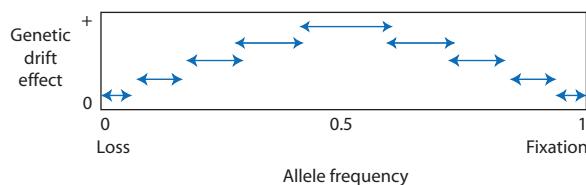
when one allele is very nearly fixed except for one copy of the alternate allele ( $p = 1 - \frac{1}{2N}$  and  $q = \frac{1}{2N}$ ), drift has a reasonable chance of only several sampling errors, such as no, one, two, or three copies of the low-frequency allele. Sampling error that causes fixation of the high-frequency allele is quite likely. However, sampling error that results in greatly increased frequencies of the low-frequency allele in one generation would be very, very unlikely.

There is a graphical metaphor to summarize the consequences of initial allele frequency for the range of outcomes in allele frequency under genetic drift. Figure 3.8 shows the range of possible allele frequencies (0–1) in a population and indicates the effects of genetic drift by the width of the arrows and the vertical scale. The range of outcomes probable under drift depends on the allele frequency in a population. When both alleles are equally frequent ( $pq = 0.25$ , its maximum), the sampling error is the largest so that the rate of genetic drift in changing allele frequencies is greatest. When the allele frequency is closer to fixation or loss ( $pq < 0.25$ ), the sampling error is smaller and the rate of genetic



**Figure 3.7** Probability distributions for binomial random variables based on samples of  $N = 20$  from populations where the allele frequency is 0.50, 0.75, or 0.95 (dark blue, white, and light blue bars, respectively). The range of probable outcomes with sampling depends on the allele frequency. As allele frequencies approach the boundaries of fixation or loss, there is a decreasing number of outcomes other than fixation or loss that are probable due to sampling error.

drift in changing allele frequencies is also smaller. This explains the tendency of populations to go to fixation or loss under genetic drift. The sampling effect is greatest when the genetic variation is greatest but also weakest when genetic variation is least (Fig. 3.8). A population is most likely to experience larger changes in allele frequencies, toward fixation or loss, due to drift when both alleles are near equal frequencies. However, a population with strongly unequal allele frequencies is less likely to experience genetic drift of a magnitude that would equalize allele frequencies.



**Figure 3.8** A schematic illustration of how the effects of genetic drift due to sampling error depend on allele frequencies in a population. The horizontal axis represents allele frequency and the width of the arrows represents the

standard error of allele frequency ( $\sigma = \sqrt{\frac{pq}{2N}}$ ) at a given allele frequency. Larger standard errors for allele frequency are another way of saying that sampling error will cause a greater range of outcomes, equivalent to a larger effect of genetic drift.

### Math box 3.1 Variance of a binomial variable

The variance in the outcome of binomial sampling over many trials is surprisingly easy to derive. Assuming a diallelic locus, let  $p$  be the fraction of successes (such as sampling the A allele) and  $q$  be the fraction of failures (such as failing to sample the A allele and getting an a allele instead) so that  $p + q = 1$ . In the Appendix, a variance is defined as the average of the squared differences between each estimate and the average. If 1 is used to represent a success and 0 a failure and  $p$  and  $q$  are used to represent the frequency of each outcome instead of the sums used in the Appendix, the variance in successes is:

$$\sigma^2 = p(1 - \bar{x})^2 + q(0 - \bar{x})^2 \quad (3.7)$$

The average of a binomial variable ( $\bar{x}$ ) is simply the probability of a success or  $p$ , just as when flipping a fair coin a large number of times the number of successes would approach the expected value of

one-half heads or tails. Substituting  $p$  for  $\bar{x}$  gives:

$$\sigma^2 = p(1 - p)^2 + q(0 - p)^2 \quad (3.8)$$

which can then be simplified by substituting  $1 - p = q$  into the left-hand term:

$$\sigma^2 = p(q)^2 + q(0 - p)^2 \quad (3.9)$$

and multiplying out the right-hand term to give

$$\sigma^2 = pq^2 + qp^2 \quad (3.10)$$

This result can then be rearranged by finding a factor common to both terms:

$$\sigma^2 = pq(q + p) \quad (3.11)$$

which simplifies after noticing that  $q + p = 1$ :

$$\sigma^2 = pq \quad (3.12)$$

### Markov chains

The next step in understanding genetic drift is to consider its effects in a large number of replicate populations. Instead of focusing on allele frequency in just a single population like in the last section, let's now explore the case where there is a collection of numerous independent but identical populations (an infinite number of replicate populations is sometimes called an *ensemble* in physics and mathematics). Using the approach of genetic drift in multiple finite populations, this section will cultivate an understanding of how drift works on average among many populations and will develop a prediction of how rapidly genetic drift causes populations to reach fixation and loss.

To get started, consider populations composed of a diallelic locus in a single diploid individual. Since there are only two alleles in a population, there are three possibilities for the numbers of one of the alleles: zero, one, and two copies. Each of these possible states in a population could be referred to by the number of A alleles, 0 through 2, which can be summarized in

notation as  $P(0)$ ,  $P(1)$ , and  $P(2)$ . With this very basic type of population, we can ask: what are chances that a population starting out in one of these three states ends up in one of these three states due to sampling error? For example, what is the chance of starting out with two copies of A and ending up with one copy of A with a sample size of one individual (two gametes) between generations? This chance is known as the **transition probability** for allelic states. The transition probability is determined with the binomial formula:

$$P_{i \rightarrow j} = \binom{2N}{j} p^j q^{2N-j} \quad (3.13)$$

where  $i$  is the initial number of alleles,  $j$  is the number of alleles after sampling, and  $N$  is the sample size of diploid individuals. As before,  $\binom{2N}{j} = \frac{2N!}{j!(2N-j)!}$  enumerates the possible draws that yield  $j$  copies of the allele and  $p^j q^{2N-j}$  is the probability of sampling  $j$  copies of the allele given the allele frequencies in the initial population.

Equation 3.13 can be used to determine the expected frequencies of populations with a given allelic state in one generation based on the frequencies of populations in each allelic state in the previous generation. To predict the frequency of populations with one allelic state, we need to add up the chances that populations in *all* states in the previous generation transition to this state. Let's work through what is essentially bookkeeping to see this. The expected frequency of populations with two A alleles in generation one is the sum of the probabilities that populations a generation before with zero, one, and two A alleles become populations with two A alleles through sampling error. This can be stated in an equation as:

$$P_{t=1}(2) = (P_{2 \rightarrow 2})P_{t=0}(2) + (P_{1 \rightarrow 2})P_{t=0}(1) + (P_{0 \rightarrow 2})P_{t=0}(0) \quad (3.14)$$

for the case of populations of a single diploid individual. In equation 3.14, the probability or frequency of a given allelic state is indicated by  $P(x)$  with subscripts to indicate the generation. In a population with one A and one a allele, the chance of sampling two A alleles,  $P_{1 \rightarrow 2}$ , is

$$P_{1 \rightarrow 2} = \frac{2!}{2!0!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^0 = \frac{2}{2} \left(\frac{1}{4}\right)(1) = \frac{1}{4} \quad (3.15)$$

using equation 3.13. For populations that are at fixation or loss, sampling cannot change the allele frequency. Therefore, populations initially fixed for A ( $P_{t=0}(2) = 1$ ) all transition to populations fixed for A ( $P_{2 \rightarrow 2} = 1$ ) but none of the populations initially lost for A ( $P_{t=0}(0) = 0$ ) can transition to any other state, so

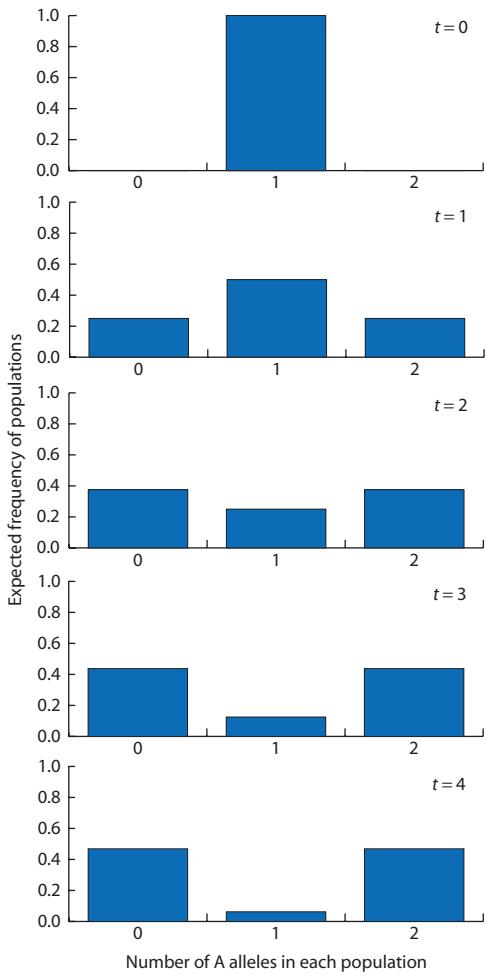
$P_{0 \rightarrow 2}$  is zero. Lastly, the  $P_{t=0}(2)$ ,  $P_{t=0}(1)$ , and  $P_{t=0}(0)$  terms each represent the frequencies of populations that possess a given allelic state. This means that the transition probabilities are multiplied by the frequency of populations in a given allelic state to determine the expected frequency of populations with a given allelic state in the next generation.

Using this same logic, the expected frequencies of populations with zero, one, and two A alleles after one generation of sampling error are shown in Table 3.2. Using these transition probabilities, the expected population frequencies over four generations of sampling are shown in Fig. 3.9. The result of the equations in Table 3.2 is a generation-by-generation prediction for the *average* behavior of populations under genetic drift when there is an infinite number of replicate populations. This method of modeling the action of genetic drift is known as a **Markov chain** model. It is important to recognize that the outcome of genetic drift for a single population cannot be predicted. Rather, we can only know the probability that a single population experiences a given change in allele frequency such as going from one to zero copies of the A allele. If many replicate populations are experiencing genetic drift, then the Markov chain predicts the proportion of populations that have a given allelic state in each generation.

Figure 3.10 shows the first two steps in the Markov chain for a population of two diploid individuals or four gametes, similar to the micro-centrifuge tube sampling experiment from the first section of this chapter. With a slightly larger population size than in Fig. 3.9, there are a larger number of allelic state transitions to account for between generations.

**Table 3.2** The equations used to calculate the expected frequency of populations with zero, one, or two A alleles in generation one ( $t = 1$ ) based on the previous generation ( $t = 0$ ). Frequencies at  $t = 1$  depend both on transition probabilities due to sampling error (constant terms like 0, 1, or  $\frac{1}{2}$ ) and population frequencies in the previous generation ( $P_{t=0}(x)$  terms). The transition probabilities are calculated with the binomial formula  $P_{i \rightarrow j} = \binom{2N}{j} p^j q^{2N-j}$ . Since sampling error cannot change the allele frequency of a population at fixation or loss,  $P_{2 \rightarrow 2} = 1$  and  $P_{0 \rightarrow 0} = 1$ , whereas the other possibilities have a probability of zero.

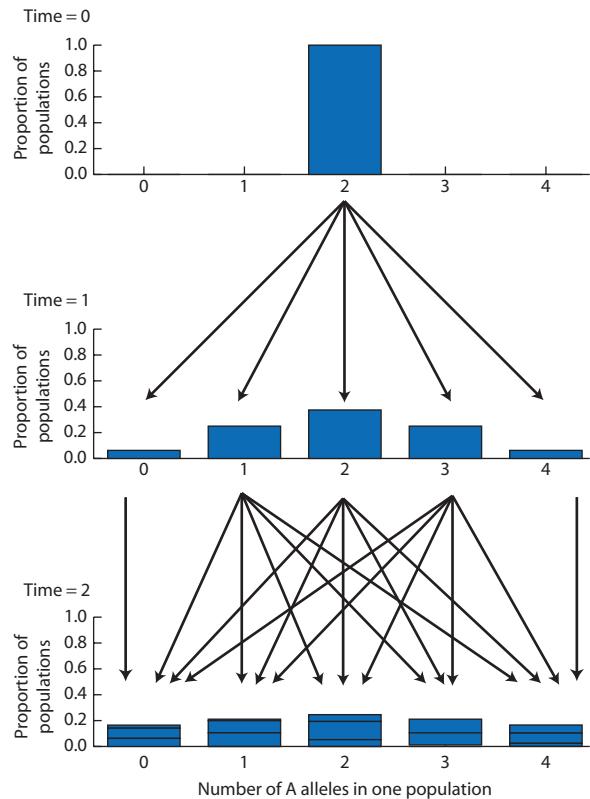
One generation later ( $t = 1$ )		Initial state: number of A alleles ( $t = 0$ )		
A alleles	Expected frequency	2	1	0
2	$P_{t=1}(2)$	=	$(P_{2 \rightarrow 2})P_{t=0}(2)$	$+ (P_{1 \rightarrow 2})P_{t=0}(1)$
1	$P_{t=1}(1)$	=	$(0)P_{t=0}(2)$	$+ (P_{1 \rightarrow 1})P_{t=0}(1)$
0	$P_{t=1}(0)$	=	$(0)P_{t=0}(2)$	$+ (P_{1 \rightarrow 0})P_{t=0}(1)$



**Figure 3.9** The expected frequencies of populations with zero, one, or two A alleles over five generations genetic drift. Initially, all populations have one A and one a allele ( $p = q = 0.5$ ). Each generation two gametes are sampled from each population under the Wright–Fisher model to found a new population. This distribution assumes a very large number of independent replicate populations.

However, the transition probabilities from each possible allelic state to each possible allelic state are still determined with the binomial formula in equation 3.13. To obtain the proportion of populations that transition from one state to any state a generation later, the binomial transition probability is multiplied by the proportion of populations in a given allelic state. Using one of the transitions in Fig. 3.10 as an example, the chance that a single population with one A allele at  $t = 1$  transitions to the same state of one A allele is

$$P_{1 \rightarrow 1} = \frac{4!}{1!3!} \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^3 = 0.422 \quad (3.16)$$



**Figure 3.10** Genetic drift modeled by a Markov chain. In this case, the sample size is two diploid genotypes ( $2N = 2$ ) or four gametes per generation. Initial allele frequencies in all populations are  $p = q = 0.5$ . In one generation, sampling error shifts some proportion of the initial populations that contain two copies of each allele to states of zero, one, two, three, or four copies of one allele. Between generations one and two, sampling error again shifts some proportion of the initial populations to states of zero, one, two, three, or four copies of one allele. However, in generation one there are populations present with all allelic states. The arrows represent the possible allelic states produced by sampling error in the third generation for each of the states in the second generation. The bars in the histogram for the third generation are divided by horizontal lines to show the contributions of each second generation allelic state to the total frequency of populations with a given allelic state (some contributions are very small and are difficult to see). As the Markov process continues, the frequency distribution accumulates more and more of the populations at states of zero and four alleles, eventually reaching fixation or loss for all populations.

The proportion of all populations with one A allele at  $t = 1$  is  $4/16$  or 0.25. Therefore,  $(0.422)(0.25) = 0.1055$  is the proportion of many replicate populations that should transition from one A allele at  $t = 1$  to one A allele at  $t = 2$ . Figure 3.10 shows how all such transitions over three generations add together to determine the overall proportions of populations with each allelic state.

### Interact box 3.2 Genetic drift simulated with a Markov chain model

PopGene.S<sup>2</sup> can be used to simulate genetic drift with a Markov chain. Launch PopGene.S<sup>2</sup> and click on the **Drift** menu and then select **Markov Process**. This simulation module requires that you enter parameter values one at a time, since values of some parameters affect the values that other parameters can take.

- Step 1 Start by entering 2 under **Population size** in the upper left corner of the simulation window. This means that there are two diploid individuals or four alleles in each population. Then click **OK**. The fields for the **Transition matrix** will now contain probabilities, with the current state given in the left column and potential future states given in the row across the top of the matrix. The **Probabilities vector** will also have a column of spaces appropriate for a diploid population of the size specified. What does each of the numbers in the **Transition matrix** mean in biological terms?
- Step 2 Next, select the initial allele frequency in individual populations using the pop-out menu above the **Transition matrix**. Each of the allele frequency values in the menu corresponds to an integer number of copies of one allele in a population. For example, a frequency of 0.2500 means one of the four alleles in the population is A. From the pop-out menu select **0.5000** for the initial allele frequency and then click the **OK** button just below the menu.
- Step 3 The histogram in the bottom right of the simulation window will now show the frequency of the A allele *in many replicate populations*. Using the **Generations to run** field you can set the number of generations that elapse between each view of the histogram as well as the **Transition matrix** and **Probabilities vector**. Enter 1 in the **Generations to run** field to be able to track changes each generation. Click the **Start** button once to simulate genetic drift in many populations over a single generation. Why did the transition matrix remain constant? The histogram and the **Probabilities vector** both changed. They give different views of exactly the same information: the proportion of populations out of many replicate populations with a given state (the number of A alleles).
- Step 4 Now press the **Start** button repeatedly, taking time at each generation of the simulation to view the histogram and the **Probabilities vector**. What is the ultimate fate of allele frequencies in all of the replicate populations? How many generations elapse to reach this equilibrium? (see generation counter under the **Start** button)

Reset the simulation using the **Cancel/Restart** button under **Generations to run**. Try the Markov model with population sizes of 4, 20, 50, and 100 with initial allele frequencies of 0.5 and compare the times to fixation and loss with those obtained in Interact box 3.1. Increase the value in the **Generations to run** field to 10 or 20 so that the simulations will progress more rapidly to equilibrium. Also try population sizes of 4, 20, 50, and 100 with initial allele frequencies of 0.1 and 0.9.

When you run the Markov model more than once for a given set of conditions what happens? Why? In contrast, what happens when the simulations in **Interact Box 3.1** are rerun with the same initial conditions? Why?

Markov chains are convenient to model genetic drift because the frequency of populations in a given allelic state depends only on the frequencies in the previous generation (a quality called the **Markov property**). Table 3.2 can be used as a matrix of transition probabilities for any one generation of genetic drift, giving the frequency populations in each allelic state based on the transition probabilities for the number of alleles

sampled and the frequencies of populations in each allelic state in the previous generation. Although a population of one diploid individual is not very interesting in biological terms, it is a convenient case to study mathematically. Using techniques of matrix algebra to determine eigenvalues for the matrix represented by Table 3.2 (see Roughgarden 1996 for a fuller explanation), its is possible to show that the

rate at which genetic variation is lost from the collection of many populations is

$$1 - \frac{1}{2N} \quad (3.17)$$

This says that genetic drift reduces genetic variation by an amount equal to the inverse of twice the effective population size every generation, due to sampling error. (This same conclusion can be reached using the approach of consanguineous mating, as shown in section 3.5.) This rate of loss of genetic variation can clearly be seen in Fig. 3.9, where the frequency of genetically variable populations (those with one A allele) halves each generation since the population size is one. This result applies to any population size and shows us that the effects of genetic drift relate directly to the size of a population.

**Markov chain** A sequence of discrete random variables in which the probability distribution of states at time  $t + 1$  depends only on the states at time  $t$ .

**Markov property** Probability of a given outcome in the next step or time interval depends only on the present state and has no “memory” of states or events before the present time.

Biological populations that closely mimic the ensemble population of Markov chain models are relatively easy to construct and maintain given the right choice of organism and some persistent effort. In fact, the first studies of allele frequencies in many identical replicate biological populations were carried out in the 1950s (for example Kerr & Wright 1954; Wright & Kerr 1954). The organisms of choice were fruit flies (*Drosophila* species) since many individuals can be raised in a small space, generation times are short, and a population can be unambiguously defined as one bottle (containing food) of flies. To rear flies, males and females are put together in a bottle and allowed to mate. The adults are removed from the bottle after the females have time to lay eggs on the food. The larvae that emerge from these eggs and become mature flies can then be sampled to found a new generation in a fresh bottle. Figure 3.11a shows the results of one such classic experiment that followed allele frequencies in 107 replicate populations for 19 generations (Buri 1956). All of the populations

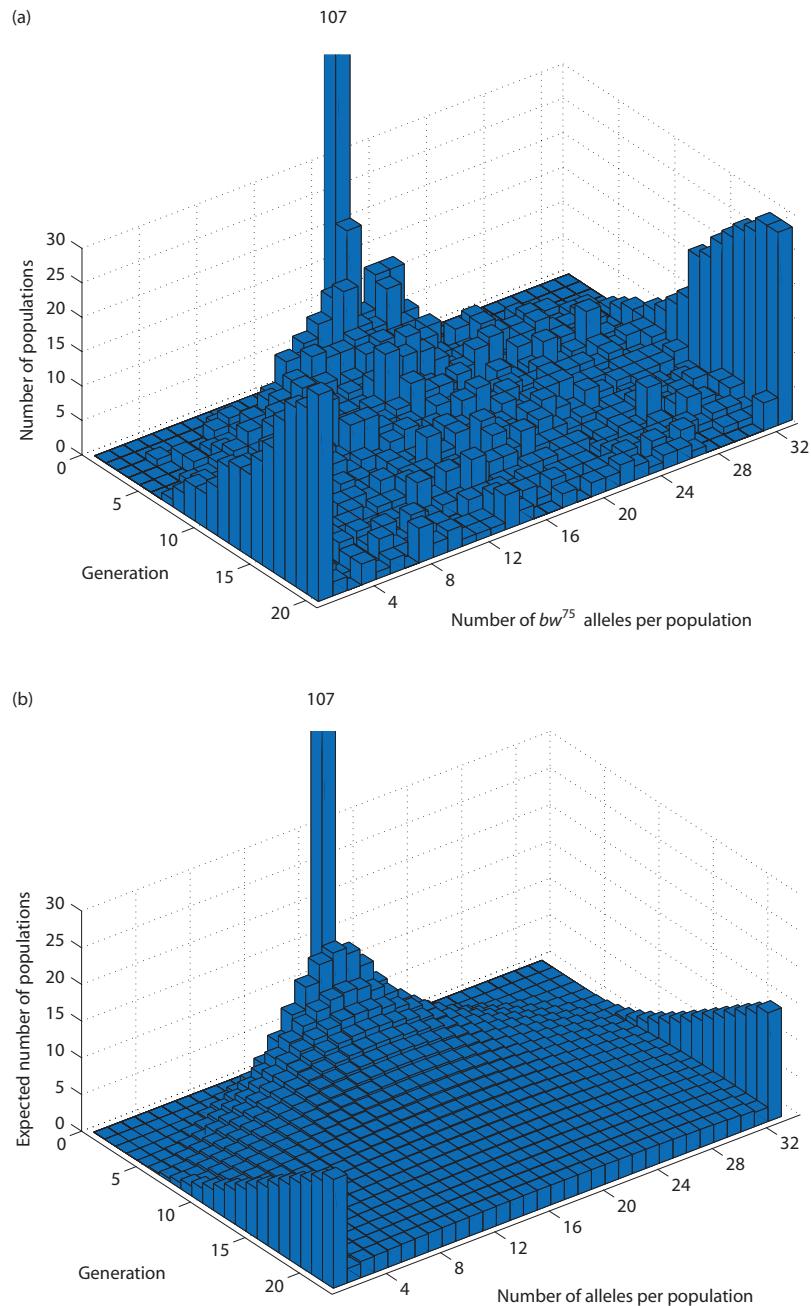
### Problem box 3.2 Constructing a transition probability matrix

Understanding Markov chains is easier with some practice constructing them. Try constructing the transition matrix for a diploid population size of two (identical to the micro-centrifuge tube sampling experiment with a sample size of four tubes). Similar to Table 3.2, set up a matrix where columns represent the initial allelic state and the terms in rows add together to determine the proportion of populations with a given state one generation later. Then use the binomial formula to calculate the chance that a single population makes each of the allelic state transitions. Indicate the frequency of populations in the initial generation ( $t = 0$ ) with a given allelic state by the variable  $P_{t=0}(x)$  where  $x$  is the number of alleles.

Think about the problem before carrying out any calculations. It is less work than it may appear at first. Two columns have probabilities of either zero or one. Two other columns have the same probabilities but in reversed order.

were constructed to have initial allele frequencies of  $p = q = 0.5$  at a diallelic locus (alleles were wild type and *bw*<sup>75</sup>). The distribution of allele frequencies in the 107 populations quickly spread out from the initial frequency. Around the fifth generation a few populations have reached either fixation or loss for the *bw*<sup>75</sup> allele. As more generations elapse, the distribution becomes flatter with more and more populations reaching fixation and loss.

The overall shape of the distribution of population allele frequencies for the fly populations closely matches the expected population frequencies according to a Markov chain model of genetic drift for a population of 16 individuals shown in Fig. 3.11b. In particular, the fly populations and the Markov chain model both show a rapid spread from the initial frequency and an equal number of populations that reach fixation or loss. However, notice that the fly populations have a less even distribution of allele frequencies due to the relatively small number of populations compared to the smoothly continuous



**Figure 3.11** (a) Allelic states (or allele frequencies) for 107 *D. melanogaster* populations where 16 individuals (eight of each sex) were randomly chosen to start each new generation. Initially, all 107 populations had equal numbers of the wild-type and *bw<sup>75</sup>* alleles (the latter causes homozygotes to have a red-orange and heterozygotes an orange body color so genotypes can be determined visually). The allelic states of the population rapidly spread out and many populations reached fixation or loss by the nineteenth generation. (b) The expected frequency of populations in each allelic state determined with a Markov chain model for a population size of 16 with 107 populations that initially have equal frequencies of two alleles. The experimental *D. melanogaster* populations show a higher rate of fixation and loss than the model populations, suggesting that the population size was actually less than 16 individuals each generation. The *D. melanogaster* data come from Table 13 in Buri (1956).

distribution of the model which assumes an infinite number of populations. Another difference is that the fly populations showed more rapid accumulation of fixation and loss compared to the model predictions. Even though the fly populations were founded with eight males and eight females every generation, perhaps not all individuals contributed to reproduction, making the effective population size smaller than it seemed. We will explore why the populations might have behaved as though they were smaller than 16 individuals in the next section of the chapter.

### The diffusion approximation of genetic drift

The Markov chain model has discrete allelic states and time advances from the initial conditions in individual, discrete generations, as is the case in actual biological populations. This discrete step process can be approximated using mathematical expressions where time and allele frequency are continuous variables. This class of model is based on the processes of molecular diffusion and so is termed the **diffusion approximation of genetic drift** (often called the

diffusion equation) first solved by Motoo Kimura (1955). The diffusion approximation is based on partial differential equations and advanced mathematical techniques beyond the scope of this text. However, the general principles behind diffusion equations can be understood, especially with the aid of a physical example. The goal of this section is to introduce the situation that diffusion equations model using a particle metaphor and then to cover some of the conclusions about the process of genetic drift that have been reached using the diffusion equation. This introduction to the physical process of diffusion relies heavily on Denny and Gaines (2000), to which readers can turn for more details and biological applications.

Diffusion is the process where particles, moving in random directions, spread out and eventually reach a uniform concentration within the physical boundaries that limit their movement. For example, imagine putting a drop of ink in the center of a Petri dish filled with water. Initially, the concentration of ink is very uneven but as time passes the ink will diffuse and eventually reach a uniform concentration everywhere in the Petri dish. The rate at which the ink spreads out depends on what is called a **diffusion coefficient**. To understand the diffusion coefficient, we have to examine random movement of particles in some detail. This will hopefully lead to an improved understanding of genetic drift, so please be patient.

Let's modify the ink-diffusion example by substituting a special Petri dish. In this imaginary dish the

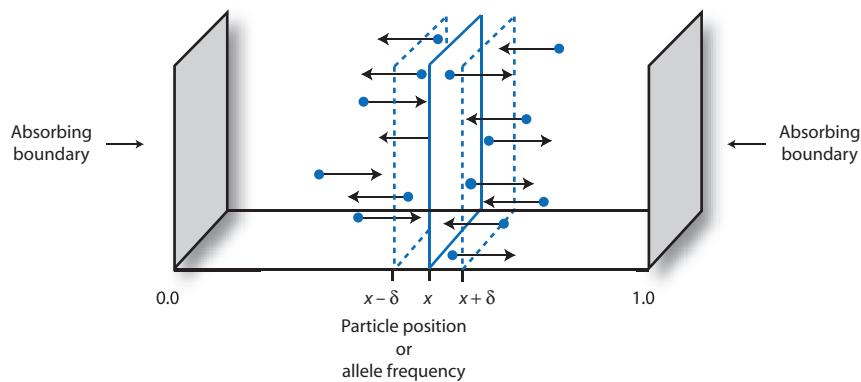
ink particles can move only to the left or to the right from their current position, a situation diagrammed in Fig. 3.12. The particles have a constant velocity, so each will move the same distance in a fixed amount of time. We can call this distance moved per unit time  $\delta$  (pronounced “delta”) since it is the change in position to the left or right. The direction of particle movement is random, with equal probability of moving to the left ( $p = 1/2$ ) or right ( $q = 1/2$ ) at any moment in time. Let's pick a point of reference somewhere along this axis of movement, call it  $x$ , and then track the movement of the ink particles relative to that point. First, what is the average movement of  $N$  particles between two time points? There are  $p$  of the particles traveling toward  $x$  that each move the distance  $+\delta$ . There are also  $q$  of the particles traveling away from  $x$  that each move the distance  $-\delta$ . The average movement of the particles is then

$$\bar{x} = p(\delta) + q(-\delta) \quad (3.18)$$

Since the particles have equal chances of moving left or right ( $p = q = 1/2$ ):

$$\bar{x} = \frac{1}{2}(\delta) - \frac{1}{2}(-\delta) = 0 \quad (3.19)$$

which means that the average or net movement of particles is zero. To relate this to genetic drift, if populations are like the ink particles but moving randomly in the one dimension of allele frequency,



**Figure 3.12** An imaginary Petri dish that confines ink particles so that they can move only to the left or to the right from their current position. The particles have a constant velocity, so each will move the distance  $\delta$  in a fixed amount of time. If the direction of particle movement is random (equal probability of moving left or right at any moment in time), the mean position of particles does not change but the variance in particle position increases with time. The frequency of particles passing through an area, such as the plane at  $x$ , depends on the net balance of particles arriving minus those that are leaving, called the flux of particles. The flux is determined by both the rate of diffusion of particles and gradients in the concentration of particles (net movement of particles is from areas of higher concentration to areas of lower concentration). If the left and right boundaries capture particles, then the diffusion coefficient drops to 0 at those points and particles will accumulate. The process of diffusion for particles is analogous to the process of genetic drift for allele frequencies in an ensemble population where allele frequencies “diffuse” because of sampling error.

then we expect equal numbers of populations to move toward fixation and toward loss. The *average* change in allele frequency among all populations is expected to be zero.

The next thing we could do is to describe the variance in the position of ink particles over time, a measure of how spread out the particles become. Intuition suggests that even though the average is zero the variance should not be zero: spreading out of particles is what occurs during diffusion, after all. Equation A.2 in the Appendix shows that the variance is the average squared deviation from the mean. We just showed the mean particle location is zero. So the variance in the location of particles is then just the average square of their positions after one time step. The location of one particle, call it particle  $i$ , at time  $t = 1$  can be expressed as its location at time  $t = 0$  plus the amount a particle moves during one time step:

$$x_{i(t=1)} = x_{i(t=0)} + \delta \quad (3.20)$$

To get an expression for the variance in particle location we need to start out by squaring this expression for particle location:

$$x_{i(t=1)}^2 = (x_{i(t=0)} + \delta)^2 \quad (3.21)$$

and expanding the right side to get

$$x_{i(t=1)}^2 = x_{i(t=0)}^2 + 2x_{i(t=0)}\delta + \delta^2 \quad (3.22)$$

Using equation 3.22, which is the squared position of one particle, we can average over all  $N$  particles to get the variance in particle position:

$$\sigma^2(x_{i(t=1)}) = \frac{1}{N} \sum_{i=1}^N x_{i(t=0)}^2 + 2x_{i(t=0)}\delta + \delta^2 \quad (3.23)$$

This expression simplifies considerably. The value of  $\delta$  for a large number of particles should be zero, using the same reasoning as when determining average particle position, since an equal number are moving left and right ( $\delta^2$  is *not* zero because the *squared* change in position will always be positive). The middle term in equation 3.23 then drops out since it is multiplied by zero. This leaves the variance in particle position as

$$\sigma^2(x_{i(t=1)}) = \bar{x}_{i(t=0)}^2 + \delta^2 \quad (3.24)$$

The first term is the average of the squared particle position at time  $t = 0$ . The second term is the square

of step length that particles take between time points. If a group of particles all started out at position zero (meaning  $\bar{x}_{i(t=1)}^2 = 0$ ), then the variance in particle position increases by  $\delta^2$  every time interval. If  $t$  is the number of time steps that have elapsed for particles that started out at position zero, the variance in particle position is  $t\delta^2$ . As intuition suggests after watching things like ink diffuse in water, the variance in particle position is not zero and increases with time.

Now we are ready to return to the diffusion coefficient. The diffusion coefficient ( $D$ ) is defined as half the rate at which the variance in particle position changes as time advances. In symbols this is

$$D = \frac{1}{2} \frac{d\sigma^2}{dt} \quad (3.25)$$

The source of the factor of  $1/2$  can be seen in Fig. 3.12. Only half of the particles near the point  $x$  (within  $\delta$  or one step of the plane) will be headed away and increasing their dispersion while the other half will be headed toward  $x$  and not dispersing. Half the variance in particle position,  $\frac{\delta^2}{2}$ , is therefore the diffusion coefficient for physical molecules. The diffusion coefficient tells us how fast particles spread out around some point due to random movement.

Allele frequency in an ensemble population has an analog of the diffusion coefficient. Allele frequency “diffusion” is the spreading out and flattening of the allelic state distributions over time as seen in Markov chain models (see Fig. 3.11). Recall from equation 3.6

that the standard error of the allele frequency is  $\sqrt{\frac{pq}{2N}}$ ,

which can also be thought of as the standard deviation of the mean allele frequency. The variance of the mean allele frequency is then the square of the standard deviation or  $\frac{pq}{2N}$ . This latter quantity is the variance per time period or per generation, which could be expressed as

$$\frac{pq}{2N} = \frac{d\sigma^2}{dt} \quad (3.26)$$

By substituting equation 3.26 into equation 3.25 we obtain an expression for the diffusion coefficient of allele frequency:

$$D = \frac{1}{2} \frac{pq}{2N} \quad (3.27)$$

along the one-dimensional axis of allele frequency. By substituting  $q = 1 - p$  we obtain

$$D = \frac{p - p^2}{4N} \quad (3.28)$$

The genetic diffusion coefficient depends on both the allele frequency and the size of the population. Diffusion of allele frequency is greatest when  $p = q = 1/2$  and declines to zero as  $p$  approaches zero or one. Populations (as with particles) tend to diffuse to areas where the diffusion coefficient is lowest and then get stuck there since the rate of spread of particles (the variance in position per time step) is reduced. The rate of diffusion also depends on the size of the population, decreasing as  $N$  increases. For particles, this is due to more frequent collision that reduces the ability to move as the concentration of particles increases (think of trying to walk in a straight line while in a large crowd of people). In biological populations, the diffusion coefficient depends on  $N$  since the population size determines the amount of sampling error from generation to generation. It is satisfying that both of these features of allele frequency diffusion agree with our previous generalizations about genetic drift obtained with distinct approaches to the problem.

Next, we would like to keep track of the chance that a particle is at a given location along the axis of diffusion. The resulting probability distribution shows how many particles out of a large number should be at each point along the axis, just as Markov chain models show the expected number of populations at each allelic state. Making such a probability distribution requires that we know the **flux** or the net number of particles moving through a defined area per time interval. Let's define the area, call it  $A$ , where we will determine the flux through the plane at the point  $x$  (Fig. 3.12). The particles that will move through plane  $A$  in one time step must be within plus or minus  $\delta$  of  $x$  because a particle travels the distance  $\delta$  in one time step. The net number of particles moving to the right through  $A$  is the same thing as the difference in the number of particles moving from the left and from the right:

$$\text{Net } N(R) = \frac{1}{2}N(L) - \frac{1}{2}N(R) \quad (3.29)$$

where  $N$  represents the number of particles moving left (L) or right (R) through  $A$ , and the  $1/2$  is because only half of the particles on each side of  $x$  will move

toward  $x$  each time step. Factoring and rearranging gives

$$\text{Net } N(R) = -\frac{1}{2}[N(R) - N(L)] \quad (3.30)$$

The flux is defined per area per time, so we need to divide by area ( $A$ ) and time ( $t$ ):

$$J_x = -\frac{1}{2} \frac{[N(R) - N(L)]}{At} \quad (3.31)$$

where  $J_x$  represents the flux at point  $x$ . We can multiply equation 3.31 by  $\delta^2/\delta^2$  (or 1) and then rearrange it to get

$$J_x = -\frac{\delta^2}{2t} \frac{1}{\delta} \left[ \frac{N(R)}{A\delta} - \frac{N(L)}{A\delta} \right] \quad (3.32)$$

Now notice that the number of particles moving left and moving right are both divided by an area times a distance along the axis of diffusion ( $A\delta$ ). As you can see in Fig. 3.12, this defines a three-dimensional volume. The number of particles per volume is equivalent to a concentration, so we can use  $C$  to represent the concentration of particles. Also notice that  $\frac{\delta^2}{2}$  is the diffusion coefficient for one time step (equation 3.25). Making these substitutions gives

$$J_x = -D \frac{C(R) - C(L)}{\delta} \quad (3.33)$$

If we say that the concentration of particles to the right of area  $A$  is taken at location  $x + \delta$  [ $C(R) = C(x + \delta)$ ] and the concentration of particles to the left is actually taken at area  $A$  [ $C(L) = C(x)$ ], the fraction in equation 3.33 takes the form of a first derivative:

$$J_x = -D \frac{C(x + \delta) - C(x)}{\delta} \quad (3.34)$$

(A straightforward refresher on methods and interpretations for derivatives can be found in Newby (1980).) As the distance between  $x$  and  $\delta$  shrinks toward zero ( $\lim_{\delta \rightarrow 0}$ ), the flux at any point  $x$  becomes

$$J_x = -D \frac{dC}{dx} \quad (3.35)$$

In total, the flux is determined by the product of the rate at which particles spread out from a given

point and the rate of change in concentration along the axis of diffusion. The sign of the flux tells the direction of net movement of particles. The flux is positive, meaning the number of particles in the area to the right of  $A$  will increase after one time step, if more particles move in from the left than move out going right. As we would expect, there is net movement of particles from areas of higher concentration into areas of lower concentration. For example, a higher concentration of particles to the left of point  $x$  in Fig. 3.12 will result in a positive flux, meaning that after one time step of diffusion there will be an increase in the number of particles at point  $x$ . When the concentration of particles is the same everywhere along the axis of diffusion, the flux must be zero since the numbers that move into and out of any area are equal.

**Diffusion coefficient ( $D$ )** Half the rate at which the variance in particle position (or allele frequency in a single population) changes as time advances.

**Flux ( $J_x$ )** The net number of particles (or populations) moving through a defined area (or allele frequency) per time interval.

Let's now take all the concepts of particle diffusion and apply them to genetic drift in an ensemble population. We want to predict the change in the chance that a population has a given allele frequency, just as we did with the Markov chain model. In this case allele frequency is a continuous variable and the chance that a population has an allele frequency between  $x$  and  $x + \delta$  at a given time  $t$  is called the probability density, symbolized by  $\phi(x,t)$ . ( $\phi$  is pronounced "phi".) Probability density for populations is just like concentration for particles, so  $\phi(x,t)$  is the analog of  $C(x,t)$ . The probability density  $\phi(x,t)$  at any point along the axis of allele frequency will depend on the net difference between populations which drift into allele frequency  $x$  and those which drift out of allele frequency  $x$ . This is the flux in allele frequency at point  $x$  and time  $t$ . To know the probability at all points along the axis of diffusion, we need the rate of change in the probability density with change in allele frequency. This is the rate of change in the flux of populations with change in allele frequency:

$$\frac{d}{dx}\phi(x,t) = -\frac{d}{dx}J(x,t) \quad (3.36)$$

The derivation for particle flux hid one detail that we now need to reveal in order to continue. The flux depends on both the mean movement and the net movement of particles. Imagine for example that the ink particles in Fig. 3.12 were positively charged and one of the boundaries was negatively charged. The ink particles would diffuse but at the same time the whole cloud of particles would be moving on average toward the negatively charged boundary. In such a case the flux at any point  $x$  would also need to account for changes to the mean position of particles  $M(x)$ , so that  $J_x = M(x)dt - D\frac{dC}{dx}$ . This mean change was neglected earlier since the mean position is zero if particles are moving left or right at random and are not influenced by some force changing the mean location of all particles. Substituting this full version of the flux (remember that  $C(x,t)$  is now  $\phi(x,t)$ ) gives

$$\frac{d}{dx}\phi(x,t) = -\frac{d}{dx}\left[M(x)\phi(x,t) - D\frac{d\phi(x,t)}{dx}\right] \quad (3.37)$$

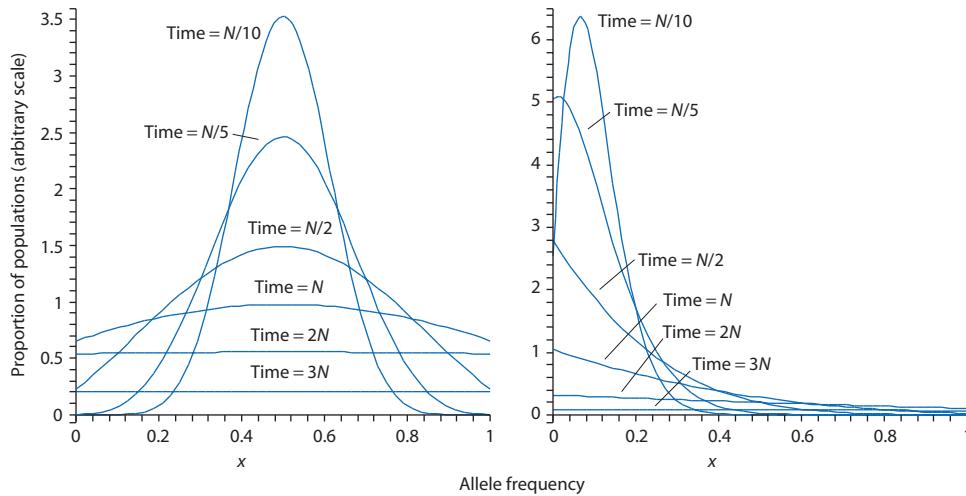
We can also substitute the flux in allele frequency from equation 3.27 (using  $x$  to represent  $p$  and  $1 - x$  to represent  $q$ ):

$$\frac{d}{dx}\phi(x,t) = -\frac{d}{dx}\left[M(x)\phi(x,t) - \left(\frac{1}{2}\frac{x(1-x)}{2N}\right)\frac{d\phi(x,t)}{dx}\right] \quad (3.38)$$

With only random sampling error acting to change allele frequency, ( $M(x)dt = 0$ ), this rearranges to the diffusion equation for genetic drift:

$$\frac{d}{dx}\phi(x,t) = \frac{1}{4N}\frac{d^2}{dx^2}[x(1-x)\phi(x,t)] \quad (3.39)$$

The diffusion equation predicts the probability distribution of allele frequencies in many populations over time and some examples are given in Fig. 3.13. Compare Fig. 3.13 with Fig. 3.10 and it is apparent that the diffusion equation and the Markov chain model both make similar predictions for the outcome of genetic drift. A final point is that the term diffusion approximation implies that the diffusion equation makes some assumptions. Noteworthy assumptions are that the number of populations is very large, approaching infinity, and the allele frequency distribution is continuous so that the distribution of allele frequencies is a smooth curve (compare these assumptions with the allelic state distribution in



**Figure 3.13** Probability densities of allele frequency for many replicate populations predicted using the diffusion equation. The initial allele frequency is 0.5 on the left and 0.1 on the right. Each curve represents the probability that a single population would have a given allele frequency after some interval of time has passed. The area under each curve is the proportion of alleles that are not fixed. Time is scaled in multiples of the effective population size,  $N$ . Both small and large populations have identically shaped distributions, although small populations reach fixation and loss in less time than large populations. The populations that have reached fixation or loss are not shown for each curve. For a color version of this image see Plate 3.13.

Fig. 3.11a with its discrete allelic states and finite number of populations).

The diffusion equation has been used to arrive at a number of generalizations about genetic drift. A widely used set of generalizations is the average time to fixation for alleles that eventually fix in a population and the average time to loss for alleles that eventually are lost from a population:

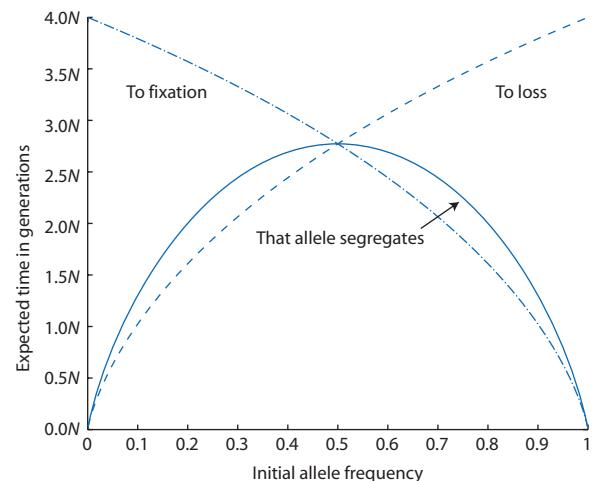
$$\bar{T}_{fix} = -4N \frac{(1-p)\ln(1-p)}{p} \quad \text{and} \quad \bar{T}_{loss} = -4N \frac{p\ln(p)}{1-p} \quad (3.40)$$

where  $p$  is the initial allele frequency (Kimura & Ohta 1969a). (Note that the natural log of a number less than one is always negative:  $\ln(1) = 0$  and  $\ln(x) \rightarrow -\infty$  as  $x$  approaches zero, so that the average time will always be a positive number.) These two expressions can be combined to obtain the weighted average time that an allele segregates in a population (the allele is neither fixed nor lost):

$$\bar{T}_{segregate} = -4N[p\ln(p) + (1-p)\ln(1-p)] \quad (3.41)$$

The predictions from these equations quantify our intuition about the action of genetic drift (Fig. 3.14). Alleles close to fixation or loss do not take long to reach fixation or loss. Alternatively, an allele initially very close to fixation (or loss) would take a long time, about  $4N$  generations if  $N$  is large, if it were to reach

the opposite condition of being very close to loss (or fixation). Also, the closer an allele is to the initial frequency of  $1/2$ , the longer it will segregate before reaching fixation or loss up to a maximum of about  $2.8N$  generations. The curves for times to fixation or loss and segregation times have an identical shape



**Figure 3.14** Average time that an allele segregates, takes to reach fixation, or takes to reach loss depending on its initial frequency when under the influence of genetic drift alone. Alleles remain segregating (persist) for an average of  $2.8N$  generations when their initial frequency is  $1/2$ . Fixation or loss takes up to an average of  $4N$  generations when alleles are initially very rare or nearly fixed, respectively. Since these are average times, alleles in individual populations experience longer and shorter fixation, loss, and segregation times. Time is scaled in multiples of the population size.

no matter what the population size is. The population size plays a role only in the absolute average number of generations that will elapse.

If you have worked your way through this section you deserve congratulations for your persistence. The basis of the diffusion equation is definitely more abstract than the basis of Markov chains, but the overall results provided by the two models are very similar. Those who would like to learn more about the diffusion equation, its assumptions, and how it can be extended to include processes such as mutation, migration, and natural selection along with genetic drift, can consult Roughgarden (1996) and Rice (2004).

### 3.3 Effective population size

- Defining genetic populations.
- Census and effective population size.
- Example of bottleneck and harmonic mean to demonstrate effective population size and census size.
- Effective population size due to unequal sex ratio and variation in family size.

Up to this point we have used the term population size without much fanfare to indicate how many individuals a population contains. We now need to focus additional attention on the idea of population size. The number of individuals in a population seems like a straightforward quantity that can be determined easily. In the context of the Wright–Fisher model the population size is an unambiguous quantity. Unfortunately, in most biological populations it is difficult or impossible to determine the number of gametes that contribute to the next generation. We need another way to define the size of populations.

The definition of the population size in population genetics relies on the dynamics of genetic variation in the population. This definition means that the size of a population is defined by the way genetic variation in the population *behaves*. The notion that “if it walks like a duck and quacks like a duck, it probably is a duck” is also applied to the size of populations. The size of a population depends on how genetic variation changes over time. If a population shows allele frequencies changing slowly over time under the exclusive influence of genetic drift, then the population has the dynamics associated with relatively large size. It “quacks” like a big population. In the same way, a population with a large number of individuals might show rapid genetic drift, indicating it is really a small population

from the perspective of genetic variation. It looks big but its “quack” gives it away as a small population.

Making a distinction between the dynamics of genetic variation in a population and the number of individuals in a population suggests that there are really two types of population size. One is the head count of individuals in a population, called the census population size, symbolized by  $N$ . The other is the genetic size of a population. This genetic size is determined by comparing the rate of genetic drift in an actual population with the rate of genetic drift in an ideal population meeting the assumptions of the Wright–Fisher model. The population size in the model that produces that same rate of genetic drift as seen in an actual population is the genetic size of the actual population. In comparing an actual population with an ideal model population, we are asking about the overall genetic effects of the census size. Thus, we also recognize the effective population size,  $N_e$ , as the size of an ideal population that experiences as much genetic drift as an actual population regardless of its census size. This concept was originally introduced by Sewall Wright (1931), who is shown in Fig. 3.15. An approximate way to think of the difference between the two population sizes is that the census size is the total number of individuals and the effective size is the number of individuals that actually contribute gametes to the next generation. We will refine this definition throughout this rest of the chapter.

**Census population size ( $N$ )** The number of individuals in a population; the head count size of a population.

**Effective population size ( $N_e$ )** The size of an ideal Wright–Fisher population that maintains as much genetic variation or experiences as much genetic drift as an actual population regardless of census size.

Let's examine several biological phenomena that cause effective population size and census population size to be different. This will help to illustrate the effective population size and make its definition more intuitive.

Actual populations often fluctuate in the number of individuals present over time. A classic example is rabbit/lynx population cycles due to predator/prey dynamics, where census population sizes of both species fluctuate over a fairly wide range on about a 10 year cycle. Another category of example



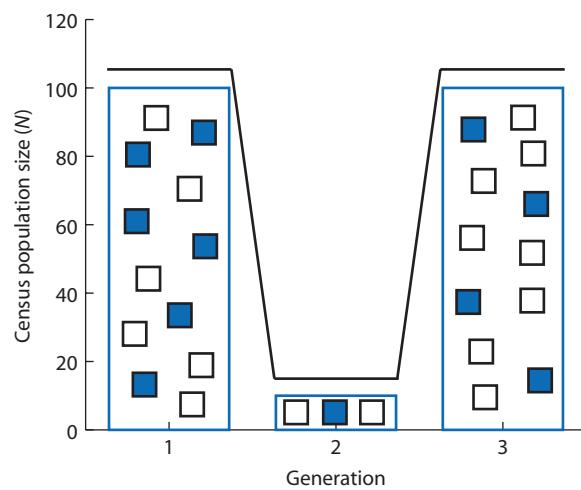
**Figure 3.15** Sewall Wright (1889–1988) with a guinea pig in an undated photograph taken during his years as a professor at the University of Chicago. Starting in 1912 and throughout his career, Wright studied the genetic basis of coat colors and physiological traits in guinea pigs. Wright, along with J.B.S. Haldane and R.A. Fisher, established many of the early expectations of population genetics using mathematical analyses. Many of the conceptual frameworks in population genetics today were originated by Wright, especially those related to consanguineous mating, genetic drift, and structured populations. An often retold (although mythical) story was that Wright, who would sometimes carry a guinea pig with him, would on occasion absent-mindedly employ the animal to erase the chalk board while lecturing. Provine's (1986) biography details Wright's manifold contributions to population genetics and his interactions with other major figures such as Fisher. Photograph courtesy of Special Collections Research Center, University of Chicago Library, Archival Photofiles, Series 1, Wright, Sewall, Informal #2.

is the establishment of a new population by a small number of individuals, called a **founder event**. One well-documented founder event was the introduction of European starlings in the New World. These birds, now very common throughout North America, can all be traced to approximately 15 pairs that survived from a larger group released in New York's Central Park in 1890. What is now a very large population descended from a sample of 60 alleles in the small number of founding individuals, which were sampled from a very large population of birds in Europe.

To model the genetic effects of this type of fluctuation in population size over time, suppose a population starts out with 100 individuals, experiences a reduction in size to 10 individuals for one generation, and then recovers to 100 individuals in the third generation (Fig. 3.16). (Recall from earlier in the chapter that this situation violates the constant population size assumption of the Wright–Fisher model of genetic drift.) This will cause an increased chance of fixation or loss of alleles (variance in allele frequency will increase) and thereby increase the rate of genetic drift in that one generation. But what is the effective size of this population after it recovers to 100 individuals? We can estimate the effect of fluctuations in populations on the overall effective size using the **harmonic mean**:

$$\frac{1}{N_e} = \frac{1}{t} \left[ \frac{1}{N_{e(t=1)}} + \frac{1}{N_{e(t=2)}} + \cdots + \frac{1}{N_{e(t)}} \right] \quad (3.42)$$

where  $t$  indicates the total number of generations (all other assumptions of Wright–Fisher populations are met). The harmonic mean gives more weight



**Figure 3.16** Schematic representation of a genetic bottleneck where census population size fluctuates across generations. The harmonic mean of census population size ( $N$ ) is 25 and provides an estimate of the effects of genetic drift over three generations, or the effective population size ( $N_e$ ). In other words, this population of fluctuating size would experience as much genetic drift as an ideal Wright–Fisher population with a constant population size of 25. The squares represent alleles present in each generation. In the first generation the alleles are equally frequent, but they end up at frequencies of 25 and 75%. Such an allelic state transition would be extremely unlikely if 200 gametes were sampled to found generations two and three. However, the observed allelic transition is expected 25% of the time in a sample of four gametes.

to small values by virtue of summing the inverses of population size. It also serves as an approximation of  $\left(1 - \frac{1}{2N_e}\right)^t$ , which we saw was the rate of decrease

in genetic variation in Markov chain models of genetic drift earlier in the chapter. In this example:

$$\frac{1}{N_e} = \frac{1}{3} \left[ \frac{1}{100} + \frac{1}{10} + \frac{1}{100} \right] \quad (3.43)$$

So,  $N_e = 1/0.04 = 25$ . Contrast this with the arithmetic mean of the census population size, which is 70. Only those alleles that actually pass through the **genetic bottleneck** of 10 individuals are represented in later generations, regardless of how large the census size is, so the mean census size is much too high to use to predict the behavior of allele frequencies since it will underestimate genetic drift. In this case, we expect allele frequencies in the population to behave similarly to allele frequencies in an ideal Wright–Fisher population with a constant size of 25 over three generations. Like finite population size, fluctuations in population size through time are a universal feature of biological populations. Populations obviously vary greatly in the degree of size fluctuations and the time scale of these fluctuations, but  $N_e < N$  caused by temporal fluctuations in  $N$  is a widespread phenomenon.

**Founder effect** The establishment of a population by one or a few individuals, resulting in small effective population size in a newly founded population.

**Genetic bottleneck** A sharp but often transient reduction in the size of a population that increases allele frequency sampling error and has a disproportionate impact on the effective population size in later generations even if census sizes increase.

Although the term bottleneck is usually associated with sharp reductions in the overall population size, there are other aspects of biological populations that have the same impact by increasing the sampling error in allele frequency across generations. Mating patterns can have a major impact on effective population size when individuals of different sexes make unequal contributions to reproduction. This occurs in populations where individuals of one sex compete

for mating access to individuals of the other sex. In such a situation the numbers of females and males that breed, or the breeding sex ratio, may not be equal (even if the population sex ratio is equal). This leads to increased genetic drift compared with the case of a breeding sex ratio of 1 : 1, since the pool of alleles passed to the next generation will be sampled from fewer individuals in one sex. Thus, the less frequent sex becomes an allelic bottleneck of sorts. The effective population size in such cases is:

$$N_e = 4 \frac{N_m N_f}{N_m + N_f} \quad (3.44)$$

where  $N_f$  is the number of females and  $N_m$  is the number of males *breeding* in the population and all other assumptions of Wright–Fisher populations are met. Equation 3.44 shows that the effective population size approaches four when the rarer sex approaches a single individual and that the effective size is maximized when there is a breeding sex ratio of 1 : 1.

Let's look at a case where  $N_m$  does not equal  $N_f$  to see the impact on the effective population size. Elephant seals (*Mirounga leonina*) are a classic example of highly unequal breeding sex ratios since the mating system is harem polygyny. In one study of breeding patterns on Sea Lion Island in the Falkland Islands, about 550 females and 75 males were observed on land where mating takes place (Fabiani et al. 2004). Using genetic markers to ascertain the parentage of pups, it was determined that only 28% of the males fathered offspring during the course of two breeding seasons. Therefore, the breeding sex ratio was about  $N_m = 21$  and  $N_f = 550$ . The effective population size during each breeding season was:

$$N_e = 4 \frac{(21)(550)}{21 + 550} = 80.91 \quad (3.45)$$

or was equivalent to an ideal population of 40 females and 40 males where breeding sex ratio is 1 : 1. The strongly unequal breeding sex ratio for elephant seals results in an effective population size an order of magnitude less than the census size of 625 individuals.

A third factor that distinguishes the census and effective population sizes is the degree to which adult individuals in the population contribute to the next generation. One of the assumptions of Wright–Fisher populations is that all individuals contribute an equal number of gametes to the infinite gamete pool. To maintain a population that is not changing in size

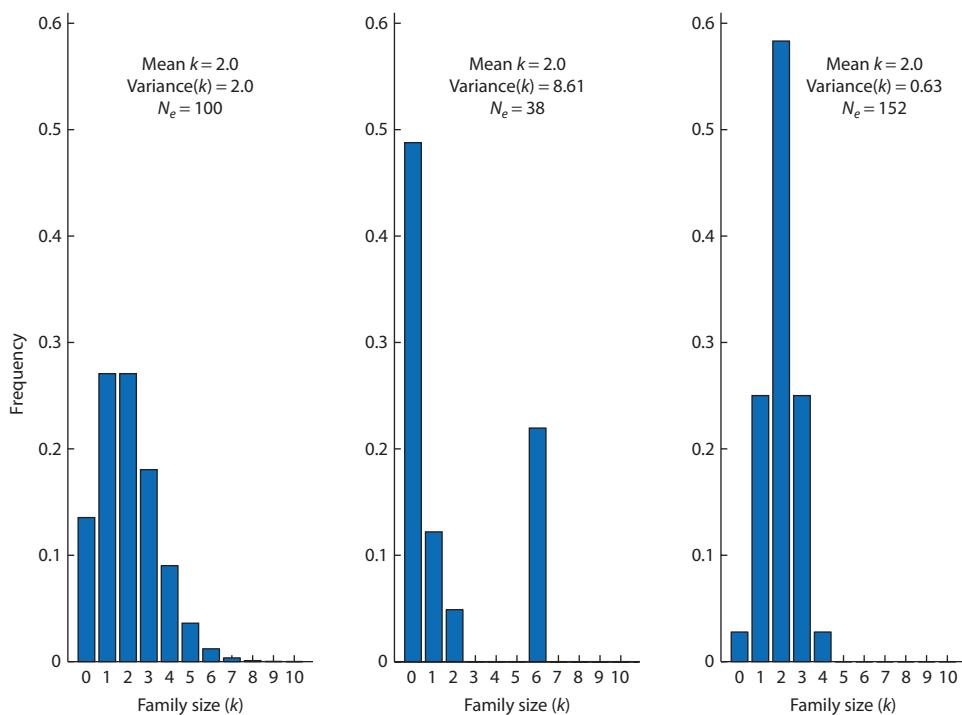
across generations, each individual must produce one surviving progeny, on average, to replace itself. In outcrossing species, when each pair of individuals produces an average of two progeny, then the population will be stable in size over time. In terms of the Wright–Fisher population, that means that each individual contributes an *average* of two gametes to the next generation from the infinite gamete pool.

There are many patterns of individual reproduction within an outcrossing population that can achieve a *mean* rate of reproduction that results in a stable population through time. It might be the case that all individuals produce exactly one progeny. Another possibility is that a few parents produce no offspring, whereas most parents produce two offspring, and a few parents produce four offspring that offset the reduction in the average caused by the parents with no progeny. In the extreme, one pair of parents could produce all  $N$  offspring and all other pairs fail to reproduce successfully. The variance in family size can be used to describe these different patterns of individual reproduction. As variance in family size increases, the alleles passed to the next generation come increasingly from those parents producing more offspring.

The effective population size due to variation in family size is:

$$N_e = \frac{4N_{t-1}}{\text{var}(k) + \bar{k}^2 - \bar{k}} \quad (3.46)$$

where  $N_{t-1}$  is the size of the parental population and  $k$  is the number of gametes that result in progeny, or family size for outcrossing organisms (Crow & Denniston 1988). The equation shows that for a stable population ( $\bar{k} = 2$ ) when the variance in family size is equal to the average family size, then there is no “bottleneck” due to family size variation: the population size of parents is the effective population size. The Wright–Fisher model assumes that production of progeny has exactly this quality of the mean family size being equal to the variance in family size. In essence the Wright–Fisher model assumes that family sizes follow a Poisson distribution (the Poisson distribution is an approximation of the binomial distribution as the sample size grows very large and the chance of a given outcome becomes very small; see Denny & Gaines 2000 or the appendix in Rice 2004).



**Figure 3.17** Distributions of family size. The variance equals the mean as expected for a Poisson distribution on the left. However, the center distribution has a few families that are very prolific while 75% of the families produce two or fewer progeny with most individuals failing to reproduce. The distribution on the right has less variance in family size than expected for a Poisson distribution with most families of size two. The Poisson distribution is taken as the standard with an effective size of 100. By comparison, the center distribution has a smaller effective population size and the distribution on the right a larger effective population size.

We can explore the consequences of variance in family size with some examples. Figure 3.17 shows three hypothetical distributions of family size. The first is an ideal Poisson distribution where the mean family size is equal to the variance in family size. This is the standard assumption used in the Wright–Fisher model of genetic drift. The next distribution is an example of highly skewed family sizes where a relatively small proportion of the population contributes most of the progeny. The final distribution shows family size variation that is less than expected for a Poisson distribution. If the Poisson distribution is used as the standard, the other populations with differing distributions of family size show more or less genetic drift, respectively, due to modification of the bottleneck-like effect of unequal family sizes. These distributions also illustrate that the effective population size based on variance in family size has the unique quality that  $N_e$  can actually be larger than  $N$  if the variance in family size is less than the mean of family size. This stands in contrast to population size fluctuations through time and unequal breeding sex ratios, where  $N_e$  can only be less than or equal to  $N$  but not greater than  $N$ .

These family size distributions are not just theoretical entities. Many annual plants show variation in reproductive success that exceeds the mean, demonstrating that variation in family size contributes to overall rates of genetic drift (Heywood 1986). In one study of salmon, the large variance in reproductive success among anadromous males had a greater impact on the effective population size than breeding sex ratio (Jones & Hutchings 2002). In contrast, Poisson-distributed male reproductive success has been observed in laboratory populations of *D. melanogaster*, partly supporting the effective population size assumption behind many genetic experiments that have used fruit flies reared in the laboratory (Joshi et al. 1999).

In the last section of the chapter, we compared allele frequencies over time in 107 *Drosophila* populations to allele frequencies expected from the Markov chain model (Fig. 3.11). The fly populations, founded each generation with eight female and eight male flies, experienced a faster rate of fixation or loss than expected for an effective population size of 16. In fact, the fly populations reached fixation and loss at a rate comparable to a population with an effective size of about 10 or 11 (this can be seen using PopGene.S<sup>2</sup>'s Markov chain module as in Interact box 3.2). The concepts in this section that distinguish between census and effective population sizes can be used to

### Problem box 3.3 Estimating $N_e$ from information about $N$

Imagine that a conservation biologist approaches you asking for assistance in estimating the genetic impacts of a recent event in a captive population of animals housed in a zoo. The zoo building where the animals were kept experienced a fire, killing some animals outright and requiring the survivors to be relocated to a new enclosure that is not ideal for breeding. Before the fire, the population was stable at 30 males and 30 females for many generations with an effective population size of 60. After the fire there were 15 females and 10 males. Due to the disruption and relocation of the animals, breeding behavior changed. Before the fire variation in family size was Poisson distributed with a mean of 2.0. In the one generation after the fire, family size has a mean of 4.0 and variance of 6.5. What are the genetic impacts of the fire on the effective population size? What are some of the assumptions specific to this case used in your estimate of  $N_e$ ?

explain Buri's results. Although there was a census size of 16 flies in each bottle, an unequal breeding sex ratio in each bottle could explain the higher rate of fixation and loss. For example, a breeding sex ratio of eight females and six males due to failure of some males to mate successfully each generation would give an effective population size of  $N_e \approx 14$  using equation 3.44. It is also possible that there was a relatively high degree of variation in reproduction among the females. For example, if variance in family size was 3.5 and was combined with the effects of the unequal breeding sex ratio (using 14 instead of 16 for  $N_{t-1}$ ), equation 3.46 estimates that  $N_e \approx 10$ . It might also be that in a few of the generations the population size was smaller than intended due to mistakes when handling and transferring flies to new bottles. However, equation 3.42 shows that the effect of a population size of 14 for one generation out of 19 is slight ( $N_e = 15.88$ ). Therefore, infrequent fluctuating population sizes would probably have had only a minor impact on the results. Thus, the difference

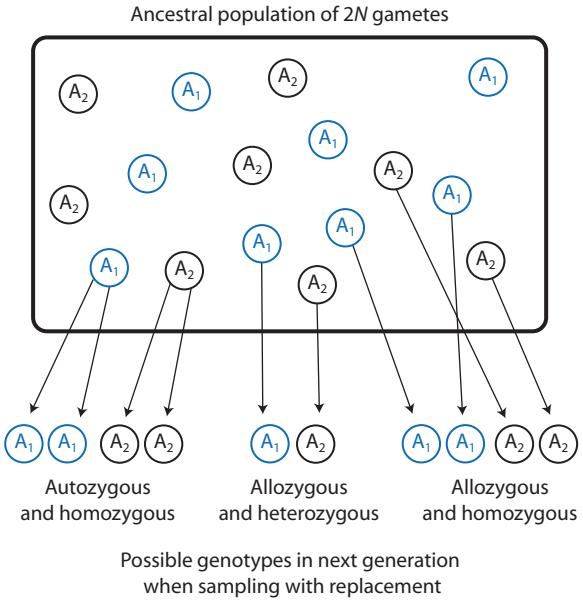
between the rate of fixation and loss in the Markov chain model and in the actual fly populations can be explained by several plausible factors that distinguish the census and effective population sizes.

### 3.4 Parallelism between drift and inbreeding

- Autozygosity due to sampling in a finite gamete population.
- The relationship between the fixation index ( $F$ ) and heterozygosity ( $H$ ).
- Decline in heterozygosity over time due to genetic drift.
- Heterozygosity in island and mainland populations.

The chapter up to this point has focused on population size and genetic drift. This section will demonstrate that finite population size can also be thought of as a form of inbreeding. In large populations with random mating, chance biparental inbreeding is unlikely to occur often. However, in small populations the chance of mating with a relative is larger since the number of possible mates is limited. As populations get smaller, the probability of chance matings between related individuals should increase. Genetic drift also occurs due to finite population size. Therefore, genetic drift and the tendency for inbreeding are interrelated phenomena connected to the size of a population. Both have the result of increasing the homozygosity in a population over time.

Before we can reach the goal of showing that genetic drift and inbreeding are really equivalent genetic processes, it is necessary to develop a bit of conceptual machinery. In Chapter 2, autozygosity was defined as the probability that two alleles are identical by descent and demonstrated using a pedigree. We now need to revisit the autozygosity from the perspective of a finite population. Figure 3.18 shows three possible ways that alleles could be sampled from a finite population of gametes when constructing diploid genotypes. This gamete population meets Wright–Fisher assumptions with the exception that it is finite and contains alleles that are identical in state but are not identical by descent. To make offspring in the next generation, alleles are sampled with replacement from the population of  $2N$  gametes. What is the probability that two alleles in a genotype in the next generation are identical by descent? Given that one allele has been sampled, say an  $A_1$  allele, what is the probability of sampling the *same* allele on the next draw? Since there is only one copy of this allele in gamete population, there is only one of the  $2N$  alleles



**Figure 3.18** Autozygosity and allozygosity in a finite population where identity by descent is related to the size of the population. Finite populations accumulate genotypes containing alleles identical by descent through random sampling in a manner akin to mating among relatives. In this example, alleles in the ancestral gamete pool identical in state are not identical by descent. Sampling of alleles takes place to form the diploid genotypes of the next generation. By chance, the same allele can be sampled twice to form an autozygous genotype with probability  $\frac{1}{2N_e}$ . The chance of not sampling the same allele twice is the probability of all other outcomes or  $1 - \frac{1}{2N_e}$ . Autozygous genotypes must be homozygous but allozygous genotypes can be either homozygous or heterozygous.

that are the same. Therefore the chances of sampling the same allele to make a genotype are  $\frac{1}{2N_e}$ , which is also the probability that the alleles in a genotype are identical by descent or are autozygous. The probability that two alleles in a genotype are not identical by descent or are allozygous is then  $1 - \frac{1}{2N_e}$ . Through the process of random sampling, populations can accumulate autozygous genotypes, as do populations where mating takes place among relatives.

We can use the probability of autozygosity in a finite population to define the fixation index (recall that it was defined as expected heterozygosity minus observed heterozygosity all over expected heterozygosity in Chapter 2) as

$$F_t = \frac{1}{2N_e} \quad (3.47)$$

for generation  $t$  under the assumption that none of the alleles in the gamete pool in generation  $t - 1$  are identical by descent. To make this more general, we could also account for the possibility that some of the gametes in the ancestral population of generation  $t - 1$  have alleles that are identical by descent from past inbreeding or random sampling. To do this we need to reexamine the probability that alleles are allozygous. Although two distinct alleles in the gamete pool of Fig. 3.18 may be sampled to form what appears to be an allozygous homozygote, it is possible that these two alleles sampled are actually identical by descent. That would mean that the gamete pool would be inbred to some degree  $F_{t-1}$  instead of containing only allozygous alleles. The fixation index then becomes

$$F_t = \frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right)F_{t-1} \quad (3.48)$$

where the first term is due to sampling between generations and the second term is the proportion of apparently allozygous alleles in the gamete population that are actually autozygous due to past sampling or inbreeding.

By definition,  $F$  is the reduction in heterozygosity as well as the increase in homozygosity compared to Hardy–Weinberg expected genotype frequencies. If  $F$  is proportional to the homozygosity and amount of inbreeding, then  $1 - F$  is proportional to the amount of heterozygosity and random mating. If we are interested in predicting levels of heterozygosity, multiplying both sides of equation 3.48 by  $-1$  and then adding one to both sides gives

$$1 - F_t = \left(1 - \frac{1}{2N_e}\right)(1 - F_{t-1}) \quad (3.49)$$

or an expression for one minus the expected homozygosity due to random sampling from a finite gamete population. To express this in terms of the heterozygosity, we can use  $H_t = 2pq(1 - F_{t-1})$  from equation 2.20 to obtain  $1 - F_{t-1} = \frac{H_t}{2pq}$ . Substituting this into equation 3.49 gives:

$$\frac{H_t}{2pq} = \left(1 - \frac{1}{2N_e}\right) \left( \frac{H_{t-1}}{2pq} \right) \quad (3.50)$$

which after multiplying both sides by  $2pq$  gives:

$$H_t = \left(1 - \frac{1}{2N_e}\right) H_{t-1} \quad (3.51)$$

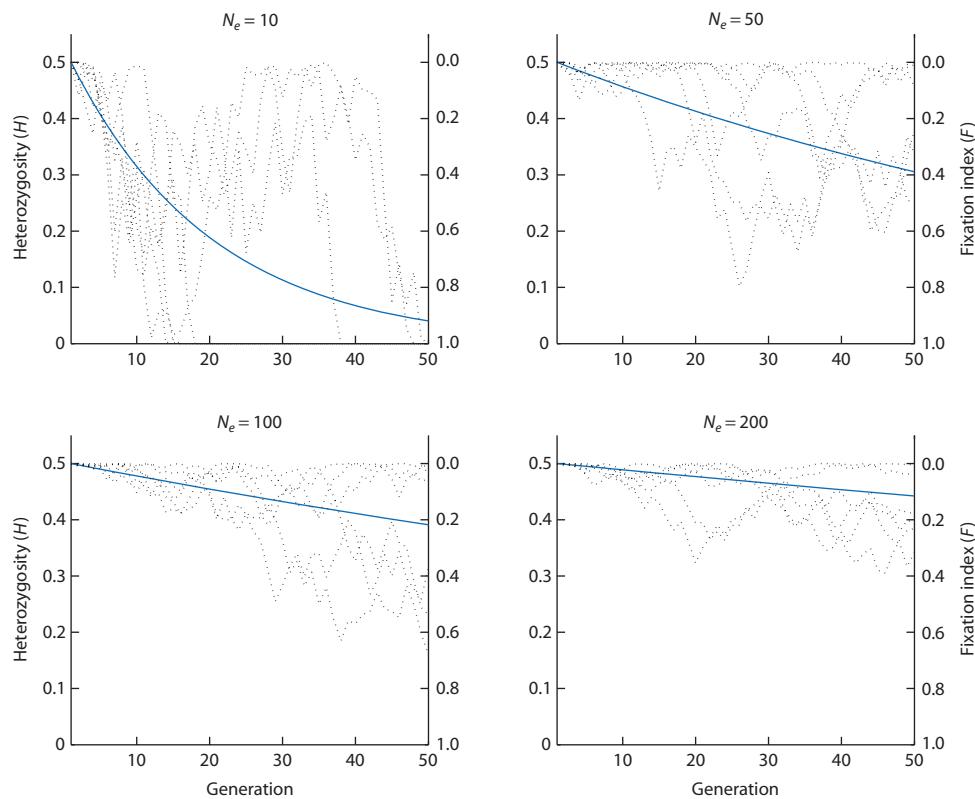
Taking this relationship over an arbitrary number of generations gives

$$H_t = \left(1 - \frac{1}{2N_e}\right)^t H_0 \quad (3.52)$$

where  $H_0$  is the initial heterozygosity and  $H_t$  is the heterozygosity after  $t$  generations have elapsed.

There is a very general and biologically meaningful relationship contained in equation 3.52. It shows that heterozygosity declines by a factor of  $1 - \frac{1}{2N_e}$  every generation caused simply by sampling from a finite population that results in some autozygous genotypes every generation. Recall that this is exactly the same result that was obtained for the rate of loss of genetic variation with the Markov chain model. The degree of sampling varies directly with the effective population size, so that the rate of increase in autozygous genotypes also depends directly on the effective population size. The expected heterozygosity from equation 3.52 is shown in Fig. 3.19 for four different effective population sizes over 50 generations. For comparison, heterozygosity in six independent replicate populations experiencing genetic drift are also plotted. The random trajectories of heterozygosity in these individual populations make clear that equation 3.52 provides an expectation for average heterozygosity taken across a large number of replicate populations or numerous independent neutral loci if applied to a single population.

There are two conclusions that can be drawn from the interrelationship between autozygosity and the effective population size. First, genetic drift causes populations to become more inbred in the sense that autozygosity and homozygosity increase *even though mating is random*. An important distinction is that genetic drift causes heterozygosity to decrease due to the fixation and loss of alleles. In contrast, consanguineous mating decreases heterozygosity by changing genotype frequencies but does not impact allele frequency. Genetic drift produces homozygosity since ultimately one allele reaches fixation while consanguineous mating produces homozygous genotypes for all alleles in the population. Second, mating systems where there is consanguineous mating cause genetic variation in populations to behave, from the perspective of heterozygosity, as if the effective population size were smaller than it would be under complete random mating. For example, when a parent self-fertilizes, the alleles transmitted to its



**Figure 3.19** The decline in heterozygosity as a consequence of genetic drift in finite populations. The solid lines show expected heterozygosity over time according to  $H_t = \left(1 - \frac{1}{2N_e}\right)H_{t-1}$ . The decrease in heterozygosity can also be thought of as an increase in autozygosity or the fixation index ( $F$ ) through time under genetic drift. The dotted lines in each panel are levels of heterozygosity ( $2pq$ ) in six replicate finite populations experiencing genetic drift. There is substantial random fluctuation around the expected value for any individual population.

progeny are sampled from a gamete pool containing only two possible alleles. Compare that with the case of outcrossing between unrelated individuals, where the alleles transmitted to progeny are sampled from a pool of four possible alleles. The probability of autozygosity is clearly higher with the smaller parental gamete pool associated with consanguineous mating, just as it is also higher in a smaller population where mating is random.

The faster decrease in heterozygosity in small compared to large populations can be seen in the wild. One study used the reasoning that organisms restricted to islands should have smaller census population sizes than the same species found on adjacent mainland areas (Frankham 1998). Based on the relationship between autozygosity and effective population size just derived, the expectation is that island populations should show lower levels of heterozygosity than mainland populations. Although there were some exceptions, the general pattern was that island

populations had lower levels of heterozygosity than did mainland populations (Table 3.3). This is exactly the pattern expected due to the faster rate of decrease in heterozygosity in smaller populations even when mating is random. The comparison assumed that the island and mainland populations remained very similar in terms of the degree of consanguineous mating as well as genetic parameters that influence the input of genetic variation such as the migration and mutation rates. Since the comparisons were between populations of the same species that are therefore very closely related, these assumptions seem likely to be met at least approximately.

### 3.5 Estimating effective population size

- Variance and inbreeding effective population size.
- Breeding effective population size in continuous populations.
- Effective population sizes for different genomes.

### Interact box 3.3 Heterozygosity and inbreeding over time in finite populations

Populus can be used to simulate genetic drift in a finite population and then track values of the inbreeding coefficient over time. In Populus, click on the **Model** menu and select **Mendelian Genetics** and then **Inbreeding**. A dialog box will open that has entry fields for the effective population size and the initial level of the inbreeding coefficient for a diallelic locus. You can also set the number of generations to run the simulation. To get started, set **Population** = 30, **Initial Frequency** = 0.0 and **Generations** = 120. A graph of the results will appear after entering the simulation parameter values. Despite the name, **Initial Frequency** is actually the initial level of inbreeding in the population. A value of zero means that the population is in Hardy–Weinberg equilibrium. Pressing the **View** button in the model dialog will generate a new data set and redraw the graph.

The graph will show three types of inbreeding coefficients.  $F_t$  is the “theoretical” inbreeding coefficient based on the decline in heterozygosity over time (equation 3.52) or  $F_t = 1 - \left(1 - \frac{1}{2N_e}\right)^t$ .

The other two lines are both based on the observed frequency of homozygous genotypes in the simulated population.  $F_a$  (the blue line) is the actual frequency of homozygous individuals in the population (often called  $F_i$  outside of Populus).  $F_f$  (the green line) is the population homozygosity.  $F_a$  and  $F_f$  can be different because the individual homozygosity tracks the combined frequency of homozygotes for either allele while the population homozygosity tracks how close the population is to fixation or loss (global homozygosity). When the population homozygosity is one there can be only one homozygous genotype in the population.

Why do the individual and population homozygosity values fluctuate? Is the amount of fluctuation related to the population size? Although the graph does not show the heterozygosity over time, what would lines for the theoretical, and individual and population heterozygosities look like? Try graphing each of these quantities on paper for a given run based on the three inbreeding coefficients.

**Table 3.3** Levels of heterozygosity found in island and mainland populations of the same species demonstrates that small population size has effects akin to inbreeding. Heterozygosity in island and mainland populations is compared using the effective inbreeding coefficient  $F_e = 1 - \frac{H_{island}}{H_{mainland}}$ .  $F_e > 0$  when the mainland population(s) exhibit more heterozygosity,  $F_e < 0$  when the island population(s) exhibit more heterozygosity, and  $F_e = 0$  when levels of heterozygosity are equal. Values given are ranges when more than one set of comparisons was reported from a single source. Data from Frankham (1998).

Species	$F_e$
Mammals	
Wolf ( <i>Canis lupis</i> )	0.552
Lemur ( <i>Lemur macaco</i> )	0.518
Mouse ( <i>Mus musculus</i> )	-0.048 to 1.000
Norway rat ( <i>Rattus rattus</i> )	-0.355 to 0.710
Leopard ( <i>Panthera pardus</i> )	0.548
Cactus mouse ( <i>Peromyscus eremicus</i> )	0.445–0.899
Shrew ( <i>Sorex cinereus</i> )	-0.241 to 0.468
Black bear ( <i>Ursus americanus</i> )	0.545
Birds	
Singing starling ( <i>Aplonis cantoroides</i> )	0.231–0.833
Chaffinch ( <i>Fringilla coelebs</i> )	-0.164 to 0.504
Reptiles	
Shingleback lizard ( <i>Trachydosaurus rugosus</i> )	0.069–0.311

The concept of effective population size should now seem comfortable. This section will focus on putting models into practice in order to estimate the effective population size. This requires a shift in perspective from using models to develop expectations about allele frequencies to using models to explain past patterns in allele frequencies. Working through a detailed example of estimating the effective population size will highlight the distinctions between two definitions of the effective population size. This section will also introduce a new definition of the effective population size used widely for species with continuous distributions that lack discrete geographic population boundaries. In addition, the effective population sizes of nuclear and organelle genomes will be compared.

### **Different types of effective population size**

Just as there are several models constructed on different foundations used to describe how genetic variation changes due to finite population size, there are several ways to define the effective population size such that it can be estimated. This is because the effective population size is really defined as a consequence of the models constructed to predict the behavior of genetic variation in populations. To see this, let's consider the **inbreeding** and **variance** effective population sizes.

#### **Inbreeding effective population size ( $N_e^i$ )**

The size of an ideal population that would show the same probability of allele copies being identical by descent as an actual population.

**Variance effective population size ( $N_e^v$ )** The size of an ideal population that would show the same sampling variance in allele frequency as an actual population.

In an ideal finite population (under the infinite alleles model), the chances of sampling two copies of the same allele depend on the size of the population and are  $\frac{1}{2N_e^i}$  (see section 3.3). This is the same as the probability that two alleles are identical by descent (IBD) in a pedigree (see section 2.6). So we have:

$$P(\text{IBD}) = \frac{1}{2N_e^i} \quad (3.53)$$

which can just as easily be restated as:

$$N_e^i = \frac{1}{2P(\text{IBD})} \quad (3.54)$$

The first equation is used to determine what to expect for the packaging of alleles into genotypes. The expected value of the probability of identity by descent depends on the size of the population. The flip side of this same coin is that we can determine what to expect for the effective population size based on the probability of identity by descent in a population, as shown in the second equation. In big populations the chances that two alleles are identical by descent is low whereas in small populations the chances of identity by descent are high. When the effective population size is defined by reference to autozygosity or inbreeding, the result is the inbreeding effective population size.

The change in allele frequencies in many replicate populations over generations was the focus of the Markov and diffusion models of genetic drift (section 3.2). The range of change in allele frequency in these models among many populations could also be expressed as a variance. Earlier in the chapter (equation 3.6) the standard error of the mean allele frequency among replicate populations was derived. This leads to the variance in the change in allele frequencies from one generation to the next ( $\Delta p = p_{t-1} - p_t$ ) taken among replicate populations:

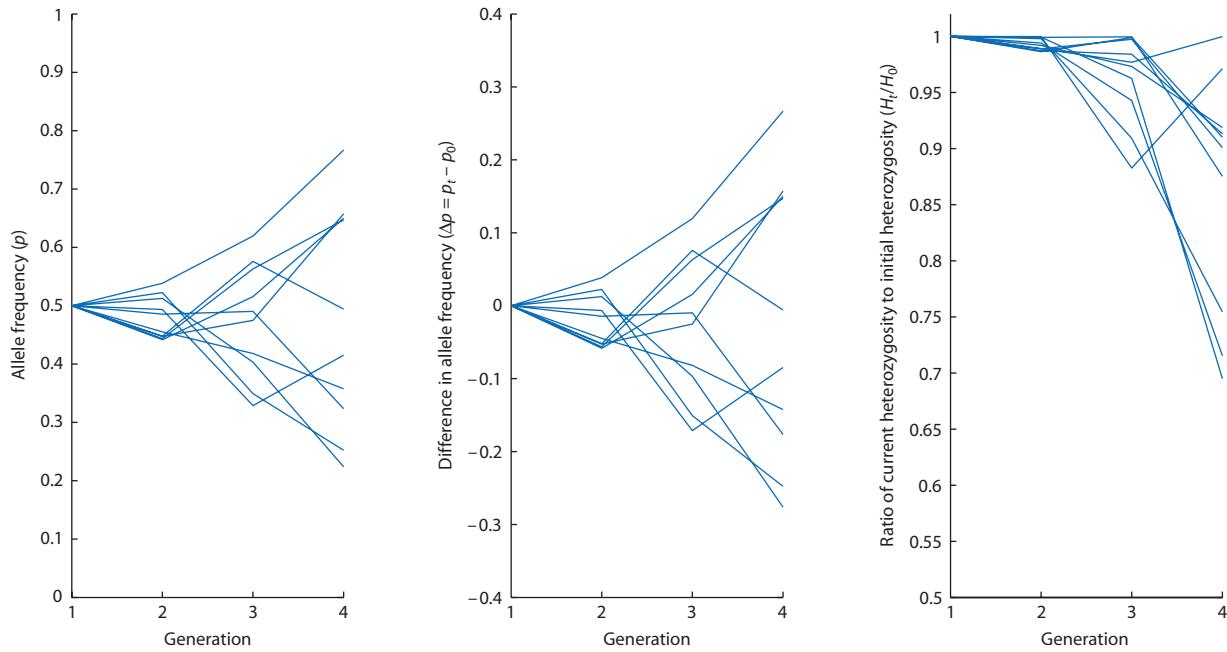
$$\text{Variance}(\Delta p) = \frac{p_{t-1}q_{t-1}}{2N_e^v} \quad (3.55)$$

where  $p$  and  $q$  are allele frequencies at a diallelic locus and the subscripts indicate the generation. As we did above, this equation can be restated by solving for the effective population size:

$$N_e^v = \frac{pq}{2 \times \text{variance}(\Delta p)} \quad (3.56)$$

Here again we see that the first equation shows that the variance in allele frequencies among many identical replicate depends in the effective population size. By turning the equation around, we can then define how big the effective size is by quantifying the variance in the change in allele frequencies among a group of replicated populations. This definition provides the variance effective population size.

Why make a big deal about distinguishing between the inbreeding and variance effective population sizes?



**Figure 3.20** Simulated allele frequencies in 10 replicate populations that experienced effective population sizes of 100, 10, 50, and 100 individuals across four generations. The variance in the change in allele frequency ( $\Delta p$ ) can be used to estimate the variance effective population size. The inbreeding effective population size can be estimated from the change in heterozygosity through time.

In the beginning of the section, the effective population size was defined by how genetic variation in a population behaves over time. In ideal populations that are not changing much in census size over time, the different effective sizes are usually equivalent. In some situations, however, the different types of effective population size are not equivalent since they measure different aspects of genetic variation (Ewens 1982; Crow & Denniston 1988; Crandall et al. 1999).

A good way to understand the variance and inbreeding effective population sizes is to employ them with genetic data taken from a finite population experiencing genetic drift. Figure 3.20 shows the allele frequencies, change in allele frequencies, and change in heterozygosity in 10 simulated populations over four generations. The effective population sizes fluctuated from 100 to 10 to 50 and then back to 100 over four generations. Using equation 3.42, we would expect the genetic variation in this simulated population to exhibit the same amount of genetic drift over four generations as a population with a constant size of about 28 or 29 individuals.

The heterozygosities and allele frequencies in the 10 simulated populations are given in Tables 3.4 and 3.5. The change in heterozygosity can be used to estimate the inbreeding effective population, since the heterozygosity can be thought of as one minus

the autozygosity. The decline in heterozygosity with time (equation 3.52) can be approximated by:

$$\frac{H_t}{H_0} \approx e^{-\frac{t}{2N_e}} \quad (3.57)$$

Taking the natural log of both sides

$$\ln\left(\frac{H_t}{H_0}\right) \approx -\frac{t}{2N_e} \quad (3.58)$$

and then solving in terms of the effective population size

$$N_e^i \approx -\frac{t}{2 \ln(H_t/H_0)} \quad (3.59)$$

gives an expression that can be applied to the simulation data. Table 3.4 shows the computational steps required to estimate  $\hat{N}_e^i$  for the 10 replicate populations. In a similar fashion, Table 3.5 works through the computations necessary to estimate  $\hat{N}_e^v$  for the 10 replicate populations using equation 3.56.

The estimates of  $\hat{N}_e^v$  and  $\hat{N}_e^i$  are close to each other but both are an order of magnitude lower than the expected population size of 28 or 29. What happened? The estimates are poor for a number of reasons that

**Table 3.4** Data from simulated allele frequencies in Fig. 3.20 used to estimate the effective population size. Here, the ratio of heterozygosity in generations three and four is used to estimate inbreeding effective population size ( $\hat{N}_e^i$ ) according to equation 3.59. Initial allele frequencies were  $p = q = 0.5$ , so  $H_{t=1} = 0.5$ . One generation of genetic drift took place, hence 1 is used in the numerator of the expression for  $N_e^i$ . The average  $N_e^i$  excludes the negative values.

$H_{t=3}$	$H_{t=4}$	$\ln\left(\frac{H_{t=4}}{H_{t=3}}\right)$	$\hat{N}_e^i = -\frac{1}{2} \frac{1}{\ln\left(\frac{H_{t=4}}{H_{t=3}}\right)}$
0.4987	0.4504	-0.1018	4.91
0.4866	0.4594	-0.0575	8.69
0.4813	0.3474	-0.3259	1.53
0.4998	0.4376	-0.1329	3.76
0.4546	0.3772	-0.1864	2.68
0.4884	0.4999	0.0232	-21.58
0.4920	0.4566	-0.0747	6.69
0.4413	0.4856	0.0957	-5.22
0.4715	0.3578	-0.2761	1.81
0.4995	0.4550	-0.0932	5.36
Average $\hat{N}_e^i = 4.43$			

**Table 3.5** Data from simulated allele frequencies in Fig. 3.20 used to estimate the effective population size. Here, the change in allele frequency between generations three and four is used to estimate variance effective population size ( $\hat{N}_e^v$ ) according to equation 3.56. Allele frequencies in the third generation were used to estimate  $pq$ .

$p_{t=4}$	$\Delta p = p_{t=4} - p_{t=3}$	$pq$	$\text{Var}(\Delta p) = \frac{1}{10} \sum (p_{t=4} - \bar{p})^2$	$\hat{N}_e^v = \frac{pq}{2 \times \text{variance}(\Delta p)}$
0.6574	0.1825	0.2494	0.0186	6.71
0.3575	-0.0606	0.2433		6.55
0.2238	-0.1795	0.2406		6.47
0.3234	-0.1668	0.2499		6.72
0.2523	-0.0970	0.2273		6.12
0.4940	-0.0819	0.2442		6.57
0.6473	0.0842	0.2460		6.62
0.4153	0.0866	0.2207		5.94
0.7667	0.1473	0.2357		6.34
0.6499	0.1343	0.2498		6.72
Average $\hat{N}_e^v = 6.48$				

all stem from using models to describe actual populations as opposed to making idealized predictions. The model assumptions, which seem reasonable in the abstract, have become critical now when faced with a realistic sized data set. It reminds us again of the difference between a parameter and a parameter estimate. The data we used to make the estimates probably do not fit the model assumptions well.

Using only 10 replicate populations to estimate the effective population size is clearly not enough given that the Markov model assumed an ensemble population that approaches infinite size. In some cases the heterozygosity actually increased between generations, although the expectation in equation 3.52 is that it should only decrease. The average  $\Delta p$  was not zero even though with many replicate populations

it is expected to be. The estimates also only utilized a single generation of genetic drift while the expectation for a fluctuating populations was over four generations. It seems that the change in allele frequencies and heterozygosity in the simulated populations were disconcertingly random! All this serves to remind us that assumptions like “many populations” or “over long time periods” may not be biologically realistic because a data set or organisms themselves may not play by the same rules. Violation of model assumptions often leads to poor or imprecise estimates of parameters, just as we have seen in this example.

### Problem box 3.4 Estimating $N_e$ from observed heterozygosity over time

Use the data provided in Table 3.5 to estimate  $\hat{N}_e^i$  over four generations for each of the 10 replicate populations. Heterozygosity in the first generation was 0.5 in all populations since initial allele frequencies were  $p = q = 0.5$ . Does the average estimate of  $\hat{N}_e^i$  better match what is expected for a fluctuating population of 100–10–50–100?

There are an array of methods that have been employed to estimate effective population size from genetic data using various estimators of the variance and inbreeding effective population sizes. Table 3.6 provides some estimates of the ratio of the effective population size to the census size. A general

conclusion is that the effective population size is often one-tenth or less, sometimes much less, of the census population size. See Waples (1989) for more details on estimating effective population size from changes in allele frequency through time. In addition, the effective population size can be estimated with gametic disequilibrium (equation 2.31; Waples 1991; Slatkin 1994), heterozygote excess in small populations of self-incompatible individuals due to random allele-frequency differences between populations of male and female breeders (Balloux 2004), and gene diversity of DNA sequences or other molecular marker loci.

#### Breeding effective population size

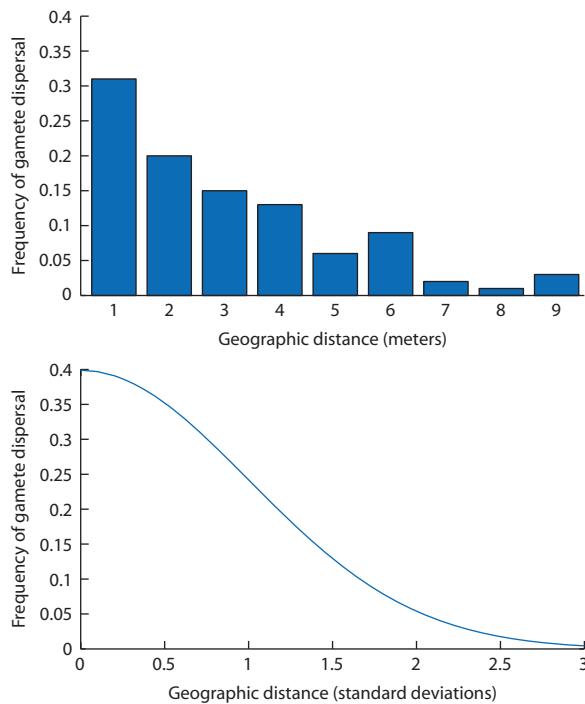
Up to this point, an implicit assumption has been that a population is an easily recognized and discrete entity. In species where individuals are more or less continuously distributed over large areas, there are no obvious physical or geographic boundaries that define populations. Instead, populations can be defined by average mating and dispersal patterns among individuals that result in limits to the movement of gametes each generation. Based on the size of the breeding and dispersal area there is the **breeding effective population size**, which is particularly suitable for populations where individuals may occur relatively uniformly over large areas and not form discrete aggregations. Imagine a large, continuous plant population that covers many hectares (plants are a good example since they stay in the same place over time, but the concept applies to all organisms). Now imagine examining all of the successful mating events for many individuals. The distances of mating

**Table 3.6** Estimates of the ratio of effective to census population size  $\left(\frac{N_e}{N}\right)$  for various species based on a wide range of estimation methods and assumptions.

Species	$\frac{N_e}{N}$	Reference
Leopard frog ( <i>Rana pipiens</i> )	0.1–1.0	Hoffman et al. 2004
New Zealand snapper ( <i>Pagrus auratus</i> )	$(0.25–16.7) \times 10^{-5}$	Hauser et al. 2002
Red drum ( <i>Sciaenops ocellatus</i> )	0.001	Turner et al. 2002
White-toothed shrew ( <i>Crocidura russula</i> )	0.60	Bouteiller and Perrin 2000
Flour beetle ( <i>Tribolium castaneum</i> )	0.81–1.02 <sup>a</sup>	Pray et al. 1996
Review of 102 species	0.10 <sup>b</sup>	Frankham 1995

<sup>a</sup>Ratios declined as census population sizes increased.

<sup>b</sup>Mean of 56 estimates in the “comprehensive data set” that included impacts of unequal breeding sex ratio, variance in family size, and fluctuating population size over time.



**Figure 3.21** Isolation by distance is characterized by the declining probability of gamete dispersal with increasing geographic distance. The specific shape of the gamete dispersal by distance curve may vary (top), but it is often modeled using a normal distribution (bottom). In a normal distribution, about 95% of observations are expected to fall within two standard deviations from the mean. Empirical estimates of mating and progeny movement in a generation can be used to estimate the variance and thereby the standard deviation of overall dispersal in gametes in order to estimate the area of a genetic neighborhood.

events would show a frequency distribution like that of Fig. 3.21. The critical feature of the distribution is that the frequency or chance of mating drops off on average as the physical distance between individuals increases. This is a widely observed phenomenon called **isolation by distance** (Wright 1943a). With enough distance separating them, two individuals have a low probability of mating and can be considered members of distinct genetic populations even if they are not located in geographically distinct populations. The distance required for reproductive isolation by distance may be on the order of meters or thousands of kilometers depending on the species.

Estimating the breeding effective population size depends on the probability distribution of gamete dispersal in space and can be approximated by a normal distribution (other distributions can be used as well). Recall that two standard deviations on either side of the mean contains about 95% of the observations in a normal distribution (see the Appendix). The standard deviation in dispersal, which is the

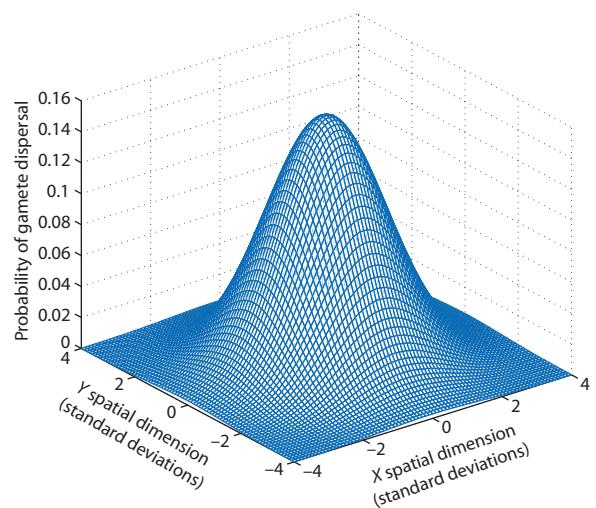
square root of the variance in dispersal, can then be used to describe the probability that a gamete disperses in one dimension (Fig. 3.21). Extending this to gamete dispersal into two dimensions, the dispersal area can be thought of as a circle with the average individual at the center. Since the circle describes a probability distribution, a radius of twice the standard deviation in dispersal in a generation will sweep an area containing about 95% of the observations in a two-dimensional normal distribution (Fig. 3.22). This two-dimensional normal distribution model of gamete dispersal probability combined with the average density of individuals is used to quantify the breeding effective population size. Since the area of a circle is  $\pi r^2$  where  $r$  is two standard deviations (or twice the square root of the dispersal distance variance):

$$N_e^b = \pi(2\sqrt{\text{dispersal variance}})^2 d \quad (3.60)$$

which simplifies to

$$N_e^b = 4\pi(\text{dispersal SD})^2 d \quad (3.61)$$

where  $d$  is the average density of individuals. The area described by equation 3.61 is known as the **genetic neighborhood** in continuous populations. Multiplying the genetic neighborhood (an area) by the density of individuals (individuals per unit area)



**Figure 3.22** An ideal two-dimensional normal distribution used to model the size of genetic neighborhoods and to estimate the breeding effective size ( $N_e^b$ ) of demes within continuous populations. The radius of the distribution is twice the standard deviation in total gamete dispersal in a generation. The actual physical dimensions of the distribution could range from just a few meters to hundreds or thousands of kilometers.

gives the breeding effective population size in individuals (see Wright 1943a; Crawford 1984).

### Breeding effective population size ( $N_e^b$ )

The number of individuals found in a genetic neighborhood defined by the variance in gamete dispersal.

**Deme** The largest area or collection of individuals where mating is (on average) random.

**Genetic neighborhood** An area or subunit of a population within which mating is random.

**Isolation by distance** Decrease in the probability of mating and dispersal of gametes as physical distance increases.

In a classic study, Schaal (1980) estimated the breeding effective population size in Texas bluebonnets (*Lupinus texensis*). This species is pollinated by bumblebees and occurs in large continuous populations that cover many hectares (see color images on the text web page in the section for Chapter 3). An experimental population was constructed using 91 plants with known genotypes at the phosphoglucomutase-1 enzyme locus (*Pgi-1*). Mating distances were estimated from seven central plants (homozygous for the *Pgi-1* fast allele) by genotyping progeny from plants without the *Pgi-1* fast allele in the experimental population. Since gametes are also dispersed in seeds every generation, the passive dispersal of seeds was tracked as well. Dispersal distances were very limited, with gametes moving via pollen an average of 1.82 m and seeds moving an average of 0.58 m. The genetic neighborhood size for the experimental population of *L. texensis* was estimated as 6.3 m<sup>2</sup>, containing a breeding effective population size of 95.4 individual plants.

The term **deme**, the largest area or collection of individuals where mating is (on average) random, is often applied to continuous populations and is closely connected with the concepts of breeding effective population size and genetic neighborhoods. The bond orbital in chemistry serves as a metaphor for the deme in population genetics. A bond orbital describes the probability that an electron will be found at some location in space. Similarly, a deme describes the probability that an average individual will move its gametes some distance in space by mating or dispersal in a generation. Members of the same deme are considered able to mate at random whereas members of two different demes have a low probability (< 5%) of mating

and are therefore members of separate populations. It is important to distinguish demes, or genetically defined population demarcations, from geographically defined populations. From the perspective of effective population size and predicting the behavior of genetic variation, deme is a more useful definition for populations than geographic or spatial definitions.

### Effective population sizes of different genomes

Plastid genomes (mitochondria and chloroplasts) are an example where the effective population size is lower than that of the nuclear genome of the same organism. The effective population size of plastid genomes is reduced by two independent factors. First, these genomes are haploid (one copy of the genome per plastid) compared to the two homologous copies of each diploid nuclear chromosome. In addition, most plastids are inherited by offspring from one parent only via the gamete cytoplasm (uniparental inheritance). In species with two equally frequent sexes, uniparental inheritance causes plastid genomes to have half the effective population size of genomes inherited from all possible parents. These two factors combined cause animal mitochondrial genomes and plant mitochondrial and chloroplast genomes to have one-quarter of the effective size of the diploid, biparentally inherited nuclear autosomes. Thus, loci in these genomes experience a higher rate of genetic drift compared to loci in the nuclear genome. Genetic marker studies frequently take advantage of this fact, using plastid genome marker loci to study phenomena such as recent population divergence due to genetic drift where nuclear marker loci would show much less divergence because of a larger effective population size.

## 3.6 Gene genealogies and the coalescent model

- Modeling the branching of lineages to predict the time to the most recent common ancestor.

At this point in the chapter we need an interlude in the discussion of genetic drift and effective population size to develop a new approach based around lineage branching or gene genealogy. Initially, it is necessary to introduce some basic terminology and concepts used in this approach. Although it may not be evident at first, the lineage-branching approach to population genetics has a great deal in common with the material in the first two sections of this chapter. The immediate goal of this section is to establish and motivate the building blocks necessary to model

lineage branching events. The next section of this chapter will then show how the concept of effective population size applies in genealogical branching models. A major advantage of coalescent models is that the action of population genetic processes (genetic drift, gene flow, and natural selection) on the branching pattern of lineages is independent of the allelic states of the lineages. Details about the lineages themselves are developed in Chapter 5, but bear in mind that each lineage represents an independent copy of an allele or DNA sequence. Once both the branching processes and the mutation processes that change on allelic states are brought together, the coalescent approach serves to make testable predictions for the evolution of DNA sequences under a combination of population genetic processes.

Tracing the pattern of ancestry for allele copies in a pedigree provides a means to understand the present patterns in those allele copies (see section 2.6). For example, the pedigree in Fig. 2.14 shows the equivalence of homozygosity in the present and the probability that two allele copies descended from a single ancestor in the past. Given the known individuals at each generation in that pedigree, we traced ancestor–descendant relationships forward in time to predict autozygosity in the most recent generation. Thus, that pedigree is an example of using a prospective or time-forward model, using knowledge of ancestors back in time and basic probability to work forward in time to predict the autozygosity at the most recent point in time.

Another type of analysis of ancestor–descendant relationships is possible based on a retrospective or time-backward model. Imagine that we have a sample of individuals taken in the present time, analogous to individual G in Fig. 2.14, but we have no knowledge of their parents or grandparents or any of their genealogical relationships. Would it be possible to learn something about the past population genetic events that lead up to that sample of individuals? The answer is yes, if we have models of ancestor–descendant relationships (genealogy) that allow us to predict identity by descent in the past based only on knowledge of the present. With such models, we look at patterns among the individuals available to us in the present and try to reconstruct versions of events such as inbreeding, gene flow, or natural selection in the past that could have lead to the individuals in the present. These models are referred to collectively as **coalescent theory** since the perspective of the models is to predict the probability of possible patterns of genealogical branching

working back in time from the present to the point of a single common ancestor in the past. When two lineages trace back in time to a single ancestral lineage it is said to be a **coalescent event**, hence the term coalescent theory.

A central concept in coalescent theory is connecting a group of lineages in the present back through time to a single ancestor in the past. This single ancestor is the first ancestor (going backward in time) of all the lineages in a sample of lineages in the present time and is referred to as the **most recent common ancestor** or **MRCA**. Section 3.2 develops a time-forward model of genetic drift that predicts that a sample of alleles (or lineages) will eventually arrive at fixation or loss. Fixation is reached by random sampling that expands the numbers of a given lineage or allele in the population. The lineage that reaches fixation can be traced back to a single ancestor at some point in the past. In the process of reaching fixation, a population loses all lineages except one, the one that was fixed by genetic drift. This same genetic drift process can be viewed from a time-backward perspective. A sample of lineages in the present must eventually be the product of a single ancestral lineage at some point back in the past that happened to become more frequent under random sampling. The coalescent model turns the random sampling process around, asking: what is the probability that two lineages in the present can be traced back to a single lineage in the previous generation? Answering this question relies on the same probability tools that were used earlier in the chapter to describe the process of genetic drift.

Before moving on with a more formal description, let's consider a metaphor for the coalescence process to set the stage. Imagine a sealed box full of bugs. Each bug moves around the box at random. Whenever two bugs meet by chance, one of them (picked at random) completely eats the other one in an instant. When a bug is eaten the population of bugs decreases by one and the remaining bugs continue to move about the box at random. The time that elapses between bug meetings tends to get longer as the number of bugs in the box gets smaller. This is because chance meetings between bugs depend on the density of bugs in the box. Eventually, the entire box that was full of bugs initially will wind up holding only a single bug after some time has passed. Each bug is analogous to a lineage and one bug eating another is analogous to a coalescent event. The very last bug is analogous to the lineage that is the most recent common ancestor.

**Coalescence or coalescent event** The point in time where a pair of lineages or genealogies trace back in time to a single common ancestral lineage (to coalesce literally means to grow together or to fuse).

**Genealogy** The record of ancestor-descendant relationships for a family or locus.

**Lineage** A line of descent or ancestry for a homologous DNA sequence or a locus (regardless of whether or not copies of the locus are identical or different).

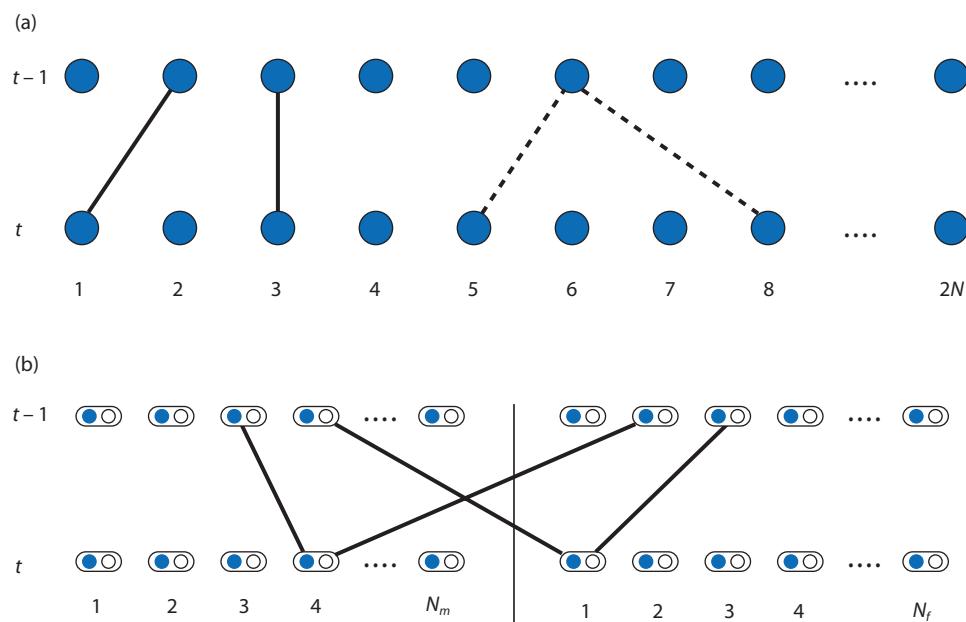
**Most recent common ancestor (MRCA)** The first common ancestor of all lineages (or gene copies) at some time in the past for a sample of lineages taken in the present.

**Gene copy or allele copy** A replicated DNA sequence that has passed from an ancestor to a descendant; used synonymously with the term lineage.

**Waiting time** The mean or expected time back in the past until a single coalescence event in a sample of lineages.

A schematic representation of the ancestor-descendant process for two generations can be seen in Fig. 3.23a for a set of haploid lineages. Using rules of random sampling based around the Wright–Fisher model (and its assumptions) we can develop a prediction for the number of generations back in time until two lineages “find” their MRCA or coalesce to a single lineage. Consider a random sample of two of the  $2N$  total lineages in the present generation. Given that one of these two sampled lineages finds its ancestor in the previous generation, what is the probability that the other lineage also shares that same common ancestor such that a coalescent event occurs? Given that one of the lineages has a given common ancestor, for coalescence to occur the other lineage must have the same ancestor among the  $2N$  possible ancestors in the previous generation. Thus the probability of coalescence is  $\frac{1}{2N}$  for two lineages whereas the probability that two lineages do not have a common ancestor in the previous generation is  $1 - \frac{1}{2N}$ .

The coalescent sampling process just developed for haploid lineages can also be extended to approximate



**Figure 3.23** Haplod (a) and diploid (b) reproduction in the context of coalescent events. In a haplod population, the probability of coalescence is  $\frac{1}{2N}$  (dashed lines) whereas the probability that two lineages do not have a common ancestor in the previous generation is  $1 - \frac{1}{2N}$  (solid lines). In a diploid population, the two gene or allele copies in one individual in the present time have one ancestor in the female population ( $N_f$ ) and one ancestor in the male population ( $N_m$ ). Coalescent events in the diploid population arise when the gene copies in males and females are identical by descent. The haplod model with  $2N$  lineages is routinely used to approximate the diploid model with  $N = N_f + N_m$  diploid individuals.

the process of reproduction for diploid lineages. In diploid reproduction, each offspring is composed of one allele copy inherited from a female parent and another allele copy from a male parent (Fig. 3.23b). In a time-backward view, this can be thought of as reproduction where one allele copy finds its ancestor in the male population of the last generation while the other allele copy finds its ancestor in the female population of the last generation. For a given male or female parent, each of their two allele copies has a probability of  $1/2$  of being the ancestral copy. As long as the number of males and females in a diploid population is equal and the haploid and diploid population sizes are large, the predictions of the coalescent model are very similar for haploid and diploid populations containing an identical total number of gene copies. The haploid model is more straightforward and so it is used throughout this section. In practice, the predictions that follow from the haploid coalescent model can be applied to samples of lineages (usually DNA sequences) from diploid organisms.

Like Markov chains, the probability of coalescence displays the Markov property since it is an independent event that depends only on the state of the population at the point of time of interest. Because of this, the basic probabilities of coalescence and non-coalescence between two generations can be used to describe the probability of coalescence over an arbitrary number of generations. If two randomly sampled lineages do not coalesce for  $t - 1$  generations, then the probability that they do coalesce to their common ancestor in generation  $t$  is:

$$\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \quad (3.62)$$

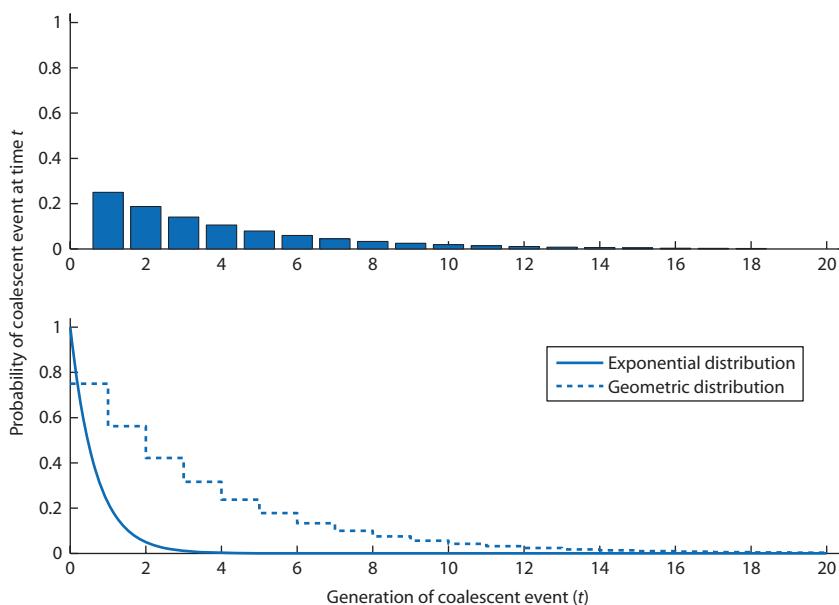
For example, in a population of  $2N = 10$  the chance that two randomly sampled lineages coalesce in four generations is the product of the probability of three generations of not coalescing ( $\left(1 - \frac{1}{10}\right)^3 = 0.729$ )

and the chance of coalescing between any two generations ( $1/10$ ), which gives a probability of coalescence of 0.0729. The distribution of the probabilities of a coalescent event occurring for two lineages in each of 30 generations for the case of  $2N = 10$  is shown in Fig. 3.24. It is important to note that only a single coalescent event is possible each generation since we are only considering two lineages.

In practice, the probabilities of coalescence are approximated using an exponential function (see Math box 3.2). As we have seen, the exact probability of coalescence for a pair of lineages is  $\frac{1}{2N}$  and the prob-

ability of not coalescing is  $1 - \frac{1}{2N}$  each generation.

The exponential approximation  $1 - e^{-\frac{1}{2N}t}$  gives the cumulative probability of a pair of lineages coalescing at or before generation  $t$ . This probability is symbolized as  $P(T_C \leq t)$  where  $T_C$  is the generation of coalescence and  $t$  is the maximum time to coalescence being considered. Let's use an example of the probability of coalescence at or before four generations have passed in a population of  $2N = 10,000$ . The



**Figure 3.24** The distribution of the probabilities of coalescence over time predicted by  $\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}$  (top) for a population of  $2N = 4$ . The bottom panel shows the approximations of these coalescence probabilities based on geometric and exponential distributions with a probability of “success” of  $1/4$ .

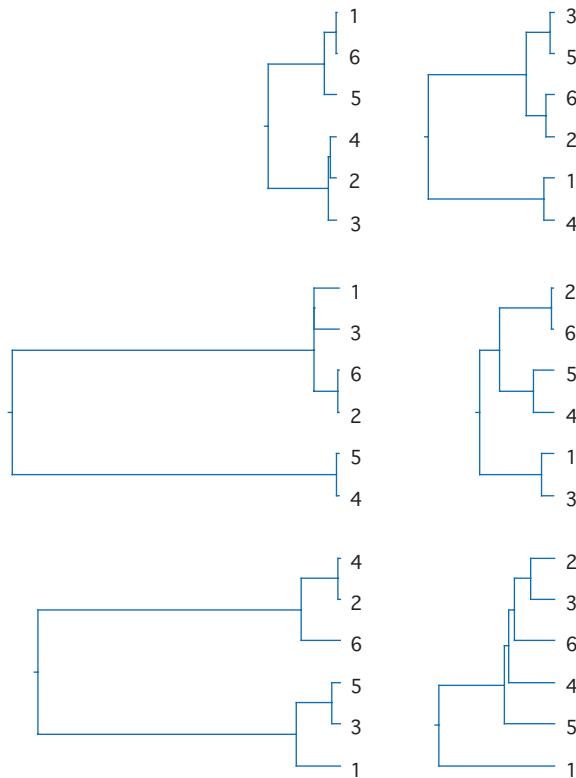
exact probability is the sum of the probabilities of coalescence in each generation,  $P(T_C \leq 4) = P(T_C = 1) + P(T_C = 2) + P(T_C = 3) + P(T_C = 4)$ . Substituting in expressions for the exact probability of coalescence at each of these four time points gives  $P(T_C \leq 4)$

$$\begin{aligned} &= \left(1 - \frac{1}{10,000}\right)^0 \frac{1}{10,000} + \left(1 - \frac{1}{10,000}\right)^1 \frac{1}{10,000} \\ &\quad + \left(1 - \frac{1}{10,000}\right)^2 \frac{1}{10,000} + \left(1 - \frac{1}{10,000}\right)^3 \frac{1}{10,000} \\ &= 0.0004. \text{ Using the exponential approximation} \end{aligned}$$

instead,  $1 - e^{-\frac{1}{10,000} \cdot 4} = 0.00039992$  as the chance that a pair of lineages experiences a coalescence at or before four generations elapses. Thus, the exact probability and the approximation are in close agreement. The exponential approximation of the exact probability improves as the population size increases. See Hein et al. (2005) and Wakeley (2008) for a more detailed discussion of the relationship between population size and the error of the exponential approximation of coalescence probabilities.

Approximating probabilities of coalescence with the exponential distribution makes computing more practical and also yields several generalizations about the coalescence process. In particular, the geometric and exponential probability distributions can be used to obtain an approximate average and variance for coalescence times. It turns out that for both types of distribution the mean time to an event is simply the inverse of the probability of an event occurring. In the coalescence process, the probability of coalescence for a pair of lineages is  $\frac{1}{2N}$  so the average time that elapses until coalescence is  $2N$  when the coalescence process is approximated with the exponential distribution. The average time to a coalescent event is often called the **waiting time**. Another generalization is that the range of individual coalescence times around that average is quite large. Based on the exponential distribution, the variance in the waiting time is  $4N^2$  so that range of coalescence times around the mean grows rapidly as the size of the population increases. Thus, the length of branches connecting lineages to their ancestors will be highly variable about their mean value. This can be seen in Fig. 3.25, which shows six independent realizations of the coalescent tree for six lineages.

It is possible to determine the average time for more than two lineages to find their MRCA. Suppose we want to determine the waiting time for  $k$  lineages,



**Figure 3.25** Six independent realizations of the coalescent tree for six lineages. All six trees are drawn to the same scale. Each genealogy exhibits coalescent events between random pairs of lineages. The differences in the height of the trees is due to random variation in waiting times. Because of this random variation, average times to coalescence for a given number of lineages are only approached in a large sample of independent trees. Because genealogies are drawn sideways here, tree “heights” are actually width, from left to right.

where  $k$  is less than or equal to the total number of lineages sampled from a population of  $2N$ . To see the problem in detail, let's consider the case of  $k = 3$  lineages. When no coalescence events occur, one lineage finds its ancestor among any of the  $2N$  individuals in the previous generation. That means the next lineage must find its ancestor among  $2N - 1$  individuals in the previous generation and the final lineage must find its ancestor among  $2N - 2$  possible parents. Thus the probability of non-coalescence is

$$\prod_{x=0}^{k-1} \left(1 - \frac{x}{2N}\right) \quad (3.63)$$

If the number of lineages sampled is much smaller than the total number of lineages in the population ( $2N$ ) then the probability of non-coalescence for  $k$  lineages can be approximated by

$$1 - \left( \frac{k(k-1)}{2} \right) \left( \frac{1}{2N} \right) \quad (3.64)$$

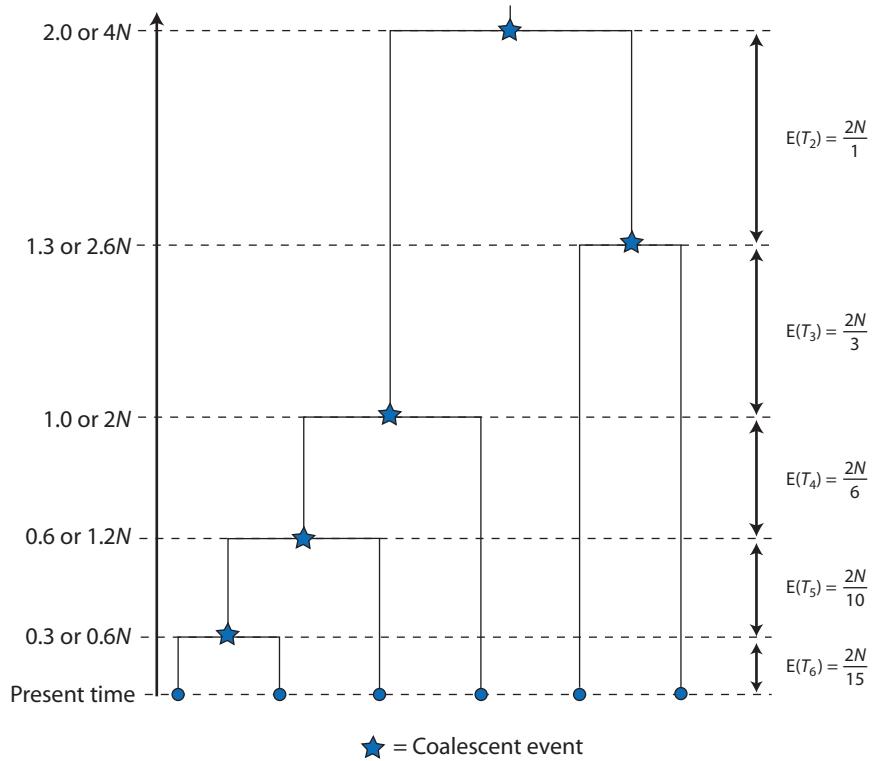
where  $\frac{k(k-1)}{2}$  enumerates the different ways to uniquely sample pairs of lineages from a total of  $k$  lineages. (Note that the number of unique pairs of lineages can also be determined with “ $k$  draw 2” or  $\binom{k}{2} = \frac{k!}{2!(k-2)!}$ .) The probability of a coalescence for any one of the unique pairs of the  $k$  lineages is then

$$\left( \frac{k(k-1)}{2} \right) \left( \frac{1}{2N} \right) \quad (3.65)$$

Bringing these two probabilities together into an equation like 3.62 gives the probability that  $k$  lineages experience a single coalescent event  $t$  generations ago:

$$\left( 1 - \left( \frac{k(k-1)}{2} \right) \left( \frac{1}{2N} \right) \right)^{t-1} \left( \frac{k(k-1)}{2} \right) \left( \frac{1}{2N} \right) \quad (3.66)$$

Since this probability also follows an exponential distribution ( $e^{-t \left( \frac{k(k-1)}{2} \right) \left( \frac{1}{2N} \right)}$ ), the average time to coalescence for  $k$  lineages in a population of  $2N$  is  $\frac{2N}{k(k-1)}$ . For example, if  $k = 3$  and  $2N = 10$ , the average time to coalescence is  $3^{1/3}$  generations. This is one-third of the average waiting time for two lineages since each of the three unique pairs of lineages (1–2, 1–3, and 2–3) can independently experience coalescence. Figure 3.26 shows the average coalescence times for six lineages based on this same logic. The general pattern is that coalescence times decrease when more lineages are present since there are a larger number of lineage pairs that can independently coalesce.



**Figure 3.26** A schematic coalescent tree that shows average coalescence times and the relationship between time scales in continuous units and units of  $2N$  generations. With a sample of  $k$  lineages, the expected or average time to coalescence is  $\frac{2N}{k(k-1)}$

since each of the independent  $\frac{k(k-1)}{2}$  pairs of lineages can coalesce. Dividing continuous time in generations ( $j$ ) by  $2N$  yields a continuous time scale ( $t = j/(2N)$ ) that can be used to describe coalescent trees. Multiplying the continuous time scale by  $2N$  gives time to coalescence in units of population size.  $E$  refers to expected and  $T$  refers to time to coalescence so that  $E(T_n)$  is the expected or average time to coalescence for  $n$  lineages. The basic patterns seen in all coalescent trees apply to populations of all sizes, although the absolute time for coalescent events does depend on  $N$ . Values on the left are one realization of coalescent waiting times. Genealogy is not drawn to scale.

## Math box 3.2 Approximating the probability of a coalescent event with the exponential distribution

The series of failures (non-coalescence) until a success (coalescence) in the genealogical process can be modeled where time is continuous (a real number) rather than discrete (an integer). The exponential distribution describes situations in which an object initially in one state can change to an alternative state with some probability that remains constant through time. The exponential distribution could be applied to the time until one of many light bulbs fails, for example, as well as to the time until a coalescent event in a population of lineages (see Fig. 3.24). The exponential distribution is described by

$$\text{Probability of change} = ae^{-at} \quad (3.67)$$

where  $a$  is the constant probability of changing states in a time interval of one,  $t$  is time, and  $e$  is the mathematical constant base of the natural logarithm ( $e = 2.71828 \dots$ ). The exponential distribution has a mean of

$$E(x = t) = \frac{1}{a} \quad (3.68)$$

and a variance

$$E(x^2) = \frac{1}{a^2} \quad (3.69)$$

Under the assumption that the effective population size  $2N$  is large (and constant) such that the probability of coalescence ( $\frac{1}{2N}$ ) is small, the probability that a coalescent event occurs,

$$\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \quad (3.70)$$

can be approximated using the exponential distribution

$$\frac{1}{2N} e^{-t \frac{1}{2N}} \quad (3.71)$$

for two lineages where  $t$  is time in generations. Notice that the constant in the exponential expression (equivalent to the  $a$  in equation 3.67)

is  $\frac{1}{2N}$ . That means that the average time to coalescence is  $\frac{1}{\frac{1}{2N}} = 2N$  and the variance in

time to coalescence is  $\left(\frac{1}{\frac{1}{2N}}\right)^2 = 4N^2$ . Setting the exact probability of coalescence approximately equal to the exponential gives an expression

$$\left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N} \approx \frac{1}{2N} e^{-\frac{t}{2N}} \quad (3.72)$$

that can be simplified by canceling the  $\frac{1}{2N}$  constant term on both sides

$$\left(1 - \frac{1}{2N}\right)^{t-1} \approx e^{-\frac{t}{2N}} \quad (3.73)$$

Therefore, the exponential distribution approximates the probability of non-coalescence at each time  $t$ .

To determine the chances that a coalescence event occurred at or before some time, call it  $T_C \leq t$  where  $T_C$  is the time to coalescence and  $t$  is time scaled in units of  $2N$  generations, the cumulative exponential distribution is used. The cumulative distribution

$$P(T_C \leq t) \approx 1 - e^{-\frac{t^{k(k-1)}}{2}} \quad (3.74)$$

effectively adds up the probabilities of coalescence (as one minus the chance of non-coalescence) as time increases to express the probability of a time interval passing without experiencing a coalescent event. The cumulative distribution approaches one as time increases since the chance of non-coalescence decreases toward zero with increasing time. The probability of coalescence increases more rapidly toward one for larger numbers of lineages  $k$ . This cumulative distribution is used to determine the waiting times needed to construct coalescent trees.

When approximating the probability of coalescent events with the exponential distribution, it is standard practice to put coalescence times on a continuous scale of units of  $2N$  generations. To see how this continuous time scale operates, let  $j$  be time measured as a real number (e.g. 1.0, 1.1, 1.2, 1.3 . . .  $j$ ) in generations. The time to coalescent events  $t$  can then be expressed as  $t = j/(2N)$ . As an example, imagine that a coalescence event occurred at  $t = 1.4$  on the continuous time scale. That coalescence event could also be thought of as occurring  $(1.4)(2N) = 2.8N$  generations in the past (see Fig. 3.26). If the population size was  $2N = 100$  lineages, then that coalescent event was  $(1.4)(100) = 140$  generations in the past. However, if the population size was  $2N = 20$  lineages, then that coalescent event was

$(1.4)(20) = 28$  generations in the past. Aside from the practical matter of interpreting a specific time value, the use of a continuous time scale makes an important biological point about the effects of population size on the coalescence process. The basic nature of the coalescent process is identical for all populations no matter what their size. For example, in populations of any size a single coalescent event will occur faster on average when there is a larger number of lineages sampled than when just a pair of lineages is sampled. Population size just serves to scale the time required for coalescent events to occur. Coalescent events occur more rapidly in small populations compared to bigger populations, a conclusion analogous to that reached for genetic drift earlier in the chapter.

### Interact box 3.4 Build your own coalescent genealogies

Building a few coalescent trees can help you to understand how the exponential distribution is put into practice to estimate coalescence times as well as give you a better sense the random nature of the coalescence process. You can use a Microsoft Excel spreadsheet at the textbook website that has been constructed to calculate the quantities necessary to build a coalescent genealogy. The spreadsheet contains the cumulative exponential distributions for a time interval passing without experiencing a coalescent event (see equation 3.74) for up to six lineages. To determine a coalescence time for a given number of lineages,  $k$ , a random number between zero and one is picked and then compared with the distribution when  $\frac{k(k - 1)}{2}$  lineages can coalesce. The time interval on the distribution that matches the random number is taken as the coalescence time.

- Step 1 Open the spreadsheet and look over all the quantities calculated. Click on cells to view the formulas used, especially the cumulative probability of coalescence for each  $k$ . This will help you understand how the equations in this section of the chapter are used in practice. View and compare the cumulative probability distributions graphed for  $k = 6$  and  $k = 2$ .
- Step 2 Press the recalculate key(s) to generate new sets of random numbers (see Excel help if necessary). Watch the waiting times until coalescence change. What is the average time to coalescence for each value of  $k$ ? How variable are the coalescence times you observe in the spreadsheet when changing the random numbers with the recalculate key?
- Step 3 Draw a coalescence tree using the coalescence times found in the spreadsheet (do not recalculate until step 6 is complete). Along the bottom of a blank sheet of paper, draw six evenly spaced dots to represent six lineages. Starting at the top of the random number table, pick two lineages which will experience the first coalescent event. Label the two left-most dots with these lineage numbers. Then, using a ruler, draw two parallel vertical lines as long as the waiting time to coalescence (e.g. if the time is 0.5, draw lines that are 0.5 cm). Connect these vertical lines with a horizontal line. Assign the lineage number of one of the coalesced lineages to the pair's single ancestor, at the horizontal line. Record the other lineage number on a list of lineages no longer present in the population (skip over these numbers if they appear again in the random number table). There are now  $k - 1$  lineages.

- Step 4** Use the random-number table to get the next pair of lineages that coalesce. Remember that the single ancestor of lineage pairs that have already coalesced will eventually coalesce with one of the remaining lineages. If one of the lineages of this random pair matches a previous pair's ancestor, begin at the horizontal line indicating that pair's coalescence, and draw a vertical line toward the top of the paper that is the length of the coalescence time for the number of lineages remaining. Draw a vertical line from lineage  $n$  to an equal height and connect the two vertical lines with a horizontal line (the line from lineage  $n$  will be as long as the sum of all coalescence times to that point). If neither number matches a previous pair's ancestor, draw the branches as in step 3, beginning at the baseline, but this time adding this pair's particular coalescence time to the sum of previous coalescence times to find the vertical branch height.
- Step 5** Repeat the process in step 4 until all lineages have coalesced.
- Step 6** Then add together all of the times to coalescence to obtain the total height of the coalescent tree and sum the height of all of the branches to obtain the total branch length of the tree. How do these compare with the average values for a sample of six lineages?
- Step 7** Press the recalculate key combination to obtain another set of coalescence times and repeat steps 3 through 5 to create another coalescence tree. Draw several coalescence trees to see how each differs from the others. You should obtain coalescence trees like those in Fig. 3.25. Your trees will differ from these, because the random coalescence times vary around their average, but the overall shape of your trees will be similar.

This section will conclude by considering several measures of coalescent trees useful to summarize general patterns of the coalescence process. The total time from the present to the point in the past where all  $k$  sampled lineages find their MRCA is called the **height of a coalescent tree**. The height of a tree for  $k$  sampled lineages is just the sum of the coalescence waiting times as coalescent events reduce the number of lineages from  $k$  to  $k - 1$  to  $k - 2$  down to one. The mean or expected value of the height of a coalescent tree is then

$$E(H_k) = \sum_{i=2}^k E(T_i) = 2 \sum_{i=2}^k \frac{1}{i(i-1)} = 2 \left( 1 - \frac{1}{k} \right) \quad (3.75)$$

where time is in continuous units of multiples of  $2N$ . Thus, the average height in continuous time of a coalescent tree starts at  $2(1 - 1/2) = 1$  for a pair of lineages and approaches 2 as  $k$  grows very large. In time units of  $2N$  generations, the average tree height starts at  $2N$  and increases to  $4N$  as  $k$  grows. The average total waiting time for all lineages to coalesce is heavily influenced by the average waiting time for the last pair of lineages to coalesce. Stated another way, average waiting times are shorter when  $k$  is larger and increase as lineages coalesce and  $k$  gets smaller. This result makes sense since there are fewer and fewer independent pairs of lineages that

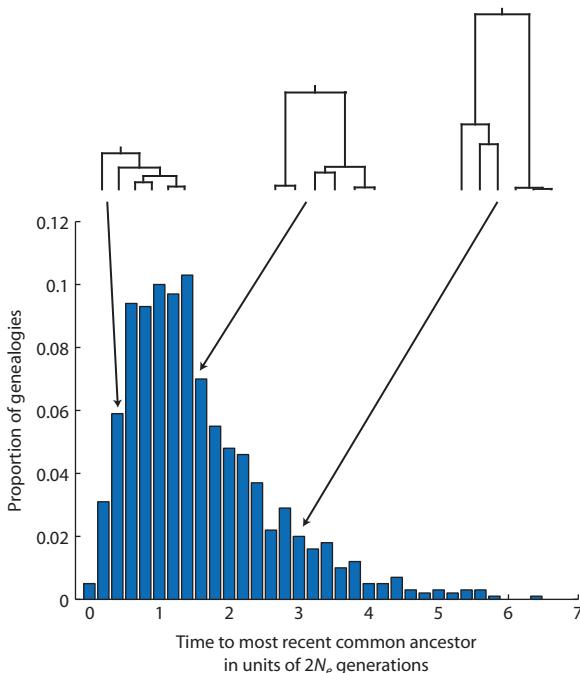
can coalesce as  $k$  decreases (think of  $\frac{k(k-1)}{2}$  as  $k$  decreases to 2).

The variance in the height of a coalescent tree is the sum of the variances in the coalescence time for each set of coalescent events since all coalescent events are independent:

$$\text{var}(H_k) = \sum_{i=2}^k \text{var}(T_i) = 4 \sum_{i=2}^k \frac{1}{i^2(i-1)^2} \quad (3.76)$$

The variance rapidly approaches the value of about 1.16 as the number of lineages sampled,  $k$ , increases to large values. The finite maximum value of the variance and the fact that the variance in tree height for  $k = 2$  lineages is 1.0 highlights the major impact of the coalescence time of the last pair of lineages on the overall height of a coalescence tree. Figure 3.27 illustrates the variance in the total height of genealogies by displaying the time to MRCA for 1000 replicate genealogies each starting with  $k = 6$ . The range of time to MRCA is large and the distribution has a very long tail representing a small proportion of genealogies that take a very long time for all coalescence events to occur.

Another useful measure is the total branch length of a coalescent genealogy of  $k$  lineages or  $L_k$ . The total branch length is the sum of the waiting times



**Figure 3.27** The distribution of times to a MRCA (or genealogy heights) for 1000 replicate genealogies starting with six lineages ( $k = 6$ ). The distribution of total coalescence times has a large variance because the range of times is large and also asymmetric with a long tail of a few genealogies that take a very long time to reach the MRCA. The genealogies shown above the distribution are those for the tenth, fiftieth, and ninetieth percentile times to MRCA. In this example  $N_e = 1000$ .

represented along each lineage in the tree. If the branches represented a system of roadways, then the sum of the distances traversed when driving over each segment of road once would be analogous to the total branch length. For example, if a pair of lineages has a waiting time  $x$  until coalescence, then the total waiting time is  $x + x = 2x$ . The average total branch length of a genealogical tree is then just twice the sum of the average waiting times (using continuous time) for each coalescence event:

$$E(L_k) = \sum_{i=1}^{k-1} 2 \frac{1}{i} = 2 \sum_{i=1}^{k-1} \frac{1}{i} \quad (3.77)$$

For example, with  $k = 2$  lineages  $i = 1$  in equation 3.77 gives an expected total branch length of two. This result makes sense because two lineages are expected to coalesce in  $2N_e$  generations or one time unit on the continuous time scale. Multiplying this expected time to coalescence by two for the two independent lineage branches gives a total branch length

of two. The expected total branch length starts at two, increasing with greater  $k$  and never reaching a maximum. However, the expected total branch length of a genealogy grows more and more slowly as  $k$  increases since the expected time to coalescence decreases with increasing  $k$ .

This section defines how coalescent theory describes the probabilistic events associated with lineage branching for a single population. This basic model can be extended to include the influence of numerous population genetic processes on coalescence. Later sections and chapters will explore how changes in population size (the next section), subdivided populations that experience gene flow (Chapter 4), mutation (Chapter 5), and natural selection (Chapter 7) influence patterns of coalescence. The application of coalescent theory to DNA sequence data, including examples based on empirical data, will be covered in Chapter 8.

### 3.7 Effective population size in the coalescent model

- The coalescent model of effective population size.
- Coalescent genealogies and population bottlenecks.
- Coalescent genealogies in growing and shrinking populations.

This section will explore the meaning of effective population size in the coalescent model. As in the models of genetic drift developed previously in this chapter, effective population size also plays a critical role in coalescence models. In the context of the coalescent model, the effective population size determines the chance that two gene copies descend from the same ancestor when working back in time from the present to the past.

In the coalescent model, two randomly sampled gene copies have the probability  $\frac{1}{2N_e}$  of finding their MRCA in the previous generation. If we call this the probability of coalescence,  $P_C$ , then for a diploid population:

$$P_C = \frac{1}{2N_e} \quad (3.78)$$

This equation can be rearranged:

$$N_e^i = \frac{1}{2P_C} \quad (3.79)$$

to give the inbreeding effective population size for the coalescent model over one generation. The effective size of a haploid population is defined identically except that the population contains  $N$  instead of  $2N$  gene copies. (The reasoning used is parallel to that used to arrive at the inbreeding effective population size in equations 3.53 and 3.54 based on the probability of identity by descent in a finite population.)

### Interact box 3.5 Simulating gene genealogies in populations with different effective sizes

Simulating lineage branching events forward in time for populations of different sizes is a direct way to understand how sampling leads to patterns of lineage coalescence. Use the link on the text web page to reach a simulation to model lineage branching events in finite populations.

Start by simulating populations of  $N : 4$  and  $N : 10$ . Before pressing the run button (the arrow that points to the right), determine the expected time for all of the gene copies in the present to find a single most recent common ancestor. Use your answer to intelligently set the number of generations (the **G:** entry field) to run the simulation to be able to see most of the lineage coalescence events (note that the maximum number of generations is 30). Run 25 simulations for each population size and tabulate the number of generations for all lineages in the present to find their most recent common ancestor. What does the distribution of coalescence times look like?

Use 100% for a **Speed** value to carry out replicate runs rapidly. The “untangle” button (the leftmost button at the bottom) rearranges the lineages so that the branches do not overlap and is very useful when tracing individual genealogies to visualize common ancestors.

We can apply the coalescent definition of the effective population size to the case of the breeding sex ratio in a population. For the coalescent model, it helps to think of the two sexes as two separate populations where gene copies can find their ancestors

(see Fig. 3.24). In a diploid population with two sexes and a 1 : 1 breeding sex ratio, half of the gene copies reside in females and half reside in males. A coalescence event requires that two gene copies descend from a single ancestor, either a single male or single female individual. The total probability that two randomly sampled gene copies coalesce in the previous generation,  $P_C$ , is the sum of the probability that the coalescence was in the population of females and the probability that the coalescence was in the population of males, or

$$P_C = P_C(\text{female population}) + P_C(\text{male population}) \quad (3.80)$$

Regardless of whether or not the two gene copies come from the male or female population, the probability of coalescence is the chance of sampling the same gene copy twice in the previous generation, or  $\frac{1}{2N_e^i}$ . To take the populations of the two sexes into account, notice that the probability that a gene copy is sampled from either the female or male population is  $\frac{1}{2}$ . The probability that *both* gene copies are sampled from the population of the same sex is therefore  $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ . Putting these probabilities together into equation 3.78 gives

$$P_C = \frac{1}{4} \frac{1}{2N_{ef}} + \frac{1}{4} \frac{1}{2N_{em}} \quad (3.81)$$

where the effective size of the female and male populations sum to the total effective size ( $N_e^i = N_{ef}^i + N_{em}^i$ ). Based on the definition of the effective population size as the probability of coalescence in equation 3.63,

$$\frac{1}{2N_e^i} = \frac{1}{4} \frac{1}{2N_{ef}} + \frac{1}{4} \frac{1}{2N_{em}} \quad (3.82)$$

which after some rearrangement, looks like

$$\frac{1}{2N_e^i} = \frac{1}{8} \left( \frac{N_{em}}{N_{ef}N_{em}} + \frac{N_{ef}}{N_{ef}N_{em}} \right) \quad (3.83)$$

$$N_e^i = 4 \frac{N_{ef}N_{em}}{N_{em} + N_{ef}} \quad (3.84)$$

This leads to an expression for the inbreeding effective population size in terms of the number of males and females. This equation shows us that the effective size of a diploid population with two sexes and an equal

breeding sex ratio ( $N_{ef}^i = N_{em}^i$ ) is equivalent to an ideal random-mating Wright–Fisher population and also gives a means to express the effective population size when the breeding sex ratio is not equal. We also see that the coalescence model allows us to reach exactly the same result that was obtained using the identity by descent approach (equation 3.44 and section 3.4).

The probability that two randomly sampled gene copies do not coalesce,  $P_{NC}$ , over some number of generations can be used to show the overall effective population size when the population size is not constant. As the basis of modeling coalescent times with the exponential distribution (the continuous time coalescent),  $e^{-\frac{t}{2N_e}}$  is used to approximate  $1 - \frac{1}{2N_e}$  as long as  $N_e$  does not get too small. This means that  $P_{NC}$  can be approximated by an exponential function:

$$P_{NC} = \left(1 - \frac{1}{2N_e}\right)^t \approx e^{-\frac{t}{2N_e}} \quad (3.85)$$

where  $t$  is the number of generations. Imagine that a population fluctuates in size over three generations as considered in section 3.3. In such a situation the probability of two randomly sampled gene copies not experiencing a coalescence event over the three generations could be approximated by

$$P_{NC} \approx e^{-\frac{t}{2N_e}} = e^{\left(-\frac{1}{2N_{e(t=1)}}\right) \left(-\frac{1}{2N_{e(t=2)}}\right) \left(-\frac{1}{2N_{e(t=3)}}\right)} \quad (3.86)$$

where each term in the exponent of  $e$  is distinct since the population sizes fluctuate over time. The exponential terms of  $e$  can be solved for the effective population size by taking the natural log of both sides to eliminate  $e$ :

$$-\frac{t}{2N_e} = \left(-\frac{1}{2N_{e(t=1)}}\right) + \left(-\frac{1}{2N_{e(t=2)}}\right) + \left(-\frac{1}{2N_{e(t=3)}}\right) \quad (3.87)$$

and then multiplying both sides by  $1/t$  and then  $-2$  to get

$$\frac{1}{N_e} = \frac{1}{t} \left( \frac{1}{N_{e(t=1)}} + \frac{1}{N_{e(t=2)}} + \frac{1}{N_{e(t=3)}} \right) \quad (3.88)$$

The term on the right side is the harmonic mean of the effective population size as given in equation 3.42.

Again, the identity by descent and coalescence approaches lead to the same conclusions about effective population size.

### **Coalescent genealogies and population bottlenecks**

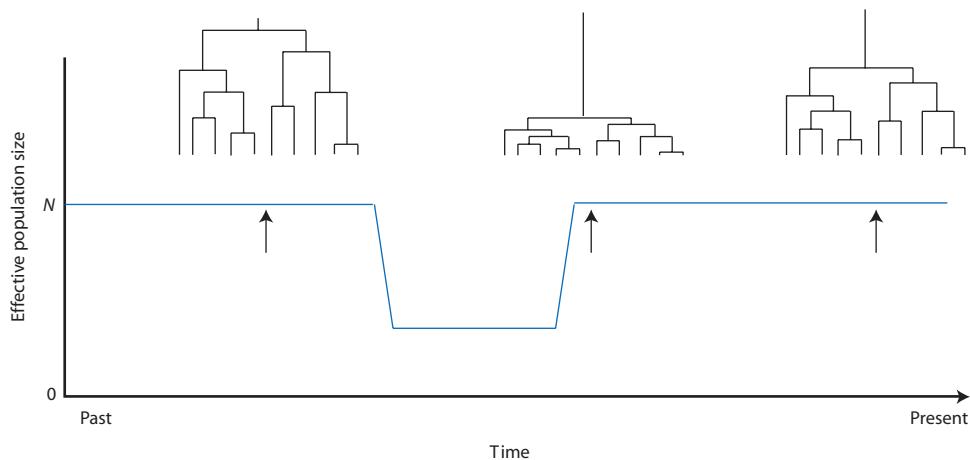
Let's examine the distribution of coalescence times in a population that experiences a bottleneck to see how genealogical branching patterns are affected. The situation in Fig. 3.28 is analogous to that in section 3.3 where a population starts out with 100 individuals, is reduced to 10 individuals for one generation, and then returns to a size of 100 individuals in the third generation (see Fig. 3.16). For a bottleneck of 100–10–100 over three generations, equations 3.85 and 3.86 can be used to determine the probability of coalescence events. A 100–10–100 population is equivalent to a population of 25 that is constant in size. Therefore, the average probability of two randomly sampled gene copies not finding their common ancestor over the three generations spanning the bottleneck is:

$$P_{NC} \approx e^{-\frac{3}{2(25)}} = 0.9418 \quad (3.89)$$

giving a chance of coalescence of  $1 - P_{NC} = 1 - 0.9418 = 0.0582$ . This probability is equivalent to  $1 - \frac{1}{2N_e}$  for three generations of 100–10–100 or  $P_{NC} = (0.995)(0.95)(0.995) = 0.9405$  to give  $1 - P_{NC} = 1 - 0.9405 = 0.0595$ . (Note the close agreement of the two answers even though a population of 10 is very small and violates the assumption that

$e^{-\frac{t}{2N_e}}$  is a good approximation of  $1 - \frac{1}{2N_e}$ .)

Compare the probability of coalescence  $P_C = 1 - 0.995 = 0.005$  for a population of 100 and  $P_C = 1 - 0.95 = 0.05$  for a population of 10. During the bottleneck there is a 10-fold greater chance each generation that two lineages find their ancestor or coalesce than before or after the bottleneck, leading to shorter branch lengths (times to coalescence). Thus, the bottleneck causes increased sampling of lineages and a greater chance that ancestral lineages are lost to sampling. After the bottleneck, the probability of coalescence returns to the same probability as in the first generation. However, the lineages present after the bottleneck are now much more likely to be recently related so the overall probability of coalescence is like that in a smaller population of constant size.



**Figure 3.28** The effects of a population bottleneck on gene genealogies. During the bottleneck the chance that two randomly sampled gene copies are derived from one copy in the previous generation  $\left(\frac{1}{2N_e}\right)$  increases. This can also be thought of as a reduction in the overall height of a genealogical tree caused by the bottleneck since lineages that find their ancestors during the bottleneck lead to short branches. The overall effect of a bottleneck on coalescence among gene copies sampled in the present depends on the reduction in the effective population size and the duration. The arrows indicate the point in time when gene copies were sampled from the population.

After a population recovers from a bottleneck, it is possible that the lineages present will all descend from a most recent common ancestor during the bottleneck period. If this occurs, it is equivalent to saying that a single lineage among those present in the pre-bottleneck population becomes the most recent common ancestor of all lineages during the bottleneck. The chance that this occurs increases as the bottleneck exhibits a smaller population size or persists for a longer period of time. The expected height of a coalescent tree (the sum of all the time periods between coalescence of pairs of gene copies until there is a single lineage, equation 3.75) can be used to show this effect quantitatively. In the continuous coalescent the expected height in units of

$$2N \text{ generations is } 2\left(1 - \frac{1}{k}\right) \text{ where } k \text{ is the number of}$$

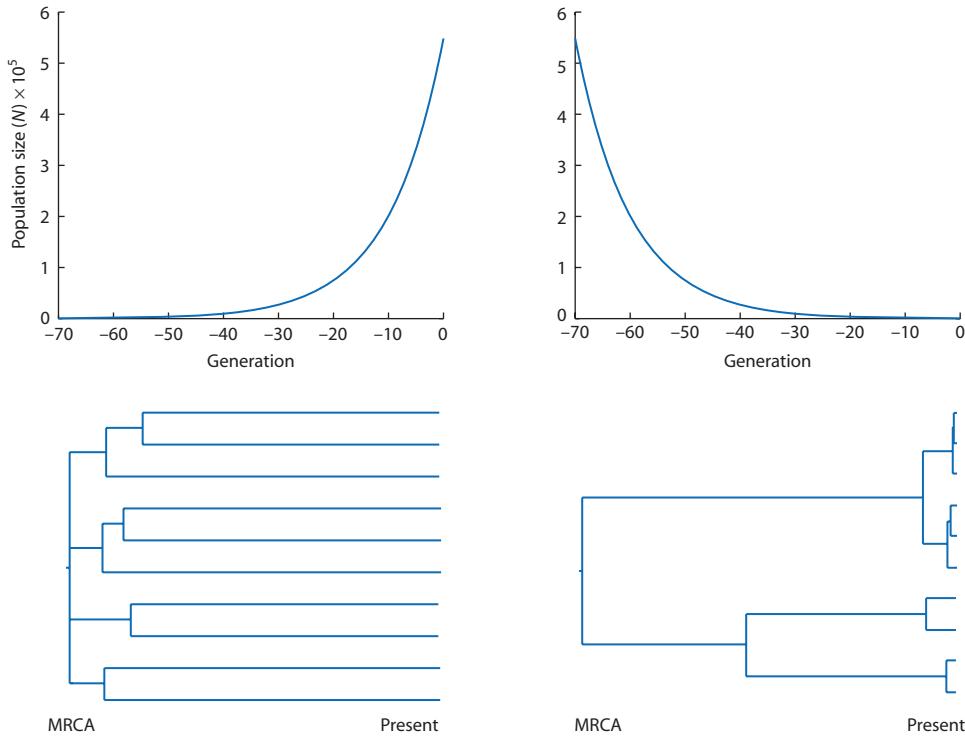
gene copies in the present. Figure 3.28 shows  $k = 10$  so the expected height of each coalescent tree is  $(1.8)(2N_e) = 3.6N_e$  generations. Before the bottleneck, a sample of 10 gene copies from a population of  $N_e = 100$  would coalesce to a single lineage in an average of 360 generations. The same sample of 10 gene copies taken from a population of  $N_e = 10$  would coalesce to a single lineage in an average of 36 generations. At the time point closest to the present in Fig. 3.28, the population has an effective size of about 25 based on the harmonic mean as well as the probability of coalescence over three generations.

Therefore, the expected height of the coalescent tree is 90 generations.

#### **Coalescent genealogies in growing and shrinking populations**

In a population growing in size over time, the probability of a coalescent event is least near the present because the population is at its largest size. Working back in time from the present to the past, the size of the population is continually shrinking. This means that the probability of a coalescence event ( $\frac{1}{2N}$ ) must also be continually increasing as we move back in time toward the MRCA, since  $N$  is shrinking. The result is that genealogies in growing populations do not follow the rule that the final coalescence time from two lineages to the MRCA is the longest on average in a genealogy established in the last section for populations of constant size. Instead, genealogies in growing populations tend to have longer times between coalescent events toward the present and shorter times between coalescent events in the past (Fig. 3.29). Genealogies from rapidly growing populations therefore tend to have longer branches toward the present and shorter branches deep in the tree.

In a population shrinking in size over time, the probability of a coalescent event is greatest near the present because the population size is at its smallest. The effect on genealogies is that coalescent waiting



**Figure 3.29** The effects of exponential population growth or shrinkage on coalescent genealogies. The upper panels show change in population size over time with exponential growth according to  $N(t) = N_0 e^{-rt}$  with  $r = \pm 0.1$ , yielding relatively slow exponential population growth. The two genealogies illustrate examples of waiting times that might be seen under strong exponential population growth (left) and shrinkage (right). With strong exponential population growth coalescent times are longest in the present when the population is the largest, leading to genealogies characterized by long branches near the present and very short branches in the past around the time of the MCRA. With exponential population shrinkage, coalescence times are greatest in the past near the MRCA when the population was larger and shortest near the present when the population is at its smallest size. The genealogy on the left was obtained using equation 3.91 with  $r = 100$ .

times in the present are even shorter than they would be under constant population size. Similarly, when population size is shrinking coalescence times in the past are relatively greater than under constant population size because the probability of a coalescence event was greater (Fig. 3.29).

A common way to model growing or shrinking populations is to assume that population size is changing exponentially over time. Under exponential growth, the population size at time  $t$  in the past is a function of the initial population size in the present  $N_0$  and the rate of population growth  $r$  according to

$$N(t) = N_0 e^{-rt} \quad (3.90)$$

Examples of population size over time under exponential population growth are shown in Fig. 3.29. With exponential growth, population size tends to change very rapidly.

The generalizations above regarding coalescent waiting times depend on rapid and sustained changes

in population size over time such as under exponential population growth with a constant rate ( $r$ ). In populations that are changing in size slowly over time, it is possible that coalescence waiting times differ very little from those expected under constant population size. Increases in population size over time cause the probability of coalescence to decrease toward the present. At the same time, the chance of a coalescence event increases toward the present simply due to a larger number of lineages (increasing  $k$ ) available to coalesce. Only in populations with rapidly changing population size will the reduction in the probability of coalescence be great enough to overcome the effect of an increasing number of lineages available to coalesce toward the present. In addition, the variance in coalescence waiting times with constant  $N$  is large so that only very rapid and sustained change in population size will noticeably impact the distribution of coalescence waiting times.

Changing population size over time complicates finding the distribution of coalescent times. In the

### Interact box 3.6 Coalescent genealogies in populations with changing size

The coalescent process can be simulated for populations experiencing exponential growth in population size through time. The simulation displays a genealogy based on values for the number of lineages (**n:**) and the growth rate of the population (**exp:**). The growth rate parameter corresponds to the rate of exponential population growth and is therefore the key variable in this simulation. Click the **Recalc** button to begin a new simulation. The resulting genealogy can be viewed as an animation going back in time (click **Animation** tab at top left and use playback controls at bottom) or as a genealogy (click **Trees** tab at top left).

Set the growth rate parameter to 0 and click **Recalc**. View the resulting genealogy. Then set the growth rate parameter to 128 and click **Recalc** and view the results. How do the genealogies in growing populations compare with genealogies when population size is constant over time? Re-run the simulation a few times for both values of the growth rate parameter. Also try other less extreme values of the growth rate parameter, such as 4, 8, and 32.

previous section, the exponential distribution of coalescence waiting times was obtained by relying on the fact that the probability of non-coalescence,

$$\text{or } \prod_{x=0}^{k-1} \left(1 - \frac{x}{2N}\right) \text{ where } k \text{ is the number of lineages}$$

available to coalesce, was a constant since  $N$  remains the same over time. If  $N$  is instead changing rapidly over time, then the probability of non-coalescence also changes rapidly over time. The result is that the distribution of coalescent times cannot be exponentially distributed when population size is changing over time as it is when population size is constant over time. The waiting time between coalescent events in an exponentially growing population is obtained from the equation:

$$t_{k \rightarrow k-1} = \ln \left[ 1 + (N_0 r) e^{-\tau_i} \left( \frac{-2}{k(k-1)} \right) \ln(U) \right] \quad (3.91)$$

where  $U$  is a uniformly distributed random variable between zero and one,  $N_0$  is the initial population size,  $r$  is the rate of population growth, and  $\tau_i$  is the sum of all the past coalescence waiting times up to the current number of lineages  $k$  that have not yet coalesced according to

$$\tau_i = \sum_{k=n}^{i+1} t_k \quad (3.92)$$

as shown by Slatkin and Hudson (1991). In equation 3.91 time is scaled in units of  $r$  since  $\tau = rt$ . As  $\tau_i$  gets larger then more time has elapsed, meaning that the size of the population has changed more.

Note that equation 3.91 only applies to populations growing though time. In populations shrinking in size toward the present, there is a chance that there will never be coalescence to a MRCA since the probability of coalescence approaches zero as the population size approaches infinity going back in time from the present to the past.

### Chapter 3 review

- In finite populations, allele frequencies can change from generation to generation since the sample of gametes that found the next generation may not contain exactly the same number of each allele as the previous generation. The chances of a large change in the number of alleles decreases as the number of gametes sampled increases. Sampling error in allele frequency causes genetic drift, the random process whereby all alleles eventually reach fixation or loss.
- The Wright–Fisher model is a simplification of the biological life cycle used to model genetic drift. It makes assumptions identical to Hardy–Weinberg in addition to assuming that each generation is founded by sampling  $2N$  gametes from an infinite pool of gametes. The binomial distribution can be used to predict the probability that a Wright–Fisher population goes from some initial number of alleles to any number of alleles in the next generation.
- The action of genetic drift in a very large number of identical replicate populations can be modeled with a Markov chain model (based on the binomial distribution). The model tracks the probabilities

of a population with any number of copies of a given allele transitioning to all possible numbers of copies of the same allele in the next generation. Markov models predict that the chances of fixation or loss are equal when  $p = q = 1/2$  and that genetic drift reduces genetic variation by  $1 - \frac{1}{2N}$  every generation.

- The process of genetic drift in many replicate populations can be thought of as analogous to the diffusion of particles in space. This leads to the diffusion approximation of genetic drift, where the rate at which individual populations reach fixation or loss depends on the diffusion coefficient: a function of population size and allele frequency. The diffusion equation predicts that an allele at an initial frequency of  $1/2$  will remain segregating for an average of about  $2.8N$  generations.
- The size of a population is defined by the behavior of allele frequencies over time. The effective population size ( $N_e$ ) is the size of an ideal Wright–Fisher population that shows the same allele frequency behavior over time as an observed biological population regardless of its census population size ( $N$ ).
- Finite population size and consanguineous mating are analogous processes since both lead to increasing homozygosity and decreasing heterozygosity over time. The distinction is that genetic drift in finite populations causes changes in both genotype *and* allele frequencies (alleles are lost and fixed) while consanguineous mating changes only genotype frequencies.
- Numerous models predict dynamics of genetic variation based on the effective population size ( $N_e$ ). As a consequence, there are several definitions of  $N_e$ , including the variance effective population size, the inbreeding effective population size, and the breeding effective population size. The effective population size can be estimated from direct observation of genetic variation over time and  $N_e$  is often less than  $N$  in actual populations.
- The average time for a pair of lineages to coalesce is the same as the population size or  $2N_e$  with a large expected range around this average (the variance is  $4N_e^2$ ). In a sample of  $k$  lineages the average time to the first coalescent event is  $2N_e$  divided by the number of unique pairs of lineages ( $\frac{k(k-1)}{2}$ ).
- Most coalescent events for a sample of lineages occur in the recent past with only a few lineages having long times to coalescence.

- The effective population size ( $N_e$ ) can be defined for lineage branching models by reference to the probability of two randomly sampled gene copies descending from the same ancestral copy. This probability decreases as the effective size of populations grows larger. The coalescent model leads to definitions of the inbreeding effective population size that are identical to those obtained using autozygosity.
- Exponential population growth changes the distribution of coalescence times relative to a population with constant population size. When population size is growing, lineages nearest the present tend to have the longest coalescence waiting times because the probability of coalescence grows steadily smaller toward the present. When population size is shrinking rapidly, the probability of coalescence grows steadily greater toward the present causing lineages nearest the present to have the shortest coalescence waiting times.

### Further reading

For an intriguing account of the role of chance in everyday affairs, see:

Mlodinow L. 2008. *The Drunkard's Walk: How Randomness Rules Our Lives*. Pantheon Books, New York, NY.

For more mathematical detail on Markov chains and the diffusion equation, see:

Ewens WJ. 2004. *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd edn. Springer-Verlag, New York.

To learn more about the diffusion equation, its assumptions, and how processes such as mutation, migration, and natural selection can be combined with genetic drift, consult:

Roughgarden J. 1996. *Theory of Population Genetics and Evolutionary Ecology: an Introduction*. Prentice Hall, Upper Saddle River, NJ.

Rice SH. 2004. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates, Sunderland, MA.

For further information on methods used to estimate the effective population size, see:

Caballero A. 1994. Developments in the prediction of effective population size. *Heredity* 73: 657–79.

More extensive background on coalescence theory along with many examples and a comprehensive list of references can be found in:

Hein J, Schierup MH, and Wiuf C. 2005. *Gene Genealogies, Variation and Evolution*. Oxford University Press, New York.

For a discussion of the coalescent effective population size, see:

Sjödin P, Kaj I, Krone S, Lascoux M, and Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics* 169: 1061–70.

## Problem box answers

### Problem box 3.1 answer

In the first generation both populations have 24 A alleles and 24 a alleles. After one generation, however, there are  $(0.458)(48) = 22$  A alleles in one population and  $(0.521)(48) = 25$  alleles in the other. The chances of observing these allele frequencies can be determined with the binomial formula under the assumptions of the Wright–Fisher model:

$$\begin{aligned} P_{A=22} &= \frac{48!}{22!(26)!}(0.5)^{22}(0.5)^{26} \\ &= 2.7386 \times 10^{13}(2.3841 \times 10^{-7})(1.4901 \times 10^{-8}) \\ &= 0.0973 \end{aligned}$$

$$\begin{aligned} P_{A=25} &= \frac{48!}{25!(23)!}(0.5)^{25}(0.5)^{23} \\ &= 3.0958 \times 10^{13}(2.9802 \times 10^{-8})(1.1921 \times 10^{-7}) \\ &= 0.110 \end{aligned}$$

In both populations, the chance of observing the allele frequency changes under genetic drift is about 10%. The chance of observing these allele frequency changes in both populations is only 0.0107 since the

probability of two independent events is the product of their individual probabilities.

### Problem box 3.2 answer

Each of the values in the transition matrix is obtained using the binomial formula. The chance that a population at fixation or loss transitions to an allele frequency different than 1 or 0, respectively, is always 0. The chance of transitioning from one to four A alleles is identical to the chance of transitioning from three to no a alleles, since the number of A alleles is four minus the number of a alleles. Using this symmetry permits two columns to be filled out after performing calculations for only one of the columns (see table below). The transition probabilities are a function of the sample size only and so are constant each generation. The total frequency of populations in a given allelic state in the next generation depends on initial frequencies of populations in each state ( $P_{t=0}(x)$ ). The expected frequencies of populations in each allelic state therefore changes each generation. This Markov chain model is available as a Microsoft Excel spreadsheet on the textbook website under the link to **Problem Box 3.2**.

	One generation later ( $t = 1$ )	Initial state: number of A alleles ( $t = 0$ )				
State	Expected frequency	0	1	2	3	4
0	$P_{t=1}(0) = (1.0000)P_{t=0}(0) + (0.3164)P_{t=0}(1) + (0.0625)P_{t=0}(2) + (0.0039)P_{t=0}(3) + (0.0000)P_{t=0}(4)$					
1	$P_{t=1}(1) = (0.0000)P_{t=0}(0) + (0.4219)P_{t=0}(1) + (0.2500)P_{t=0}(2) + (0.0469)P_{t=0}(3) + (0.0000)P_{t=0}(4)$					
2	$P_{t=1}(2) = (0.0000)P_{t=0}(0) + (0.2109)P_{t=0}(1) + (0.3750)P_{t=0}(2) + (0.2109)P_{t=0}(3) + (0.0000)P_{t=0}(4)$					
3	$P_{t=1}(3) = (0.0000)P_{t=0}(0) + (0.0469)P_{t=0}(1) + (0.2500)P_{t=0}(2) + (0.4219)P_{t=0}(3) + (0.0000)P_{t=0}(4)$					
4	$P_{t=1}(4) = (0.0000)P_{t=0}(0) + (0.0039)P_{t=0}(1) + (0.0625)P_{t=0}(2) + (0.3164)P_{t=0}(3) + (1.0000)P_{t=0}(4)$					

### Problem box 3.3 answer

This captive population has experienced a reduction in population size due to three

factors. This case can be thought of as a triple bottleneck. First, the breeding sex ratio became unequal. The effective size based on the number of males (10) and females (15) is:

$$N_e = 4 \frac{(10)(15)}{10 + 15} = 24$$

from equation 3.44. The mean family size among the 15 females was four, returning the population back to a census size of 60. However, the variance in family size was 6.5 and thus greater than expected for a Poisson distribution. The effective population size based on the variance in family size is:

$$N_e = \frac{4(24)}{6.5 + 16 - 4} = 5.2$$

from equation 3.46. Notice that the effective population size used in the numerator is 24, as determined for the unequal sex ratio. In total, the population has fluctuated from  $N_e = 60$  before the fire, to census sizes of 25, and then 60 over three generations. The effective size in generation two was 24 due to unequal sex ratio. The effective size in generation three was five (after rounding) due to a growing population with high variance in family size. Therefore, the effective population size over three generations is:

$$\frac{1}{N_e} = \frac{1}{3} \left[ \frac{1}{60} + \frac{1}{24} + \frac{1}{5} \right]$$

$H_{t=1}$

0.50  
0.50  
0.50  
0.50  
0.50  
0.50  
0.50  
0.50  
0.50  
0.50  
Average  $\hat{N}_e^i$

$H_{t=4}$

0.4504  
0.4594  
0.3474  
0.4376  
0.3772  
0.4999  
0.4566  
0.4856  
0.3578  
0.4550

or  $N_e = 11.6$ . Thus the fire caused a major reduction in the overall effective population size. The population experienced about the same amount of genetic drift over these three generations as an ideal Wright–Fisher population with an effective size of 12 individuals. One assumption in this analysis is that these animals approximate organisms that breed only once during their lifetimes. This assumption is often unmet for animals and additional reductions in the effective population size can arise if mating occurs between members of overlapping generations since related individuals can mate (see Nunney 1993 and an application in Ollivier & James 2004).

### Problem box 3.4 answer

We can use a table format just like that of Table 3.5 to set up the calculations (see below). In this case three generations of sampling have elapsed so that  $t = 3$  in the numerator of the formula for  $\hat{N}_e^i$ . The mean is now much higher than we expect if we include the sixth replicate, or 5.2 if we exclude this value. Either way, the estimate of  $\hat{N}_e^i$  has not improved much. We are still limited by too little elapsed time or too few replicate populations.

$$\ln\left(\frac{H_{t=4}}{H_{t=3}}\right) \quad \hat{N}_e^i = -\frac{1}{2} \frac{3}{\ln\left(\frac{H_{t=4}}{H_{t=3}}\right)}$$

−0.1044	4.79
−0.0847	5.90
−0.3641	1.37
−0.1332	3.75
−0.2816	1.78
−0.0001	3471.97
−0.0908	5.50
0.0291	17.17
−0.3348	1.49
−0.0942	5.31
	351.9

## CHAPTER 4

# Population structure and gene flow

### 4.1 Genetic populations

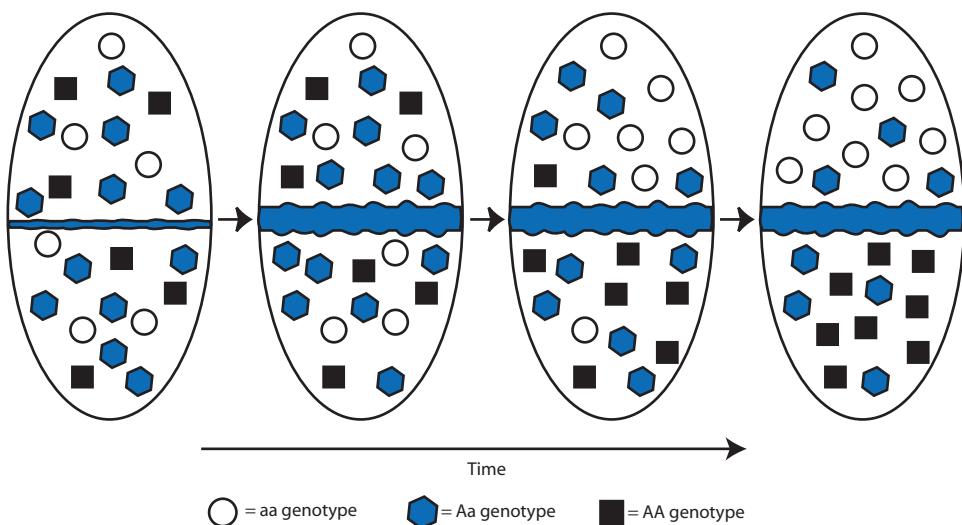
- Genetic versus geographic organization of populations.
- Isolation by distance and divergence of populations.
- Gene flow and migration.
- Direct and indirect measures of gene flow.

The expectation that genotypes will be present in Hardy–Weinberg frequencies, covered in detail in Chapter 2, depends on the assumption of random mating throughout a population. Implicit is the view that a population is a single entity where processes such as mating and movement of individuals are uniform throughout, a condition often called **panmixia**. Several processes and features at work in actual populations make this initial perspective of population uniformity unlikely to hold true for many populations. It is often the case that within large populations the chances of mating are not uniform as assumed by Hardy–Weinberg. Instead, the chance that two individuals mate often depends on their location within the population. This leads to what is called **population structure**, or heterogeneity across a population in the chances that two randomly chosen individuals will mate. The first section of this chapter will introduce biological phenomena that contribute to population structure in mating and migration that can lead to differences in allele and genotype frequencies in different parts of a population. The goal of the entire chapter is to develop expectations for the impact of population structure on genotype and allele frequencies along with methods to measure patterns of population structure.

To get an initial idea of how a population might be divided into smaller units that behave independently, consider the hypothetical population in Fig. 4.1. Initially, all individuals in the population have equal chances of mating regardless of their location. Since mating is random, genotype frequencies in the

entire population match Hardy–Weinberg expectations and allele frequencies are equal on both sides of the creek. Then imagine that the creek bisecting the population changes permanently into a large river that serves as a barrier to movement of individuals from one side to the other side. Although some individuals still cross the river on occasion, the rate of genetic mixing or **gene flow** between the two subpopulations bisected by the river is reduced. Lowered levels of gene flow mean that the two subpopulations have allele and genotype frequencies that tend to be independent through time. At the later time points in Fig. 4.1, the two subpopulations have increasingly different allele frequencies over time due to genetic drift, even though there are Hardy–Weinberg expected genotype frequencies within each subpopulation. In the last time period in Fig. 4.1, the allele frequencies in the subpopulations separated by the river are quite different and the genotype frequencies in the total population no longer meet Hardy–Weinberg expectations. In this example, a reduction in gene flow allows the two subpopulations to be acted on independently by genetic drift, ultimately resulting in population differentiation of allele frequencies. The appearance of a geographic barrier that restricts gene flow among populations like that in Fig. 4.1 is sometimes called a **vicariance** event. Subpopulations – entities recognized with names such as herds, flocks, prides, schools, and even cities – can be formed by a wide range of temporal, behavioral, and geographic barriers that ultimately result in subpopulation allele frequencies that differ from the average allele frequency of the total population.

Another cause of population structure is more subtle, but easy to understand with a thought experiment. Think of one common species of animal or plant that you encounter regularly at home or work. Think of individuals of this species seeking out mates completely at random. Where would individuals likely find mates? They would probably find mates among the other individuals nearby rather than far away. I thought of the trees that are near my home and also

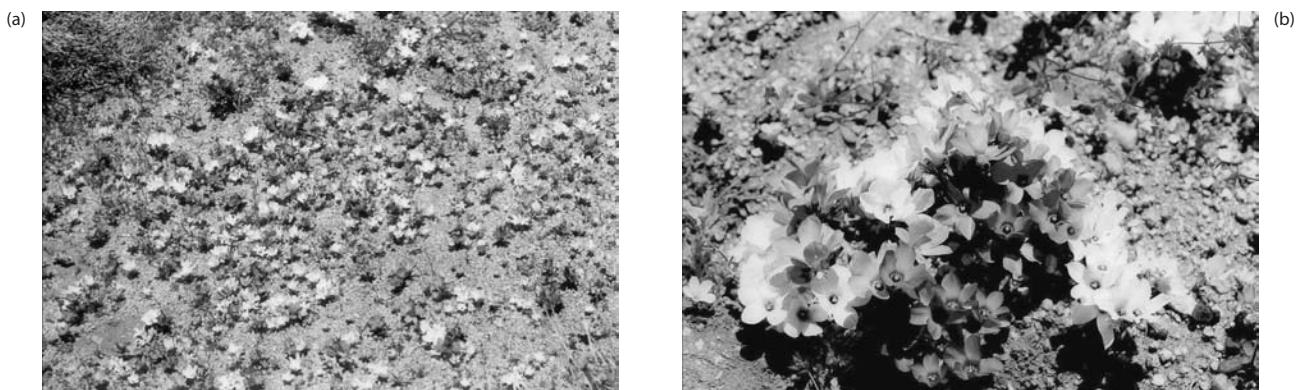


**Figure 4.1** An example of population structure and allele-frequency divergence produced by limited gene flow. The total population (large ovals) is initially in panmixia and has Hardy–Weinberg expected genotype frequencies. Then the stream that runs through the population grows into a large river, restricting gene flow between the two sides of the total population. Over time allele frequencies diverge in the two subpopulations through genetic drift. In this example, you can imagine that the two subpopulations drift toward fixation for different alleles but neither reaches fixation due to an occasional individual that is able to cross the river and mate. Note that there is random mating (panmixia) within each subpopulation so that Hardy–Weinberg expected genotype frequencies are maintained within subpopulations. However, after the initial time period genotype frequencies in the total population do not meet Hardy–Weinberg expectations.

on the university campus where I work. When these trees flower and mate via the movement of pollen, it seems likely to me that trees that are closer together are more likely to be mates. I would not expect two trees that are tens or hundreds of kilometers apart to have a good chance of mating. Imagine the species that you thought of and the distances over which mating events might take place. Even if individuals can find mates very far away, there is usually some spatial scale at which the chances of mating are

limited. This varies with the species and could be distances as small as a few meters or as large as thousands of kilometers depending on the range of movement of individuals and their gametes.

This phenomenon of decreasing chances of mating with increasing distance separating individuals is termed **isolation by distance** (Wright 1943a, 1943b, 1946). Sewall Wright was motivated by data on the spatial frequencies of blue and white flowers of the plant *Linanthus parryae* (Fig. 4.2) to develop



**Figure 4.2** The plant *Linanthus parryae*, or desert snow, is found in the Mojave Desert region of California. (a) *L. parryae* can literally cover thousands of hectares of desert during years with rainfall sufficient to allow widespread germination of dormant seeds present in the soil. This tiny plant has either blue or white flowers. (b) In some locations most plants have blue or most have white flowers whereas in other locations more equal frequencies of the two flower colors are found. Reproduced with permission of Barbara J. Collins. For a color version of this image see Plate 4.2.

expectations for populations experiencing isolation by distance. The patchwork spatial pattern of flower color frequencies in *L. parryae* was considered by Wright as a prime example of the consequences of isolation by distance in continuous populations. However, as flower-color frequencies were followed over more and more years, the interpretation that flower-color frequencies were primarily due to genetic drift made possible by isolation by distance has been challenged (e.g. Turelli et al. 2001). Wright carried out a series of detailed analyses of *L. parryae* data (Wright 1978) but the nature of the processes that affect the spatial distributions of *L. parryae* flower colors is a controversy that has continued for more than 50 years (see Schemske & Bierzychudek 2001). Regardless of the specific situation in *L. parryae*, the phenomenon of isolation by distance is ubiquitous in natural populations. One biologist described isolation by distance as a given in the genetics of natural populations, likening it to the force of gravity in physics, with the only question being the geographic scale at which it impacts genotype and allele frequencies.

Computer simulations are a convenient way to explore how isolation by distance influences allele and genotype frequencies. Figure 4.3 shows two simulated populations where each point on a grid represents the geographic location of a diploid individual. In one case, the population exhibits panmixia and individuals find a mate at random from all individuals within a  $99 \times 99$  individual mating area. In the contrasting case where there is strong isolation by distance, each individual mates at random within a much smaller  $3 \times 3$  individual area. Both populations start off looking very similar, with Hardy–Weinberg expected genotype frequencies and randomly scattered locations of the three genotypes. After 200 generations the population with a  $99 \times 99$  individual mating area (Fig. 4.3a) still shows random locations of the three genotypes. However, the population with a  $3 \times 3$  mating area (Fig. 4.3b) has distinct clumps of identical genotypes and fewer heterozygotes (represented by blue squares). One effect of isolation by distance is clearly local changes in allele frequency in a population, with local regions approaching fixation or loss, akin to the impact of reducing the effective population size (see the breeding effective population size in section 3.5). Alternatively, isolation by distance can be thought of as a form of inbreeding, since restricted mating distances cause homozygosity within subpopulations to increase. The patterns of genotypes in the simulated populations bear this out, with an obvious decline

in the overall frequency of heterozygotes over time with isolation by distance (Fig. 4.3b) but no such decline when there is panmixia (Fig. 4.3a).

**Isolation by distance** Decreasing chances of mating or gene flow as the geographic distance between individuals or populations increases.

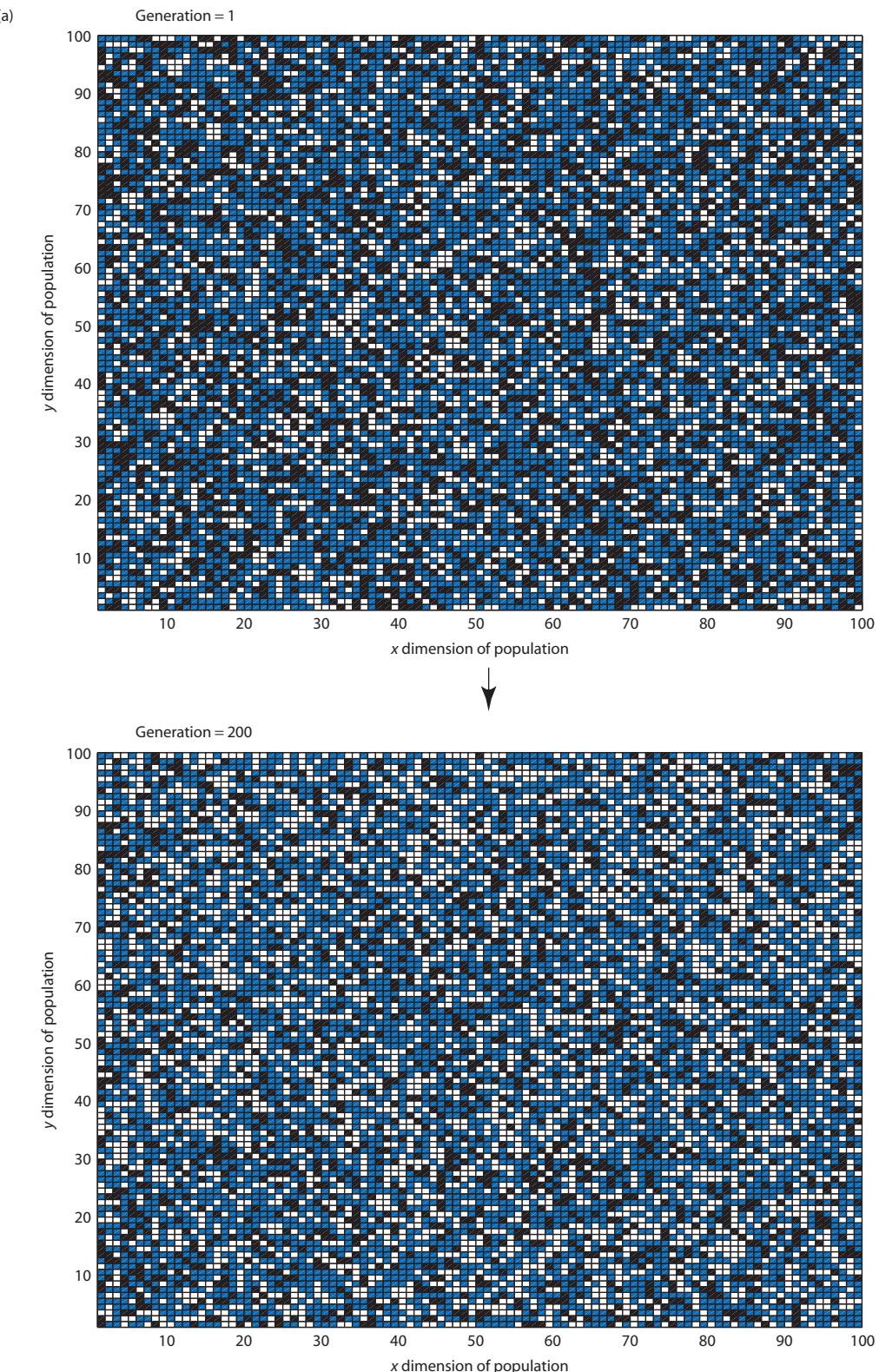
**Gene flow** The successful movement of alleles into populations through the movement of individuals (migration) or the movement of gametes.

**Panmixia** Random mating, literally meaning “all mixed.”

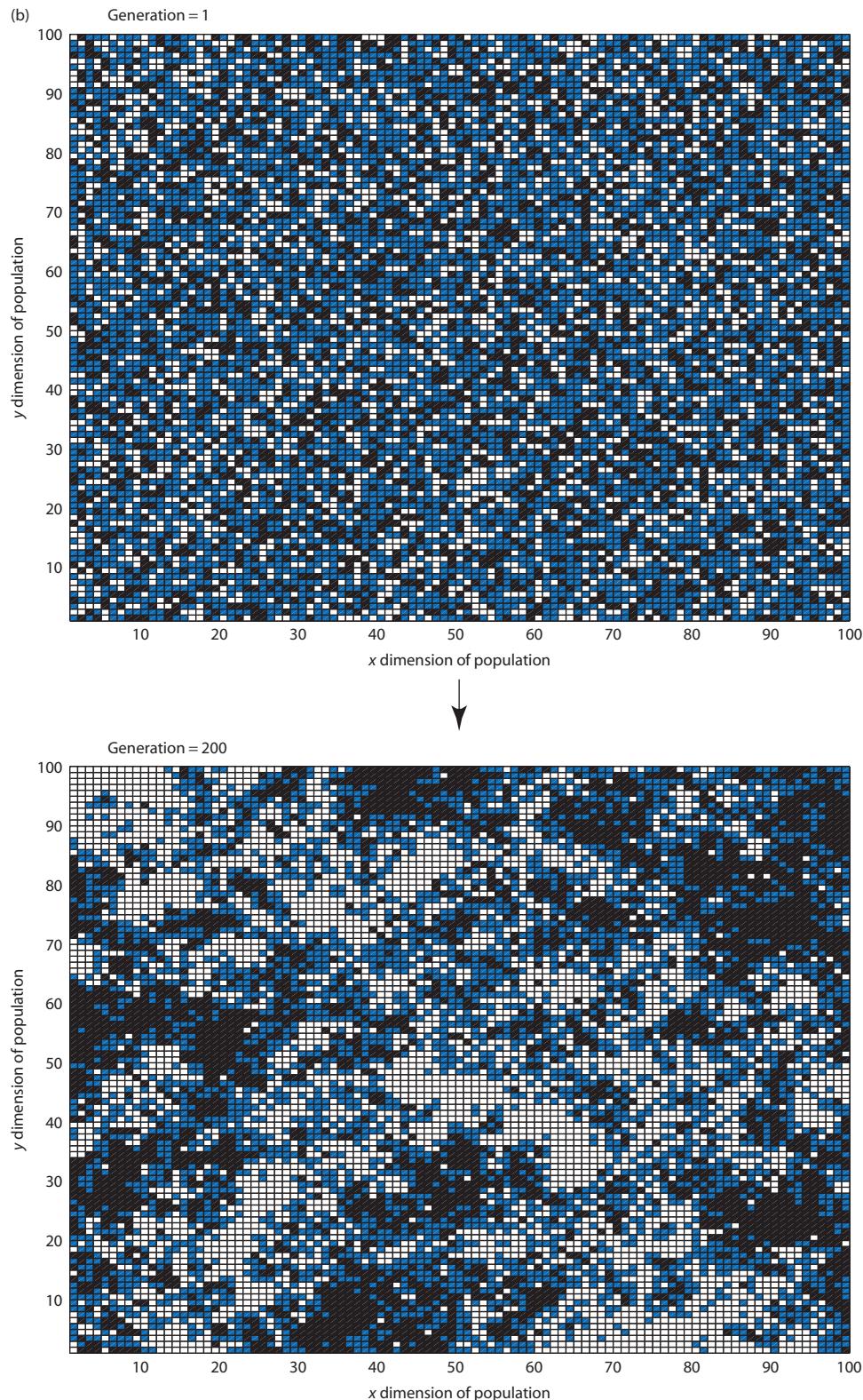
**Population structure** Heterogeneity in allele frequencies across a population caused by limited gene flow.

**Subpopulation** A portion of the total population that experiences limited gene flow from other parts of the total population so that its allele frequencies evolve independently to some degree; synonymous with **deme**.

Population structure has profound implications for genotype and allele frequencies. Subdivision breaks up a population into smaller units that are each genetically independent to some degree. One consequence is that each subpopulation has a smaller effective population size than the effective size of the entire population if there were random mating. The genetic variation found in a single large panmictic population and a population subdivided into many smaller demes is organized in a different manner. Think of the simple case of a diallelic locus. A single large population may take a very long time to experience fixation or loss due to genetic drift and thus maintain both alleles. In a highly subdivided population each deme may quickly reach fixation or loss, but both alleles can be maintained in the overall population since half of the subpopulations are expected to reach fixation and half loss for a given allele. Processes that cause population structure can also be thought of as both creative and constraining in evolutionary change (Slatkin 1987a). The genetic isolation among demes caused by subdivision can prevent novel and even advantageous alleles from spreading throughout a population. But, at the same time, genetic isolation allows subpopulations to evolve independent allele frequencies and maintain unique



**Figure 4.3** Isolation by distance causes spatial structuring of allele and genotype frequencies. In these pictures, a population is represented in two dimensions with each point on a grid representing one diploid individual. The colors represent an individual with a heterozygous (blue) or homozygous (black and white) genotype at each point. In (a) there is random mating over the entire population (the mating neighborhood is  $99 \times 99$  squares) whereas in (b) there is strong isolation by distance (the mating neighborhood is  $3 \times 3$  squares). The population with isolation by distance (b) develops and maintains spatial clumping of



genotypes and therefore spatial clumping of allele frequencies. There is no such spatial structure in the population with random mating (a). In the simulation that produced these pictures, the grid is initially populated at random with genotypes at Hardy–Weinberg expected frequencies and  $p = q = 1/2$ . Each generation every individual chooses a mate at random within its mating neighborhood and replaces itself with one offspring. The offspring genotypes are determined by Hardy–Weinberg probabilities for each combination of parental genotypes.

### Method box 4.1

#### Are allele frequencies random or clumped in two dimensions?

How can genetic variation in space be described to look for evidence of isolation by distance or other processes that cause spatial genetic differentiation in populations? The general approach is to compare pairs of individuals or populations, looking at both the similarity of their genotypes and how far apart they are located. Isolation by distance is a form of inbreeding due to non-random mating, so it causes individuals that are located near to each other to be more related on average.

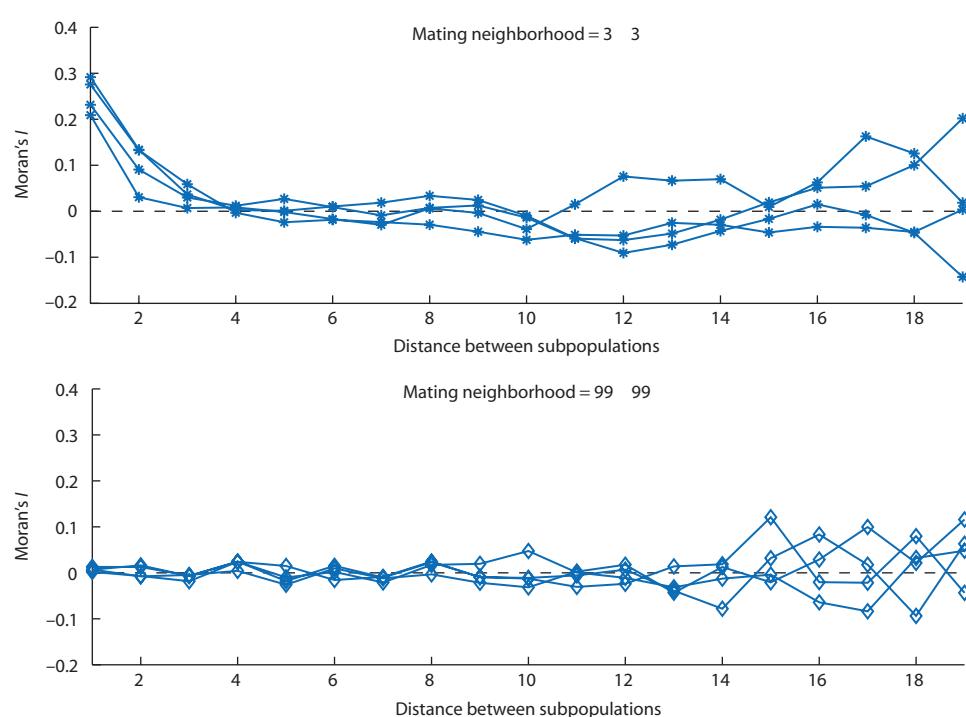
One classic statistic used to estimate spatial genetic structure is a correlation measure called Moran's  $I$ :

$$I_k = \frac{n \sum_{i=1}^n \sum_{j=1, j \neq i}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{W_k \sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.1)$$

where  $k$  represents a distance class (e.g. all populations two distance units apart) so that  $w_{ij}$  equals one if the distance between location  $i$  and  $j$  equals  $k$ , and zero otherwise. Within a distance class  $k$ ,  $n$  is the number of

populations,  $y$  is the value of a genetic variable such as allele frequency for location  $i$  or  $j$ ,  $\bar{y}$  is the mean allele frequency for all populations, and  $W_k$  is the sum of the weights  $w_{ij}$  or  $2nk$ . The numerator is larger when pairs of populations have similar allele frequencies that both show a large difference from the mean allele frequency.

The formula for Moran's  $I$  might be a bit daunting at first, but the results it produces are easy to understand and interpret biologically. Like correlations in general, Moran's  $I$  takes on values from  $-1$  to  $+1$  when calculated with a large number of samples. A positive value of  $I$  means that that allele frequencies between pairs of locations are similar on average while a negative value means that allele frequencies between pairs of locations tend to differ on average. A value of zero indicates that differences in subpopulation allele frequencies are not related to the distance between locations or that genetic variation is randomly distributed in space. The spatial locations of genotypes like those shown in Fig. 4.3 are the perfect situation to use Moran's  $I$  (see Fig. 4.4).



*(continued)*

alleles as is required for genetic adaptation to local environments under natural selection, for example.

It is worth noting that there are some important biological distinctions between gene flow and migration. Migration is simply the movement of individuals from one place to another. As such, migration may or may not result in gene flow. Gene flow requires that migrating individuals successfully contribute alleles to the mating pool of populations they join or visit. Thus, migration alone does not necessarily result in gene flow. Similarly, gene flow can also occur without migration of individual organisms. Plants are a prime example, with gene flow that takes place via movement of pollen grains (male gametes) but individuals themselves cannot migrate except as seeds. Gene flow can also occur without easily detected migration of individuals, such as cases where individuals move briefly to mate and then return to their original geographic locations. To confuse matters, the variable  $m$  (for migration rate) is almost universally used to indicate the rate of gene flow in models of population structure. Even though models do not normally make the distinction, it is wise to remember the biological differences between the processes of migration and gene flow in actual populations.

This chapter is devoted to expectations for allele and genotype frequencies in subdivided populations. The next section will cover so-called direct measures of gene flow that can be used in natural populations to determine the extent of population subdivision based on patterns of parentage determined with genetic markers. Then in the third section, we will return to the fixation index (or  $F$ ) from Chapter 2 and extend it for the case of structured populations in order to serve as a measure of population subdivision. The fourth section will consider how genotype frequencies are impacted by population structure. The fifth section will return to fixation index estimates to show how they can be compared with an idealized population model to arrive at an indirect measure of historical gene flow. The final section of the chapter incorporates population subdivision into coalescent models.

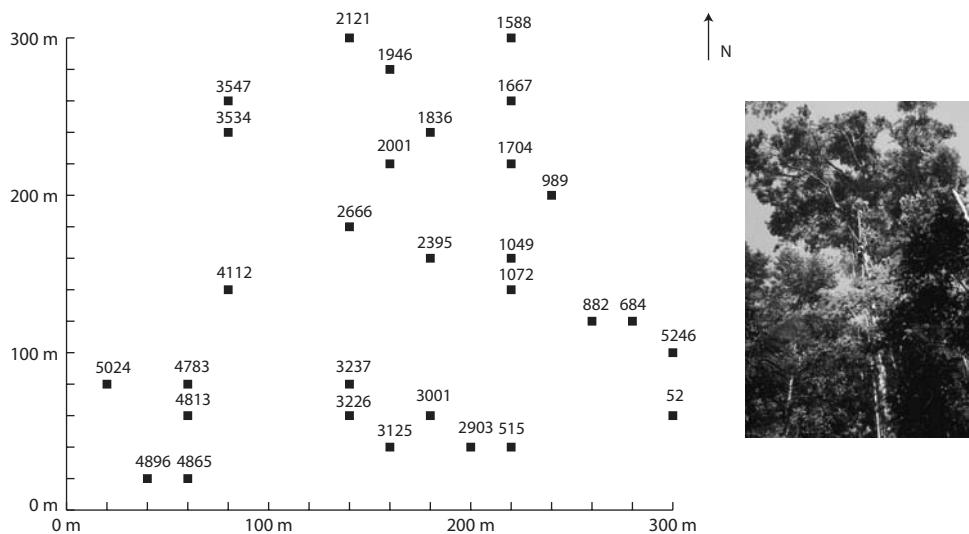
## 4.2 Direct measures of gene flow

- Genetic-marker-based parentage analysis.

This section of the chapter will introduce and explain the use of molecular genetic markers to identify the unknown parent or parents of a sample of progeny or juveniles and thereby describe the patterns of mating that took place among the parents. Parentage analyses are considered direct measures of gene flow since they reveal and measure the pattern of gamete movement at the scale over which the candidate parents are sampled. Parentage analyses are also commonly used to test hypotheses about what factors influence patterns of mating among individuals. For example, animal parentage studies can test for correlations between mating success and phenotypes or behaviors. Parentage analysis is most often performed in the case where one parent is known and the other parent is unknown and could potentially be any one of a number of individuals or **candidate parents**. Genetic analyses that attempt to identify unknown fathers or unknown mothers from a population of candidate parents are called **paternity analysis** or **maternity analysis**, respectively (see Meagher 1986; Dow & Ashley 1996; Devlin & Ellstrand 1990; reviewed by Jones & Ardren 2003). Although not detailed here, it is also possible to attempt to infer both unknown parents within a population of candidate parents to estimate the minimum number of parents that contributed to a group of progeny (see Jones & Arden 2003). This section will review some of the basic concepts required to understand the methods and results of parentage analyses by means of an example paternity analysis. One focus in particular will be the distinction between identifying the true parent of an offspring and identifying a candidate parent that appears to be the true parent due to chance.

To understand the steps carried out in parentage analysis, let's work through an example based on genotype data from the tropical tree *Corythophora alta*, a member of the Brazil nut family (Fig. 4.5). All

**Figure 4.4** (opposite) Moran's  $I$  for simulated populations like those in Fig. 4.3. To estimate Moran's  $I$ , the  $100 \times 100$  grid was simulated for 200 generations and was then divided into square subpopulations of  $10 \times 10$  individuals. The frequency of the A allele within each subpopulation is  $y_i$ , and the mean allele frequency over all subpopulations is  $\bar{y}$  in equation 4.1. The distance classes are the number of subpopulations that separate pairs of subpopulations. As expected, the simulations with strong isolation by distance ( $3 \times 3$  mating neighborhood) show correlated allele frequencies in subpopulations that are close together. However, the simulations with panmixia ( $99 \times 99$  mating neighborhood) show no such spatial correlation of allele frequency. The fluctuation of  $I$  at the largest distances classes in both figures is random variation due to very small numbers of individuals compared. Each line is based on an independent simulation of the  $100 \times 100$  population.



**Figure 4.5** An individual *Corythophora alta* tree found at the Biological Dynamics of Forest Fragments Project field sites north of Manaus, Brazil. The plot shows the relative locations of the individual trees that make up the candidate parent population within a 9 ha forest plot at Cabo Frio. All trees can be both maternal parents and candidate paternal parents since the trees are hermaphrodites capable of self-fertilization.

**Table 4.1** Microsatellite genotypes (given in base pairs) for some of the 30 mature individuals of the tropical tree *Corythophora alta* sampled from a 9 ha plot of continuous forest in the Brazilian Amazon. Progeny are seeds collected from known trees. Missing data are indicated by a 0.

Microsatellite locus . . .	Genotype									
	A	B	C	D	E					
<b>Candidate parents</b>										
684	333	339	97	106	169	177	275	305	135	135
989	330	336	97	106	165	181	275	275	135	153
1072	315	333	103	106	169	179	296	302	138	138
1588	318	327	106	106	165	167	272	293	135	150
1667	324	333	0	0	165	185	275	284	141	159
1704	318	327	103	106	0	0	284	296	144	147
1836	333	339	97	97	181	183	275	296	138	144
1946	327	333	91	106	167	187	284	287	135	147
2001	321	336	0	0	177	181	284	302	138	144
2121	318	333	100	106	179	181	284	302	144	144
2395	327	333	103	103	179	187	275	296	150	159
3001	324	333	91	106	167	183	284	302	147	159
3226	327	327	103	106	163	181	275	275	135	144
3237	324	324	91	103	179	187	284	305	144	159
3547	321	321	103	106	177	179	275	296	0	0
4112	327	327	97	106	169	181	296	302	144	144
4783	321	327	0	0	183	185	290	308	144	156
4813	327	333	106	106	177	179	284	302	135	138
4865	321	327	106	106	167	179	284	296	144	153
4896	315	333	100	106	181	189	275	284	162	162
5024	318	327	100	103	165	167	275	284	147	147
<b>Seed progeny</b>										
989 seed 1-1	327	336	91	106	165	185	275	287	153	153
989 seed 2-1	327	330	103	106	165	181	275	275	135	135
989 seed 3-1	330	336	97	106	165	181	0	0	135	153
989 seed 25-1	321	330	106	106	167	181	275	296	135	153

*C. alta* individuals 10 cm in diameter or greater at breast height were sampled from a 9 ha area inside of a large tract of continuous forest. These trees 10 cm in diameter or greater at breast height are the candidate parents. A sample of seeds was also collected from some of these trees. The genotypes of both the trees and the seeds were determined for 10 nuclear microsatellite loci (see Box 2.1 for an introduction to this type of genetic marker). A subset of these data is shown in Table 4.1. The goal of the parentage analysis in this case is to determine the fathers of the seeds given the known mothers in order to estimate the proportion of seeds that resulted from pollen transport within the sampled plot compared with pollen transport from outside the sampled plot.

The first step in a parentage analysis is to examine the genotypes of an individual progeny and its known parent for allelic matches. *C. alta* seed genotypes are grouped with their known parent in Table 4.2. For example, in Table 4.2 the genotype of seed 1-1 from tree 989 is given in the first row and the genotype of the known maternal parent tree (989) is given in the second row. At each locus, one (or sometimes both) of the alleles found in the known parent genotype is observed in the progeny genotype. For seed 1-1 from tree 989, the known parent contributed the 336 allele at locus A, the 106 allele at locus B, the 165 allele at locus C, the 275 allele at locus D, and the

153 allele at locus E. Given those alleles came from the known parent, the true father must have contributed alleles 327, 91, 185, 287, and 153 at loci A through E. This set of single alleles at each diploid locus is called the paternal **haplotype**. We can now scan the genotypes of the candidate parents to see whether there is any individual with a haplotype that contains all of those alleles (this is normally done with the assistance of a computer program). All candidate parents that have a matching haplotype are possible fathers of seed 1-1 from tree 989. In this case, tree 1946 is the only individual with the required haplotype and so 1946 is possibly the father while all of the other candidate parents are **excluded** as fathers due to a genetic mismatch at one or more loci in the paternal haplotype. Repeating this process for the next two seeds also excludes all but a single individual as the father.

With the exclusion of all but a single candidate parent, it would seem like certain identification of the true parent has been accomplished. Unfortunately, it is always possible that any non-excluded candidate parent is not the actual parent. There is the possibility, by chance alone, that an individual possesses a genotype with the same haplotype as the true parent. Evaluating the chance that a non-excluded candidate parent (sometimes called an **inclusion** or an **included parent**) is not the true parent requires

**Table 4.2** Seed progeny genotypes (top row of every three) given with the known maternal parent genotype (middle row of every three) along with the genotype of the most probable paternal parent (bottom row of every three) from the pool of all possible candidate parents. Alleles in the seed progeny that match those in the known maternal parent are underlined. The known maternal parent is also a candidate paternal parent since this species can self-fertilize. Missing data are indicated by zero.

Microsatellite locus . . .	Genotype									
	A	B	C	D	E					
989 seed 1-1	327	<u>336</u>	91	<u>106</u>	<u>165</u>	185	<u>275</u>	287	<u>153</u>	153
989	330	<u>336</u>	97	<u>106</u>	<u>165</u>	181	<u>275</u>	275	<u>135</u>	153
1946	327	333	91	106	167	185	284	287	135	147
989 seed 2-1	327	<u>330</u>	103	<u>106</u>	<u>165</u>	<u>181</u>	<u>275</u>	275	<u>135</u>	135
989	330	<u>336</u>	97	<u>106</u>	<u>165</u>	<u>181</u>	<u>275</u>	275	<u>135</u>	153
3226	327	327	103	106	163	181	275	275	135	144
989 seed 3-1	<u>330</u>	<u>336</u>	<u>97</u>	<u>106</u>	<u>165</u>	<u>181</u>	0	0	<u>135</u>	<u>153</u>
989	330	<u>336</u>	97	<u>106</u>	<u>165</u>	<u>181</u>	275	275	<u>135</u>	<u>153</u>
989	330	<u>336</u>	97	106	165	181	275	275	135	153
989 seed 25-1	321	<u>330</u>	<u>106</u>	106	167	<u>181</u>	<u>275</u>	296	<u>135</u>	<u>153</u>
989	330	<u>336</u>	97	106	165	<u>181</u>	<u>275</u>	275	<u>135</u>	<u>153</u>
4865	321	327	106	106	167	179	284	296	144	153

Table 4.3 Allele frequencies for five *Corythophora dita* microsatellite loci used for paternity analysis.

Microsatellite locus . . .	A		B		C		D		E	
	Allele	Frequency								
315	0.0405	91	0.0735	163	0.0217	272	0.0238	135	0.2917	
318	0.0541	97	0.3088	165	0.2283	275	0.4167	138	0.0625	
321	0.1216	100	0.0735	167	0.0761	281	0.0357	141	0.0313	
324	0.0541	103	0.1471	169	0.0435	284	0.1429	144	0.2188	
327	0.2703	106	0.3971	171	0.0217	287	0.0119	147	0.0625	
330	0.1892		177	0.0543	290	0.0119	150	0.0938		
333	0.1216		179	0.1304	293	0.0238	153	0.1250		
336	0.1216		181	0.2065	296	0.1905	156	0.0208		
339	0.0270		183	0.0652	299	0.0119	159	0.0521		
			185	0.0435	302	0.0833	162	0.0417		
			187	0.0326	305	0.0357				
			189	0.0109	308	0.0119				
			193	0.0109						
			197	0.0543						

one to determine the chance of such a random match. Let the frequency of an allele in the matching haplotype be  $p_i$  where  $i$  indicates the locus. At each locus the chance of matching at random is simply the probability that an individual is either homozygous ( $p_i^2$ ) or heterozygous ( $2p_i(1 - p_i)$ ) for the allele in question. Thus, the total probability of a random match for one locus is:

$$P(\text{random match}) = p_i^2 + 2p_i(1 - p_i) \quad (4.2)$$

under the assumptions of random mating and panmixia. Assuming that all of the loci used in a parentage analysis are independent, the probability of a random match for all loci in a given haplotype is the product of the locus by locus frequency of a random match, or:

$$\begin{aligned} & P(\text{multilocus random match}) \\ &= \prod_{i=1}^{\text{loci}} (p_i^2 + 2p_i(1 - p_i)) \end{aligned} \quad (4.3)$$

where  $\Pi$  indicates chain multiplication over all loci.

Returning to our *C. alta* example, we can calculate the chances of a random match for each of the paternal haplotypes. The haplotypes, allele frequencies (see Table 4.3), probability of a random match at each locus, and probability of a random match at all five loci are given in Table 4.4. Focus first on the haplotype for tree 1946. Given that allele 327 at

locus A has an observed frequency of 0.2703 in the population of candidate parents (which is an estimate of the allele frequency in the entire population), the chance of any genotype having one copy of this allele is  $(0.2703)^2 + 2(0.2703)(1 - 0.2703) = 0.4675$ . We therefore expect 46.75% of individuals in the population to have a genotype with either one or two copies of the 327 allele. This is the same as the probability that an individual taken at random from the population (and not necessarily included in the sample of candidate parents) could provide the correct haplotype to be included as a possible father of seed 989 1–1 in Table 4.2. The chances of a random match at each of the five loci is calculated in the same fashion. We see that a genotype that would compliment the known parent's haplotype and explain the seed genotype is expected to occur between about 2 and 47% of the time for any single locus. When these probabilities are combined across all five loci the expected frequency of a random match becomes very small. As shown in Table 4.4, the expected frequency of a random match at all five loci is between 44 in 1000 and 66 in 1,000,000 genotypes under the assumption of random mating. This is a demonstration of the general principle that the ability to distinguish true parentage from apparent parentage due to random matches depends on both the allele frequencies at each locus as well as the total number of loci available. Random matches become less likely as allele frequencies decrease and as the number of independent loci increases.

**Table 4.4** The chance of a random match for the included fathers in Table 4.2. The probability of a random match at each locus is  $p_i^2 + 2p_i(1 - p_i)$ . The combined probability of a random match for all loci in the haplotype is the product of the probabilities of a random match at each independent locus. Paternal haplotype data are treated as missing (0) for the purposes of probability calculations when progeny genotype data are missing. In the cases where the paternal haplotype has multiple possible alleles at some loci, the highest probability of a chance match is given. The allele frequencies for each locus are given in Table 4.3.

Included father	Microsatellite haplotype					$P(\text{multilocus random match})$
	A	B	C	D	E	
1946 (seed 1-1)	327	91	185	287	135	
Allele frequencies	0.2703	0.0735	0.0435	0.0119	0.2917	
$P(\text{random match})$	0.4675	0.1416	0.0851	0.0237	0.4983	0.0000665
3226 (seed 2-1)	327	103	106	181	275	135
Allele frequencies	0.2703	0.0735	0.3971	0.2065	0.4167	0.2917
$P(\text{random match})$	0.4675	0.1416	0.6365	0.3704	0.6598	0.4983
						≤0.03624
989 (seed 3-1)	330	336	97	106	165	181
Allele frequencies	0.1892	0.1216	0.3088	0.3971	0.2283	0.2065
$P(\text{random match})$	0.3426	0.2284	0.5222	0.6365	0.4045	0.3704
						≤0.0440

**Candidate parent** An individual in the pool of possible parents that shares one or both alleles found in an offspring genotype at all loci.

**Cryptic gene flow** Gene-flow events incorrectly assigned to candidate parents but actually due to unobserved parents outside the area where candidate parents were sampled, leading to an underestimate of gene-flow distances.

**Exclusion** Rejection of an individual as a possible parent due to genetic mismatch (neither allele in the individual's genotype is identical to one of the alleles in the progeny genotype).

**Exclusion probability** The chance that an individual can be rejected as a candidate parent due to genetic mismatch; depends on allele frequencies and increases with the number of loci and the numbers of alleles per locus employed in a parentage analysis.

We can express the probability that an individual taken at random from a population would be ruled out as a parent due to genetic mismatch. Equation 4.2 gives the probability of a random match at a single locus, or the probability that a genotype has a matching allele by chance alone. If a genotype does *not* match by chance then it is excluded from possibly being the parent. This means that the **exclusion probability** for a single individual sampled at random from a population is just one minus the probability of a random match:

$$P(\text{exclusion}) = 1 - P(\text{random match}) \quad (4.4)$$

If more than one candidate parent is sampled from a population, the probability of exclusion for each individual is independent (the genotype of each individual represents a random sampling of the alleles present in the population). Therefore, the total probability of ruling out or excluding all candidate parents is the product of the exclusion probabilities for each individual. For a sample of  $n$  individuals from a population, the total probability of exclusion is then:

$$\begin{aligned} P(\text{exclusion for } n \text{ individuals}) \\ = (1 - P(\text{random match}))^n \end{aligned} \quad (4.5)$$

This means that the chance of exclusion decreases as more individuals are sampled from a population. This is the same as saying that the chances of samp-

ling an individual that matches a parental haplotype just by chance increases as more candidate parents are sampled.

Based on the exclusion probability in a population of  $n$  candidate parents, we can estimate the chances that a random match *does* occur. Since the exclusion probability is the chance of *not* matching at random, the probability of a haplotype match between a candidate parent and an offspring in a population of  $n$  individuals is just one minus the probability of exclusion for  $n$  individuals, or

$$\begin{aligned} P(\text{random match in } n \text{ individuals}) \\ = 1 - P(\text{exclusion for } n \text{ individuals}) \quad (4.6) \\ = 1 - (1 - P(\text{random match}))^n \end{aligned}$$

This is the probability that a haplotype matching the true parent will occur at random in a sample of  $n$  candidate parents.

The probability of a random match in a sample of  $n$  candidate parents (equation 4.6) can be thought of as the chance that a candidate parent is mistakenly assigned as the true parent since its genotype provides the matching haplotype by chance, while the true parent is not identified since it is not included in the sample of candidate parents. This phenomenon is referred to as **cryptic gene flow** in paternity analysis since the true gene flow event is not identified, even though a parent has been mistakenly inferred for the progeny. If the true parent was not included in the sample of candidate parents because it was outside the sampling area, incorrectly inferred parentage results in an underestimate of gene flow distances. Equation 4.6 shows that the probability of incorrectly assigning parentage due to random matches increases as the number of candidate parents increases for a given expected genotype frequency.

Returning to the *C. alta* example in Table 4.2, we can determine the chances that one of the candidate parents is incorrectly inferred to be a father while the true father remains undetected as well as the chances of paternity exclusion with the 30 candidate parents in the study. For seed 3-1 the maternal and paternal parents are the same (Table 4.4), indicating a self-fertilization event. Based on the paternal parent haplotype expected frequency, the chance of paternity exclusion is  $(1 - 0.044)^{30} = 0.259$  and the probability of a random match is therefore 0.741. Since the four-locus inferred paternal haplotype is expected to occur very frequently (74% of the time) by chance in a sample of 30 candidate parents, there is also a good chance that the seed could appear

self-fertilized even though it was actually sired by an individual not in the sample of candidate parents. For seed 989 1-1 where tree 1946 was the only included candidate parent, the chance of paternity exclusion is  $(1 - 0.0000665)^{30} = 0.9980$  and the probability of a random match is therefore 0.0020. The five-locus inferred paternal haplotype for seed 989 1-1 is expected to appear by chance in only two of 1000 samples of 30 candidate parents given the estimated allele frequencies.

There are four general outcomes for each offspring-known-parent pair in parentage analysis, as follows.

- 1 A single candidate parent is identified as the parent. Such single parentage assignments need to be interpreted in light of the exclusion probability or likelihood of parentage.

### Problem box 4.1 Calculate the probability of a random haplotype match and the exclusion probability

The seed 25-1 from maternal tree 989 shows exact haplotype matches with candidate paternal tree 4865 (see Table 4.2). Using the allele frequencies provided in Table 4.3, calculate the probability of a random match for the paternal haplotype. Then use this probability of a random match to calculate the exclusion probability for the sample of 30 candidate parents. What loci are most and least useful in determining paternity for these two seed progeny? Why?

### Interact box 4.1 Average exclusion probability for a locus

In planning a parentage analysis study, it is necessary to determine whether a set of genetic markers will have a sufficiently small probability of exclusion (this is called the power of the genetic markers). As shown in equation 4.4, the exclusion probability will depend on the expected genotype frequency for a single parental haplotype. This expected genotype frequency is in turn a function of the number of alleles and the allele frequencies at each locus. Since there are many possible genotypes for a locus with three or more alleles, the *average* probability of exclusion is used to estimate the power of a set of genetic markers to demonstrate nonpaternity (see Chakraborty et al. 1988; Weir 1996).

You can use an Excel spreadsheet that has been set up on the textbook website to calculate the average probability of exclusion (abbreviated as  $P_E$  in the spreadsheet) for a case of one locus with six alleles and one locus with 12 alleles. The spreadsheet uses the allele frequencies (that you can modify) to calculate (i) the expected frequencies of each maternal-parent-offspring genotype combination and (ii) the exclusion probabilities for the paternal haplotype(s) for each maternal parent-offspring genotype combination. The average exclusion probability is then the average of the exclusion probabilities where each exclusion probability is weighted by the expected frequency of the maternal-parent-offspring genotype combination. The spreadsheet follows the derivation for a locus with three alleles given in Table 1 of Chakraborty et al. (1988). The maximum average exclusion probability occurs when all alleles at a locus have identical allele frequencies (e.g. each allele has a frequency of  $1/6$  when there are six alleles). The maximum average exclusion probability is computed in each spreadsheet according to:

$$\text{Max.prob.exclusion} = \frac{(k-1)(k^3 - k^2 - 2k + 3)}{k^4} \quad (4.7)$$

where  $k$  is the number of alleles at the locus (Selvin 1980).

Compare the average probability of exclusion for cases where the frequencies of each allele are very similar to cases where one or a few alleles are very common and the remaining alleles are rare. How does the evenness of the frequencies for the alleles influence the average exclusion probability? How do you combine the average exclusion probabilities for multiple loci? What is the average exclusion probability of two loci with 12 alleles or two loci with six alleles when allele frequencies are all equal for each locus? How many independent loci with 12 equally frequent alleles would be required for a probability of exclusion of 90% when there are 50 candidate parents?

- 2 Multiple candidate parents are identified for a single progeny. In these cases one commonly used criterion is to assign as parent the candidate parent with the lowest probability of matching by chance. Additional criteria might also include spatial separation from the known parent, degree of reproductive overlap with the known parent, or reproductive dominance, if such information is available.
- 3 None of the candidate parents have a genotype that could have combined with the known parent to yield the progeny genotype. In this case the actual parent may not be present in the sample of candidate parents. Such an outcome is often used to infer that the gene-flow event leading to that progeny was from a relatively long distance from a parent outside the sample area of candidate parents (so-called off-plot gene flow). However, it is also possible that the actual parent is in the population of candidate parents but has a genetic mismatch at one or more loci due to a genotyping error or mutation. An additional alternative is that the actual parent was inside the sampling area of candidate parents when mating took place, but the individual either died or migrated before sampling of the candidate parents was carried out.
- 4 Parentage is assigned to a candidate parent but the true parent is an individual not included in the sample of possible parents. When making paternity assignments, the chance of incorrectly assigning paternity within a group of sampled individuals when the father is actually outside the population, or missing a “cryptic gene flow” event, will be related to the expected frequency of a given multilocus genotype.

Parentage analyses measure gene flow by inferring numerous mating events within the population of candidate parents that lead to each sampled progeny or juvenile in a population. This provides estimates of quantities such as the average distance between parents or the number of matings where both parents were within a sample area compared to the number of matings where a parent was outside that area. This means that resulting estimates of gene flow do not rely on any model of population structure or gene flow other than the assumptions that are used to construct the parentage assignments themselves. The resulting estimates of gene flow are therefore considered “direct.” The clear strength of parentage analyses is that much can be learned about

patterns of mating since parental pairings that lead to a specific offspring can often be identified with medium to high confidence.

Parentage analyses have been a critically important tool used to learn about mating and relatedness patterns in wild populations. An example is the numerous studies of parentage among bird nestlings that overturned the long-held idea that birds were usually monogamous breeders. Instead, birds have variable and complex mating patterns where mating outside of nesting pairs by both females and males can be common and juveniles in the nest may not be related to one or both of the nest-attendant “parents” (Westneat & Stewart 2003). Parentage analyses have also been used in a wide variety of plant and animal species to produce detailed descriptions of mating and gene-flow patterns.

Although the term direct has connotations of precision and ready insight, it is important to recognize that parentage analyses do have limitations when used to infer patterns of gene flow. A major limitation stems from the fact that most parentage studies cover a time scale of only a few generations at most. In all organisms with population sizes that are stable through time, each parent produces just one offspring on average that survives to reproduce successfully. The other progeny die or do not reproduce. This means that many, perhaps even most, of the progeny included in parentage studies ultimately do not reproduce. This problem is particularly acute in long-lived organisms, where parentage studies examine only a very small fraction of progeny produced over a period much less than the average individual lifetime. Gene flow can be thought of as the long-term average of the matings that lead to individuals that survive and contribute progeny to the next generation. How effective parentage studies are at estimating longer-term patterns of gene flow then depends on the sampling duration of parentage studies relative to generation time and how variable parentage patterns are over the short term compared to their long-term averages.

#### 4.3 Fixation indices to measure the pattern of population subdivision

- Extending the fixation index to measure the pattern of population structure through  $F_{IS}$ ,  $F_{ST}$ , and  $F_{IT}$ .

The first section of the chapter reviewed the processes that contribute to the formation of allele frequency

**Table 4.5** The mathematical and biological definitions of heterozygosity for three levels of population organization. In the summations,  $i$  refers to each subpopulation 1, 2, 3 . . .  $n$  and  $p_i$  and  $q_i$  are the frequencies of the two alleles at a diallelic locus in subpopulation  $i$ .

$H_I = \frac{1}{n} \sum_{i=1}^n \hat{H}_i$	The average observed heterozygosity within each subpopulation.
$H_S = \frac{1}{n} \sum_{i=1}^n 2p_i q_i$	The average expected heterozygosity of subpopulations assuming random mating within each subpopulation.
$H_T = 2\bar{p}\bar{q}$	The expected heterozygosity of the total population assuming random mating within subpopulations and no divergence of allele frequencies among subpopulations.

differences among populations. Given that these processes might be acting, it is necessary to develop methods to measure and quantify population structure. The parentage analyses such as those described in the last section can be carried out when genotype data are available both for a sample of candidate parents and a sample of progeny. An alternative situation is where genotype data are determined for individuals sampled within and among a series of geographic locations. This type of sampling is very commonly carried out in empirical studies and requires methods to quantify the pattern of population structure present among the subpopulations as well as the genotype frequencies found within subpopulations. It would be advantageous if such measures could be readily compared to reference situations, as is expected with no population structure. This was our approach when comparing observed and expected heterozygosity using the fixation index ( $F$ ) in Chapter 2. We can now extend the fixation index to apply to cases where there are multiple subpopulations. In this more complex case there can be deviations from Hardy–Weinberg expected frequencies of heterozygotes at two levels: within each subpopulation due to non-random mating and among subpopulations due to population structure. This section of the chapter will develop and explain fixation-index-based measures of departure from expected heterozygosity commonly used to quantify population structure.

Let's look in detail at the case where we can measure the genotypes at a diallelic locus for a sample of individuals located in several different subpopulations. Recall that the heterozygosity in a population is just one minus the homozygosity ( $H = 1 - F$ ) so the heterozygosity can be related to the fixation index.

With such genotype data it is possible to compute the observed and expected frequencies of the heterozygote genotype in several ways (Table 4.5). The first way is to simply take the average:

$$H_I = \frac{1}{n} \sum_{i=1}^n \hat{H}_i \quad (4.8)$$

where  $\hat{H}$  is the *observed* frequency of heterozygotes in each of the  $n$  subpopulations. We could call this  $\hat{H}$  since it is the average of the observed heterozygote frequencies in all subpopulations. This is just the probability that a given individual is heterozygous or the average observed heterozygosity. As shown in Chapter 2, the heterozygosity within populations can be increased or decreased relative to Hardy–Weinberg expectations by non-random mating.

Next, we can determine the expected heterozygosity *assuming* each subpopulation is in Hardy–Weinberg equilibrium. This assumption means that the frequency of the heterozygous genotype is expected to be  $2pq$  for a locus with two alleles. The average expected heterozygosity of subpopulations is then

$$H_S = \frac{1}{n} \sum_{i=1}^n 2p_i q_i \quad (4.9)$$

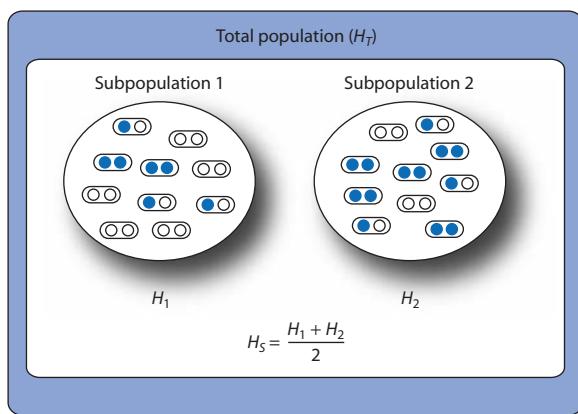
where  $p_i$  and  $q_i$  are the allele frequencies in subpopulation  $i$  and there are  $n$  subpopulations. We could use the notation  $\overline{2pq}$  since the expected heterozygosity is determined for each subpopulation and then averaged. Here the observed allele frequency is used to estimate the Hardy–Weinberg expected heterozygosity for each subpopulation.

At the most inclusive level in a subdivided population, we can calculate the expected heterozygosity of the total population:

$$H_T = 2\bar{p}\bar{q} \quad (4.10)$$

where  $\bar{p}$  and  $\bar{q}$  are average allele frequencies for all the subpopulations. The average allele frequency for all subpopulations is equivalent to combining all alleles for all subpopulations into a single population and then simply estimating allele frequencies. In other words, it is the allele frequency for the total population without any divergence among subpopulations taken into account.  $H_T$  is therefore the Hardy–Weinberg expected frequency of heterozygotes in the entire population if there were no population structure of allele frequencies.

These different levels of observed and expected heterozygosity are diagrammed in Fig. 4.6 for the case of a total population composed of two subpopulations that each contain 10 diploid individuals. In both subpopulations, three of the 10 individuals are heterozygotes giving observed heterozygote



**Figure 4.6** Illustration of the hierarchical nature of heterozygosity in a subdivided population.  $H_1 = \frac{3}{10}$  and  $H_2 = \frac{3}{10}$  to give an average observed heterozygosity of  $H_S = \frac{1}{2}\left(\frac{3}{10} + \frac{3}{10}\right) = 0.30$ . If  $p$  is the frequency of one allele (open circles) and  $q$  the frequency of the alternate allele (filled circles), then  $p_1 = 13/20 = 0.65$  and  $q_1 = 1 - p_1 = 0.35$  while  $p_2 = 7/20 = 0.35$  and  $q_2 = 1 - p_2 = 0.65$ . The average expected heterozygosity in the two subpopulations is then  $H_S = \frac{1}{2}[2(0.65)(0.35) + 2(0.35)(0.65)] = 0.455$ . In the total population the average allele frequencies are  $\bar{p} = \frac{1}{2}(0.65 + 0.35) = 0.50$  and  $\bar{q} = \frac{1}{2}(0.35 + 0.65) = 0.50$  giving an expected heterozygosity in the total population of  $H_T = 2\bar{p}\bar{q} = 2(0.5)(0.5) = 0.5$ .

frequencies of  $H_1 = \frac{3}{10}$  and  $H_2 = \frac{3}{10}$ . Together these yield an average observed heterozygosity of

$$H_S = \frac{1}{2}\left(\frac{3}{10} + \frac{3}{10}\right) = 0.30. \text{ To determine the aver-}$$

age expected heterozygosity of the subpopulations requires observed allele frequencies for each subpopulation. In subpopulation one 13 of the 20 alleles are white and seven of the 20 alleles are blue. If  $p$  is the frequency of the white allele and  $q$  the frequency of the blue allele, then  $p_1 = 13/20 = 0.65$  and  $q_1 = 1 - p_1 = 0.35$ . In subpopulation two the situation is the exact opposite with  $p_2 = 7/20 = 0.35$  and  $q_2 = 1 - p_2 = 0.65$ . The average expected heterozygosity in the two subpopulations is then

$$H_S = \frac{1}{2}[2(0.65)(0.35) + 2(0.35)(0.65)] = 0.455. \text{ In the total population average allele frequencies are } \bar{p} = \frac{1}{2}(0.65 + 0.35) = 0.50 \text{ and } \bar{q} = \frac{1}{2}(0.35 + 0.65) = 0.50. \text{ (Notice that obtaining the average of the subpopulation allele frequencies is equivalent to combining all the alleles in the total population and then estimating the allele frequency, as in } \bar{p} = \frac{13+7}{40} = 0.50.) \text{ The expected heterozygosity of the total population is then } H_T = 2(0.5)(0.5) = 0.5.$$

After calculating the different observed and expected heterozygosities in Fig. 4.6, it is apparent that they are not all equivalent. There are differences between the observed and expected heterozygosities at the different hierarchical levels of the population. Recall from section 2.5 that the difference between observed and Hardy–Weinberg expected genotype frequencies was used to estimate the fixation index or  $F$ . In that case there was only a single population and we were only concerned with how alleles combined into diploid genotypes compared with the expectation under random mating. The fixation index can be extended to accommodate multiple levels of population organization, thereby creating measures of deviation from Hardy–Weinberg expected genotype frequencies caused by two distinct processes. With multiple subpopulations there is a possible excess or deficit of heterozygotes due to non-random mating within subpopulations and a possible deficit of heterozygotes among subpopulations compared to panmixia. In the latter case the fixation index will show how much allele frequencies have diverged among subpopulations due to processes that cause population structure compared with the ideal of

**Table 4.6** The mathematical and biological definitions of fixation indices for two levels of population organization.

$F_{IS} = \frac{H_S - \bar{H}_I}{H_S}$	The average difference between observed and Hardy–Weinberg expected heterozygosity within each subpopulation due to non-random mating. The correlation between the states of two alleles in a genotype sampled at random from any subpopulation.
$F_{ST} = \frac{H_T - H_S}{H_T}$	The reduction in heterozygosity due to subpopulation divergence in allele frequency. The difference between the average expected heterozygosity of subpopulations and the expected heterozygosity of the total population. Alternately, the probability that two alleles sampled at random from a single subpopulation are identical given the probability that two alleles sampled from the total population are identical.
$F_{IT} = \frac{H_T - H_I}{H_T}$	The correlation between the states of two alleles in a genotype sampled at random from a single subpopulation given the possibility of non-random mating within populations and allele frequency divergence among populations.

uniform allele frequencies among subpopulations expected with panmixia.

Accounting for non-random mating and divergence of subpopulation allele frequency necessitates several new versions of the fixation index. The definitions of these new fixation indices are shown in Table 4.6. Let's employ and interpret each of these versions of the fixation index for the example in Fig. 4.6.  $F_{IS}$  compares average observed heterozygosity of individuals in each subpopulation and the average Hardy–Weinberg expected heterozygosity for all subpopulations (the  $I$  stands for individuals and the  $S$  for subpopulations).  $F_{IS}$  is identical to the single-population  $F$  used in section 2.5 except that it is now an average for all subpopulations. Using the heterozygosities determined above,

$$F_{IS} = \frac{0.455 - 0.30}{0.455} = 0.341 \quad (4.11)$$

This is a result that makes biological sense since there are fewer heterozygotes in each subpopulation than would be expected under random mating given the subpopulation allele frequencies. Thus, there is more homozygosity or fixation within the two subpopulations than expected under random mating. The subpopulations on average have a deficit of heterozygosity as expected if there is consanguineous mating taking place.

The next level in the hierarchy is the average expected heterozygosity for subpopulations compared with expected heterozygosity for the total population or  $F_{ST}$  (the  $S$  stands for subpopulations and the  $T$  for

the total population). Based on the heterozygosities determined previously,

$$F_{ST} = \frac{0.50 - 0.455}{0.50} = 0.09 \quad (4.12)$$

This result says that there is somewhat less heterozygosity on average for the two subpopulations compared with the heterozygosity expected in the ideal case where the entire population is panmictic. This is consistent with the fact that the two subpopulations have slightly different allele frequencies and each has an expected heterozygosity of slightly less than  $\frac{1}{2}$ . However, the total population would have a heterozygosity of  $\frac{1}{2}$  (the maximum) if there was no allele frequency divergence between the two subpopulations.

The final level in the hierarchy is  $F_{IT}$ , the comparison of the average observed heterozygosity for subpopulations with the heterozygosity expected for the total population:

$$F_{IT} = \frac{0.50 - 0.30}{0.50} = 0.40 \quad (4.13)$$

This gives the combined departure from Hardy–Weinberg expected genotype frequencies due to the combination of non-random mating within subpopulations and divergence of allele frequencies among subpopulations. For this example, homozygosity is 40% greater or heterozygosity 60% less than would be expected in an ideal, randomly mating panmictic population with the same allele frequencies.

### Problem box 4.2

#### Compute $F_{IS}$ , $F_{ST}$ , and $F_{IT}$

Levin (1978) used allozyme electrophoresis to estimate genotype frequencies for the phosphoglucomutase-2 gene (*Pgm-2*) in *Phlox cuspidata*, a plant capable of self-fertilization. Genetic data were collected from 43 populations across the species range in southeast Texas. Using starch gel electrophoresis, the frequencies of two alleles (fast and slow running) and the frequencies of the heterozygous genotype were recorded for each population. A portion of the data is given in the table below (population numbers match Table 2 in Levin (1978)).

	Subpopulation			
	1	9	43	68
Frequency of <i>Pgm-2</i> fast	0.0	0.93	0.17	0.51
Frequency of <i>Pgm-2</i> slow	1.0	0.07	0.83	0.49
Heterozygote frequency	0.0	0.14	0.34	0.40

Using the heterozygote and allele frequencies, compute the hierarchical heterozygosities  $H_I$ ,  $H_S$ , and  $H_T$  and use these to calculate  $F_{IS}$ ,  $F_{ST}$ , and  $F_{IT}$ . Is there evidence that *P. cuspidata* individuals engage in selfing? Are the populations panmictic or subdivided?

The individual, subpopulation, and total population heterozygosities are identical in populations after compensating for the degree to which observed and expected heterozygosities are not met at different levels of population organization. The average observed heterozygosity is greater or less than the average expected heterozygosity for subpopulations:

$$H_I = H_S(1 - F_{IS}) \quad (4.14)$$

to the extent that there is non-random mating ( $F_{IS} \neq 0$ ). Similarly, the average expected heterozygosity for subpopulations is less than the expected heterozygosity of the total population under panmixia:

$$H_S = H_T(1 - F_{ST}) \quad (4.15)$$

to the extent that subpopulations have diverged allele frequencies ( $F_{ST} > 0$ ). The total deviation from expected heterozygosity within and among subpopulations is then

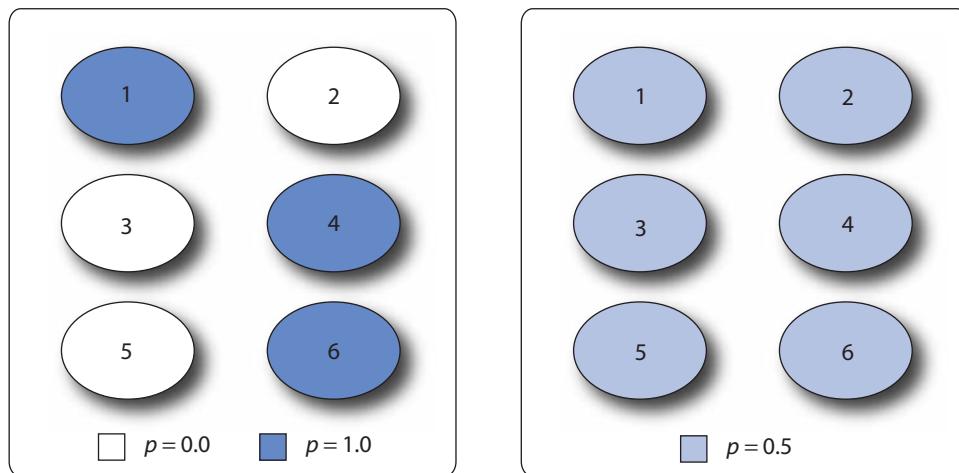
$$H_I = H_T(1 - F_{IT}) \quad (4.16)$$

Although equations 4.14–4.16 can be considered as rearrangements of equations 4.11–4.13, they also represent a different way to articulate and think of the biological impacts of allele frequency divergence among subpopulations and non-random mating within subpopulations. Each fixation index expresses the degree to which random mating expectations for the frequency of heterozygous genotypes are not met. Using these equations it is also possible to show how the total reduction in heterozygosity relates to the combined fixation due to non-random mating and subpopulation divergence:

$$1 - F_{IT} = (1 - F_{ST})(1 - F_{IS}) \quad (4.17)$$

Since using the fixation index to measure allele frequency divergence among subpopulations is the novel concept in this section, let's consider an additional example that focuses exclusively on  $F_{ST}$ . Figure 4.7 shows allele frequencies for a diallelic locus in two populations that are both composed of six subpopulations. The pattern of allele frequencies among the subpopulations is very different. On the right all subpopulations have the same allele frequencies, while on the left each subpopulation is at either complete fixation or complete loss for one allele. In both sets of populations  $H_T = 2(0.5)(0.5) = 0.5$ . The only difference between the two sets of populations is how allele frequencies are organized, or  $H_S$ . In the right-hand population, all six subpopulations have allele frequencies of  $\frac{1}{2}$ , giving  $H_S = (6(2)(0.5)(0.5))/6 = 0.5$ . In the left population, three subpopulations have an allele frequency of zero and three subpopulations have an allele frequency of one. This pattern gives  $H_S = (3(2)(1.0)(0) + 3(2)(0)(1.0))/6 = 0.0$ . Using these expected heterozygosities for the subpopulations and total population gives  $F_{ST} = 0.0$  on the right and  $F_{ST} = 1.0$  on the left.

The average allele frequency in the total population is the same in both cases. However, there is a major difference in the way that allele frequencies are organized. On the right all the subpopulations have identical allele frequencies, as would be expected if the subpopulations were really not subdivided at all. On the left the subpopulations are highly diverged in



**Figure 4.7** Allele frequencies at a diallelic locus for populations that consist of six subpopulations. Allele frequencies within subpopulations are indicated by shading. On the left, individual subpopulations are either fixed or lost for one allele. On the right, all subpopulations have identical allele frequencies of  $p = q = 0.5$ . In both cases, the total population has an average allele frequency of  $\bar{p} = 0.5$  and an expected heterozygosity of  $H_T = 2\bar{p}\bar{q} = 0.5$ . In contrast, the average expected heterozygosity for subpopulations is  $H_S = \overline{2pq} = 0.5$  on the right and  $H_S = \overline{2pq} = 0.0$  on the left.  $F_{ST} = 1.0$  on the left since the subpopulations have maximally diverged allele frequencies.  $F_{ST} = 0.0$  on the right since the subpopulations all have identical allele frequencies. Divergence of allele frequencies among subpopulations produces a deficit of heterozygosity relative to the Hardy–Weinberg expectation based on average allele frequencies for the total population.

allele frequencies as expected with strong population subdivision. Therefore, the different values of  $F_{ST}$  reflect the different degrees of allele frequency divergence among the sets of subpopulations. When all of the subpopulations are well mixed and have similar allele frequencies,  $H_S$  and  $H_T$  are identical. Biologically, an  $F_{ST}$  value of zero says that all subpopulations have alleles at the same frequencies as the total population and any single subpopulation has as many heterozygotes as any other subpopulation. As the populations diverge in allele frequency due to whatever process,  $H_S$  will decrease and  $F_{ST}$  will approach one. Biologically, an  $F_{ST}$  value of one says that the genetic variation is partitioned completely as allele-frequency differences among the subpopulations with an absence of segregating alleles within subpopulations.

An alternative way to think about the pattern of population differentiation in allele frequency is using the variance in allele frequency relative to the amount of genetic variation in the total population to estimate  $F_{ST}$ . The estimate of allele-frequency differentiation among subpopulations is then

$$F_{ST} = \frac{\text{var}(p)}{\bar{p}\bar{q}} \quad (4.18)$$

where the variance in allele frequency among  $n$  subpopulations is  $\text{var}(p) = \frac{1}{n} \sum_{i=1}^n (p_i - \bar{p})^2$  and there

are a very large number of subpopulations (Wright 1943a). If there is more variance in allele frequency then subpopulations differ more in allele frequency and the resulting  $F_{ST}$  is larger. For example, in Fig. 4.7 the average allele frequency or  $\bar{p}$  in both sets of six subpopulations is 0.5. On the left the variance in  $p$  is  $\frac{3((0 - 0.5)^2) + 3((1 - 0.5)^2)}{6} = 0.25$  while on the right the variance in  $p$  is  $\frac{6((0.5 - 0.5)^2)}{6} = 0.0$ . This leads to  $F_{ST} = 1.0$  on the left where there is maximum variance in allele frequencies among the subpopulations given the allele frequencies, and  $F_{ST} = 0.0$  on the right where there is no variance in allele frequencies among the subpopulations. Measuring the variance in allele frequencies among subpopulations is the basis of several widely employed methods used to estimate  $\hat{F}_{ST}$  from actual genetic marker data (see Method box 4.2).

The relationship between the allele frequencies of subpopulations, the hierarchical measures of heterozygosity, and the fixation indices can be seen in a simulation of gene flow and genetic drift in a subdivided population (Fig. 4.8). When gene flow is relatively strong and maintains similar allele frequencies in the subpopulations, the expected heterozygosity of the subpopulations and the total population are also similar and result in low values of the fixation indices (Fig. 4.8a). When gene flow is weaker and the

### Method box 4.2 Estimating fixation indices

Throughout this section a single locus with two alleles has been used to demonstrate hierarchical heterozygosities and fixation indices. These conditions are highly idealized, meaning that this section is really presenting the conceptual derivation of a parameter. In practice, there are numerous details involved in obtaining fixation-index parameter estimates  $\hat{F}_{IS}$ ,  $\hat{F}_{ST}$ , and  $\hat{F}_{IT}$ . Loci commonly have more than two alleles, so each allele provides a different estimate of the fixation index that is averaged within a locus. Multiple loci are also used, requiring computation of a multilocus average. It is also common that sample sizes are variable among subpopulations and that some genotype data may be missing at some loci for some individuals. Additionally, different types of genetic markers may require adjustments for dominance (inability to distinguish dominant homozygotes and heterozygotes) or patterns and rates of mutation that can be incorporated into fixation-index estimators. Finally, empirical studies may have more than three levels of hierarchy, may wish to test for variation in allele frequency associated with

subpopulations as well as other variables, and require statistical estimates of uncertainty in parameter estimates such as confidence intervals.

$G_{ST}$  is an estimator based on multilocus versions of the observed and expected heterozygosities or gene diversities (Nei 1973).  $\theta$  or  $\theta_{ST}$  (pronounced "theta"; Weir 1996) and  $\Phi_{ST}$  (Excoffier et al. 1992) are estimators based on analysis of variance of allele frequencies within and among subpopulations. The estimator  $\rho_{ST}$  (pronounced "roe") or  $R_{ST}$  is frequently used with microsatellite or simple sequence repeat (SSR) loci to account for high rates of stepwise mutation that can obscure population structure (Slatkin 1995; see Chapter 5). An estimator of  $F_{ST}$  is also available for DNA sequence data based on mean number of pairwise differences between sequences taken either from the same subpopulation or from different subpopulations (Hudson et al. 1992). Many commonly employed estimators are implemented in computer programs or software applications that can be obtained over the internet.

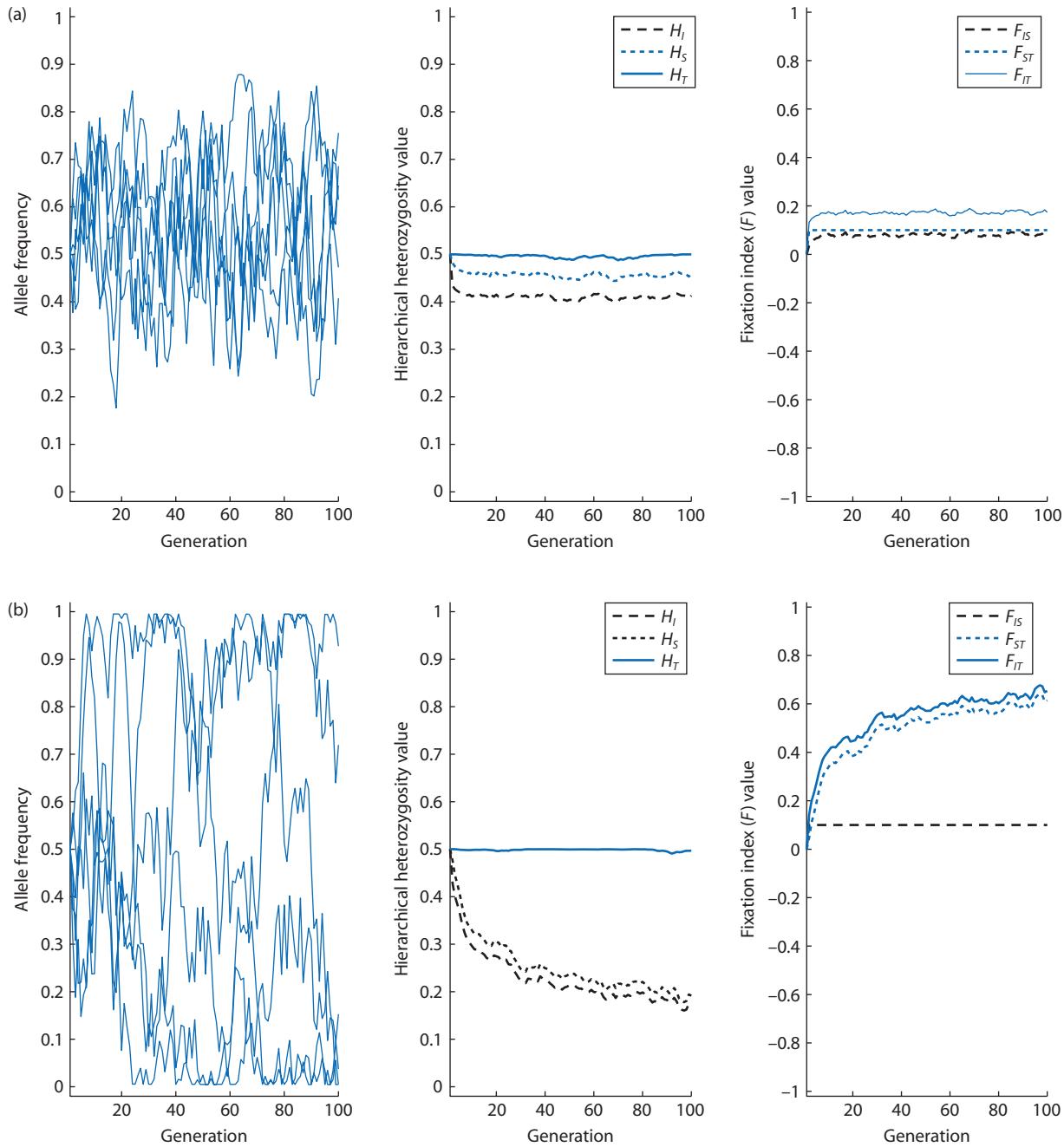
subpopulations diverge in allele frequency, the expected heterozygosity of the subpopulations is less than the expected heterozygosity of the total population, resulting in high values of the fixation indices (Fig. 4.8b). An additional point is that  $F_{ST}$  can vary considerably among independent replicate loci sampled in an identical fashion from the same subpopulations. Figure 4.9 shows the range of  $F_{ST}$  values obtained for 1000 independent loci in a simulation of the finite island model (see section 4.5) where genetic drift and gene flow among subpopulations were the processes acting to change allele frequencies. The range of  $F_{ST}$  values for individual loci under the influence of identical population genetic processes is due to the random nature of genetic drift. Each locus has experienced random fluctuations in allele frequencies that has resulted in a range of allele frequency variance among subpopulations. This random variation in  $F_{ST}$  due to genetic drift seen in the simulation underscores the need for estimates of  $F_{ST}$  to be obtained from the average of multiple loci.

### 4.4 Population subdivision and the Wahlund effect

- Genetic variation can be present as heterozygosity within a panmictic population or as differences in allele frequency among diverged subpopulations.

The last section of the chapter showed how departures from Hardy–Weinberg expected frequencies of heterozygotes can be used to quantify departures from random mating within demes and allele-frequency divergence among demes. This section will further explore heterozygosity within and among several demes, with two main goals. The first is to explore the consequences of population subdivision on expected genotype frequencies. The second is to show why  $F_{ST}$  functions to estimate allele frequency divergence among demes.

Consider the case of a diallelic locus for two randomly mating demes. The expected heterozygosity for each deme is:



**Figure 4.8** Allele frequencies, hierachal heterozygosities, and fixation indices from a simulation of a subdivided population. Each subpopulation contained 10 individuals. The rate of gene flow was  $m = 0.2$  in (a) and  $m = 0.01$  in (b). The allele frequencies are shown for six randomly chosen subpopulations out of the 200 subpopulations in the simulation. The heterozygosities and fixation indices were calculated from all 200 subpopulations.

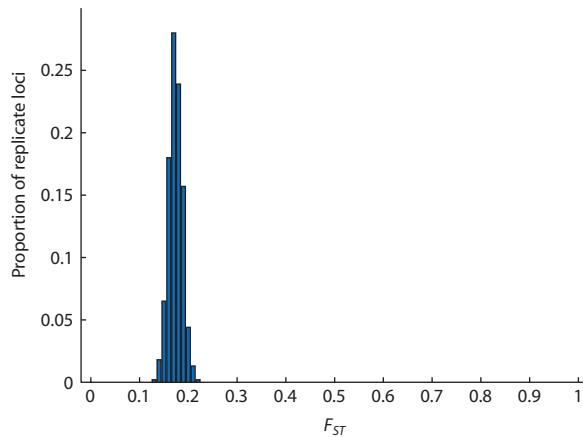
$$H_I = 2p_i q_i \quad (4.19)$$

where  $i$  indicates a single subpopulation. The average heterozygosity of the two demes is based on taking the heterozygosity within each subpopulation and then averaging:

$$H_S = \frac{2p_1 q_1 + 2p_2 q_2}{2} \quad (4.20)$$

In contrast, the heterozygosity in the total population is

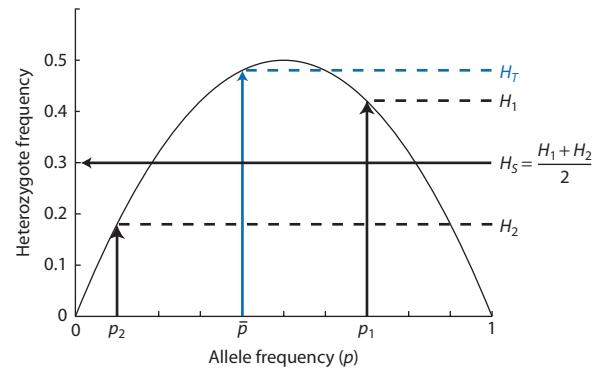
$$H_T = 2\bar{p}\bar{q} \quad (4.21)$$



**Figure 4.9** The distribution of  $F_{ST}$  values for 1000 replicate neutral loci in a finite island model of 200 subpopulations where each subpopulation contains 10 individuals and the rate of gene flow is 10% of each subpopulation ( $m = 0.10$ ). In the distribution, 95% of the replicate loci show  $F_{ST}$  values between 0.1459 and 0.2002 whereas the average of all 1000 replicate loci is 0.1586 (based on the average of  $H_1$  and  $H_2$  then used to calculate  $F_{ST}$ ). Replicate loci exhibit a range of  $F_{ST}$  values since allele frequencies among subpopulations are partly a product of the stochastic process of genetic drift. In an infinite island model with  $N_e m = 1.0$  the expected value of  $F_{ST}$  is 0.2.

based on the product of subpopulation average allele frequencies.  $H_T$  and  $H_S$  both cannot exceed 0.5, the maximum heterozygosity for a diallelic locus. In addition,  $H_S$  is an average of  $H_1$  and  $H_2$ , so when subdivided populations have different allele frequencies  $H_S$  will always be less than the expected heterozygosity of the total population. These conditions assure that  $H_T \geq H_S$  when there is random mating within the subpopulations. This relationship between  $H_T$  and  $H_S$  is shown graphically in Fig. 4.10. This phenomenon is called the **Wahlund effect** after the Swedish geneticist Sten Gosta William Wahlund who first described it in 1928. One result is that  $F_{ST}$  will be greater or equal to zero since the numerator in the expression for  $F_{ST}$  is  $H_T - H_S$ .

**Wahlund effect** The decreased expected frequency of heterozygotes in subpopulations with diverged allele frequencies compared with the expected frequency of heterozygotes in a panmictic population of the same total size with the same average allele frequencies.



**Figure 4.10** A graphical demonstration of the Wahlund effect for a diallelic locus in two demes. If there is random mating within subpopulations ( $H_1$  and  $H_2$ ) and in the total population ( $H_T$ ), the heterozygosity of each falls on the parabola of Hardy-Weinberg expected frequency. The average heterozygosity of subpopulations ( $H_S$ ) is at the mid-point between the deme heterozygosities. Therefore,  $H_S$  can never be greater than  $H_T$  based on the average allele frequency (the mid-point between the deme allele frequencies  $p_1$  and  $p_2$ ). Greater variance in allele frequencies of the demes is the same as a wider spread of deme allele frequencies in the two-deme case.

The Wahlund effect can also be shown in another fashion that more clearly connects it to variation in allele frequencies among subpopulations. The goal will now be to show that the difference between the expected heterozygosity in the total population ( $H_T$ ) and the average expected heterozygosity of the subpopulations ( $H_S$ ) depends on the variance in allele frequencies among the subpopulations.

The variance in allele frequencies among a set of subpopulations is

$$\text{Var}(p) = \frac{\sum(p_i - \bar{p})^2}{n} = \frac{\sum p_i^2}{n} - \bar{p}^2 \quad (4.22)$$

where  $p_i$  is the allele frequency in subpopulation  $i$ . It also turns out that for a diallelic locus,  $\text{var}(p)$  equals  $\text{var}(q)$  since  $p = 1 - q$ . This result will be used later.

The average expected heterozygosity of the subpopulations,

$$H_S = \frac{1}{n} \sum_{i=1}^n 2p_i q_i \quad (4.23)$$

can also be expressed as

$$H_S = \sum 2 \left( \frac{p_i}{n} - \frac{p^2}{n} \right) \quad (4.24)$$

by noticing that  $p_i q_i = p_i(1 - p_i) = p_i - p_i^2$  since  $p = 1 - q$ . This equation can be rearranged to

$$H_S = 2 \left( \frac{\sum p_i}{n} - \frac{\sum p_i^2}{n} \right) \quad (4.25)$$

The right-hand term inside the parentheses is identical to a term in the expression for the variance in allele frequency. Rearranging equation 4.22 so that  $\frac{\sum p_i^2}{n} = \text{var}(p) + \bar{p}^2$  and then substituting gives

$$H_S = 2 \left( \frac{\sum p_i}{n} - \text{var}(p) - \bar{p}^2 \right) \quad (4.26)$$

This too can be simplified by noting that  $\frac{\sum p_i}{n}$  is just the average allele frequency, or  $\bar{p}$ , and then making the substitution so that

$$H_S = 2(\bar{p} - \bar{p}^2 - \text{var}(p)) \quad (4.27)$$

The next step is again to use the fact that  $p = 1 - q$  to replace  $\bar{p} - \bar{p}^2$  with its equivalent expression  $\bar{p}\bar{q}$  and multiply the terms inside the parentheses by 2 to finally obtain

$$H_S = 2\bar{p}\bar{q} - 2 \text{ var}(p) \quad (4.28)$$

Recall from equation 4.21 that  $H_T = 2\bar{p}\bar{q}$ . Making this substitution gives

$$H_S = H_T - 2 \text{ var}(p) \quad (4.29)$$

Using an equivalent set of substitutions and algebraic rearrangements, it is also possible to show that the expected frequencies of homozygote genotypes in the total population are

$$\text{Freq}(AA)_T = \bar{p}^2 + \text{var}(p) \quad (4.30)$$

and

$$\text{Freq}(aa)_T = \bar{q}^2 + \text{var}(p) \quad (4.31)$$

The changes to homozygosity and heterozygosity caused by allele-frequency divergence among populations are exactly analogous to the consequences of consanguineous mating in a single population. In section 2.5 it was shown that  $\text{freq}(AA) = p^2 + fpq$

where  $f$  is the probability of identity by descent. The Wahlund effect describes a similar phenomenon where allele-frequency divergence of populations leads to an increase of homozygosity in subpopulations compared to the heterozygosity expected based on total population allele frequencies.

These equations show that in a subdivided population, the expected genotype frequencies in the total population are a function of the average allele frequencies as well as the variance in allele frequencies among subpopulations. A set of subpopulations in panmixia is equivalent to a situation where there is no variance in allele frequency ( $\text{var}(p) = 0$ ). In that case,  $H_T = H_S$  and  $F_{ST}$  is zero since  $H_T - H_S$  is also zero. This result is consistent with the intuitive expectation that extensive gene flow homogenizes allele frequencies among subpopulations. However, when subpopulations have diverged in allele frequencies

### Interact box 4.2 Simulating the Wahlund effect

The Wahlund effect can be seen readily with a simple simulation that allows you to set allele frequencies for five subpopulations and the degree of non-random mating within populations. Launch PopGeneS<sup>2</sup>, and select **Wahlund effect** from the **Gene Flow and Subdivision** menu. The simulation computes genotype frequencies within each population depending on allele frequencies and the degree of non-random mating. It also shows the expected and observed genotype frequencies averaged over the subpopulations along with the average allele frequency and variance in allele frequency among subpopulations. The fixation indices  $F_{IS}$ ,  $F_{ST}$ , and  $F_{IT}$  are also calculated. Try the first simulation with the default values of  $f = 0$  and allele frequencies of 0.9, 0.8, 0.7, 0.5, and 0.5.

What is the relationship between the variance in allele frequency and the difference between the average observed and expected heterozygosities? How does  $F_{ST}$  change with variance in allele frequency? What happens to the observed genotype frequencies when you try different levels of non-random mating within subpopulations? (Set  $f$  to a value other than zero.)

and  $\text{var}(p) > 0$  then the total population will have a deficit of heterozygotes and an excess of homozygotes compared to the case of panmixia. This method also provides the prediction that the total deficit of heterozygotes will equal the total excess of homozygotes when  $\text{var}(p) > 0$ .

One consequence of the Wahlund effect is termed **isolate breaking** to describe the increase of heterozygote genotypes that occurs when previously subdivided populations with diverged allele frequencies experience random mating. In human populations, disease phenotypes caused by recessive alleles expressed in homozygotes include cystic fibrosis, Tay–Sachs disease, and sickle-cell anemia. These disorders are more common in relatively insular populations such as Ashkenazi Jews, native American groups, and the Amish, but rarer in human populations that have experienced greater amounts of genetic mixing and thereby have less of a heterozygosity deficit due to subdivision. To see the impact of isolate breaking, imagine two randomly mating populations of squirrels that initially do not share any migrants and have allele frequencies that have diverged over time (Fig. 4.11). Suppose that the population on the left has albino individuals and the basis of the albino phenotype is the completely recessive allele  $a$  with frequency  $q$ . The albino allele is completely absent in the population on the right. The average frequency of albino squirrels in the subdivided population is

$$\begin{aligned} \text{Average frequency } (aa) &= \overline{q^2} = \frac{0.16 + 0}{2} \quad (4.32) \\ &= 0.08 \end{aligned}$$

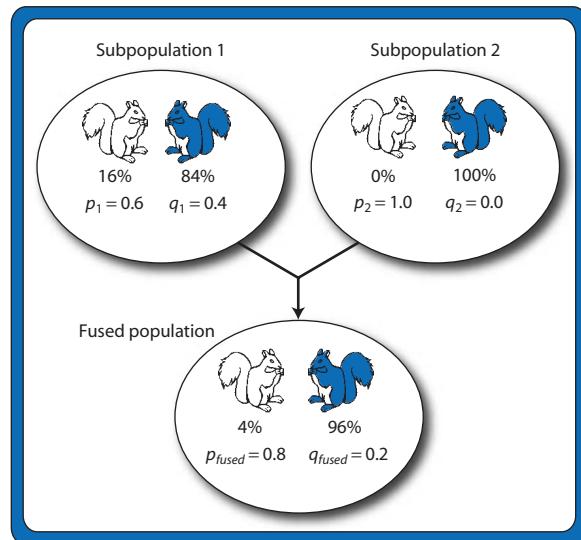
Relying on Hardy–Weinberg, we can similarly determine the average frequency of dominant

$$\begin{aligned} \text{homozygotes } \left( \overline{q^2} = \frac{0.36 + 1.0}{2} = 0.68 \right) \text{ and hetero-} \\ \text{zygotes } \left( \overline{2pq} = \frac{0.48 + 0.0}{2} = 0.24 \right) \text{ for the two} \end{aligned}$$

subpopulations.

Next, imagine that the two populations of squirrels fuse into one randomly mating population. What are the frequency of the recessive allele and expected frequencies of albino squirrels in this fused population after random mating occurs? First, determine the allele frequencies in the fused population:

$$q_{\text{fused}} = \frac{0.4 + 0.0}{2} = 0.2 \quad (4.33)$$



**Figure 4.11** A hypothetical example of how the Wahlund effect relates variation in allele frequency between subdivided populations and genotype frequencies in a single panmictic population. Initially, the two subpopulations have different allele frequencies and therefore different frequencies of homozygous recessive albino phenotypes. The average frequency of the albino phenotype is 8% in the subpopulations. When the populations fuse, the allele frequencies become the average of the two subpopulations. However, the genotype frequencies are not the average of the two subpopulations. Rather, homozygotes become less frequent and heterozygotes more frequent than their respective subpopulation averages. In the fused population, the degree to which the frequencies of both homozygotes combined and the heterozygotes differ from their subpopulation averages is the same as the variance in allele frequency between the two subpopulations.

and then use that result to determine the expected frequency of homozygous recessive genotypes in the fused population:

$$q_{\text{fused}}^2 = \left( \frac{0.4 + 0.0}{2} \right)^2 = (0.2)^2 = 0.04 \quad (4.34)$$

There are fewer albino squirrels in the fused population (4%) than there were for the average of the two subdivided populations (8%). You can verify that the other homozygote also decreases in frequency in the fused population. The frequencies of both homozygotes have decreased by 4% in the fused population compared to their average frequencies in the subdivided populations. In contrast, the frequency of heterozygotes in the fused population,

$$2pq_{\text{fused}} = 2(0.2)(0.8) = 0.32 \quad (4.35)$$

**Table 4.7** Allele and genotype frequencies for the hypothetical example of albino squirrels in Fig. 4.11 used to demonstrate Wahlund's principle. Initially, the total population is subdivided into two demes with different allele frequencies. These two populations are then fused and undergo one generation of random mating.

	Initial subpopulations	Fused population
Allele frequency $q$	0.4 and 0.0	$\frac{0.4 + 0.0}{2} = 0.2$
Variance in $q$	$\frac{(0.4 - 0.2)^2 + (0.0 - 0.2)^2}{2} = 0.04$	0
Frequency of aa	$\overline{q^2} = \frac{0.16 + 0}{2} = 0.08$	$(0.2)^2 = 0.04$
Frequency of Aa	$\overline{2pq} = \frac{0.48 + 0.0}{2} = 0.24$	$2(0.2)(0.8) = 0.32$
Frequency of AA	$\overline{p^2} = \frac{0.36 + 1.0}{2} = 0.68$	$(0.8)^2 = 0.64$

is greater than the average frequency of heterozygotes in the subdivided populations (see Table 4.7).

Now let's determine the variance in allele frequency among the two populations before and after fusion. Initially, the variance in allele frequency for the two subdivided populations is

$$\text{Var}(q) = \frac{(0.4 - 0.2)^2 + (0.0 - 0.2)^2}{(2 - 1)} = 0.08 \quad (4.36)$$

whereas  $\text{var}(q)$  is zero after fusion because there is no longer any subdivision for allele frequencies. Take note of the fact that the initial variance in the allele frequencies (0.08) is exactly twice the difference between the average frequency of albinos before fusion (equation 4.32) and the expected frequency of albinos in the fused population (equation 4.34)! With fusion of the subdivided populations, each homozygote has decreased by 4% and the heterozygote has increased by exactly the same total amount or 8%.

This example shows that removing the allele frequency differences between the two subpopulations by making them into a panmictic population has changed the total population heterozygosity. The result is exactly what is predicted by the Wahlund effect, with more total population heterozygosity under panmixia than under subdivision. Subdivided populations store some genetic variation as differences (variation) in allele frequency among populations at the expense of heterozygosity in the total population. Another way to think of this is that population subdivision is equivalent to inbreeding that increases the

total population homozygosity (or reduces the total population heterozygosity). A fused or panmictic population has a larger effective size than individual subdivided populations with restricted gene flow. In the subpopulations, mating is most probable within the subpopulation rather than with a migrant from the total population. The subpopulations therefore have more autozygosity compared to a panmictic population of equivalent size, analogous to the decline in heterozygosity seen in a single finite population due to genetic drift (see section 3.4).

A more realistic application of Wahlund's principle can be found in forensic DNA profiling. As covered in section 2.4, the use of DNA markers to determine the expected frequency of a given genotype occurring by chance relies on estimates of allele frequencies in various racially defined human populations. Although allele frequencies at loci used in DNA profiles have been estimated in many populations, there are a limited number of these reference allele-frequency databases available. It is therefore possible that population-specific allele-frequency estimates are not available for some individuals depending on their racial, ethnic, or geographic background. A further complication is that many individuals have ethnically diverse ancestry that may not be represented by any single set of available reference allele frequencies. None of this would be a problem in DNA profiling if human populations exhibited panmixia, since there would then be uniform allele frequencies among all racially defined human populations. However, racially and geographically defined human populations like

**Table 4.8** Expected frequencies for individual DNA-profile loci and the three loci combined with and without adjustment for population structure. Calculations assume that  $F_{IS} = 0$  and use the upper-bound estimate of  $F_{ST} = 0.05$  in human populations. Allele frequencies are given in Table 2.3.

Locus	Expected genotype frequency	
	With panmixia	With population structure
D3S1358	$2(0.2118)(0.1626) = 0.0689$	$2(0.2118)(0.1626)(1 - 0.05) = 0.0655$
D21S11	$2(0.1811)(0.2321) = 0.0841$	$2(0.1811)(0.2321)(1 - 0.05) = 0.0799$
D18S51	$(0.0918)^2 = 0.0084$	$(0.0918)^2 + 0.0918(1 - 0.0918)(0.05) = 0.0126$
All loci	$(0.0689)(0.0841)(0.0084) = 0.000049$	$(0.0655)(0.0799)(0.0126) = 0.000066$

those used to construct allele-frequency reference databases show up to 3–5% population divergence of allele frequencies (Rosenberg et al. 2002).

We can use Wahlund's principle to adjust DNA-profile odds ratios for the effects of population structure. This requires a method to adjust the expected genotype frequency at each locus to account for the increased frequency of homozygotes and the decreased frequency of heterozygotes caused by the divergence of allele frequencies among populations. The adjusted expected frequencies for homozygote genotypes are

$$f(A_i A_i) = p_i^2 + p_i(1 - p_i)F_{IT} \quad (4.37)$$

and the adjusted expected frequencies for heterozygote genotypes are

$$f(A_i A_j) = 2p_i p_j - (2p_i p_j)F_{IT} = 2p_i p_j(1 - F_{IT}) \quad (4.38)$$

where  $i$  and  $j$  represent different alleles at the A locus and  $F_{IT}$  measures the total departure of genotype frequencies from frequencies expected under panmixia due both to non-random mating within populations and allele-frequency divergence among populations (National Research Council, Commission on DNA Forensic Science 1996). If mating within populations is random ( $F_{IS} = 0$ ) then  $F_{IT}$  is equivalent to  $F_{ST}$  in these two equations. In that case, applying these corrections increases the frequency of homozygotes and decreases the frequency of heterozygotes in proportion to the degree of allele frequency divergence among populations.

In section 2.4, the expected frequency of a three-locus DNA profile was determined under the assumptions of Hardy–Weinberg and panmixia. Let's return to that example and adjust the expected genotype frequency and odds ratio to compensate for popula-

tion structure in human populations. The expected genotype frequencies are given in Table 4.8 based on the upper bound estimate of  $F_{ST} = 0.05$  in human populations. The adjustment reduces the expected frequencies of the two heterozygous loci and increases the expected frequency of the homozygous locus. The odds ratio for chance match of this three-locus genotype was one in 20,408 under the assumption of panmixia and becomes one in 15,152 after adjusting for population structure. Thus, population structure increases the expected frequency of this three-locus genotype by about 35% of its expected frequency under panmixia. A random match for this three-locus genotype is more probable after adjustment for population structure. Compensating for population structure in determining DNA-profile odds ratios is required to obtain an accurate estimate

### Problem box 4.3 Account for population structure in a DNA-profile match probability

Return to section 2.4 and Problem box 2.1 to determine the expected genotype frequency and probability of a random match after compensation for population structure seen in human populations. Assume that  $F_{ST} = 0.05$  for human populations. How does the expected genotype frequency change at individual loci when there is population structure and why? Is the 10-locus genotype still rare enough that the chance of a random match is low?

of how often DNA profiles match by chance alone (National Research Council, Commission on DNA Forensic Science 1996). Using equations 4.37 and 4.38 to adjust for population structure is necessary when an appropriate reference allele-frequency database is not available, the ethnicity of the individual is not known, or the genotype comes from a person of mixed ancestry and therefore the choice of the appropriate database is not obvious.

The next section explores models of population structure that can be used to infer the causes of a given pattern of population structure.

#### 4.5 Models of population structure

- Continent-island, two-island, and infinite island models.
- Stepping-stone and metapopulation population models.
- General expectations and conclusions from the different migration models.

The various models of population structure attempt to approximate various gene-flow properties likely to be found in actual populations. However, these models do not necessarily capture the exact mixture of gene-flow features in actual populations. It is likely, in fact, that gene flow within and among actual subpopulations of real organisms is not as easily categorized nor as invariant as is assumed in these models. Nonetheless, these models of population structure are useful tools to study the general principles that cause population differentiation. The utility of these different models of population structure is their ability to show basic and somewhat general features of the impact of rates of gene flow, the size of subpopulations, and the patterns of genetic connectedness among subpopulations on the evolution of genotype and allele frequencies within and among populations.

##### **Continent-island model**

Perhaps the simplest model of gene flow is called the continent-island model (Fig. 4.12a). It assumes that there is one very large population where allele frequency changes very little over short periods of time and a smaller population that receives migrants from the large continent population each generation. The island population experiences the replacement of a proportion  $m$  of its individuals through migration, with  $1 - m$  of the original individuals remaining each generation. (We assume that the proportion

$m$  of island individuals replaced by gene flow each generation either die or emigrate to the continent population, which is so large that immigrants do not impact allele frequencies.) The continent-island model assumes no genetic drift, no natural selection, random mating in both populations, that migrants are a random sample of genotypes, and no mutation.

**Continent-island model** An idealized model of population subdivision and gene flow that assumes one very large population where allele-frequency changes only slowly over time (like a continent full of many individuals) connected by gene flow with a small population where migrants make up a finite proportion of the individuals present each generation. Gene flow from the island to the continent occurs but the continent population is assumed to be large enough that immigration has a negligible effect on allele frequencies.

Based on this situation along with its assumptions, it is possible to predict how gene flow changes allele frequency at a diallelic locus in the island population over one generation. Allele frequency in the island population one generation in the future (call it  $p_{t+1}^{island}$ ) is a function of (i) the allele frequency in the proportion of the island population that are not migrants and (ii) the allele frequency in the proportion of the island population that arrives via gene flow from the continent population. This can be stated in an equation as

$$p_{t=1}^{island} = p_{t=0}^{island}(1 - m) + p^{continent}m \quad (4.39)$$

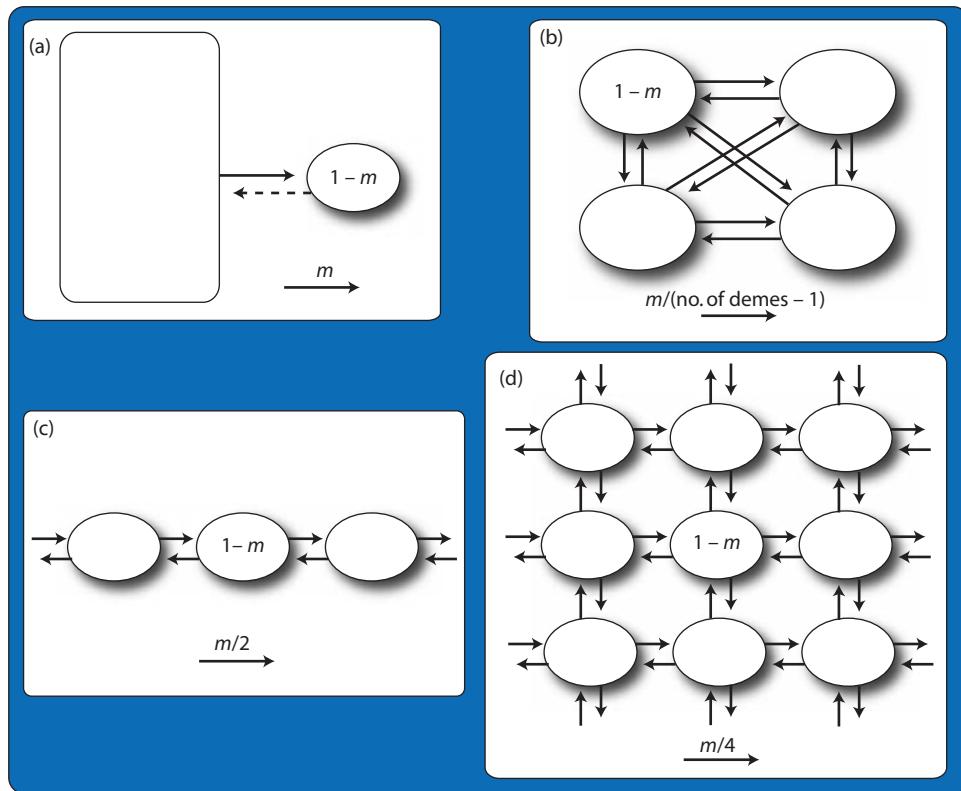
and used to predict the island population allele frequency after one generation of gene flow from the continent. Expanding the right side of this equation gives

$$p_{t=1}^{island} = p_{t=0}^{island} - p_{t=0}^{island}m + p^{continent}m \quad (4.40)$$

which can be rearranged to an equation that gives the change in allele frequency in the island population over one generation:

$$p_{t=1}^{island} - p_{t=0}^{island} = -m(p_{t=0}^{island} - p^{continent}) \quad (4.41)$$

in a form readily interpreted in biological terms.



**Figure 4.12** Classic models of population structure make different assumptions about the paths and rates of gene flow among subpopulations. (a) In the continent-island model, gene flow is essentially unidirectional from a very large population to a smaller population. The continent population is so large that allele frequencies are not impacted by emigration or drift whereas allele frequencies in the small population(s) are strongly influenced by immigration. (b) The island model has equal rates of gene flow exchanged by all populations regardless of the number of populations or their physical locations. (c, d) Stepping-stone models restrict gene flow to populations that are either adjacent or nearby in one (c) or two (d) dimensions and thereby incorporate isolation by distance. Gene-flow models can also incorporate the extinction and re-colonization of subpopulations, a feature commonly added to stepping-stone model populations. Each panel shows the rate of gene flow indicated by the arrows if  $m$  percent of each population is composed of migrants and  $1 - m$  is composed of non-migrating individuals each generation.

Equation 4.41 predicts that the degree of difference between allele frequencies in the island and continent populations ( $p_{t=0}^{island} - p_{t=0}^{continent}$ ) will determine the direction as well as the rate of change in the island allele frequency as long as the rate of gene flow is not zero ( $m \neq 0$ ). For example, if  $p_{t=0}^{island} > p_{t=0}^{continent}$  then the island allele frequency should decrease. Likewise, the island allele frequency is expected to increase if  $p_{t=0}^{island} < p_{t=0}^{continent}$ . To use a numerical example, suppose that  $p_{t=0}^{island} = 0.1$  and  $p_{t=0}^{continent} = 0.9$ . The difference between the island and continent allele frequencies is  $-0.8$ , so according to equation 4.41 the island allele frequency should increase for any amount of gene flow. If  $m = 0.1$ , then the island allele frequency will increase by  $0.08$  to  $p_{t=1}^{island} = 0.18$  in one generation.

The expected change in allele frequency due to a single generation of gene flow can also be extended to predict allele frequency in the island population

over an arbitrary number of generations. If there is a second generation of gene flow, the allele frequency in the island population is then

$$p_{t=2}^{island} = p_{t=1}^{island}(1 - m) + p_{t=0}^{continent}m \quad (4.42)$$

Substituting  $p_{t=1}^{island}$  as defined in equation 4.41 into this equation,

$$p_{t=2}^{island} = (p_{t=0}^{island}(1 - m) + p_{t=0}^{continent}m)(1 - m) + p_{t=0}^{continent}m \quad (4.43)$$

and rearranging terms,

$$p_{t=2}^{island} = p_{t=0}^{island}(1 - m)^2 + p_{t=0}^{continent}(m(1 - m) + m) \quad (4.44)$$

to eventually give an expectation for the island allele frequency after two generations of gene flow

$(p_{t=2}^{island})$  in terms of the initial island allele frequency  $(p_{t=0}^{island})$ :

$$p_{t=2}^{island} = p_{t=0}^{island}(1 - m)^2 + p^{continent}(1 - (1 - m)^2) \quad (4.45)$$

Notice that the exponents are equal to the number of generations that have elapsed. Changing these exponents to an arbitrary number leads to the allele frequency in the island population after  $t$  generations have elapsed starting from an initial allele frequency gives

$$p_t^{island} = p_{t=0}^{island}(1 - m)^t + p^{continent}(1 - (1 - m)^t) \quad (4.46)$$

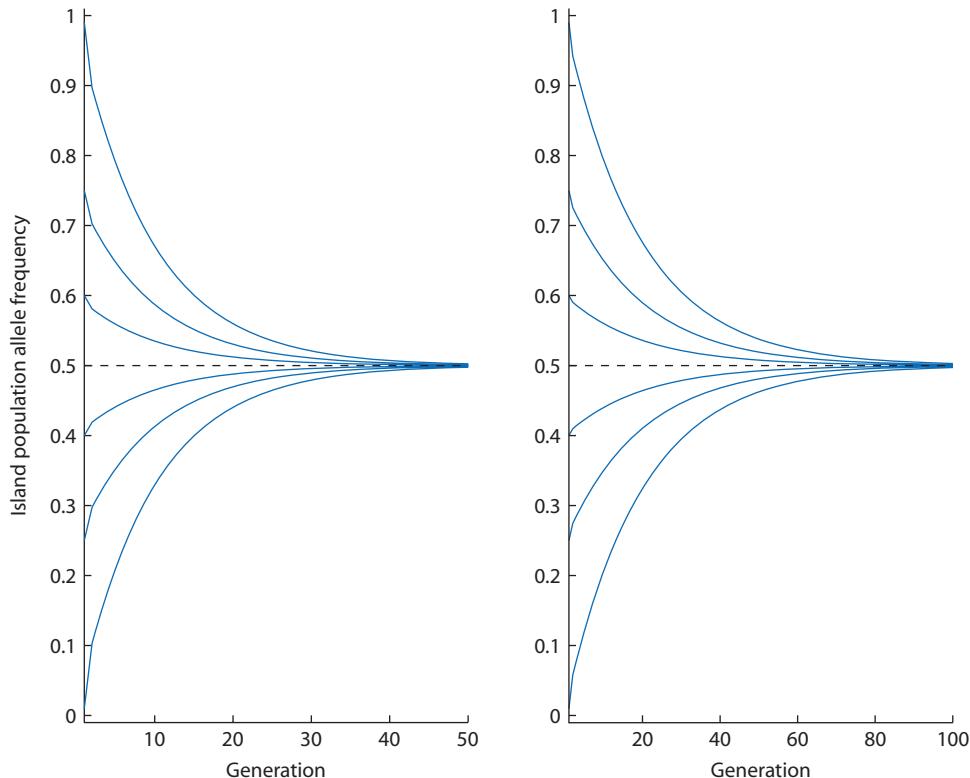
which can be rearranged to

$$p_t^{island} = p^{continent} + (p_{t=0}^{island} - p^{continent})(1 - m)^t \quad (4.47)$$

The rate of allele frequency change in the island population can also be seen in this equation. The

proportion of the island population that made up its initial allele frequency decreases by  $(1 - m)^t$ , approaching zero as time passes. This means that the island population is increasingly composed of immigrants from the continent. Therefore, the allele-frequency difference between the island and continent decreases toward zero over time and the allele frequency of the island approaches the allele frequency of the continent. Figure 4.13 shows how the island allele frequency approaches the continent allele frequency over time for a range of initial island allele frequencies. Notice the smooth approach to the continent allele frequency: this is a consequence of the fact that the outcome is completely determined by a constant rate of gene flow and has no random processes such as genetic drift to introduce chance variation.

These predictions of the continent-island model are consistent with intuition. Given that the continent population has a constant allele frequency over time, the island population should eventually reach an identical allele frequency when the two are mixed. How long it takes for the two populations



**Figure 4.13** Allele frequency in the island population for a diallelic locus under the continent-island model of gene flow. The island population allele frequencies ( $p_{island}^{island}$ ) over time are shown for six different initial values (solid lines). The continent population has an allele frequency of  $p^{continent} = 0.5$  shown by the dashed line. In the left-hand panel  $m = 0.1$  and in the right-hand panel  $m = 0.05$ . Equilibrium is reached more slowly when the rate of gene flow is lower. In contrast, the difference in allele frequencies between the island and continent does not affect time to equilibrium for a given rate of gene flow. Note that the time scales in the two graphs differ.

### Interact box 4.3 Continent-island model of gene flow

PopGene.S<sup>2</sup> contains a module to simulate the continent-island model of gene flow. In PopGene.S<sup>2</sup> click on the **Gene Flow and Subdivision** menu and then select **Continent-Island** model of migration. The simulation window allows you to set allele frequencies in the island and continent, the rate at which island alleles are replaced by continent alleles (or the migration rate) and the number of generations to simulate. Enter the parameters of **pC** = 0.9, **pl** = 0.1, **m** = 0.1, and 100 generations to run. Before clicking the **OK** button, predict the equilibrium allele frequency in the island population.

Keeping the same values for initial allele frequencies, try a series of values of the migration rate to see how it affects time to equilibrium. Click the **Clear Screen** button to clear the graph window. Increase the **Generations to run** to 300. Run the simulation with **m** = 0.1, **m** = 0.05, **m** = 0.001, and **m** = 0.0001 without clearing the graph window. This should give a plot with four blue lines (one for each value of the migration rate). What is the relationship between the migration rate and time to equilibrium?

to converge on the same allele frequency depends on the proportion of continent individuals moving to the island each generation. In contrast, the difference in allele frequencies between the island and continent does not alter the time to equilibrium for a given migration rate (see Fig. 4.13). This occurs since the rate of change in the island allele frequency is determined by the difference in allele frequencies. Greater differences lead to greater rates of change toward the continent allele frequency. Thus, the continent-island model shows that the process of gene flow alone is capable of bringing populations to the same allele frequency. Identical allele frequencies between or among populations is really a lack of population structure or panmixia. So the continent-island can be thought of as a demonstration that gene flow acting in the absence of other processes will eventually result in panmixia.

#### *Two-island model*

One simple adjustment to the continent-island model is to consider the two subpopulations as being equal in size, removing the assumption that one population (the continent) serves as an unchanging source of migrants. The model then represents gene flow between two islands which can each exhibit changes in allele frequency over time. The switch to a two-island model also allows an independent rate of gene flow for each subpopulation,  $m_1$  and  $m_2$ . Using reasoning similar to that for the continent-island model, the allele frequency in a subpopulation one generation in the future is the sum of the allele frequency in the proportion of individuals that do not migrate ( $1 - m$ )

plus the allele frequency in the immigrants. Assuming that  $m_1 = m_2 = m$ , the allele frequency in either subpopulation is

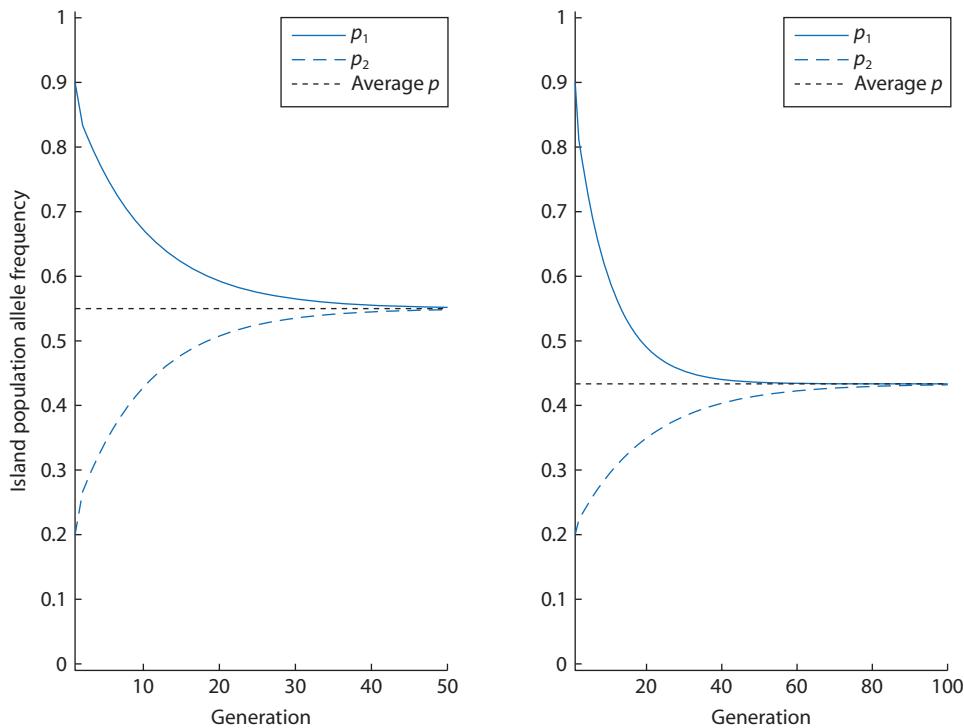
$$p_{t=1} = p_{t=0}(1 - m) + \bar{p}m \quad (4.48)$$

where  $\bar{p} = \frac{p_1 + p_2}{2}$ . The allele frequency in the migrants is now the average of the two subpopulations rather than just a constant like the continent allele frequency. This happens because both subpopulations receive immigrants so the allele frequencies of each subpopulation are approaching the allele frequency in the total population as gene flow mixes the subpopulations. Similar to the result for the continent-island model, the allele frequency in either of the two islands is

$$p_t = \bar{p} + (p_{t=0} - \bar{p})(1 - m)^t \quad (4.49)$$

after  $t$  generations have elapsed. Figure 4.14 shows allele frequencies in the two-island model over time.

When the rates of gene flow are not equal then the average allele frequency is  $\bar{p} = \frac{p_1 m_2 + p_2 m_1}{m_1 + m_2}$ , or the gene-flow-weighted average of the allele frequencies in the two subpopulations. When  $m_1 \neq m_2$  the equilibrium allele frequencies will be closer to the initial allele frequency of the subpopulation with the lower migration rate. This happens because the subpopulation with the lower migration rate experiences less immigration and remains closer to its initial allele frequency, yet it supplies migrants to the other subpopulation. As seen in Fig. 4.14, the time to equilibrium



**Figure 4.14** Allele frequency in the two-island model of gene flow for a diallelic locus. Dashed lines in each panel highlight gene-flow-weighted average or equilibrium allele frequencies. Starting from allele frequencies of 0.9 and 0.2 and with equal rates of gene flow ( $m = 0.1$ ), the subpopulations approach an equilibrium allele frequency of  $\bar{p} = (0.9 + 0.2)/2 = 0.55$  (left panel). With initial allele frequencies of 0.9 and 0.2 but asymmetric rates of gene flow ( $m_1 = 0.1$  and  $m_2 = 0.05$ ), the subpopulations approach an equilibrium allele frequency of  $\bar{p} = (0.9 \times 0.05 + 0.2 \times 0.1)/0.15 = 0.433$  (right panel). Equilibrium is reached more slowly in the case of asymmetric rates of gene flow on the right because the average rate of gene flow is lower. Note that the time scales in the two graphs differ.

is also longer when the migration rates are asymmetric. Consider the example where migration rates are unequal ( $m_1 = 0.01$  and  $m_2 = 0.1$ ) and the initial allele frequencies in the two subpopulations are  $p_1 = 0.9$  and  $p_2 = 0.1$ . The weighted average allele frequency is then  $\bar{p} = \frac{(0.9)(0.1)}{0.11} + \frac{(0.1)(0.01)}{0.11} = 0.827$ . This is also the expected allele frequency in both subpopulations at equilibrium.

The main conclusion of the two-island model is that the equilibrium allele frequencies in the two subpopulations are the average allele frequency of the total population when the two migration rates are equal. This conclusion holds when there are a larger number of subpopulations, a result that will be useful to remember when considering the process of gene flow in an island model in combination with another process such as genetic drift.

### Infinite island model

One of the most widely used models of the process of gene flow among a set of subpopulations is Wright's

(1931, 1951) infinite island model. Gene flow takes the form of all subpopulations being equally likely of exchanging migrants with any other subpopulation, equivalent to a complete absence of isolation by distance. In addition, the size and migration rate of each subpopulation is most commonly assumed to be equal. The total population is made up of an infinite set of subpopulations each of size  $N_e$  with  $m$  percent of each subpopulation's gene copies exchanged at random with the rest of the population every generation (see Fig. 4.12b). Using this model it is possible to approximately relate the degree of differentiation among subpopulations to a function of the effective population size and the amount of migration.

**Infinite island model** An idealized model of population subdivision and gene flow that assumes an infinite number of identical subpopulations (demes) and that each subpopulation experiences an equal probability of gene flow from all other subpopulations.

### Interact box 4.4 Two-island model of gene flow

The two-island model of gene flow can be simulated in PopGene.S<sup>2</sup>. Launch PopGene.S<sup>2</sup> and select **Island-island** model of migration from the **Gene Flow and Subdivision** menu. The simulation window has entry fields to set initial allele frequencies in each island subpopulation, the rate at which each island receives immigrants from the other island, and the number of generations to run the simulation. Enter parameters of **p1** = 0.9, **p2** = 0.1, **m1** = 0.1, **m2** = 0.1, and 100 generations to run. Before clicking the **OK** button, predict the equilibrium allele frequency in each subpopulation.

A major conclusion from the two-island model is that the allele frequencies in each subpopulation approach the average allele frequency in the total population. Confirm that equal migration rates for both subpopulations give an equilibrium allele frequency of the average of the initial allele frequencies (for example, try **p1** = 0.6, **p2** = 0.4, **m1** = 0.1, and **m2** = 0.1, and then **p1** = 0.99, **p2** = 0.01, **m1** = 0.1, **m2** = 0.1). Also simulate cases when the migration rates are not equal (such as **p1** = 0.9, **p2** = 0.1, **m1** = 0.01, **m2** = 0.1) and use the gene-flow-weighted average of the allele frequencies to predict allele frequencies at equilibrium.

Let's first consider what will happen in the infinite island model when there is no gene flow among the subpopulations ( $m = 0$ ). Since each subpopulation is a finite island, allele frequencies will vary from one generation to the next simply due to genetic drift. The expected value of the fixation index for subpopulations compared to the total population is:

$$F_{ST} = 1 - e^{-\frac{1}{2N_e}t} \quad (4.50)$$

where  $t$  is time in generations and  $N_e$  is the effective size of a single subpopulation (Wright 1943a). In

equation 4.50 the  $e^{-\frac{1}{2N_e}t}$  term gets smaller as time increases. This serves to approximate what happens to  $F_{ST}$  as  $t$  increases: the average expected heterozygosity of subpopulations ( $H_S$ ) decreases and eventually reaches zero, with the consequence that  $F_{ST}$  reaches 1. This results from the process of genetic drift, causing all subpopulations to eventually reach fixation or loss. Note, however, that the total population heterozygosity ( $H_T$ ) is not impacted by genetic drift since each subpopulation has equal chances of fixation or loss and there are an infinite number of subpopulations.

Next we will move on to consider what occurs in the infinite island model when both gene flow and genetic drift are acting at the same time. In Chapter 3, the fixation index:

$$F_t = \frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right) F_{t-1} \quad (4.51)$$

was developed as a measure of the probability that two alleles in a genotype are autozygous or identical by descent in a single finite population. We can extend this equation to include the influence of migration on autozygosity when there are numerous subpopulations that experience limited gene flow each generation. The goal is to develop an expression for the fixation index that accounts for both population size and migration. Finite population size causes autozygosity to increase over time in individual subpopulations. Migration counteracts this trend, bringing in alleles from other subpopulations that are not identical by descent, thereby decreasing the autozygosity. Therefore, in general in subdivided populations, the net autozygosity is the balance of the processes of genetic drift and migration.

When there is gene flow, two modifications need to be made to the probabilities of autozygosity given in the two terms of equation 4.51. The first modification involves the probability of autozygosity or  $\frac{1}{2N_e}$ . With migration, some proportion  $m$  of the alleles in a subpopulation arrived via gene flow from other subpopulations while  $1 - m$  of the alleles are contributed by individuals and gametes that did not leave their subpopulation. Therefore, there is some chance that one or both of a pair of alleles was introduced to a subpopulation by migration. A randomly sampled pair of alleles in a subpopulation with zero, one, or two alleles due to gene flow each generation have probabilities of  $(1 - m)^2$ ,  $2m(1 - m)$ , and  $m^2$ , respectively. Only the  $(1 - m)^2$  proportion of genotypes with no alleles introduced by gene flow can contribute to the pool of alleles that may become identical by descent

due to finite sampling. We can also see this by noting that  $2m$  genotypes heterozygous and  $m^2$  genotypes homozygous for alleles entering the subpopulation by gene flow are expected each generation. Together, these two classes of genotypes bearing alleles that entered the population by gene flow reduce the autozygosity by a factor of  $1 - 2m - m^2 = (1 - m)^2$ . This gives  $\frac{1}{2N_e}(1 - m)^2$  as the autozygosity adjusted for gene flow. Using the same reasoning, the chances that a randomly sampled pair of alleles in a subpopulation are autozygous due to past inbreeding

(the  $\left(1 - \frac{1}{2N_e}\right)F_{t-1}$  term in equation 4.51) also needs to be adjusted by a factor of  $(1 - m)^2$ .

Bringing these two changes to the autozygosity together to account for gene flow leads to:

$$F_t = \frac{1}{2N_e}(1 - m)^2 + \left(1 - \frac{1}{2N_e}\right)F_{t-1}(1 - m)^2 \quad (4.52)$$

As seen by examining this equation, when  $m$  is between zero and one, the effect of gene flow is to reduce the expected value of the fixation index by reducing the probability of identity both in the present (time  $t$ ) and in the past (time  $t - 1$ ). This makes intuitive sense: if gene flow introduces an allele copy into a subpopulation, it has not been present for sampling events between time  $t - 1$  and time  $t$ . Therefore an allele copy introduced by gene flow has not yet had the opportunity to become identical by descent at time  $t$  and it cannot contribute to the frequency of autozygous genotypes gauged by the fixation index.

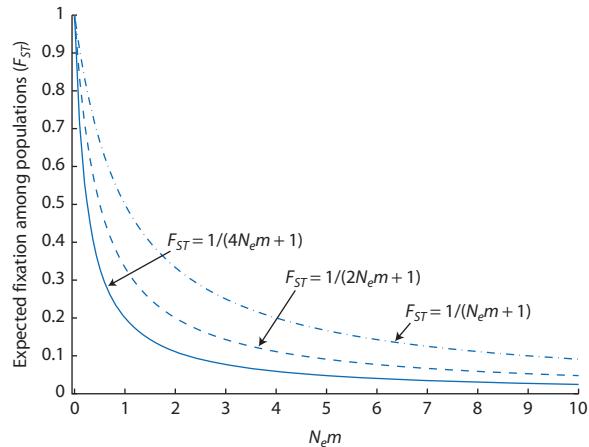
Equation 4.52 is really an expression for the balance of gene flow and genetic drift among multiple subpopulations so  $F$  is identical to  $F_{ST}$ . We can make this equation more general by using it to get an expected value of the fixation index among populations ( $F_{ST}$ ) in the infinite island model when allele frequency differentiation among subpopulations by genetic drift and allele frequency homogenization among subpopulations by gene flow, reach a net balance. With the assumption that the migration rate is small and much, much less than the effective population size (see Math box 4.1), an *approximation* for the expected amount of fixation among subpopulations at equilibrium in an infinite island population is:

$$F_{ST} \approx \frac{1}{4N_e m + 1} \quad (4.53)$$

as shown by Wright (1931, 1951).

Based on these assumptions, Fig. 4.15 shows the expected levels of genetic differentiation among infinitely many island model subpopulations for different levels of the product of the effective population size of each deme and the migration rate among demes ( $N_e m$ ). When  $N_e m$  – sometimes called the effective migration rate – is large, very little differentiation among demes is expected for a diploid locus since the combination of the effective size of demes and the migration rate is large enough to overcome the genetic differentiation caused by genetic drift. As the effective migration rate declines from a relatively large value such as  $N_e m = 10$ , genetic differentiation among demes increases slowly at first and then rapidly once the effective migration rate is less than about 1. An effective migration rate of one individual

every other generation  $\left(N_e m = \frac{1}{2N_e}\right)$  is often cited as sufficient to prevent substantial genetic differentiation at a diploid locus in the infinite island model



**Figure 4.15** Expected levels of fixation among subpopulations depend on the product of the effective population size ( $N_e$ ) and the amount of gene flow ( $m$ ) in the infinite island model of population structure. Each line represents expected  $F_{ST}$  for loci with different probabilities

of autozygosity (from bottom to top  $\frac{1}{2N_e}$ ,  $\frac{1}{N_e}$ , and  $\frac{2}{N_e}$ ).

Marked divergence of allele frequencies among subpopulations ( $F_{ST} \geq 0.2$ ) are expected when  $N_e m$  is below 1 for biparentally inherited nuclear loci with an

autozygosity of  $\frac{1}{2N_e}$ . Y-chromosome or mitochondrial

loci (autozygosity =  $\frac{2}{N_e}$ ) are examples where marked divergence among populations is expected at higher levels of  $N_e m$ .

### Math Box 4.1 The expected value of $F_{ST}$ in the infinite island model

At equilibrium when the differentiating effects of genetic drift and the homogenizing effects of gene flow have come into balance, the value of  $F_{ST}$  does not change from one generation to the next so that  $F_{ST(t)} = F_{ST(t-1)} = F_{ST(\text{equilibrium})}$ . If a population is at equilibrium then we can set both  $F_t$  and  $F_{t-1}$  equal to  $F_{eq}$ . Making this substitution in equation 4.52:

$$F_{eq} = \frac{1}{2N_e}(1-m)^2 + \left(1 - \frac{1}{2N_e}\right)F_{eq}(1-m)^2 \quad (4.54)$$

This equation can be solved for  $F_{eq}$  most transparently by restating it as

$$F_{eq} = ac + bcF_{eq} \quad (4.55)$$

where  $a = \frac{1}{2N_e}$ ,  $b = 1 - \frac{1}{2N_e}$ , and  $c = (1-m)^2$ .

Then using algebraic manipulation

$$F_{eq} - bcF_{eq} = ac \quad (4.56)$$

$$F_{eq}(1 - bc) = ac \quad (4.57)$$

$$F_{eq} = \frac{ac}{1 - bc} \quad (4.58)$$

Substituting the full expressions for  $a$ ,  $b$ , and  $c$  gives

since this rate is enough to balance the expected rate of loss of heterozygosity by genetic drift  $\left(1 - \frac{1}{2N_e}\right)$  in an isolated population (see section 3.4).

The expected relationship between the fixation index and the effective number of migrants relies on the infinite island model for two reasons. First, in the island model all subpopulations have an identical rate of migration from all other populations so there is only a single migration rate ( $m$ ) that applies to all subpopulations. Second, since there are an infinite number of subpopulations the entire ensemble population will never reach fixation or loss due to genetic drift. In an island model of gene flow where the number of subpopulations is finite, called the **finite island model**, the entire set of populations

$$F_{eq} = \frac{\frac{1}{2N_e}(1-m)^2}{\left(1 - \left(1 - \frac{1}{2N_e}\right)\right)(1-m)} \quad (4.59)$$

which when multiplied by  $\frac{2N_e}{2N_e}$  gives the simpler equation

$$F_{ST} = \frac{(1-m)^2}{2N_e - (2N_e - 1)(1-m)^2} \quad (4.60)$$

The terms in the numerator and denominator can be multiplied out to give an expression that is fairly complex (feel free to do the expansion if you are curious). However, if we again invoke the assumption that the migration rate is small and much, much less than the effective population size, then terms in the expansion of equation 4.60 containing  $m$  or powers of  $m$  can be ignored since they are very small (e.g. if  $m = 0.01$  then  $2m = 0.02$  and  $m^2 = 0.0001$ ). This then leads to the approximation for the expected value of the fixation index

$$F_{ST} \approx \frac{1}{4N_e m + 1} \quad (4.61)$$

will eventually reach fixation or loss and  $F_{ST}$  will eventually decline to zero since the entire set of subpopulations will eventually reach fixation or loss due to genetic drift (Nei et al. 1977; Varvio et al. 1986). The expected amount of genetic differentiation in the finite island model is

$$G_{ST} = \frac{1}{4N_e m \left(\frac{n}{n-1}\right)^2 + 1} \quad (4.62)$$

where  $G_{ST}$  is an estimator of  $F_{ST}$  for loci with any number of alleles (Nei 1973) and  $n$  is the number of subpopulations (Latter 1973; Crow & Aoki 1984; Takahata & Nei 1984). This version of  $G_{ST}$  corrects the expected amount of differentiation among subpopulations for a finite number of subpopulations.

The correction term  $\left(\frac{n}{n-1}\right)^2$  is at a maximum of 4 with two subpopulations and approaches one as  $n$  gets large. For example, when  $N_e m = 0.1$  and  $n = 10$ , the expected value of  $G_{ST}$  is about 94% of that expected for an infinite number of demes. This implies that a given level of gene flow is more effective at homogenizing allele frequencies among fewer subpopulations than among a very large number of subpopulations. For  $n$  greater than about 50 the adjustment for a finite number of demes makes little difference and the finite number of subpopulations behave essentially as an infinite number of subpopulations.

Given that the infinite island models leads to an expected level of genetic differentiation among demes for some level of the effective migration rate, it is natural to reverse the relationship:

$$N_e m \approx \frac{1}{4} \left( \frac{1}{F_{ST}} - 1 \right) \quad (4.63)$$

to get an expected effective migration rate given an amount of genetic differentiation among subpopulations in the infinite island model. This equation

### Problem box 4.4 Expected levels of $F_{ST}$ for Y-chromosome and organelle loci

What is the expected value of  $F_{ST}$  at equilibrium in the island model for Y-chromosome loci or mitochondrial and chloroplast (organelle) loci? A hint at how to approach the problem is to think about the autozygosity of loci other than diploid autosomes and then make adjustments to equation 4.52 that result in different versions of equation 4.53. What level of fixation among populations ( $F_{ST}$ ) would be expected for these types of loci compared to biparentally inherited diploid loci? What causes the difference in levels of  $F_{ST}$  for the different types of loci?

again reinforces that the level of allele frequency differentiation among subpopulations ( $F_{ST}$ ) is a function of the balance between the processes of gene flow tending to homogenize allele frequencies among

### Interact box 4.5 Finite island model of gene flow

Use PopGene.S<sup>2</sup> to simulate the island model of gene flow among a large (but finite) number of demes. Launch the program, click on the **Gene Flow and Subdivision** menu, and select the **F-Statistics** module. Run simulations with the following values for the effective population size, the migration rate, and the initial allele frequency.

$N_e$	$m$	Initial allele frequency $p$
10	0	0.5
10	0.001	0.5
10	0.1	0.5
50	0	0.5
50	0.001	0.5
50	0.1	0.5
100	0	0.5
100	0.001	0.5
100	0.1	0.5

For each simulation run, examine the allele frequencies over time and record the different hierarchical heterozygosity measures ( $H_I$ ,  $H_S$ , and  $H_T$ ) and fixation indices ( $F_{IS}$ ,  $F_{ST}$ , and  $F_{IT}$ ).

What is happening when the allele frequency lines sometimes hit the top or bottom axis (go to fixation or loss) and then reappear? What are the units of the migration values you entered in the model parameters box? Why does increasing  $m$  maintain lower  $F_{ST}$  and  $F_{IT}$  values? How does migration counteract genetic drift? Is migration always able to do this?

**Table 4.9** Estimates of the fixation index among subpopulations ( $\hat{F}_{ST}$ ) for diverse species based on molecular genetic marker data for nuclear loci. Different estimators were employed depending on the type of genetic marker and study design. Each  $\hat{F}_{ST}$  was used to determine the effective number of migrants ( $\widehat{N_e m}$ ) that would produce an identical level of population structure under the assumptions of the infinite island model according to equation 4.63.

Species	$\hat{F}_{ST}$	$\widehat{N_e m}$	Reference
<b>Amphibians</b>			
<i>Alytes muletensis</i> (Mallorcan midwife toad)	0.12–0.53	1.8–0.2	Kraaijeveld-Smit et al. 2005
<b>Birds</b>			
<i>Gallus gallus</i> (broiler chicken breed)	0.19	1.0	Emara et al. 2002
<b>Mammals</b>			
<i>Capreolus capreolus</i> (roe deer)	0.097–0.146	2.2–1.4	Wang and Schreiber 2001
<i>Homo sapiens</i> (human)	0.03–0.05	7.8–4.6	Rosenberg et al. 2002
<i>Microtus arvalis</i> (common vole)	0.17	1.2	Heckel et al. 2005
<b>Plants</b>			
<i>Arabidopsis thaliana</i> (mouse-ear cress)	0.643	0.1	Bergelson et al. 1998
<i>Oryza officinalis</i> (wild rice)	0.44	0.3	Gao 2005
<i>Phlox drummondii</i> (annual phlox)	0.17	1.2	Levin 1977
<i>Prunus armeniaca</i> (apricot)	0.32	0.5	Romero et al. 2003
<b>Fish</b>			
<i>Morone saxatilis</i> (striped bass)	0.002	11.8	Brown et al. 2005
<i>Sparisoma viride</i> (stoplight parrotfish)	0.019	12.4	Geertjes et al. 2004
<b>Insects</b>			
<i>Drosophila melanogaster</i> (fruit fly)	0.112	2.0	Singh and Rhomberg 1987
<i>Glossina pallidipes</i> (tsetse fly)	0.18	1.1	Ouma et al. 2005
<i>Heliconius charithonia</i> (butterfly)	0.003	79.8	Kronforst and Flemming 2001
<b>Corals</b>			
<i>Seriatopora hystrix</i>	0.089–0.136	2.6–1.6	Maier et al. 200

subpopulations and genetic drift causing subpopulations to diverge as they individually approach fixation or loss ( $N_e m$ ) in the context of the infinite island model. This relationship has been used in literally thousands of studies to estimate  $\widehat{N_e m}$  from empirical estimates of  $\hat{F}_{ST}$  in wild populations like the examples in Table 4.9. This equation (or expectations like it but based on different population models) is the basis of so-called indirect estimates of the number of effective migrants ( $\widehat{N_e m}$ ) that cause a given pattern of allele frequency differentiation among populations ( $\hat{F}_{ST}$ ).

It is important to recognize that employing equation 4.63 to estimate  $\widehat{N_e m}$  is really using the infinite island model as an ideal standard rather than actually estimating the long-term effective number of migrants for a specific population. Because of this dependence on the infinite island model, using  $\hat{F}_{ST}$  to obtain an estimate of  $\widehat{N_e m}$  should be interpreted as “the observed level of population differentiation ( $\hat{F}_{ST}$ )

would be *equivalent* to the differentiation expected in an infinite island model with a given number of effective migrants ( $N_e m$ ).” Such a comparison of actual and ideal populations is identical to that used in the definition of effective population size (see section 3.3). Despite this dependence on a highly idealized model, Slatkin and Barton (1989) concluded that using observed levels of population differentiation to estimate  $\widehat{N_e m}$  under island model assumptions should be roughly accurate, even when the actual population structure deviates from the island model. In contrast, Whitlock and McCauley (1999) review the many ways in which actual populations will deviate from the infinite island model and the assumptions used to approximate the relationship between  $\hat{F}_{ST}$  and  $N_e m$ , generally invalidating the indiscriminate use of equation 4.63.

The estimate of the effective number of migrants or  $\widehat{N_e m}$  obtained through the island model is referred to as an indirect estimate of the rate of gene flow. The

term indirect is used because the observed pattern of allele frequency differences among subpopulations is used in a model (containing many assumptions) to produce a parameter estimate. This is in contrast to a direct estimate of gene flow from a method like parentage analysis (although section 4.2 suggests direct methods also depend on assumptions). Such indirect estimates of gene flow have the effect of averaging across all of the past events that lead up to the current pattern of allele frequency differentiation among subpopulations. In contrast, direct estimates apply only to those periods of time when parentage or movement is observed. Slatkin (1987a) considers an example where mark–recapture methods suggest movement of a butterfly among different geographic locations is extremely limited, yet a multilocus estimate of  $\hat{F}_{ST}$  suggests almost no allele frequency differentiation among the butterfly populations. One possible explanation is that gene flow was extensive in the past and has very recently decreased, but not enough time has elapsed to witness increased population differentiation. Another possibility is that the infrequent gene-flow events required to prevent differentiation are not well measured by the mark–recapture technique.

### **Stepping-stone and metapopulation models**

Although the island model assumes that the amount of gene flow among all subpopulations is identical, a population organized into discrete subpopulations or demes can also experience isolation by distance. The stepping-stone model, inspired by the flat stones that form a walking path in a Japanese garden, approximates the phenomenon of isolation by distance among discrete subpopulations by allowing most or all gene flow to be only between neighboring subpopulations (Kimura 1953; see Fig. 4.12). This gene-flow pattern produces an allele-frequency clumping effect among the subpopulations qualitatively very similar to that seen in the first section of the chapter for isolation by distance in a continuous population of individuals (Fig. 4.3). A classic analysis of the stepping-stone model was carried out by Kimura and Weiss (1964), who showed that the correlation between the states of two alleles sampled at random from two subpopulations depends on (i) the distance between the subpopulations and (ii) the ratio of gene flow between neighboring colonies and long-distance gene flow where alleles are exchanged among subpopulations at random distances. As expected for isolation by distance, the correlation between allelic states decreases with increasing distance between

subpopulations. Interestingly, the correlation between allelic states drops off more rapidly with distance when subpopulations occupy two dimensions than when they occupy one dimension. In a two-dimensional stepping-stone model,  $F_{ST}$  is expected to grow like the logarithm of the number of colonies for fixed values of gene-flow parameters (see Cox & Durrett 2002; Slatkin & Barton 1989). Another way of saying this is that increasing levels of gene flow are required to maintain the same level of population structure as the number of colonies increases.

A logical extension of the stepping-stone model is the **metapopulation** model. Metapopulation models approximate the continual extinction and recolonization seen in many natural populations in addition to the process of gene flow. These models are motivated by organisms like pioneer plants and trees that colonize and grow in newly created clearings but eventually disappear from a patch as succession introduces new species and changes the environmental and competitive conditions. Even though each subpopulation of a pioneer species eventually goes extinct, there are other subpopulations in existence at any given time and new subpopulations are continuously being formed by colonization. A metapopulation is then just a collection of a number of smaller subpopulations or habitat patches (see various definitions of metapopulation and related concepts in Hanski & Simberloff 1997), conceptually similar to the stepping-stone model. However, in metapopulations the individual subpopulations have some probability of going extinct and these unoccupied locations that become available can also be colonized to found a new subpopulation.

Gene flow in metapopulation models is of two types. First, there is gene flow among the existing subpopulations like that in the continent-island or island models. Second, there is the gene flow that occurs when an open patch is colonized to replace a subpopulation that went extinct. The pattern of gene flow that takes place during colonization may take different forms (Slatkin 1977). The first form is where colonists are sampled at random from all subpopulations, called migrant-pool gene flow. The second form is where colonists are sampled at random from only a single random subpopulation, called propagule-pool gene flow. Migrant-pool gene flow is identical to the pattern of gene flow in the island model where migrants represent the average allele frequencies of all subpopulations. In contrast, propagule-pool gene flow can introduce a genetic bottleneck when a new subpopulation is founded

because the colonists are only drawn from a single existing subpopulation.

The impact of the form of colonization on heterozygosity in newly established subpopulations within a metapopulation is described by

$$F_{ST}^{colony} = \frac{1}{2k} + \phi \left( 1 - \frac{1}{2k} \right) F_{ST} \quad (4.64)$$

where  $F_{ST}^{colony}$  is the expected allele frequency differentiation in newly established subpopulations,  $k$  is the number of diploid colonists,  $F_{ST}$  is the degree of allele frequency differentiation among the existing subpopulations, and  $\phi$  is the probability that the two alleles in a newly established population come from the same or different subpopulations (Whitlock & McCauley 1990). Colonization corresponds to the propagule pool for  $\phi = 1$  (the chance of one that two founding alleles come from the same subpopulation) and the migrant pool for  $\phi = 0$  (no chance that two founding alleles come from the same subpopulation so all alleles must come from different subpopulations). In the equation, all newly founded subpopulations have a chance of being established with alleles that are identical by descent due to sampling from the total population, hence the  $\frac{1}{2k}$  term (see equation 3.47). For those subpopulations that are founded by individuals from a propagule pool (or  $\phi = 1$ ), the chances of alleles being identical by descent and homozygous is greater to the degree that existing subpopulations are differentiated in their allele frequencies. With colonization from the propagule pool, newly founded populations inherit the average level of homozygosity of existing subpopulations plus some additional homozygosity due to sampling from a finite population. With colonization from the migrant pool ( $\phi = 0$ ), founding alleles are always drawn from a different subpopulation, so the heterozygosity is the same as the total population heterozygosity ( $2\bar{p}\bar{q}$ ) except for sampling error from a finite number of founders.

The general conclusion is that extinction and recolonization can be an additional source of gene flow or an additional restriction on gene flow in metapopulations (Maruyama & Kimura 1980; Wade & McCauley 1988). Propagule pool colonization tends to increase overall population differentiation for all values of the number of diploid colonists ( $k$ ). In contrast, the change in overall differentiation with the migrant model depends on the rate of gene flow among existing subpopulations. When the number

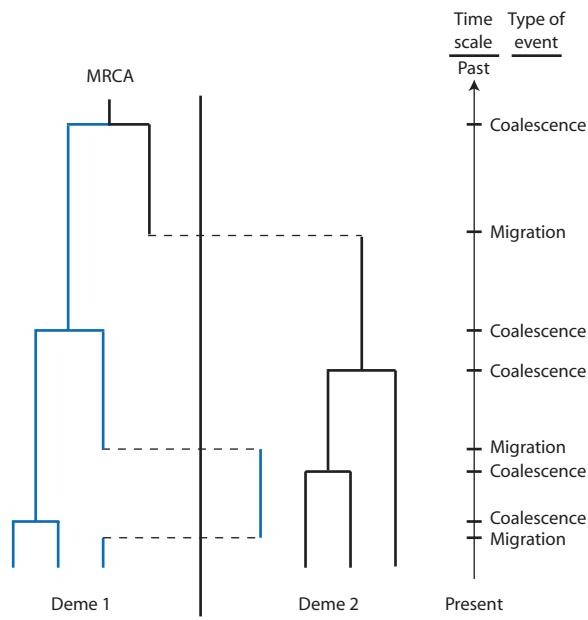
of diploid colonists ( $k$ ) exceeds twice the effective number of migrants ( $2N_e m$ ) then differentiation tends to decrease since colonization accomplishes additional mixing of alleles. Using newly established populations of the plant *Silene alba*, McCauley et al. (1995) estimated  $\phi$  between 0.73 and 0.89, suggesting that new populations do experience some additional sampling during their formation that increases population differentiation.

#### 4.6 The impact of population structure on genealogical branching

- Bugs in many boxes.
- Event times with population subdivision.
- Sample configurations.
- Mean and variance of waiting time in two demes.

In structured populations with gene flow, lineages can move from deme to deme. In a retrospective view, two lineages sampled in the present can experience either coalescence or migration going back in time (Fig. 4.16). Determining the mean and variance of time to coalescence in structured populations will show the overall impact of population structure on genealogical trees. In particular, we would like to know whether population structure will alter the average and variance of the height of genealogical trees in comparison with the basic coalescent process in a single panmictic population. We will again utilize the properties of the exponential distribution to approximate the time to an event (see section 3.6).

Let's begin by thinking about the coalescent process when there is gene flow among several demes in terms of the bugs-in-a-box metaphor used to describe the basic coalescent process. With population subdivision the bugs are located in multiple boxes with each box representing a deme. Bugs move about within a box at random and eat each other, reducing their numbers. There is also the possibility of migration where a bug is chosen at random and moved to another box. If migration events are very rare, then the individual boxes have a good chance of being reduced to a single bug before a migrant bug enters or leaves the box. It will then take a long time for enough migration events to happen such that the entire group of boxes is reduced to a single bug. When migration events are common, migrant bugs move among the boxes frequently and the boxes are effectively interconnected. Therefore, there should be little or no time spent waiting for migration events as the bugs in all the boxes eat their way to a single bug.



**Figure 4.16** A hypothetical genealogy for two demes. Initially there are three lineages in each deme. The very first event going back in time is the migration of a lineage from deme one into deme two. Immediately after this migration occurs, the chance of coalescence in deme two increases since there are more lineages and the chance of coalescence in deme one decreases since there are fewer lineages. Continuing back in time, a coalescence event occurs in deme one and then a coalescence event occurs in deme two. The lineage that migrated out of deme one migrates back into deme one by chance. Coalescence to the single most recent common ancestor (MRCA) of all lineages cannot occur until the final two lineages are brought together in a single deme by migration.

### Combining coalescent and migration events

Describing genealogies with gene flow can be accomplished by adding another type of possible event that can occur working from the present to a time in the past where all lineages find their most recent common ancestor. We will assume that both coalescent and migration events are rare (or that  $N_e$  is large and the rate of migration is small), so that when an event does occur going back in time it is either coalescence or migration. In other words, we will assume that migration and coalescence events are mutually exclusive. The fact that events are mutually exclusive is an important assumption. When two independent processes are operating, the coalescence model becomes one of following lineages back in time and waiting for an event to happen. When events are independent but mutually exclusive, the probability of each event is added over all possible events to obtain the total chance that an event occurs. For example,

the chance that a diploid genotype for a diallelic locus is a heterozygote under random mating is  $2pq$ . This is the sum of the independent chance of sampling  $Aa$  and the chance of sampling  $aA$  since a heterozygote results from one of the two ways of sampling of two different alleles (the probability of a heterozygote under random mating is not  $(pq)^2$ , which is the chance of sampling  $Aa$  and  $aA$  simultaneously). Therefore, if we can find an exponential approximation for the chance that a lineage migrates to a different deme each generation, we can just add this to the exponential approximation for the chance of coalescence.

In a subdivided population, each generation there is the chance that a lineage in one deme migrates to some other deme. The rate of migration,  $m$ , is the chance that a lineage migrates each generation. The chance that a lineage does not migrate is therefore  $1 - m$  each generation. The chance that  $t$  generations pass before a migration event occurs is then the product of the chances of  $t - 1$  generations of no migration followed by a migration, or

$$P(T_{migration} = t) = (1 - m)^{t-1}m \quad (4.65)$$

This is in an identical form to the chances that a coalescent event occurs after  $t$  generations given in Chapter 3. Like the probability of coalescence, the probability of a migration through time is a geometric series that can be approximated by the exponential distribution (see Math box 3.2). To obtain the exponent of  $e$  (or the intensity of the migration process), we need to determine the rate at which migration is expected to occur in a population.

Now consider migration events in the context of an island model of gene flow where there are  $d$  demes and each deme contains  $2N_e$  lineages. The total population size is the sum of the sizes of all demes or  $2N_e d$  lineages. When time is measured on a continuous scale with  $t = \frac{j}{2N_e d}$ , one unit of time is equivalent to  $2N_e d$  generations. If  $2N_e d$  generations elapse and  $m$  is the chance of migration per generation, then  $2N_e d m$  migration events are expected in the total population during one unit of continuous time. If we define  $M = 4N_e m$ , then  $M/2$  is equivalent to  $2N_e m$  or the chance that a lineage in one deme migrates (the per deme migration rate). The chance of migration is independent in all of the demes, so the expected number of migration events in the total population is the sum of the per deme chances of migration or  $\frac{M}{2}d$ . This leads to the exponential

approximation for the chances that a single lineage in any of the demes migrates at generation  $t$ :

$$P(T_{migration} = t) = e^{-t \frac{Md}{2}} \quad (4.66)$$

on a continuous time scale. When there is more than one lineage, each lineage has an independent chance of migrating but only one lineage will migrate. So we

add the  $e^{-t \frac{Md}{2}}$  chance of migration for each lineage over all  $k$  lineages to obtain the total chance of migration:

$$P(T_{migration} = t) = e^{-t \frac{Md}{2}} \quad (4.67)$$

for  $k$  ancestral lineages of the  $d$  demes. The chance that one of  $k$  lineages migrates at or before a certain time can then be approximated with the cumulative exponential distribution:

$$P(T_{migration} \leq t) = 1 - e^{-t \frac{Md}{2}} \quad (4.68)$$

in exactly the same fashion that times to coalescent events are approximated.

When two independent processes are operating, the genealogical model becomes one of following lineages back in time and waiting for an event to happen. The possible events in this case are migration or coalescence, so the total chance of *any* event is the sum of the independent probabilities of each type of mutually exclusive event. Since lineages cannot coalesce unless they are in the same deme, the chance of a coalescent event is

$$P(T_{event} \leq t) = 1 - e^{-td \sum_{i=1}^d \frac{k_i(k_i-1)}{2}} \quad (4.69)$$

when there are  $k_i$  ancestral lineages in deme  $i$ , a slightly modified version of the basic coalescent model that takes into account the  $d$  demes and time scaling by  $2N_e d$ . (Note that when  $d = 1$  the expected time to coalescence reduces to  $\frac{k(k-1)}{2}$  on a continuous time scale.) The total chance of an event occurring when going back in time (increasing  $t$ ) is then

$$P(T_{event} \leq t) = 1 - e^{-t \left[ dk \frac{M}{2} + d \sum_{i=1}^d \frac{k_i(k_i-1)}{2} \right]} \quad (4.70)$$

where the exponent is the sum of the intensities of migration and coalescence. In the simplest case of

two demes ( $d = 2$ ) with  $k_1$  and  $k_2$  ancestral lineages in each deme, equation 4.70 reduces to

$$P(T_{event} \leq t) = 1 - e^{-t \left[ (k_1+k_2) \frac{M}{2} + \frac{k_1(k_1-1)}{2} + \frac{k_2(k_2-1)}{2} \right]} \quad (4.71)$$

(the example given in Hudson 1990) where time is scaled in units of the total population size  $2N_e d$  or the sum of number of lineages in all of the demes.

When an event does occur at a time given by this exponential distribution in equation 4.70, it is then necessary to decide whether the event is a coalescence or a migration. The total chance that the event is either a migration or a coalescence event is  $dk \frac{M}{2} + d \sum_{i=1}^d \frac{k_i(k_i-1)}{2}$ . Therefore the chance the event is a migration is

$$\frac{dk \frac{M}{2}}{dk \frac{M}{2} + d \sum_{i=1}^d \frac{k_i(k_i-1)}{2}} = \frac{kM}{k(M-1) + \sum_{i=1}^d k_i^2} \quad (4.72)$$

whereas the chances that the event is a coalescence is

$$\frac{d \sum_{i=1}^d \frac{k_i(k_i-1)}{2}}{dk \frac{M}{2} + d \sum_{i=1}^d \frac{k_i(k_i-1)}{2}} = \frac{\sum_{i=1}^d (k_i^2 - k_i)}{k(M-1) + \sum_{i=1}^d k_i^2} \quad (4.73)$$

When the event is a coalescence, the deme is picked at random given that demes with more ancestral lineages have a greater chance of experiencing a coalescent event (the chance that deme  $j$  experiences

$$\text{the coalescence is } \frac{\frac{k_j(k_j-1)}{2}}{\sum_{i=1}^d \frac{k_i(k_i-1)}{2}} = \frac{k_j(k_j-1)}{\sum_{i=1}^d k_i(k_i-1)}.$$

Figure 4.19 shows two realizations of the combined coalescence and migration process when the migration rate is either relatively high or relatively low. The times to each event are determined by the exponential distributions specified by equation 4.70.

### **The average length of a genealogy with migration**

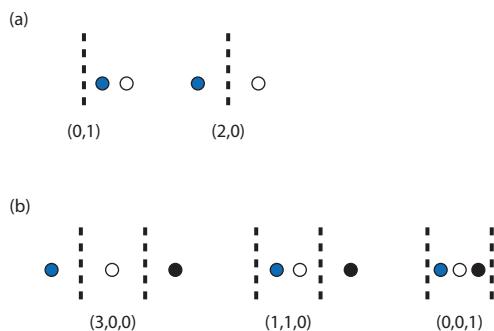
Before determining the average time to coalescence for a genealogy in a structured population, it is first necessary to introduce some notation that is useful to describe the different possible locations of lineages

### Interact box 4.6 Coalescent events in two demes

A coalescent genealogy that includes the possibility of migration among demes can be constructed using the cumulative exponential distribution specified by equation 4.70 to determine the waiting time to an event. Once the waiting time is obtained, then determining whether each event is a migration or a coalescence can be accomplished with equations 4.72 or 4.73. If the event is coalescence, a random pair of lineages in a random deme is picked to coalesce and the number of ancestral lineages in that deme ( $k_i$ ) is reduced by 1. If the event is migration, a random lineage is picked and moved into a random deme. It is possible to construct a coalescent genealogy that includes the possibility of mutations occurring along each branch.

A natural way to see the results of these steps is to simulate genealogies in a simple case with two demes. Using the link on the text web page, go to the BiRC Animators web page and click on the **Migration** link under the heading **Hudson Animator**. This will take you to a web page containing a Java simulation. There are three parameters that can be set in the simulation: **n**: sets the number of lineages sampled in *both* demes in the present time (or  $k_1 + k_2$  in equation 4.71), while **M1:** and **M2:** set the expected number of migrants in deme 1 and deme 2 each time period (or  $M$  in equations 4.68 and 4.70). Pressing **Recalc** will calculate the waiting times for a new genealogy. The animation process can be controlled with the buttons below the figure. Waiting times can be seen in at the lower right when the pointer is placed over a circle in the tree. Click on the **Trees** tab at the top left to see how population structure impacts the genealogical tree itself.

Initially set **n** to 10 and both **M1** and **M2** to the low migration rate of 0.1. Simulate 10 independent trees, in each case recording the number of migration events (the light blue circles in the animation) and the total waiting time until coalescence to a single most recent common ancestor. Increase both **M1** and **M2** to a higher migration rate of 1.0 and again simulate 10 independent trees and record the number of migration events and total waiting time to an MRCA. How do the genealogies compare on average when migration rates are lower or higher?



**Figure 4.17** Sample configurations for two lineages and two demes (a) and three lineages and three demes (b). Lineages are represented by the circles and the separation between demes is represented by a dotted vertical line. Only one possibility is given for each sample configuration even though some configurations can occur in multiple ways. For example, (0,1) can occur when both lineages are in the left-hand deme or when both lineages are in the right-hand deme.

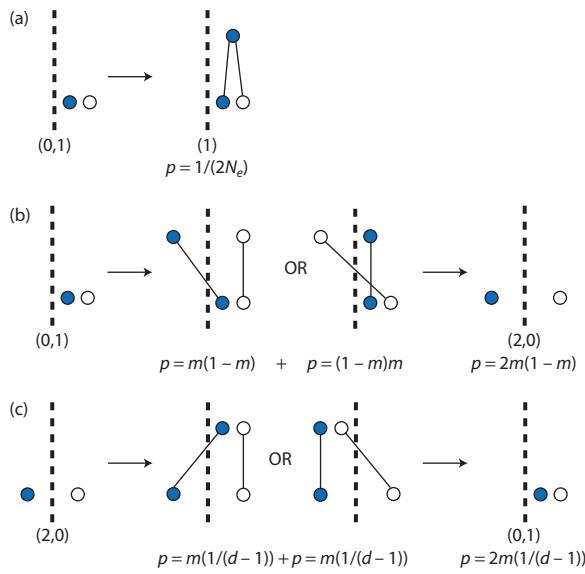
within and among demes. We can define a list (or row vector) that tracks the way lineages are partitioned among all demes as

$$d = (d_1, d_2, d_3, \dots, d_n) \quad (4.74)$$

where each  $d_i$  is the number of demes with  $i$  lineages and  $n$  is the total number of demes. The total number of lineages is then the product of the number of demes containing  $i$  lineages and the number of lineages  $i$  summed over all possible numbers of lineages

per deme or  $\sum_{i=1}^n id_i$ . With a sample of two lineages

taken from a total population composed of two demes, there are two possible ways the lineages could be sampled. The two lineages could either be sampled from different demes to give  $d = (2,0)$  or sampled from a single deme to give  $d = (0,1)$ . This notation specifies what is called the **sample configuration** of a number of lineages drawn from some number of demes. Figure 4.18 gives several examples of sample configurations for two or three demes. With coalescence to a single ancestral lineage the sample configuration becomes (1). This sample configuration notation is useful because the mean and variance of coalescence times in a structured population depends on whether lineages are located in the same or different demes.



**Figure 4.18** The possible events that can occur when two lineages are in the same deme (0,1) or when two lineages are in two different demes (2,0) along with their probabilities of occurring. The separation between demes is represented by a dotted vertical line. Two lineages can coalesce only when they are in the same deme. The probability of coalescence (a), migration of one lineage such that the two lineages are in different demes (b), and migration that places both lineages in the same deme (c) determine the overall chances that two lineages coalesce. The chance that both lineages migrate (with probability  $m^2$ ) is not shown in (b) and applies when there are three or more demes.

With that background on sample configurations, let's move on to derive the average coalescence time and expected total length of a genealogy in a structured population. We will focus on the simplest case of two lineages in the context of two demes. We need to determine the chances that two lineages in either of the two possible sample configurations ((2,0) or (0,1)) experience coalescence. Figure 4.18 shows these possible transitions between sample configuration states. As in the basic coalescent process, the chance of coalescence is the product of one over the population size and the number of unique pairs of lineages that can coalesce. If each deme contains  $2N_e$  lineages the probability of coalescence is  $\frac{1}{2N_e}$  for two lineages in one deme. However, two lineages cannot coalesce unless they are together in the same deme and restricted gene flow will make this less likely to happen.

For two lineages that are together in the same deme, or in sample configuration (0,1), there are two possible events that eventually lead to coalescence. The first possible event is simply that the two lineages

coalesce with probability  $\frac{1}{2N_e}$ . The second possible event is that one or both of the two lineages migrates into another deme before they can coalesce. If the proportion of migrants per generation in any deme is  $m$  then the chance that a single lineage is an emigrant is  $m$  and the chance that it is not an emigrant is  $1 - m$ . The chance that one lineage migrates and the other lineage does not is  $m(1 - m) + (1 - m)m = 2m(1 - m)$ . The chance that both lineages migrate is  $m^2$ . The total chance that one or both lineages migrate is then  $2m(1 - m) + m^2$ , which is approximately  $2m$  if  $m$  is small and  $m^2$  terms can be ignored. For two lineages in the same deme or (0,1), the total chance that *any* event occurs in the previous generation, either coalescence or migration, is therefore  $2m + \frac{1}{2N_e}$ .

For two lineages that are in different demes, or in sample configuration (2,0), the total chance that one lineage migrates is  $2m$ , following the same logic as when two lineages are in a single deme. However, to transition from (2,0) to (0,1), the migration event is not into any random deme but must be into the one other deme where the second lineage is found. The chance that migration into a specific deme occurs

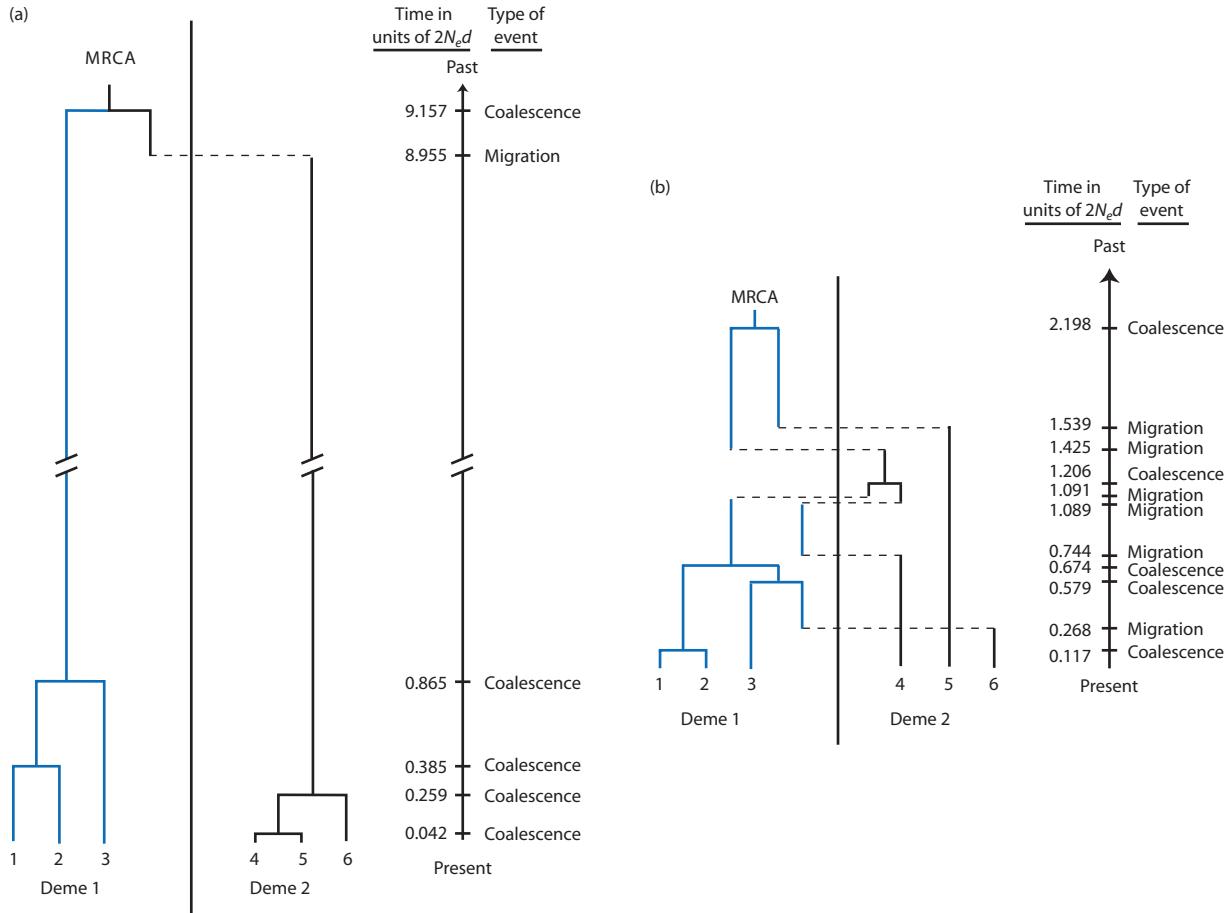
is  $\frac{1}{d-1}$  where  $d$  is the number of demes. The total

probability that two lineages initially in separate demes end up in a single deme with the possibility that they

can later coalesce is therefore  $2m\frac{1}{d-1}$ .

To determine the average time to coalescence in two demes we can use the fact that the average time to an event is one over the probability of each event in a process where waiting times are exponentially distributed. Let  $\bar{T}_{(0,1)}$  represent the average time until coalescence for two lineages in the same deme and  $\bar{T}_{(2,0)}$  the average time to coalescence for two lineages in different demes. For two lineages in the same deme, the average time to coalescence is the average time to coalescence plus the average time spent in two different demes if there is a migration event. The average time either of coalescence or migration is one over the total chance of an event, or  $\frac{1}{2m + \frac{1}{2N_e}}$ .

When an event occurs, there is a  $\frac{1}{2N_e}$  chance that it is a coalescence and a  $2m$  chance that it is a migration event. Bringing this all together leads to an expression



**Figure 4.19** Genealogies for six lineages initially divided evenly between two demes when the migration rate is low (a) and when the migration rate is high (b). When migration is unlikely, coalescent events within demes tend to result in a single lineage within all demes before any migration events take place. There is then a long wait until a migration event places both demes in one deme where they can coalesce. When migration is likely, lineages regularly move between the demes, and lineages originally in the same deme are as likely to coalesce as lineages initially in different demes. These two genealogies are examples and substantial variation in coalescence times is expected. In (a)  $M = 4N_e m = 0.2$  and in (b)  $M = 4N_e m = 2.0$ . The two genealogies are not drawn to the same scale. MRCA, most recent common ancestor.

for the average time to coalescence for two lineages in the same deme:

$$\bar{T}_{(0,1)} = \frac{1}{2N_e} + \frac{2m}{2m + \frac{1}{2N_e}} \bar{T}_{(2,0)} \quad (4.75)$$

For two lineages in different demes, the average time to coalescence is the sum of the average time needed to migrate into the same deme and the average time until coalescence once the lineages are in the same deme. Since the chance of migration into the same deme is  $2m \frac{1}{d-1}$ , the average time for two lineages to migrate into the same deme is  $\frac{d-1}{2m}$ . The average

time to coalescence for two lineages in different demes this then

$$\bar{T}_{(2,0)} = \frac{d-1}{2m} + \bar{T}_{(0,1)} \quad (4.76)$$

Solving these two equations (see Math box 4.2) leads, respectively, to:

$$\bar{T}_{(0,1)} = 2N_e d \quad (4.77)$$

and

$$\bar{T}_{(2,0)} = 2N_e d + \frac{d-1}{2m} \quad (4.78)$$

(see Slatkin 1987b; Strobeck 1987; Nordborg 1997; Wakeley 1998).

### Math box 4.2 Solving two equations with two unknowns for average coalescence times

We can restate equations 4.75 and 4.76 as

$$\bar{T}_{(0,1)} = x = a + by \quad (4.79)$$

$$\bar{T}_{(2,0)} = y = c + x \quad (4.80)$$

where  $a = \frac{1}{2N_e}$ ,  $b = \frac{2m}{2m + \frac{1}{2N_e}}$ , and  $c = \frac{d-1}{2m}$ . When time is scaled in units of  $2N_e$  then  $a = \frac{1}{2m + \frac{1}{2N_e}}$ .

We can then substitute the equation for  $x$  into the equation for  $y$  to get

$$y = c + a + by \quad (4.81)$$

which rearranges to

$$y - by = c + a \quad (4.82)$$

and then

$$y = \frac{c + a}{1 - b} \quad (4.83)$$

Substituting the values for  $a$ ,  $b$ , and  $c$  gives

$$y = \frac{\frac{d-1}{2m} + \frac{1}{2m + \frac{1}{2N_e}}}{1 - \left( \frac{2m}{2m + \frac{1}{2N_e}} \right)} \quad (4.84)$$

The denominator above can be rearranged to

$$\frac{2m + \frac{1}{N_e}}{2m + \frac{1}{N_e}} - \frac{2m}{2m + \frac{1}{N_e}} = \frac{\frac{1}{2N_e}}{2m + \frac{1}{2N_e}}. \text{ Next let}$$

$f = 2m + \frac{1}{2N_e}$  and then substitute it into the

equation 4.82 with the rearranged denominator to get

$$y = \frac{\frac{d-1}{2m} + \frac{1}{f}}{\frac{1}{\frac{2N_e}{f}}} \quad (4.85)$$

which when the numerator and denominator are multiplied by  $f$  gives

$$y = \frac{f \frac{d-1}{2m} + 1}{\frac{1}{2N_e}} \quad (4.86)$$

Substituting the full expression for  $f$  and then expanding gives

$$y = \frac{\left( 2m + \frac{1}{2N_e} \right) \frac{d-1}{2m} + 1}{\frac{1}{2N_e}} \quad (4.87)$$

$$= \frac{\frac{1}{2N_e} \left( \frac{2N_e 2m(d-1) + (d-1) + 2N_e 2m}{2m} \right)}{\frac{1}{2N_e}}$$

and then multiplying by  $2N_e$  instead of dividing by  $\frac{1}{2N_e}$  cancels the  $\frac{1}{2N_e}$  term in the numerator, and then expanding gives

$$y = \frac{2N_e 2md}{2m} - \frac{2N_e 2m}{2m} + \frac{d-1}{2m} + \frac{2N_e 2m}{2m} \quad (4.88)$$

which after addition and canceling terms finally gives

$$y = 2N_e d + \frac{d-1}{2m} \quad (4.89)$$

Equations 4.76 or 4.80 for  $\bar{T}_{(2,0)}$  can then be solved by substituting this expression for  $y$  and similar methods of algebraic rearrangement.

These average times to coalescence for two lineages in the context of two demes are both simple expressions that are easy to interpret. Equation 4.77 is a bit surprising since it says that the average time to coalescence for two lineages in the same deme is independent of the migration rate and is simply a function of the total population size as in a panmictic population (note that if each of  $d$  demes contains  $2N_e$  lineages the total population size is  $N_T = 2N_e d$ ). We can understand why this is the case by imagining what happens as the migration rate changes. If the migration rate decreases, the probability that a lineage migrates into another deme decreases with the effect of shortening the time to coalescence. However, in those cases when a migration event does occur, the lineage would take a longer time to migrate back before coalescence. The average time to coalescence is independent of the migration rate since these two factors exactly balance as the migration rate changes. When two lineages are in different demes the average time to coalescence increases as the migration rate decreases and as the number of demes increases. The average coalescence time is inversely proportional to the migration rate since migration is required to put two lineages into the same deme by chance. As the number of demes increases there are an increasing number of places for two lineages to be apart so that more migration events will have to occur until two lineages are together in the same deme.

Average coalescence times within demes and in the total population can also be used to express the degree of population structure. Earlier in the chapter we used probabilities of autozygosity to express population structure as a difference between the chance that two alleles sampled from the total population are different in state ( $H_T$ ) and the chance that two alleles sampled from a subpopulation are different in state ( $H_S$ ), or  $F_{ST} = \frac{H_T - H_S}{H_T}$ . For two lineages drawn at random from the total population of  $d$  demes there is a  $\frac{1}{d}$  chance they are from the same deme and a  $\frac{d-1}{d}$  chance they are from different demes. Therefore, the average coalescence time for two lineages sampled at random from a subdivided population is

$$\bar{T} = \frac{1}{d} \bar{T}_{(0,1)} + \frac{d-1}{d} \bar{T}_{(2,0)} = 2N_e d + \frac{(d-1)^2}{2md} \quad (4.90)$$

and equation 4.77 provides the average coalescence time for two lineages sampled from the same deme

( $\bar{T}_{(0,1)} = 2N_e d$ ). Putting these two average coalescence times together,

$$F_{ST} = \frac{\bar{T} - \bar{T}_{(0,1)}}{\bar{T}} \quad (4.91)$$

gives an expression for the pattern of population structure from the perspective of coalescence times (Slatkin 1991). Population structure can then be thought of as the difference in average coalescence times for a pair of lineages in the total population and a pair of lineages in a subpopulation.

In general, population subdivision is expected to increase the time required for lineages to coalesce to a single most recent common ancestor. When gene flow is limited, coalescent events within demes occur much as they would in an isolated panmictic population. However, the single ancestor for each deme must wait for a relatively rare migration event until two lineages in different demes can coalesce to a single ancestor. This tends to produce genealogical trees that have long branches connecting the individual ancestors of different demes. As rates of migration increase, the genealogical tree branch lengths approach the patterns expected in a single panmictic population of the same total size since migration events frequently move lineages among the demes.

## Chapter 4 review

- Spatial and temporal separation of discrete subpopulations as well as isolation by distance in continuous populations both result in mating that is not random throughout a population. Without enough gene flow to maintain random mating (panmixia), genetic drift causes divergence of allele frequencies among subpopulations.
- Parentage analyses use genotypes of progeny and one known parent to infer the haplotype of the unknown parent. This unknown parent haplotype is then used to exclude possible parents from the pool of candidate parents. The power of this procedure to identify the true parent depends on the chance that a given haplotype will occur at random in a population.
- The Wahlund effect demonstrates that genetic variation can be stored as variance in allele frequencies among subpopulations or as heterozygosity within a panmictic population. Fusion of diverged subpopulations or subdivision of a

panmictic population converts one type of genetic variation into the other type of genetic variation.

- $F_{IS}$  measures the average excess or deficit of heterozygous genotypes compared with random mating.  $F_{ST}$  measures the deficit of heterozygosity in subpopulations due to population structure compared to heterozygosity expected with panmixia.  $F_{IT}$  measures the total excess or deficit of heterozygous genotypes due to both non-random mating within and allele-frequency divergence among subpopulations.
- Levels of gene flow can be measured by directly tracking parentage in contemporary populations (a direct estimate) or by measuring the pattern of allele-frequency differentiation among a set of subpopulations and then comparing the result to what is expected in an ideal standard such as the infinite island model (an indirect estimate).
- Genealogical trees in subdivided populations can be modeled with an exponentially distributed waiting time where the chance of migration and the chance of coalescence are combined.
- In two demes, the average time to coalescence for two lineages in the same deme is the total population size and is independent of the migration rate. For two lineages in different demes, the average time to coalescence gets longer as the number of demes increases and as the migration rate decreases, since two lineages can only coalesce when they are in the same deme.
- Population structure and limited gene flow lengthen the average coalescence time of two lineages sampled at random from the population compared to the average coalescence time of two lineages sampled from the same subpopulation.

### Further reading

A review of spatial patterns of genetic variation within and among populations, methods to measure spatial aspects of genetic variation, and discussion of the processes causing these patterns can be found in

Epperson BK. 2003. *Geographical Genetics*. Princeton University Press, Princeton, NJ.

To learn more about the role that the plant *Linanthus parryae* played in the development of the theory of isolation by distance, as well as the personalities associated with competing interpretations of the spatial distributions of blue and white flower colors, see chapters 11 and 13 in

Provine WB. 1986. *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago, IL.

A review of probability theory for parentage assignment along with a detailed listing of available analysis software available can be found in

Jones AG and Ardren WR. 2003. Methods of parentage analysis in natural populations. *Molecular Ecology* 12: 2511–23.

An older yet still valuable review of concepts and empirical estimates of population structure and indirect estimates of gene flow can be found in

Slatkin M. 1985. Gene flow in natural populations. *Annual Review of Ecology and Systematics* 16: 393–430.

The impacts of population subdivision or isolate breaking on genotype frequencies is more complicated for loci with more than two alleles. For a treatment of this topic consult:

Li CC. 1969. Population subdivision with respect to multiple alleles. *Annals of Human Genetics* 33: 23–9.

A review of the conceptual bases and methodological approaches to estimation of population structure using analysis of variance can be found in

Weir BS. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

A review of the impacts of population structure in the context of the coalescent model can be found in

Charlesworth B, Charlesworth D, and Barton NH. 2003. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology and Systematics* 34: 99–125.

## Problem box answers

### Problem box 4.1 answer

The allele frequencies for each of the alleles in the paternal haplotype are obtained in Table 4.3. For tree 4865 there is only one possible paternal allele at each locus. The chance of any genotype having one copy of each paternal allele at each locus is:

- A:  $(0.1216)^2 + 2(0.1216)(1 - 0.1216) = 0.2284$
- B:  $(0.3971)^2 + 2(0.3971)(1 - 0.3971) = 0.6365$
- C:  $(0.0761)^2 + 2(0.0761)(1 - 0.0761) = 0.1464$
- D:  $(0.1905)^2 + 2(0.1905)(1 - 0.1905) = 0.3447$
- E:  $(0.1250)^2 + 2(0.1250)(1 - 0.1250) = 0.2344$

The paternal allele is expected to occur in between 14 and 64% of possible genotypes for any individual locus. The probability of a random match at all five loci is  $0.2284 \times 0.6365 \times 0.1464 \times 0.3447 \times 0.2344 = 0.0017$ , or in 17 out of 10,000 random genotypes. The probability of exclusion is then  $1 - 0.0017 = 0.9983$  while the probability of exclusion for a sample of 30 candidate parents is  $(0.9983)^{30} = 0.9502$ . There is about a 95% chance that there would not be a random match in a sample of 30 candidate parents; therefore, we have high confidence that 4865 is the true father of seed 25-1 from tree 989. For this offspring–maternal parent combination, the B locus is the least useful in resolving paternity since the frequency of the 106 allele is almost 40%. The 167 allele at the C locus is the most useful with a frequency of just over 7%.

### Problem box 4.2 answer

$H_I = \frac{1}{n} \sum_{i=1}^n \hat{H}_i$ , where  $\hat{H}$  is the observed frequency of heterozygotes in each of the  $n$  subpopulations.

$$H_I = (0.0 + 0.14 + 0.34 + 0.40)/4 \\ H_I = 0.22$$

The average observed heterozygote frequency is 0.22 or 22%.

$H_S = \frac{1}{n} \sum_{i=1}^n 2p_i q_i$  where  $p_i$  and  $q_i$  are the allele frequencies in subpopulation  $i$ .

$$H_S = (2(0.0)(1.0) + 2(0.93)(0.07) + 2(0.17)(0.83) + 2(0.51)(0.49))/4$$

$$H_S = (0.0 + 0.1302 + 0.2822 + 0.4998)/4$$

$$H_S = 0.228$$

$H_T = 2\bar{p}\bar{q}$  where  $\bar{p}$  and  $\bar{q}$  are average allele frequencies for all the subpopulations. Let  $f$  be the frequency of the fast allele and  $s$  the frequency of the slow allele so that  $f + s = 1$ . Then estimate the average allele frequency for the fast allele in the total population (the slow allele could be averaged as well):

$$\bar{f} = (0.0 + 0.93 + 0.17 + 0.51)/4$$

$$\bar{f} = 0.4025$$

whereas the frequency of the other allele is found by subtraction:  $\bar{s} = 1 - 0.4025 = 0.5975$ .

$$H_T = 2(0.4025)(0.5975)$$

$$H_T = 0.481$$

We can now calculate the  $F$  statistics using  $H_I$ ,  $H_S$ , and  $H_T$ .

$$F_{IS} = \frac{H_S - H_I}{H_S}$$

$$F_{IS} = (0.228 - 0.220)/0.228$$

$$F_{IS} = 0.035$$

There is no evidence for self-fertilization since these four populations have observed heterozygosities very close to that expected under random mating. Comparing the observed and expected heterozygosity for each population shows that subpopulations 9 and 43 have a slight excess of heterozygotes while subpopulation 68 has about a 10% deficit. These three deviations along with the zero deviation in subpopulation 1 all average out to approximately zero.

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$$F_{ST} = (0.481 - 0.228)/0.481$$

$$F_{ST} = 0.526$$

There is less heterozygosity within the subpopulations than we expect under Hardy–Weinberg based on allele frequencies for the total population. This value reflects the substantial differences in subpopulation allele frequencies.

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

$$F_{IT} = (0.481 - 0.220)/0.481$$

$$F_{IT} = 0.543$$

This is the deficit of heterozygosity caused by both non-random mating within populations and allele-frequency divergence among subpopulations. In this case almost all of the deficit in heterozygosity is due to allele frequency divergence among the subpopulations.

The three fixation measures are related by

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

Using the values for  $F_{IS}$  and  $F_{ST}$  and then solving for  $F_{IT}$  gives the same value that was determined by direct computation:

$$(1 - F_{IT}) = (1 - 0.035)(1 - 0.526)$$

$$(1 - F_{IT}) = (0.965)(0.474)$$

$$(1 - F_{IT}) = 0.4574$$

$$F_{IT} = 0.543$$

Based on the data from all 43 subpopulations, Levin (1978) estimated  $F_{IS} = 0.70$ ,  $F_{ST} = 0.80$ , and  $F_{IT} = 0.80$  in *P. cuspidata*.

### Problem box 4.3 answer

The Wahlund effect shows that population structure causes heterozygotes to become less

frequent and homozygotes to become more frequent by a factor proportional to the amount of allele frequency divergence among populations. Using the allele frequencies in Table 2.3 we can calculate the expected genotype frequencies for each locus with adjustment for population using equation 4.37 for homozygous loci and equation 4.38 for heterozygous loci:

D3S1358	$2(0.2118)(0.1626)(0.95) = 0.0655$
D21S11	$2(0.1811)(0.2321)(0.95) = 0.0799$
D18S51	$(0.0918)^2 + (0.0918)(0.9082)(0.05) = 0.0126$
vWA	$(0.2628)^2 + (0.2628)(0.7372)(0.05) = 0.0788$
FGA	$2(0.1378)(0.0689)(0.95) = 0.0181$
D8S1179	$2(0.3393)(0.2015)(0.95) = 0.1299$
D5S818	$2(0.3538)(0.1462)(0.95) = 0.0942$
D13S317	$2(0.0765)(0.3087)(0.95) = 0.0448$
D7S820	$2(0.2020)(0.1404)(0.95) = 0.0539$

Assuming that the Amelogenin locus should not be affected by population structure, the expected frequency of the 10-locus genotype after adjustment for population structure is  $0.0655 \times 0.0799 \times 0.0126 \times 0.0788 \times 0.0181 \times 0.1299 \times 0.0942 \times 0.0448 \times 0.0539 \times 0.5 = 1.514 \times 10^{-12}$  with an odds ratio of one in 660,501,981,506. Compare that with the expected genotype frequency of  $1.160 \times 10^{-12}$  and odds ratio of one in 862,379,847,814 assuming panmixia. After accounting for population structure this genotype appears more likely to occur by chance, although its expected frequency is still extremely rare.

### Problem box 4.4 answer

The joint effects of drift and migration in the fixation index are expressed by

$$F_t = \frac{1}{2N_e}(1-m)^2 + \left(1 - \frac{1}{2N_e}\right)F_{t-1}(1-m)^2$$

This can be made more general if  $x$  is used to represent the probability of autozygosity and  $y$  is used to represent the probability of allozygosity ( $y = 1 - x$ ):

$$F_t = x(1-m)^2 + yF_{t-1}(1-m)^2$$

For the case of a diploid nuclear locus we used  $x = \frac{1}{2N_e}$  and  $y = 1 - \frac{1}{2N_e}$  to obtain the relationship between  $F_{ST}$  and  $N_e m$  at equilibrium. Both Y-chromosome and organelle loci are haploid and uniparentally inherited, giving an effective population size of one-quarter of that for nuclear loci. For example, in humans the mitochondrial genome is inherited from the maternal parent only, or half of the population, and it is also haploid or present in half the number of copies of the nuclear genome. For

such loci we would use  $x = \frac{1}{N_e} = \frac{2}{N_e}$  and

$y = 1 - \frac{2}{N_e}$  to obtain

$$F_{ST} \approx \frac{1}{N_e m + 1}$$

The result is that  $F_{ST}$  is expected to be higher for Y-chromosome and organelle loci because their effective population size is smaller (see Fig. 4.15). Compared to diploid nuclear loci, Y-chromosome and mitochondrial loci have levels of  $F_{ST}$  four-fold higher when all types of loci share a common migration rate. The greater level of divergence among subpopulations for the Y-chromosome and organelle loci comes about strictly because of differences in the autozygosity for the loci that produced increased rates of fixation or loss due to genetic drift. See Hu and Ennos (1999) and Hamilton and Miller (2002) for more details and references.

# CHAPTER 5

## Mutation

### 5.1 The source of all genetic variation

- Types of mutations and rates of mutation.
- How can a low-probability event like mutation account for genetic variation?
- The spectrum of fitness for mutations.

The previous four chapters have discussed in detail genotype frequencies under random and non-random mating, the relationship between genetic drift and the effective population size, as well as population subdivision and gene flow. These and all other processes in populations act to shape or change existing genetic variation. But where does genetic variation come from in the first place? The Hardy–Weinberg expectation shows clearly that particulate inheritance itself does not alter genotype or allele frequencies and so it is not a source of genetic variation. Any form of non-random mating alters only genotype frequencies and leaves allele frequencies constant. Genetic drift serves to erode genetic variation as sampling error leads to allele frequency change and eventually to fixation and loss. Gene flow just serves to partition genetic variation among subpopulations, thereby altering patterns of population structure. The process of **mutation**, the permanent incorporation of random errors in DNA that results in differences between ancestral and descendant copies of DNA sequences, is the ultimate source of all genetic variation. This chapter will cover the process of mutation starting out with a description of the patterns and rates of mutation. The following sections will present classical population genetic models for the fate of a new mutation, the impact of mutation on allele frequencies in a population, and the predicted balance between removal of genetic variation by genetic drift and its replacement by mutation. This chapter will also cover several models of the way new alleles are introduced by mutation commonly employed in population genetics, illustrated with applications that highlight the consequences of these models. The final section of the chapter will show

how the process of mutation can be incorporated into genealogical branching models.

Mutation is a broad term that encompasses a wide variety of events that lead to alterations in DNA sequences. **Point mutations** lead to the replacement of a single base pair by another nucleotide. Point mutations to chemically similar nucleotides (purine to purine ( $A \leftrightarrow G$ ) or pyrimidine to pyrimidine ( $C \leftrightarrow T$ )) are called **transitions**, while point mutations to chemically dissimilar nucleotides (purine to pyrimidine or pyrimidine to purine) are called **transversions**. Base substitutions that occur within coding genes may or may not alter the protein produced by that gene. **Synonymous** or silent mutations result in the same translation of a DNA sequence into a protein due to the redundant nature of the genetic code, while **nonsynonymous** or missense mutations result in a codon that does change the resulting amino acid sequence.

Mutation can take the form of **insertion** or **deletion** of DNA sequences, often referred to as **indels**. Indels within coding regions result in frameshift mutations if the change in sequence length is not an even multiple of three, altering the translation of a DNA sequence and possibly creating premature stop codons. Indels may range in size from a single base pair to segments of chromosomes containing many thousands of base pairs. Arrays of multiple copies of homologous genes called **multigene families** are formed by duplication events. Some copies of such duplicated genes may lose functions due to the accumulation of mutations, becoming **pseudogenes**. **Gene conversion** may result in the homogenization of the sequences of multiple loci within multigene families. Gene conversion occurs because of inappropriate mismatch repair that takes place during meiosis. Sections of two homologous chromosomes may anneal when they are single stranded during DNA replication. If these regions differ slightly in sequence, the annealed stretch will contain single nucleotide mismatches. These mismatches will then

be repaired to proper Watson–Crick base pairing by enzymes normally involved in proofreading during DNA replication. The process of annealing between two sister chromosomes tends to happen frequently when the same gene has been repeated many times, because the gene copies have very similar sequences and the chromosomes can anneal anywhere along the length of the gene array. The result is that all gene copies within multigene regions tend to converge on one random version of DNA sequence without recombination taking place.

Mutation may also take the form of rearrangements where a chromosomal region forms a loop structure that results in a segment breaking and being repaired in reversed orientation, called an **inversion**. **Translocations** are mutations where segments of chromosome break free from one chromosome and are incorporated by repair mechanisms into a non-homologous chromosome. **Transposable elements**, segments of DNA that are capable of moving and replicating themselves within a genome, are frequent causes of translocation mutations. **Lateral or horizontal gene transfer**, the movement and incorporation of DNA segments between different individuals and even different species, is another possible avenue of mutation that occurs relatively frequently in prokaryotes. For more detail on the molecular mechanisms that underlie these different types of mutations consult a text such as Lewin (2003).

The probability that a locus or base pair will experience a mutation is a critical parameter in population genetics since the rate of mutation describes how rapidly novel genetic variation is added to populations. Although it seems counterintuitive, mutation rates are actually quite difficult to estimate with precision in many types of organism (see Drake et al. 1998; Fu & Huai 2003). Consider the case of mutation rates at a single locus that has a well-understood effect on the phenotype of an organism, like coat color in mice. The data available to estimate mutation rates are numbers of progeny that have a different coat color than expected based on the known coat-color genotypes of the parents. It is simple to divide the number of progeny with unexpected coat colors by the total number of progeny examined. However, that calculation estimates the frequency of detectable changes to coat color due to some molecular change at the coat color locus. That is an estimate of the frequency of all types of mutation anywhere at the locus rather than an estimate of the mutation rate. Such an estimate of mutation frequency could also

be biased since only mutational changes that caused an obvious change in coat color are included. Not all mutations will be reflected in coat colors, like changes to the third position nucleotide of a codon that are silent, or synonymous, and do not change the resulting amino acid sequence of a gene. Additionally, mutations may vary in their effect on coat color with some mutations having little or no easily observable effect on the phenotype. Therefore, the frequency of observable changes to the phenotype is not equivalent to the mutation rate.

An estimate of the mutation rate requires more information. One critical detail is the number of replications a locus or genome experiences, because mutational changes usually occur during the replication process. Different cell types and different species experience different numbers of cell replications during growth and reproduction. For example, in mammals mutations are more frequent in male gametes than in female gametes because there are many more cell divisions before the production of a sperm than there are before production of an egg. However, the underlying mutation rate could be identical for male and female gametes with the difference due only to the different number of genome replications that occur. Another set of considerations is the size of a locus or genome available to mutate. In the hypothetical mouse coat color example, the number of base pairs at the coat color locus is a critical piece of information. The rate of mutation per base pair estimated from the frequency of coat color changes is very different if the locus has 900 or 90 base pairs.

The distinction between mutation frequency and mutation rate highlights the fact that mutation rates in population genetics are expressed in a variety of terms depending on experimental methods and the life cycle of an organism. The target of mutation can be an entire genome, a locus, or a single base pair, while the rate can be expressed in time units per DNA replication or per sexual generation. Comparisons of mutation rates only make sense when the target size and time period are expressed in identical units. Generally, for population genetic predictions involving sexual eukaryotes, mutation rates per sexual generation are the relevant units. Predictions for prokaryotes such as *Escherichia coli* or yeast would more naturally use mutation rates per cell division.

The most general rule of mutation is that it is a rare event with a low probability of occurrence. In a classic experiment involving literally millions of

**Table 5.1** Per-locus mutation rates measured for five loci that influence coat-color phenotypes in inbred lines of mice (Schlager & Dickie 1971). Dominant mutations were counted by examining the coat color of F1 progeny from brother–sister matings. Recessive mutations required examining the coat color of F1 progeny from crosses between an inbred line homozygous for a recessive allele and a homozygous wild-type dominant allele. The effort to obtain these estimates was truly incredible, involving around 7 million mice observed over the course of 6 years.

Locus	Gametes tested	Mutations observed	Mutation rate per locus $\times 10^{-6}$ (95% CI)
Mutations from dominant to recessive alleles			
Albino	150,391	5	33.2 (10.8–77.6)
Brown	919,699	3	3.3 (0.7–9.5)
Dilute	839,447	10	11.9 (5.2–21.9)
Leaden	243,444	4	16.4 (4.5–42.1)
Non-agouti	67,395	3	44.5 (9.2–130.1)
All loci	2,220,376	25	11.2 (7.3–16.6)
Mutations from recessive to dominant alleles			
Albino	3,423,724	0	0 (0.0–1.1)
Brown	3,092,806	0	0 (0.0–1.2)
Dilute	2,307,692	9	3.9 (1.8–11.1)
Leaden	266,122	0	0 (0.0–13.9)
Non-agouti	8,167,854	34	4.2 (2.9–5.8)
All loci	17,236,978	43	2.5 (1.8–3.4)

95% CI, 95% confidence interval.

mice, mutation rates were estimated from five genes with observable effects on the phenotype of coat color (Table 5.1; Schlager & Dickie 1971). Rates of mutation per gene were between 1.8 and 16.6 mutations per 1 million gametes produced. This is equivalent to a mutation rate of  $(1.8\text{--}16.6) \times 10^{-6}$  per locus per sexual generation. Very similar mutation rates for mice have been reported from more recent irradiation studies as well (Russell & Russell 1996). The rates of mutation from wild type to a novel allele (called **forward** mutations) are nearly a factor of 10 more common than mutations from a novel allele to wild type (termed **reverse** mutations). This asymmetry of forward and backward mutation rates per locus is a common observation in mutation experiments. It is a product of the fact that there are more ways mutation can cause a normal allele to malfunction than there are ways to exactly restore that function once it is disrupted. In this sense, forward and reverse mutation rates exist only because mutations are detected via their phenotypic effect.

Mutation rates can also be estimated in terms of the chance that a genome or a base pair mutates per replication or per sexual generation. At least in principle, the mutation rates for viruses and microbes can be estimated from direct examination of nucleotide sequences after correction for the error rate of the

DNA sequencing techniques, avoiding the underestimate of mutations that comes from detecting only those with a phenotypic effect. Table 5.2 summarizes mutation rate estimates presented in a comprehensive review of mutation rate data (Drake et al. 1998). These rates of mutation differ greatly across the taxonomic groups surveyed but tend to be very similar within the taxonomic groups. For microbes with DNA-based genomes, the rate of mutation per genome per replication clusters around 1/300. Since these organisms vary greatly in their genome sizes, the rate of mutation per base pair per generation ranges widely from  $7.2 \times 10^{-7}$  to  $7.2 \times 10^{-11}$ . In eukaryotes, mutation rates per base pair per sexual generation all fall within a narrow range from  $3.4 \times 10^{-10}$  to  $5.0 \times 10^{-11}$ . If mutation rates per genome in eukaryotes are expressed in terms of the portion of the genome that contains coding genes (termed “effective genome size” by Drake et al. 1998), then the rate of mutation per effective genome in eukaryotes is statistically indistinguishable from the 1/300 rate of DNA-based microbes. The per-base-pair mutation rate estimates are in rough agreement with older per-locus estimates based on visible phenotypes. If the coat-color loci in Table 5.1 have around 1000 or  $10^3$  base pairs, then the per-base-pair mutation rates would be on the order of  $10^{-8}$ .

**Table 5.2** Rates of spontaneous mutation expressed per genome and per base pair for a range of organisms. The most reliable estimates come from microbes with DNA genomes whereas estimates from RNA viruses and eukaryotes have greater uncertainty. Full explanation of the assumptions and uncertainties behind these estimates can be found in Drake et al. (1998).

Organism	Mutation rate per replication	
	Per genome	Per base pair
Lytic RNA viruses		
Bacteriophage Q $\beta$	6.5	
Poliiovirus	0.8	
Vesicular stomatitis virus	3.5	
Influenza A	$\geq 1.0$	
Retroviruses		
Spleen necrosis virus	0.04	
Rous sarcoma virus	0.43	
Bovine leukemia virus	0.027	
Human immunodeficiency virus	0.16–0.22	
DNA-based microbes		
Bacteriophage M13	0.0046	$7.2 \times 10^{-7}$
Bacteriophage $\lambda$	0.0038	$7.7 \times 10^{-8}$
Bacteriophages T2 and T4	0.0040	$2.4 \times 10^{-8}$
<i>Escherichia coli</i>	0.0025	$5.4 \times 10^{-10}$
<i>Neurospora crassa</i>	0.0030	$7.2 \times 10^{-11}$
<i>Saccharomyces cerevisiae</i>	0.0027	$2.2 \times 10^{-10}$
Eukaryotes		
<i>Caenorhabditis elegans</i>	0.018	$2.3 \times 10^{-10}$
<i>Drosophila</i>	0.058	$3.4 \times 10^{-10}$
Human	0.49	$1.8 \times 10^{-10}$
Mouse	0.16	$5.0 \times 10^{-11}$

<sup>a</sup>Estimate from Nachman and Crowell (2000) based on pseudogene divergence between humans and chimpanzees.

The mutation rates at microsatellite or simple sequence repeat (SSR) loci are also of interest since such loci are widely employed as selectively neutral genetic markers to study a wide range of population genetic processes. These repeated DNA regions have very high rates of mutation between  $1 \times 10^{-2}$  and  $6 \times 10^{-6}$  per sexual generation (Ellegren 2000; Steinberg et al. 2002; Beck et al. 2003).

How can such a low-probability event like mutation add more than a trivial amount of genetic variation to populations? Let's calculate an initial answer to that question using humans as our example. Averaging over coding and non-coding parts of the genome, an approximate nuclear genome mutation rate in humans is about  $1 \times 10^{-9}$  mutations per base pair per generation. The haploid genome (one sperm or egg) contains about  $3.2 \times 10^9$  base pairs (bp). Each genome of each diploid individual will have:

$$(1 \times 10^{-9} \text{ mutations bp}^{-1} \text{ generation}^{-1}) (2 \times 3.2 \times 10^9 \text{ bp}^{-1}) = 6.4 \text{ mutations} \quad (5.1)$$

where the factor of 2 is due to a diploid genome. Each of us differs from one of our parents by half this amount, or about three mutations on average. If all mutations are random events evenly distributed throughout the genome, every pair of individuals differs by twice this number of mutations or about 13 mutational differences on average. The overall effect of mutation on available genetic variation depends on the size of a population. The human population is currently about 6.486 billion people (see [www.census.gov/main/www/popclock.html](http://www.census.gov/main/www/popclock.html)). Based on this population size, there are a total of

$$(6.4 \text{ mutations individual}^{-1} \text{ generation}^{-1})(6.486 \times 10^9 \text{ individuals}) = 41.5 \times 10^9 \text{ mutations} \quad (5.2)$$

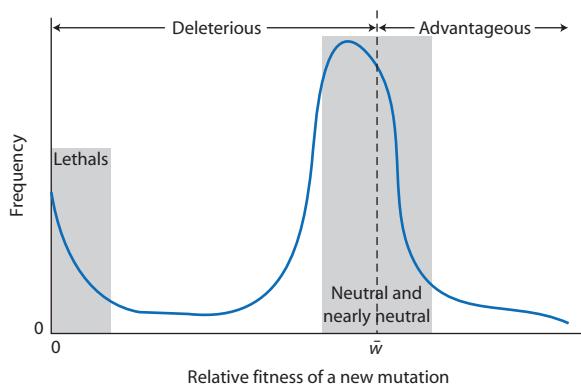
or over 41 billion mutations expected each generation! This means that the absolute numbers of mutations per generation are potentially high and depend on the rate of mutation, the size of the population, and the size of the genome. We will revisit this topic

later in the chapter to make a more formal prediction about the levels of heterozygosity expected when the input of genetic variation due to mutation and the loss of genetic variation caused by genetic drift are at equilibrium.

The impact that a mutant allele (as part of a heterozygous or homozygous genotype) has on the phenotype of an individual can vary greatly. Since natural selection along with genetic drift are critical processes that determine the fate of new mutations, the phenotype is most often considered in the context of its survivorship and reproduction, or fitness. The range of the possible fitnesses for an individual mutant allele can be thought of as a **mutation fitness spectrum** like that shown in Fig. 5.1. The fitness of all mutation effects on the phenotype is relative to the average fitness of a population (see Chapter 6 for explanations of fitness and average fitness). **Detrimental** or **deleterious** mutations reduce survival and reproduction while mutations that improve survival and reproduction are **advantageous**. Severely deleterious mutations such as those that result in death (called **lethals**) or failure to reproduce viable offspring are acted strongly against by natural selection and will likely not last for a single generation. Mutations that

are strongly deleterious and nearly lethal are sometimes called **sublethals**. Mutations that have small positive or negative effects on fitness (the shaded zone around the mean fitness in Fig. 5.1) are called **neutral** or **nearly neutral** since their fate will be dictated either totally or mostly by sampling error of genetic drift. The final type are **beneficial** mutations that increase survival and reproduction above the average fitness of the population. It is important to note that the fitness effects of mutations may depend greatly on environmental context (see Fry & Heinsohn 2002) and the genotype at other loci. These different types of mutation will be the subject of models later in this chapter that show how fitness relates to the chance that a new mutation is lost or reaches fixation in a population.

**Mutation fitness spectrum** The frequency distribution of the average fitness of new mutations measured relative to the average fitness of a reference population.



**Figure 5.1** A hypothetical distribution of the effects of mutations on phenotypes that ultimately impact the Darwinian fitness of genotypes. Mutations that have a mean fitness less than the mean fitness of the population ( $\bar{w}$ ) are decreased in frequency by natural selection. The shaded area around  $\bar{w}$  indicates the zone where mutations have small effects on fitness relative to the effects of genetic drift (the width of the neutral zone depends on the effective population size). The shaded area near zero mean fitness indicates mutations that cause failure to reproduce or are lethal. Lethals are more common since it is a category that includes many degrees of severity resulting from diverse causes. The fitness effects of mutations are inherently difficult to measure because of the rarity of mutation events, the small effect of many mutations, and the dependence of fitness on environmental context.

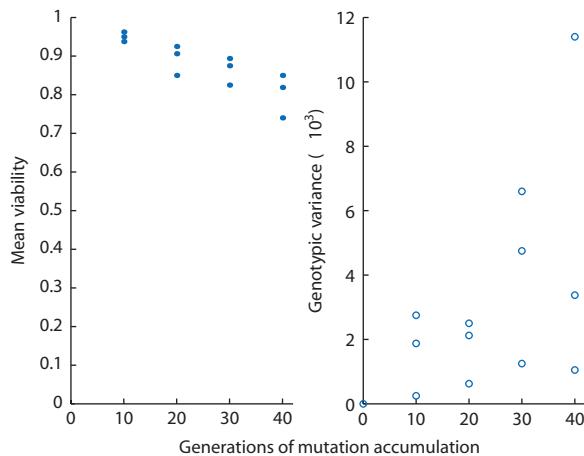
The mutation fitness spectrum plays a central role in a wide range of hypotheses to explain a multitude of phenomena in population genetics and evolution (see Charlesworth & Charlesworth 1998; Lynch et al 1999; Orr 2003; Estes et al. 2004). Explanations for phenomena as general and diverse as inbreeding depression, the evolution of mating systems, the evolution of sex and recombination, and the rate of adaptation depend in part on the nature of the mutation fitness spectrum. Strongly deleterious or strongly beneficial mutations will be steadily and predictably driven to loss or to fixation, respectively, by natural selection. However, fixation and loss of mutations that have a small impact on fitness (relative to the effective population size) is due in whole or in part to random genetic drift. A consequence is that mildly deleterious mutations may reach fixation by chance and accumulate in a population over time. Similarly, some mildly beneficial mutations may be lost from populations by chance. An accumulation of mildly deleterious mutations reduces individual fitness and may increase the risk of extinction, resulting in natural selection for processes that reduce the load of deleterious mutations in a population. The frequency of beneficial mutations may also place limits on the rate of evolution by positive natural selection. Thus, the specific shape of the frequency

distribution shown schematically in Fig. 5.1 provides crucial information about the fate of individual mutations as well as the long-term consequences of continual mutation in populations.

A commonly employed method to estimate the shape of the mutation fitness spectrum relies on founding a series of genetically identical populations and then allowing some to experience mutations for many generations while maintaining a control population that does not experience mutation. Viability and reproduction phenotypes of the mutated populations are then compared with the control population at intervals to estimate the average change in fitness caused by the mutations. Such comparisons are called **mutation-accumulation** experiments since mutations are repeatedly fixed by genetic drift over time in the mutation populations.

If there was absolutely no mutation, the replicate populations in a mutation-accumulation experiment would all maintain identical viability over time since each population started out being genetically identical. Mutation, however, will occur at random and cause independent genetic changes in the different populations, causing the populations to diverge in viability. Imagine that the mutation fitness spectrum is symmetric around the mean fitness so that the frequency of deleterious and beneficial mutations of the same magnitude is equal. That would produce no change in the *average* viability of lines in a mutation-accumulation experiment since there would be equal chances of beneficial or deleterious mutations of the same size that would cancel each other out in a sample of many mutations. However, there would be an increase in the *variance* in viability because the range of viabilities among the populations would increase with more and more mutations. Next imagine a mutation fitness spectrum like that shown in Fig. 5.1 where deleterious mutations are more common than beneficial mutations. As mutations accumulate, the average viability of lines should decrease since deleterious mutations are more common than beneficial mutations. The more skewed the distribution is toward deleterious mutations the faster the average viability should decrease in the replicate populations.

The results of several classic mutation accumulation studies that estimated the frequency distribution for mutations that affect viability in *Drosophila melanogaster* have had a major impact on perceptions of the mutation fitness spectrum (Mukai 1964; Mukai et al. 1972). Mutation-accumulation experiments in *Drosophila* rely on special breeding designs



**Figure 5.2** The results of the classic *Drosophila melanogaster* mutation accumulation experiment carried out by Mukai et al. (1972). The experiment maintained three distinct sets of mutation-accumulation populations with 25 lines each. The left-hand panel shows the change in mean viability over time and the right-hand panel shows the change in the variance among replicate independent lines. Each point is the value obtained from one set of mutation-accumulation populations. Mutation of any type makes the lines diverge genetically and increases the variance. Mean viability declines over time as deleterious mutations are more common than advantageous mutations. Redrawn from Figure 2 in Mukai et al. (1972).

that maintain the second chromosome without recombination in many replicate homozygous families or lines over many generations. Mutations of all types occur on this non-recombining chromosome and are fixed by genetic drift within each line of flies due to a single male founder for each generation. At intervals of 10 generations, the flies in all of the different independent lines were assayed for viability in comparison with a control line that did not experience any mutation due to chromosomal inversions (again accomplished with special breeding techniques). The change in average viability and variance in viability found by Mukai et al. (1972) is shown in Fig. 5.2. The variance in viability among the replicate lines has increased because the second chromosome of each line diverged due to the occurrence and fixation of mutations. In addition, the average viability has declined as expected if deleterious mutations are more common than beneficial mutations. The results are consistent with deleterious mutations that cause an average reduction in viability of 5% or less when homozygous. Therefore, this experiment and others like it motivated a view of the mutation fitness spectrum as drawn in Fig. 5.1. However, mutation-accumulation experiments have been carried out in only a relatively few organisms

and are inherently unable to detect mutations with very small effects or that do not affect the phenotype within the laboratory environment where the phenotype is measured. There is also the possibility that the distribution of mutation fitnesses varies among taxa. For example, a mutation-accumulation experiment in the plant *Arabidopsis thaliana* showed that deleterious and beneficial mutations were about equally frequent (Shaw et al. 2000).

Inherent in the mutation fitness spectrum in Fig. 5.1 is that beneficial mutations are rarer than deleterious mutations. This makes estimating of the frequency distribution of beneficial mutations even more difficult than it is for deleterious mutations. Nonetheless, a number of studies have directly measured the effects of advantageous mutations. Bacterial populations have been used to study mutations due to their short generation times and the ease of constructing and maintaining replicate populations. Using *E. coli*, several studies have shown that beneficial mutations with small effects on fitness are much more common than mutations with larger effects (Imhof & Schlotterer 2001; Rozen et al. 2002). Using an RNA virus, Sanjuan et al. (2004) used site-directed mutagenesis to make numerous single-nucleotide mutations. Beneficial mutations were much rarer than deleterious mutations, but the eight beneficial mutations had an average of a 7% improvement in fitness and small mutation effects were more common. An important caveat to these studies is that the beneficial mutations detected are biased toward those of larger effects because: very small mutation effects cannot be measured; beneficial mutations with larger effects increase in frequency more rapidly under natural selection making them more likely to reach a high enough frequency to be detected; and there is the possibility of competition among lineages with different beneficial mutations in asexual organisms (a phenomenon called clonal interference).

## 5.2 The fate of a new mutation

- The chance a neutral or beneficial mutation is lost due to Mendelian segregation.
- Mutations fixed by natural selection.
- Frequency of a mutant allele in a finite population.
- Accumulation of deleterious mutations by Muller's Ratchet without recombination.

How does the frequency of a new mutation change over time after it is introduced into a population?

This simple question is central to understanding the chance of fixation and loss for new mutations and therefore their ultimate fate in a population. The mutation rate dictates how often a new mutation will appear in a population. But once a mutation has occurred, population genetic processes acting on it will determine whether it increases or decreases in frequency. This section will consider four distinct perspectives on the frequency of a new mutation based on the processes of genetic drift and natural selection. Each of the four perspectives makes different assumptions about the population context in which a new mutation is found, considering different effective population sizes, levels of recombination, and whether mutations are neutral, advantageous, or deleterious. Naturally, these four perspectives do not cover all possible situations but are meant to explore a range of possibilities and communicate several distinct approaches to determining the fate of a new mutation. Although this section will consider the action of natural selection on mutations, the simple forms of selection assumed should be accessible to most readers. Natural selection and fitness are defined and developed rigorously in Chapter 6.

### **Chance a mutation is lost due to Mendelian segregation**

The fate of a new mutation can be tracked by considering its pattern of Mendelian inheritance, as shown by R.A. Fisher in 1930 (see Fisher 1999). Call all the existing alleles at a locus  $A_x$  where  $x$  is an integer 1, 2, 3 . . . ,  $x$  to index the different alleles, and a new selectively neutral mutation  $A_m$ . Any new mutation appears initially as a single-allele copy and it therefore must be found in a heterozygous genotype ( $A_xA_m$ ). To form the next generation, this  $A_xA_m$  heterozygote experiences random mating with the other  $A_xA_x$  genotypes in the population. For each progeny produced by the  $A_xA_m$  genotype, there is a  $1/2$  chance that the mutant allele is inherited and a  $1/2$  chance that the mutant allele is not inherited (the  $A_x$  allele is transmitted instead). The total chance that an  $A_xA_m$  heterozygote passes the mutant allele on to the next generation depends on the number of progeny produced. If  $k$  is the number of progeny parented by the  $A_xA_m$  heterozygote and there is independent assortment of alleles, then

$$P(\text{mutant lost}) = \left(\frac{1}{2}\right)^k \quad (5.3)$$

**Table 5.3** The expected frequency of each family size per pair of parents ( $k$ ) under the Poisson distribution with a mean family size of 2 ( $\bar{k} = 2$ ). Also given is the expected probability that a mutant allele  $A_m$  would not be transmitted to any progeny for a given family size. Note that  $0!$  equals one.

Family size per pair of parents ( $k$ ) . . .	0	1	2	3	4 . . . $k$
Expected frequency	$e^{-2}$	$2e^{-2}$	$2e^{-2}$	$\frac{4}{3}e^{-2}$	$\frac{2}{3}e^{-2} \dots \frac{2^k}{k!}e^{-2}$
Chance that $A_m$ is not transmitted	1	$\frac{1}{2}$	$\left(\frac{1}{2}\right)^2$	$\left(\frac{1}{2}\right)^3$	$\left(\frac{1}{2}\right)^4 \dots \left(\frac{1}{2}\right)^k$

is the probability that the mutant allele is not transmitted to the next generation in any of the progeny. As you would expect, the probability that no mutant alleles are transmitted to the next generation declines as the number of progeny produced increases.

In a population that is constant in size over time, each pair of parents produces two progeny on average that take their places in the next generation. A key phrase here is “on average,” meaning that not every pair of parents will produce two progeny: some parents will produce more progeny and some parents will produce fewer. As shown in the context of the variance effective population size in Chapter 3, the Poisson distribution is commonly used to model variation in reproductive success. Here too, we can use a Poisson distribution to determine the expected frequencies of each family size when the average family size is two progeny (Table 5.3). The reason we need to know the expected proportion of the parental pairs that produce a given number of progeny is that each family size has a different probability of not transmitting the mutant allele. For a given family size  $k$ , the probability that a mutant allele is not transmitted to the next generation is the product of the expected frequency of parental pairs and the chance of not transmitting the mutant allele:

$$P(\text{mutant lost}) = \left(\frac{2^k}{k!}\right) e^{-2} \left(\frac{1}{2}\right)^k \quad (5.4)$$

or the product of the two terms in each column of Table 5.3. The total probability that the mutant allele is not transmitted to the next generation is the sum over all possible family sizes from zero to infinity:

$$P(\text{mutant lost}) = \sum_{k=0}^{\infty} \left(\frac{2^k}{k!}\right) e^{-2} \left(\frac{1}{2}\right)^k \quad (5.5)$$

Although this sum looks daunting to calculate, the equation actually simplifies to a very neat result. The  $e^{-2}$  term is a constant so that it can be moved in front of the summation

$$P(\text{mutant lost}) = e^{-2} \sum_{k=0}^{\infty} \left(\frac{2^k}{k!}\right) \left(\frac{1}{2}\right)^k \quad (5.6)$$

and the  $2^k$  and  $\left(\frac{1}{2}\right)^k$  terms cancel to give

$$P(\text{mutant lost}) = e^{-2} \sum_{k=0}^{\infty} \frac{1}{k!} \quad (5.7)$$

The final trick is to notice that the series determined by the summation  $(1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \dots + \frac{1}{k!})$  approaches  $e$  ( $e = 2.718 \dots$ ) as  $k$  goes to infinity. The summation term can then be replaced with  $e$  to give

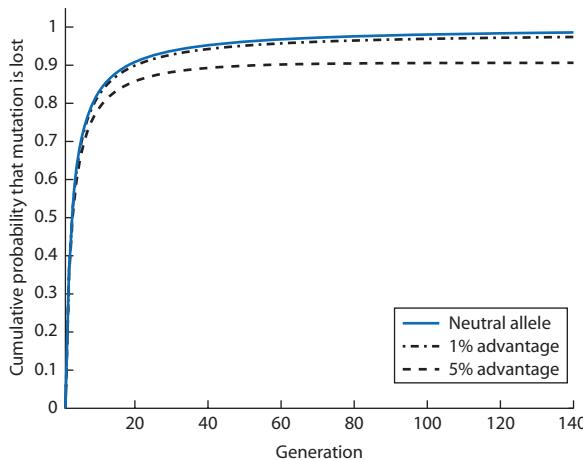
$$P(\text{mutant lost}) = e^{-2}e = e^{-1} \quad (5.8)$$

As promised, the tidy conclusion is that the chance a newly occurring mutant is lost simply due to Mendelian segregation after one generation is  $e^{-1} = 0.3679$ . Therefore, a new mutation has about a 36% chance of being lost within one generation of its introduction into a population. The world is tough for a new mutation!

This result can be extended to determine the probability that a mutation is lost over multiple generations of Mendelian segregation. A general expression for the *cumulative* probability of a mutation being lost from the population over time is

$$P(\text{mutant lost generation } t) = e^{x-1} \quad (5.9)$$

where  $x$  is the probability of loss in the generation before or at time  $t - 1$ . (The series determined by



**Figure 5.3** The probability that a novel mutation is lost from a population due to Mendelian segregation. A neutral allele is eventually lost from the population while a beneficial mutation has a probability of about twice its selective advantage of fixation. The cumulative probability over time is described by  $e^{c(x-1)}$  where  $x$  is the probability of loss in the generation before and  $c$  is the degree of selective advantage, if any. This expected probability assumes an infinitely large population that has a Poisson-distributed variance in family size.

the summation in equation 5.7 is really  $(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^k}{k!})$  and approaches  $e^{1+x}$  as  $k$  goes to infinity to give  $(e^{-2})(e^{1+x}) = e^{x-1}$ . When a mutant first appears in a population  $x = 0$ .)

Using this result shows that the probability of a new mutation being lost in two generations is  $e^{-0.6321} = 0.5315$  or the probability of being lost in three generations is  $e^{-0.4685} = 0.6295$ . Based on this progression, Fig. 5.3 shows the probability that a new mutation is lost over the course of 140 generations. The conclusion from this graph is that a new mutation must eventually be lost from a population given enough time.

We can also ask what impact natural selection might have on this prediction that a new mutation will eventually be lost. Let's imagine that a new mutation is slightly beneficial instead of being neutral. Natural selection will then improve the chances that the new mutation is transmitted to the next generation, giving it a slight advantage over any of the other alleles in the population. Let  $c$  be the selective advantage of a new mutation so that a value of 1.0 would indicate neutrality and a value of 1.01 would mean a transmission advantage of 1%. The cumulative probability that an allele is lost at generation  $t$  is then

$$P(\text{mutant lost generation } t) = e^{c(x-1)} \quad (5.10)$$

This version of the equation multiplies the exponent for the neutral case by the selective advantage of a beneficial allele. This makes very little difference to the probability that a mutant is lost if only a few generations elapse, but makes a larger difference after more generations have passed (Fig. 5.3). In general, the chance that a new beneficial mutation is not lost is approximately twice its selective advantage, still a very low probability for realistically small values of the selective advantage. However, as Fisher pointed out, if something like 250 independent beneficial mutations occur singly over time then there is a very small chance ( $0.98^{250} = 0.0064$ ) that all of them would be lost during Mendelian segregation. This suggests that at least some beneficial mutations will be established in populations as mutations continue to be introduced.

The conclusion that a new neutral mutation must always be lost from a population seems at odds with the possibility of random fixation of a new mutation due to genetic drift. Fisher's method of modeling the fate of a new mutation makes the assumption that population size is very large. This assumption allows use of the expected values for the proportion of parental pairs for each family size under the Poisson distribution and the chance of an allele being lost for each family size, probabilities that should only be met in the limit of many parental pairs that span a wide range of family sizes. Finite numbers of parental pairs would likely not meet these expected values due to chance deviations from the expected value. The assumption of infinite population size is justified because it is used to reveal that particulate inheritance by itself can lead to loss of new mutations even in the complete absence of genetic drift. Next, we will take up the fate of a new mutation in the context of a finite population.

#### Fate of a new mutation in a finite population

A second perspective on new mutations is to consider their fate as an allele in a finite population in the absence of natural selection. We can then employ the concepts and models of genetic drift developed in Chapter 3 to predict the frequency of new mutations over time in a population. The first critical observation is to recognize that the initial frequency of any new mutation is simply

$$p_0(\text{new mutation}) = \frac{1}{2N_e} \quad (5.11)$$

because a new mutation is present as a single-allele copy in a population of  $2N_e$  allele copies. If the frequency of a new mutation is determined strictly by genetic drift, then each new mutation has a probability of  $\frac{1}{2N_e}$  of going to fixation and a probability of  $1 - \frac{1}{2N_e}$  of going to loss. This result makes intuitive sense, since a new mutation is very rare and is close to loss but very far from fixation. This result also shows that the chance of fixation or loss of a new mutation depends on the effective population size.

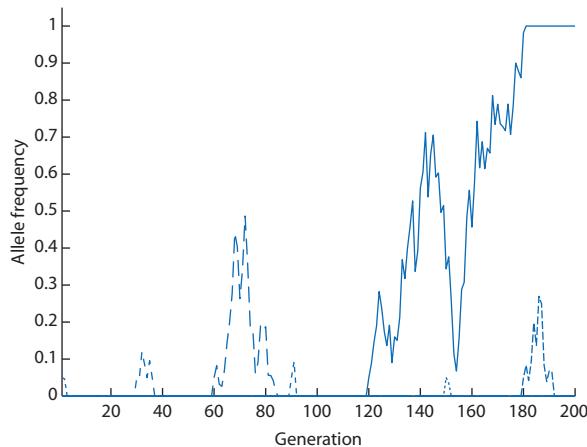
Using the diffusion approximation of genetic drift, it is possible to estimate the average number of generations before a new mutation is either fixed or lost (Kimura & Ohta 1969a). Figure 3.14 and equation 3.40 give the average number of generations until fixation or loss for an allele depending on the effective population size and initial allele frequency. Under the assumption that the effective population size is large, those alleles that eventually fix do so in an average of  $4N_e$  generations. Those alleles that are lost go to fixation in many fewer generations, approaching zero generations as the population size gets larger and the initial frequency of  $\frac{1}{2N_e}$  therefore gets smaller. However, since genetic drift is a stochastic process we expect that the variance around the average time to fixation or loss will be large. In other words, the allele frequency of each new mutation will take a random walk between zero and one. Although many mutations may be lost quickly others may segregate for several or many generations before being lost or fixed.

The fate of new mutations can be seen readily in a simulation. Figure 5.4 shows the frequency of new mutations introduced every 30 generations into a population of  $N_e = 10$ . Of the seven mutations introduced into the population, six go to loss and only one goes to fixation. This is roughly consistent with the prediction that one in 20 new mutations will fix in a population where  $N_e = 10$ . Most of the mutations that go to loss do so in fewer than 10 generations, although in one case the mutation segregates for about 25 generations. Equation 3.40 predicts that mutations go to loss in that an average of about six generations, roughly consistent with the simulation. The mutation that goes to fixation does so in 60 generations, taking a zig-zag trajectory of allele frequency. Equation 3.40 predicts that an average of about 39 generations will elapse for those mutations that go to fixation when  $N_e = 10$ ,

suggesting that the simulation result is somewhat greater than the expected average time to fixation.

### Interact box 5.1 Frequency of neutral mutations in a finite population

PopGene.S<sup>2</sup> can be used to simulate the fate of new mutations in a finite population. Click on the **Mutation** menu and then select **Neutral Mutation**. Initially, try the simulation using the default values of mutations introduced every 30 generations, a population of  $N_e = 10$  and 200 generations. Tabulate the number of fixations and losses over several separate runs (the fate and frequency of each new mutation is independent). What are the expected number of new mutations that go to fixation and loss and how do these expectations compare with the results of the simulation? Then increase the population size to  $N_e = 50$  and view 400 generations. How does the number of new mutations going to fixation and loss change? How does the time that new mutations segregate change?



**Figure 5.4** The frequencies over time of new mutations that each have an initial frequency of  $\frac{1}{2N_e}$ . In this example, one new mutation is introduced into the population every 30 generations and  $N_e = 10$ . All of the mutations except one (solid line) go to loss within a few generations. The one allele that does go to fixation takes a relatively long time to do so compared with the time to loss. At the start of the simulation the ancestral allele has a frequency of 1 (not shown). When a new mutation reaches fixation, the original ancestral allele is lost and the new mutation becomes the ancestral allele.

These predictions for the frequency and fate of new neutral mutations under genetic drift suggest that at least some genetic variation is maintained in populations simply due to the random allele-frequency walk that new mutations take before reaching either fixation or loss. If the population shown in Fig. 5.4 were observed at a single point in time, it is possible that it would be polymorphic since a new mutation happened to be somewhere between fixation and loss. Observing many such loci at one point in time, it would be very likely that at least some of them would be polymorphic. This observation forms the basis of the neutral theory of molecular evolution, the hypothesis that genetic variation in populations is caused by genetic drift, covered in Chapter 8.

### ***Geometric model of mutations fixed by natural selection***

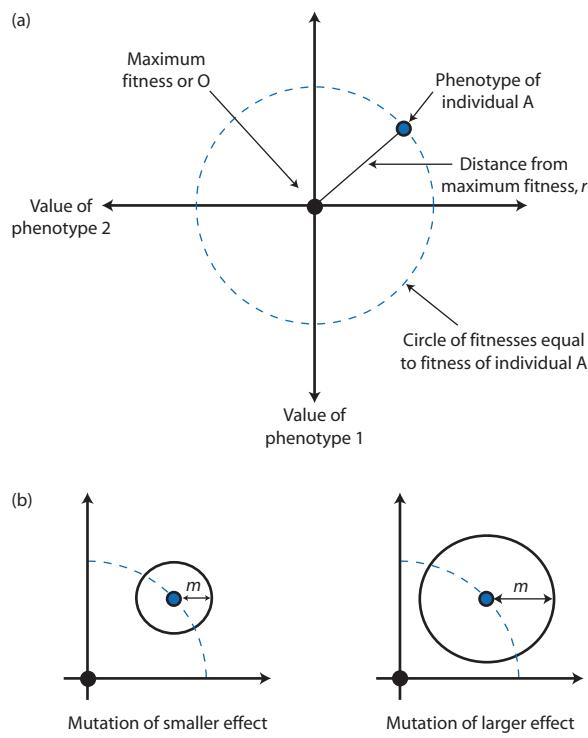
The third perspective on the fate of new mutations will focus on beneficial mutations, looking first at mutations fixed by natural selection alone and then at mutations fixed by the combined processes of natural selection and genetic drift. In addition to considering how new mutations are lost during segregation, in 1930, Fisher (see Fisher 1999 variorum edition) constructed another model of the fate of mutations that are acted on by the process of natural selection. As discussed earlier in the chapter, mutations may have a range of effects on fitness as well as on any phenotype with variation that has a genetic basis. The model Fisher developed sought to determine the range or distribution of the effect sizes (the amount of change in the phenotype caused by each mutation) of the beneficial mutations that are fixed by natural selection over time. Are the mutations that are fixed by natural selection all of large effect or all of small effect, or do they have effect sizes that fall into some type of distribution? It is quite likely that you are aware of the generalizations of this model without being aware of where these conclusions came from or what assumptions are involved. The generalization is that beneficial mutations have small effects: we do not expect to see beneficial changes taking place in single big leaps. This view of evolution is called **micromutationalism**, a concept that has been profoundly influential in evolutionary biology and population genetics (see Orr 1998a and references therein). The model that leads to this conclusion is called the **geometric model of mutation** and is developed in this section.

**Micromutationalism** The view that beneficial mutations fixed by the process of natural selection have small effects and therefore that the process of adaptation is marked by gradual genetic change.

Fisher imagined a situation where the values of two phenotypes determined the survival and reproduction, or fitness, of an individual organism (fitness is defined rigorously in Chapter 6). An example of two phenotypes might be the number of leaves and the size of leaves to achieve the maximum light capture for photosynthesis for a species of plant. However, the exact nature of the phenotypes is not important in the model as long as they contribute to the fitness of individuals. The critical point to understand is that phenotypic values closer to the maximum fitness value (Fisher called this the “optimum”) are favored by natural selection, causing genotypes conferring higher fitness phenotypes to increase in frequency and fix in a population over time. Figure 5.5 shows the model. The values of the two traits are represented by two axes and the optimum fitness value for the combination of the two traits is at the center, the point labeled O for optimum.

Let's say an individual has values of the two phenotypes that put it at point A on the phenotypic axes, some distance  $r$  from the optimum fitness. All the points on a circle of radius  $r$  centered at the optimum have the same fitness as the fitness of the individual at point A (the dashed circle in Fig. 5.5a). Next, imagine that random mutations can occur to one allele of the genotype of the individual at point A. If the effects of mutations are random, then a mutation could move the phenotype in any direction from point A, and these moves could be of any distance large or small away from point A. Some mutations would move the phenotype a short distance whereas others cause a long-distance move; some mutations move the phenotype toward the optimum whereas others move the phenotype away from the optimum.

From this graphical model, can we determine what types of mutational change are likely to be fixed by natural selection and contribute to adaptive change? One conclusion from the model is that mutations with a very large effect on phenotype (change in phenotype greater than  $2r$ ) cannot get the phenotype any closer to the optimum even if they are in the right direction. Since mutations of very large effect always move the phenotype further away from the



**Figure 5.5** R.A. Fisher's geometric model of mutations fixed by natural selection. (a) Axes for two hypothetical phenotypes that determine fitness with maximum fitness when both phenotypes have the values at the center point marked with the black dot. An individual (or the mean phenotype of a population) with a phenotypic value is some distance from the maximum fitness. The dashed circle shows a perimeter of equal fitness around the point of maximum fitness. Although only two phenotypes define fitness in this example, the dashed circle of equal fitness would be a sphere with three phenotypes and an  $n$ -dimensional hyperspace if  $n$  phenotypes contribute to fitness. (b) Two mutations with smaller or larger phenotypic effects. The phenotypic effect of the mutations could be in any direction around the current phenotype (circles with radius  $m$ ). Mutations with smaller effects have a better chance of moving the phenotype toward the maximum fitness (more of the area of the mutation-effect circle is to the left of the dashed line of equal fitness).

optimum fitness outside the dashed circle, these will never be fixed by natural selection.

What about mutations with smaller effects? Figure 5.5b shows two situations where the phenotypic effect of a mutation is smaller (change in phenotype less than  $2r$ ). On the right there is a mutation with a larger effect on the phenotype and on the left a mutation with a smaller effect on the phenotype. Both of these mutations could be in any direction, specified by the circle around A with a radius  $m$  to indicate the magnitude of the mutational effect. Notice that as the mutation gets larger in effect, less of the circle describing the effect on the phenotype falls inside of the dashed circle that describes

the current fitness of the individual at point A. As the phenotypic effect of a mutation approaches zero ( $m \rightarrow 0$ ), its effect circle will approach being half inside the arc of current fitness and half outside the arc of current fitness. Said the other way, as the phenotypic effect of a mutation gets larger and larger its effect circle encompasses more and more area outside the arc of current fitness. As mutations increase in their phenotypic effect they have an increasing probability of being in a direction that will make fitness worse rather than better. Therefore, natural selection should fix more mutations of small effect than of large effect since smaller mutations have a greater probability moving the phenotypic value toward the optimum. Mutations with almost zero effect have close to a  $1/2$  chance of being favorable while large mutations have a diminishing chance of being favorable.

This can be described in an equation:

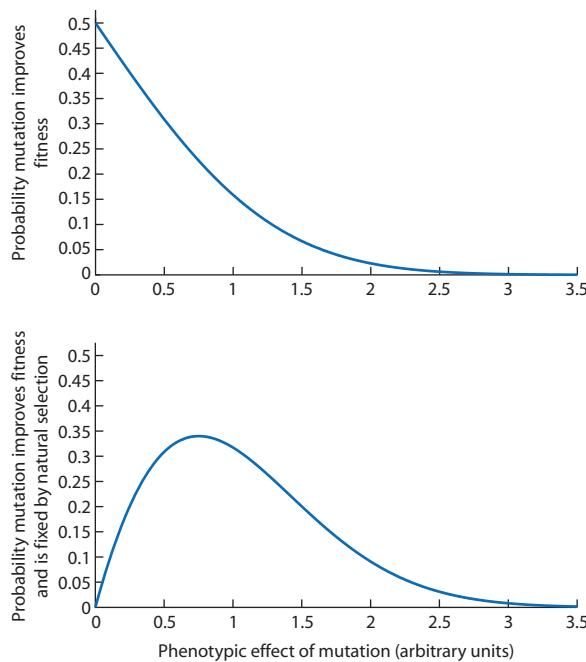
$$P(\text{mutation improves fitness}) = \frac{1}{2} \left( 1 - \frac{m}{2r} \right) \quad (5.12)$$

where  $m$  is the radius of the phenotypic effect of a mutation and  $r$  is the distance to the optimum from the current phenotypic value. As  $m$  goes to zero, the probability that a mutation moves the phenotype closer to the optimum approaches  $1/2$ . For mutations of increasing effect, there is a diminishing probability that they improve fitness. When  $m$  is equal to twice the value of  $r$  there is no chance that the mutation will improve fitness: a mutation of effect  $2r$  could just reposition A on the opposite side of the equal fitness circle around the optimum at best.

Fisher also reasoned that the fitness of organisms depends on many independent traits since the phenotypes of organisms must meet many requirements for successful growth, feeding, avoidance of predation, mating, and so forth. He therefore assumed that the dashed circle of equal fitness shown in Fig. 5.5 for illustration was really better represented by a space of many dimensions. In  $n$  dimensions, the measure of whether or not a mutation is large or small relative to the distance to the point of maximum fitness ( $r$ )

is gauged by  $\frac{r}{\sqrt{n}} = r \frac{\sqrt{n}}{2r}$  instead of  $m/2r$  in equa-

tion 5.12. The main point is that increasing phenotypic dimensions cause the probability that a mutation improves fitness to decline more rapidly as its phenotypic effect gets larger. The top panel of Fig. 5.6 plots



**Figure 5.6** The probability that a mutation is fixed by natural selection depends on the magnitude of its effect on fitness. Using the geometric model of mutation and assuming that fitness is determined by many phenotypes, Fisher showed the probability that a mutation improves fitness approaches  $1/2$  as the effect of a mutation approaches 0 (top panel). This result comes about because smaller mutations have a better chance of moving the phenotype toward the optimum than do larger mutations (see Fig. 5.5). Kimura pointed out that mutations with small effects on fitness are also the most likely to be fixed or lost due to genetic drift rather than by natural selection. Combining the chance that a mutation moves the phenotype toward higher fitness *and* the chance that a mutation has a large-enough fitness difference to escape genetic drift suggests that mutations with intermediate effects are most likely to be fixed by natural selection (bottom panel). Both models assume that mutations of any effect on fitness are equally likely to occur.

a version of equation 5.12 which assumes that fitness is determined by many independent phenotypes. This shows the distribution of the probability that a mutation improves fitness as the multi-dimensional phenotypic effect of a mutation increases.

The conclusion from the geometric model of mutation is evident in the top panel of Fig. 5.6. Mutations with small effects are most likely to bring an organism closer to its fitness optimum and are therefore most likely to be fixed by natural selection. Mutations of larger effect have a lower probability of improving fitness and are therefore less likely to be fixed by natural selection. Fisher compared the situation to the focus adjustment on a microscope. If a microscope is close to being in focus, then large random

changes to the adjustment are likely to make things worse while small random changes are more likely to make the focus better. A logical consequence of Fisher's model is that the mutation fitness spectrum approaches 50% deleterious and 50% beneficial mutations as mutation effects approach zero. This prediction is not consistent with the general picture of the mutation fitness spectrum in Fig. 5.1.

Many years later, Kimura (1983a) reevaluated the predictions of the geometric model of mutation by relaxing Fisher's implicit assumption of an infinite effective population size. This change allows genetic drift to operate on the frequency of mutations along with natural selection. In a finite population, allele frequency is determined by a combination of sampling error and the effect of natural selection to fix alleles with higher average fitness. Natural selection will only determine the fate of an allele if it is stronger than the randomizing effect of genetic drift. The pressure of natural selection also depends on the phenotypic effect of a mutation – mutations with a larger effect experience a stronger push toward fixation. Thus, the push toward fixation by natural selection is strongest for those new mutations that have the largest effects. In other words, new mutations with small effects are likely to experience random fixation or loss by genetic drift. The bottom panel of Fig. 5.6 shows the probability that a new mutation is fixed by natural selection in a finite population. The mutations with the smallest phenotypic effects are still most likely to move the phenotype toward higher fitness. However, this is now balanced by the effect of genetic drift, which has the greatest impact on new mutations with small effects on fitness. The modified result is that new mutations with an intermediate effect on fitness are the most likely to fix under natural selection in a finite population.

Orr (1998) provides an analysis of the effect sizes of mutations that are fixed by natural selection in a finite population that compensates for the fact that the effect of mutations must shrink as a population gets closer to the maximum fitness over time. The net balance of natural selection and genetic drift is considered in detail in later chapters and the phenotypic effects of loci and alleles are treated in detail in Chapter 9 on quantitative genetics.

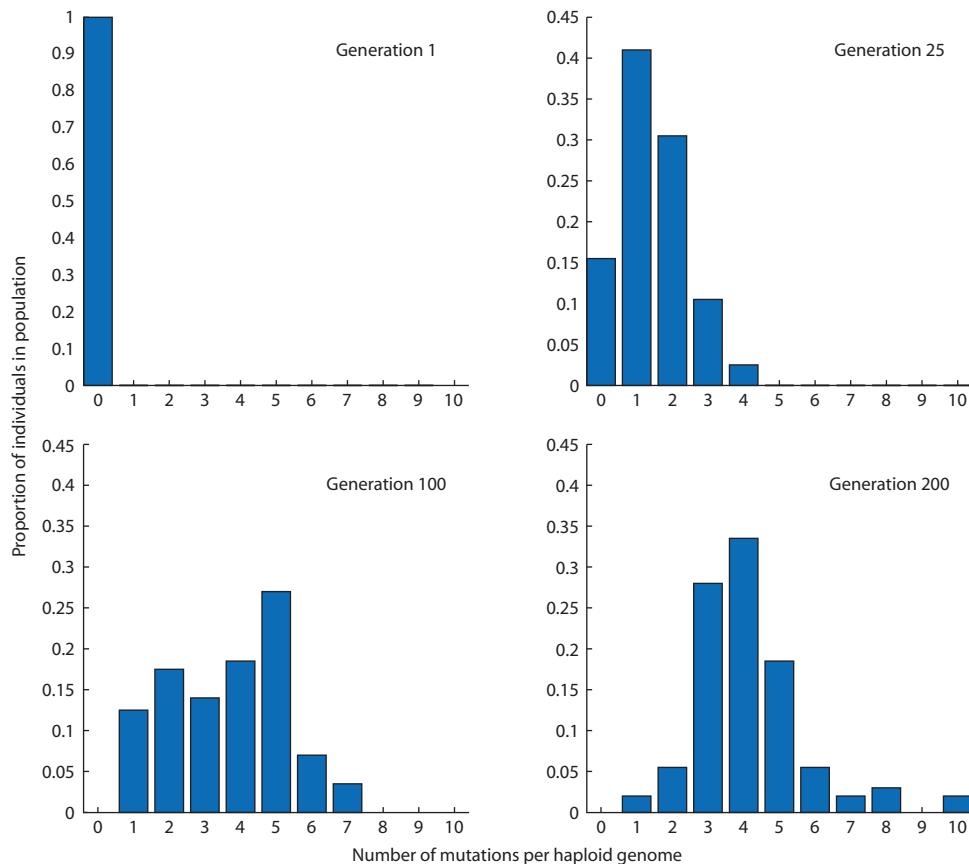
#### **Muller's Ratchet and the fixation of deleterious mutations**

The fourth and final perspective on the fate of a new mutation focuses on deleterious mutations that

occur within genomes lacking recombination. The combination of mutation, genetic drift, and natural selection results in the progressive loss of the class of individuals in a population with the fewest mutations, in a phenomenon called **Muller's Ratchet** (Muller 1964; Maynard Smith 1978; Charlesworth & Charlesworth 1997). The name is an analogy to a mechanical device like a ratchet wrench that permits rotation in only one direction. Muller's Ratchet results in the accumulation of more and more mutations in a population, which leads to ever-declining average fitness in populations if most mutations are deleterious. Thus, Muller's Ratchet demonstrates a selective advantage of recombination under some conditions.

To see how Muller's Ratchet works in detail, consider a finite population of haploid individuals that

reproduce clonally. Assume that all mutations at all loci are equally deleterious and acted against by natural selection to the same degree. The selective disadvantage is  $s$  at each locus with a mutation and the total selection coefficient against an individual with  $n$  mutated loci is  $(1 - s)^n$ . Further, assume that mutation is irreversible and can only make deleterious alleles from wild-type alleles but not wild-type alleles from deleterious ones. Initially, all individuals in the population start off with no mutations. Mutations that occur decrease the proportion of individuals with no mutations and increase the frequencies of individuals with 1, 2, 3 . . .  $n$  mutations. Over time, the frequency of the zero mutation category declines while the frequencies of individuals with one or more mutations increases. This process can be seen in the top two panels of Fig. 5.7.



**Figure 5.7** Simulation results show the action of Muller's Ratchet in increasing the number of deleterious mutations in the absence of recombination. Initially, all haploid individuals in the population have zero mutations. Mutations occur randomly over time and continually reduce the frequency of individuals with fewer mutations. Genetic drift causes sampling error and the stochastic loss of mutation classes with few individuals. Individuals with more mutations are less likely to reproduce, due to natural selection against deleterious alleles. Once the category with fewest mutations (e.g. the zero mutations class) is lost due to genetic drift and mutation, there is no process that can repopulate it. Therefore, the distribution of the number of mutations continually moves to the right but can never move back to the left. The simulation parameters were  $N_e = 200$  and  $\mu = 0.06$ , each mutation reduced the chance of reproduction by 1%, and each individual had 100 loci.

### Interact box 5.2 Muller's Ratchet

Selecting the **Muller's Ratchet** module from the **Mutation** menu of PopGene.<sup>S<sup>2</sup></sup> opens a simulation of the fate of new deleterious mutations in a finite population of chromosomes that lack recombination. The simulation starts out with a population of haploid, clonal individuals that have no mutations and then lets mutation, genetic drift, and natural selection act. The fitness of each individual determines its chances of contributing progeny to the next generation. The number of progeny produced by each individual is Poisson-distributed with a mean of one. The effective size of the population, the coefficient of selection against deleterious mutations, and the mutation rate can be set in the simulation. The results are given in terms of the proportion of individuals in the population with a given number of mutations.

Initially, run the simulation using the default values. Then try independently increasing the effective population size (or the population size of haploid chromosomes), the selection coefficient against the deleterious mutations, and the chance of a deleterious mutation. Predict the impact of each simulation parameter on the frequency distribution of the number of mutations per genome before you change each parameter.

Genetic drift and natural selection are also acting along with mutation on the frequencies of individuals with different numbers of mutations. The sampling error of genetic drift can result in the stochastic loss of mutation categories with a low frequency in the population. This effect of genetic drift works regardless of the number of mutations. Any category of mutations lost by drift can be reestablished via mutation from individuals with fewer mutations. However, when all the individuals with the lowest number of mutations are lost from the population that fewest-mutation category is gone forever. This is because mutation cannot make wild type-alleles that would reduce the number of deleterious mutations. Also, the fewest-mutation category cannot be reconstituted because there is no recombination. The overall effect of genetic drift is to push the frequency distribution of the number of mutations toward higher numbers. In contrast, natural selection tends to push the distribution of the number of mutations toward lower numbers since individuals with more mutations are increasingly disfavored by natural selection.

If the effective population size is small, Muller's Ratchet also leads to accelerated rates of fixation to a single allele within the category of individuals with fewest mutations. This occurs since the category of fewest mutations is not renewed by mutation. It is also finite and consists of alleles that have identical fitness, so that genetic drift will eventually cause fixation of a single allele within that mutation category. This effect has implications for genomes with low levels of recombination or in diploid populations with mating systems that lead to high levels of homo-

zygosity that effectively nullify recombination. In these situations, fixation may occur at higher rates than for deleterious mutations in genomes with free recombination and the same effective population size (see Charlesworth & Charlesworth 1997).

### 5.3 Mutation models

- The infinite alleles,  $k$  alleles, and stepwise mutation models.
- Understanding the implications of mutation models using the standard genetic distance and  $R_{ST}$ .
- The infinite sites and finite sites mutation models for DNA sequences.

Mutation acts in diverse ways and can produce a wide range of changes at the level of alleles and DNA sequences. To study the allele frequency consequences of mutation it is helpful to construct some simplifying models of the mutation process itself. **Mutation models** attempt to capture the essence of the genetic changes caused by mutation while at the same time simplifying the process of mutation into a form that permits generalizations about allele frequency changes. There is no single model of the process of mutation, but rather a series of models that serve to encapsulate different features of the mutation process for different classes of loci and different types of alleles. Often, mutation models are motivated by molecular methods such as allozyme electrophoresis or DNA sequencing used to assay genetic variation in actual populations. This section introduces and describes the major classes of mutation models. Two types of mutation model for

discrete alleles are applied in measures of genetic difference between populations to show the role mutation models play in the interpretation of genetic differences. Mutation models for DNA sequences are applied in the last section of the chapter on mutation in genealogical branching models.

### **Mutation models for discrete alleles**

A repeated theme in earlier chapters was determining expected levels of homozygosity and heterozygosity (autozygosity and allozygosity) under different population genetic processes. A key assumption in many of these expectations is that identity in state can be treated as identity by descent. In other words, alleles identical in state look alike because they descended from a common ancestral allele copy at some point in the past. The **infinite alleles model** of mutation (see Kimura & Crow 1964) is an assumption used to guarantee that identity in state is equivalent to identity by descent. Under the infinite alleles model, each mutational event creates a new allele unlike any other allele currently in the population. Once a given allelic state is made by mutation the first time it can never be made by mutation ever again. In essence, the allelic state is crossed off the list of possible mutations. The infinite alleles model, serves to avoid the possibility that two alleles are identical in state but not identical by descent, as can occur if the same allele can be made by mutation repeatedly over time. Under the infinite alleles model, mutation produces the original copy of each allele but is not an ongoing process influencing the frequency of any allele already in the population. Processes other than mutation are responsible for allele and genotype frequencies after an allele exists in a population. An additional consequence is that the evolutionary “distance” or number of transition events between all alleles is the same, since all alleles are produced by a single mutational event and alleles can never accumulate multiple mutations. This means that all alleles can be treated as equivalent when estimating heterozygosity or fixation indices.

The infinite alleles model might roughly approximate the mutational process for molecular markers like allozymes since alleles take discrete states (e.g. fast or slow migration on a gel) and allozyme loci are generally observed to have low mutation rates so it is likely that most alleles in a sample are not recently the product of mutation. A length of DNA sequence might also approximate the infinite alleles model. In a sequence of 500 nucleotides there are

$4^{500} = 1.072 \times 10^{301}$  unique combinations of the nucleotides. If mutation is purely random and mutation changes an existing nucleotide to any other nucleotide with equal probability, many mutations could occur in a population of DNA sequences without producing a duplicate allele since there are a truly staggering number of possible alleles.

**Homoplasy** The condition where allelic states are identical without the alleles being identical by descent.

**Infinite alleles model** A model where each mutational event creates a new allele unlike any other allele currently in the population so that identity in state for two or more alleles is always a perfect indication of identity by descent.

**k alleles model** A mutation model where each allele can mutate to each of the other  $k - 1$  possible allelic states with equal probability.

**Stepwise mutation model** A mutation model where the allelic states produced by mutation depend on the initial state of an allele. Alleles with a greater difference in state are therefore more likely to be separated by a greater number of past mutational events.

There are several features of the mutational process that the infinite alleles model does not account for, and therefore there are a number of other mutation models. Obviously, there are not an infinite number of alleles possible at actual genetic loci. The **k alleles model** of mutation is an alternative to the infinite alleles model where  $k$  refers to a finite integer representing the number of possible alleles. In this model each allele can mutate with equal probability to each of the other  $k - 1$  possible allelic states. With the  $k$  alleles model the same allele can be created by mutation repeatedly, blurring the equivalence of identity in state and identity by descent. As the number of possible alleles or  $k$  decreases and as the mutation rate increases, allelic state becomes a poorer and poorer measure of identity by descent since an increasing proportion of alleles with identical states have completely independent histories. The term **homoplasy** refers to allelic states that are identical in state without being identical by descent.

The infinite alleles and  $k$  alleles models both assume that the allelic state produced by mutation is

independent of the current state of an allele. With these models each allele has an equal probability of mutating to any of the other allowable allelic states. It is also possible that the state of a new allele produced by mutation is not independent of the initial state of an allele. An example is the common observation that transitions are more common than transversions in diverged DNA sequences. The **stepwise mutation model** accounts for cases where allelic states are somehow ordered and the allelic states produced by mutation depend on the initial state of an allele (Kimura & Ohta 1978). Mutations by slipped-strand mispairing at microsatellite or simple sequence repeat loci produce new allelic states within one or a few repeats of the initial allelic state much more often than mutations that are many repeats different than the initial allelic state. Microsatellite loci are therefore a prime example of ordered, stepwise mutation where alleles closer in state are more likely to be recently identical by descent than alleles that are very different in state. See evidence for human microsatellite loci in Valdes et al. (1993).

The role of mutation models is illustrated in summary measures that express the genetic similarity or dissimilarity of individuals or populations, called **genetic distances**. The **standard genetic distance** or **D** measure developed by Nei (1972, 1978) has been widely employed. Given allele frequencies for several populations, **D** (not to be confused with the measure of gametic disequilibrium) expresses the probability that two alleles each randomly sampled from two different subpopulations will be identical in state relative to the probability that two alleles randomly sampled from the same subpopulation are identical in state. Table 5.4 gives hypothetical allele frequencies at one locus in two subpopulations that can be used to compute **D**. With random mating, the total probability that two identical alleles are sampled from subpopulation 1 is

$$\begin{aligned} J_{11} &= \sum_{k=1}^{\text{alleles}} p_{1k}^2 = (0.6)^2 + (0.3)^2 + (0.1)^2 \\ &= 0.46 \end{aligned} \quad (5.13)$$

and the total probability that two identical alleles are sampled from subpopulation 2 is

$$\begin{aligned} J_{22} &= \sum_{k=1}^{\text{alleles}} p_{2k}^2 = (0.4)^2 + (0.6)^2 + (0.0)^2 \\ &= 0.52 \end{aligned} \quad (5.14)$$

where  $p_{ik}$  indicates the frequency of allele  $k$  in population  $i$ . The total probability of sampling an identical allele from subpopulation 1 and subpopulation 2 is

$$\begin{aligned} J_{12} &= \sum_{k=1}^{\text{alleles}} p_{1k} p_{2k} = (0.6)(0.4) + (0.3)(0.6) \\ &\quad + (0.1)(0.0) = 0.42 \end{aligned} \quad (5.15)$$

The normalized genetic identity for this locus is then

$$I = \frac{J_{12}}{\sqrt{J_{11}J_{22}}} = \frac{0.42}{\sqrt{(0.46)(0.52)}} = 0.8589 \quad (5.16)$$

which is used to compute the genetic distance as

$$D = -\ln(I) = -\ln(0.8589) = 0.152 \quad (5.17)$$

When two subpopulations have identical allele frequencies  $J_{11}$  and  $J_{22}$  are equal,  $I$  is then one and the natural logarithm of one is zero, giving a genetic distance of zero.  $D$  has no upper limit. Although this genetic distance can be calculated for any pair of populations,  $D$  for completely isolated populations where divergence is due exclusively to mutation is expected to increase linearly with time under the infinite alleles model. This expectation relies on mutation not causing any homoplasy so that alleles

**Table 5.4** Hypothetical allele frequencies in two subpopulations used to compute the standard genetic distance,  $D$ . This example assumes three alleles at one locus, but loci with any number of alleles can be used.  $D$  for multiple loci uses the averages of  $J_{11}$ ,  $J_{22}$ , and  $J_{12}$  for all loci to compute the genetic identity  $I$ .

Allele	Subpopulation 1		Subpopulation 2	
	Frequency	$p_{ik}^2$	Frequency	$p_{ik}^2$
1	0.60	$p_{11}^2 = 0.36$	0.40	$p_{21}^2 = 0.16$
2	0.30	$p_{12}^2 = 0.09$	0.60	$p_{22}^2 = 0.36$
3	0.10	$p_{13}^2 = 0.01$	0.00	$p_{23}^2 = 0.00$

identical in state are always identical by descent. If the infinite alleles model is not met,  $D$  underestimates the true genetic distance because the mutational events that record the history of the populations will not be reflected perfectly in the allele frequencies.

With an awareness of mutation models, we can also reflect back on measures of genetic divergence among populations covered earlier in the book. Chapter 4 gave the expression for the fixation index among subpopulations relative to the total population

$$\left( F_{ST} = \frac{H_T - H_S}{H_T} \right).$$

It turns out that this assumes the infinite alleles model since it treats all alleles as being an equal number of mutational steps apart with all heterozygotes considered equally distant. There is an alternative method to compute the fixation index that depends instead on the stepwise mutation model where the fixation index is measured by

$$\hat{R}_{ST} = \frac{S_T - S_W}{S_T} \quad (5.18)$$

where  $S_T$  is twice the variance in allelic sizes in the total population and  $S_W$  is twice the average of the within subpopulation variance in allelic sizes (Slatkin 1995; Goodman 1997). The states of the alleles then influence the perceived amount of population subdivision. Alleles with states further apart (greater variance in state) are counted more heavily in the estimate of population structure since they are less likely to be recently identical by descent (multiple stepwise mutations would be required to make a large change in state). In contrast, alleles with very similar states (less variance in state) make a smaller contribution to the estimate of population subdivision since they are more likely to be recently identical by descent but changed in state due to mutation. Using the stepwise mutation model and  $R_{ST}$  accounts for high rates of mutation that can give the appearance of more or less gene flow than has actually occurred. Table 5.5 gives hypothetical genetic data from two subpopulations, illustrating the degree of population subdivision under the infinite alleles and stepwise mutation models.

**Table 5.5** A comparison of hypothetical estimates of population subdivision assuming the infinite alleles model using  $F_{ST}$  or assuming the stepwise mutation model using  $R_{ST}$ . Allelic data expressed as the number of repeats at a hypothetical microsatellite locus are given for two subpopulations in each of two cases. In the case on the left, the majority of alleles in both populations are very similar in state. Under the stepwise mutation model the two alleles are separated by a single change that could be due to mutation. The estimate of  $R_{ST}$  is therefore less than the estimate of  $F_{ST}$ . In the case on the right, the two populations have alleles that are very different in state and more than a single mutational change apart under the stepwise mutation model. In contrast, all alleles are a single mutational event apart in the infinite alleles model. The higher estimate of  $R_{ST}$  reflects greater weight given to larger differences in allelic state.

	Case 1	Case 2
Subpopulation 1 (number of repeats)	9, 10, 10, 10, 10, 10, 10, 10, 10, 10	9, 10, 10, 10, 10, 10, 10, 10, 10, 10
Subpopulation 2 (number of repeats)	12, 11, 11, 11, 11, 11, 11, 11, 11, 11	19, 20, 20, 20, 20, 20, 20, 20, 20, 20
Allele size variance in subpopulation 1, $S_1$	0.10	0.10
Allele size variance in subpopulation 2, $S_2$	0.10	0.10
Allele size variance in total population, $S_T$	0.947	52.821
$R_{ST}$	0.789	0.996
Expected heterozygosity in subpopulation 1, $H_1$	0.18	0.18
Expected heterozygosity in subpopulation 2, $H_2$	0.18	0.18
Average subpopulation expected heterozygosity, $H_S$	0.18	0.18
Expected heterozygosity in total population, $H_T$	0.59	0.59
$F_{ST}$	0.695	0.695

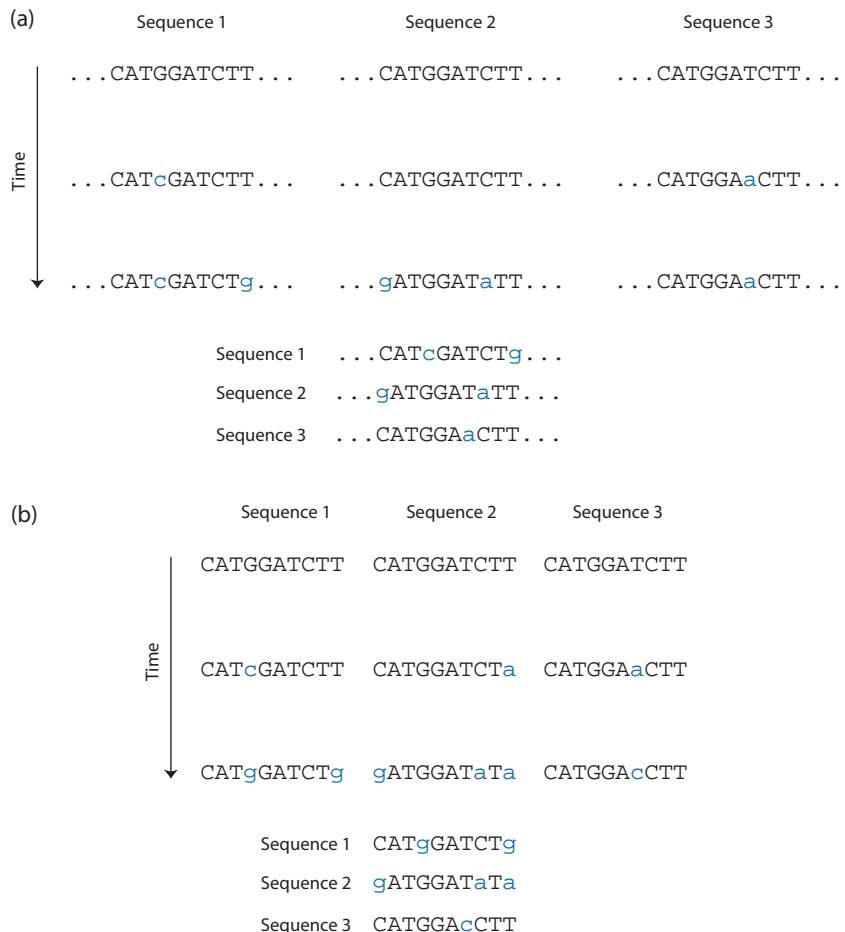
### Interact box 5.3

$R_{ST}$  and  $F_{ST}$  as examples of the consequences of different mutation models

Under the infinite alleles model allelic state is irrelevant in estimating population structure. However, in the stepwise mutation model, allelic states are weighted in the total estimate of population structure. Computing  $R_{ST}$  and  $F_{ST}$  for two subpopulations in a Microsoft Excel spreadsheet will help you develop a better understanding of how mutational models influence the perception of population structure. Enter your own allelic state values in the Excel spreadsheet to explore how allelic state differences as well as allele frequencies produce different estimates of the amount of population structure.

### Mutation models for DNA sequences

There are two widely used conceptual models of the process of mutation operating on DNA sequences (note that these types of models also apply in principle to amino acid sequences). One approximation for the process of mutation with DNA sequences is the **infinite sites model**. Each allele is an infinite DNA sequence and each mutation occurs at a different position along the DNA sequence. The infinite sites model can be thought of as an infinite alleles model built specifically for DNA sequences. A key distinction is that the infinite sites model permits the process of mutation to act on each allele in a population any number of times. As a consequence, the evolutionary “distance” between pairs of alleles can vary since a few or many sites differ between pairs of alleles depending on how many mutations have occurred for each allele. Figure 5.8a shows an example of how mutations might occur for DNA sequences under the infinite sites model. For example, after the fourth base-pair position (or site)



**Figure 5.8** Patterns of mutational change in DNA sequences under the infinite sites (a) and finite sites (b) models. Base-pair states created by a mutation are in blue lower-case letters. In the infinite sites model sequences that are identical in state at the same site are identical by descent because mutations only occur once at each site. In contrast, the finite sites model shows how multiple mutations at the same site act to obscure the history of identity based on comparisons of site differences among DNA sequences. The ellipses (...) that surround the sequences in (a) indicate that each sequence has infinitely many sites of which only 10 are displayed.

of sequence 1 mutates from G to C, no more mutations can take place at that site. The sites where mutations took place can therefore all be distinguished in alignment of the sequences since each site only experiences a mutation once. Although other processes such as genetic drift and natural selection may influence the frequency of the sequences, we can conclude that sequences sharing the same base at a site are identical by descent.

Although no DNA sequence is infinite, the infinite sites model is a reasonable approximation if not too much time has passed since sequences shared a common ancestor. If mutation occurs randomly and with equal probability at each site, then any single site has a small chance of experiencing a mutation twice (e.g. the rate of mutation per site squared is small). Over a relatively short period of time, say 1000s of generations, only a few mutations are likely to occur and so it is unlikely that one site mutates more than once.

However, actual DNA sequences are finite and the time period for mutations to occur can be very long, so a mutation model taking these facts into account is useful. The **finite sites model** is used for DNA sequences of a finite length. It is similar to the infinite sites model except that now the number of sites is finite and each site can experience a mutation more than once. Multiple mutations have the potential to obscure past mutational events as shown in Fig. 5.8b. For example, two sequences are either identical or different at each site even though a site where they differ may have mutated more than once in the past. The fourth site in sequence 1 is such a case. Although there have been two mutations at that site, the second mutation leads to the same nucleotide that was originally in that position. However, in the alignment of all three sequences the fourth site is identical and it is not possible to detect the two mutation events that occurred for sequence 1. Consider a similar example of site 7 in sequence 3 and what happens when we compare pairs of sequences. Sequences 2 and 3 differ by four sites (1, 7, 8, and 10) but there are actually five mutational events that separated them in the past. Sequences 1 and 3 differ at three sites (4, 7, and 10) but there are actually five mutation events separating them. Thus, multiple mutational changes at the same site work to obscure the complete history of mutational events that distinguish DNA sequences.

The possibility of multiple mutational changes at the same site, often called **multiple hits**, leads to saturation of mutational changes over time as mutations

**Infinite sites model** A model for the process of mutation acting on infinitely long DNA sequences where each mutation occurs at a different position along the DNA sequence and the same position cannot experience a mutation more than once.

**Finite sites model** A model for the process of mutation acting on DNA sequences of finite length so that the same site may experience a mutation more than once.

occur more times at the same sites. Saturation can be “corrected” using nucleotide substitution models that estimate and adjust for multiple mutations at the same site to estimate the “true” number of events that separate two sequences. One such correction called the Jukes–Cantor model is covered in Chapter 8.

One way to understand the impacts of multiple hits is to imagine a situation similar to the beakers containing micro-centrifuge tubes in Chapter 3. Now the beakers contain a very large number of nucleotides (A, C, T, and G) at equal frequencies. Imagine composing two DNA sequences by drawing nucleotides from the beaker. The chance that a given nucleotide is sampled at random is 25%. Therefore, given one random DNA sequence there is a 25% chance that another random DNA sequence shares an identical base pair at the same site. Therefore, DNA sequences that have experienced many mutations at the same site are expected to be identical for 25% of their base pairs. Therefore, when there is the possibility of multiple hits, identity in state is not a perfect indicator of identity by descent.

#### 5.4 The influence of mutation on allele frequency and autozygosity

- Irreversible and bi-directional mutation models.
- The parallels between the processes of mutation and gene flow.
- Expected autozygosity at equilibrium under mutation and genetic drift.
- Expected heterozygosity and the biological interpretation of  $\theta$ .

In developing expectations for allele and genotype frequencies to this point, all processes served only to shape existing genetic variation. To understand the consequences of mutation requires models that predict allele and genotype under the constant input

of genetic variation by mutation. This section presents three models for the process of mutation. The first two models are related and ask how recurrent mutation is expected to change allele frequencies over time in a population. The third model predicts genotype frequencies when genetic drift and mutation are both operating, showing how the combination of these processes influences autozygosity in a population.

Let's develop two simple models to predict the impact of constant mutation on allele frequencies (sometimes called **mutation pressure**) in a single panmictic population that is very large. Both models will focus only on the process of mutation and leave out other processes such as genetic drift or natural selection. Consider one locus with two alleles, A and a, where the frequency of A is represented by  $p$  and the frequency of a is represented by  $q$ . For the first model, assume that mutation operates to change A alleles into a alleles but that a alleles cannot mutate into A alleles. This is called the **irreversible mutation** model. The chance that mutation changes the state of each A allele every generation is symbolized by  $\mu$  (pronounced "mu"). The frequency of the A allele after one generation of mutation is then

$$p_{t+1} = p_t(1 - \mu) \quad (5.19)$$

where the  $(1 - \mu)$  term represents the proportion of A alleles that do not mutate to a alleles at time  $t$ . As long as  $\mu$  is not zero, then the frequency of A alleles will decline over time because  $1 - \mu$  is less than one. This also must mean that the proportion of the a alleles increases by  $\mu$  each generation. If the mutation

rate is constant over time, then the allele frequency after an arbitrary number of generations is

$$p_t = p_0(1 - \mu)^t \quad (5.20)$$

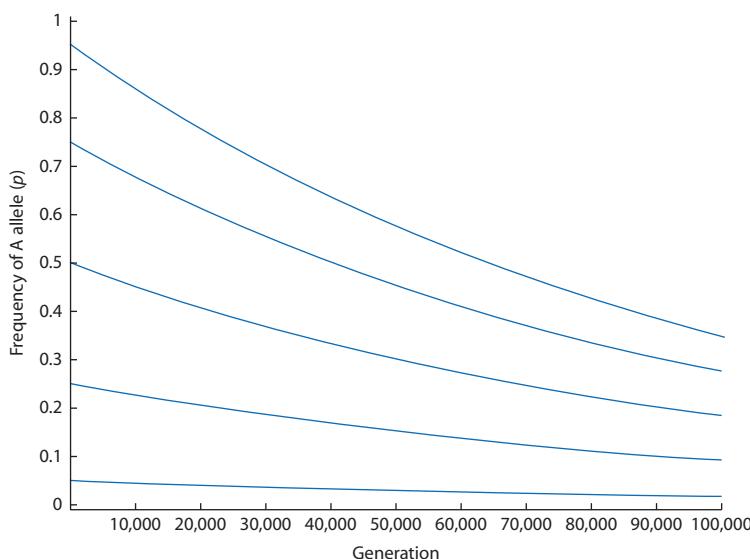
where  $p_0$  is the initial allele frequency and  $t$  is the number of generations that have elapsed.

With irreversible mutation, eventually all A alleles will be transformed into a alleles by mutation since there is no process that replaces A alleles in the population. Figure 5.9 shows the expected frequencies of the A allele over time starting at five different initial allele frequencies when the mutation rate is  $\mu = 1 \times 10^{-5}$  or 0.00001. Notice that the time scale to reduce the frequency of the A allele is very long. In this example, the equilibrium allele frequency of  $p = 0$  has not been reached even after 100,000 generations. In fact, it takes 69,310 generations to halve the frequency of A with this mutation rate (the halving time is determined by setting  $(1 - \mu)^t = \frac{1}{2}$ )

**Irreversible mutation** For a locus with two alleles, a process of mutation that changes A alleles into a alleles but does not change a alleles to A alleles.

**Mutation pressure** The constant occurrence of mutations that add or alter allelic states in a population.

**Reversible or bi-directional mutation** For a locus with two alleles, a process of mutation that changes A alleles into a alleles and also changes a alleles to A alleles.



**Figure 5.9** Expected change in allele frequency due to irreversible or one-way mutation for a diallelic locus for five initial allele frequencies. Here the chance that an A allele mutates into an a allele (or the per locus rate of mutation) is 0.00001. This rate of mutation is high compared with estimates of the per-locus mutation rate (see Table 5.1). The expected equilibrium allele frequency is  $p = 0$  since there is no process acting to replace A alleles in the population. The population has not reached equilibrium even after 100,000 generations have elapsed. Changes in allele frequency due to mutation alone occur over very long time scales.

even using a mutation rate at the high end of the observed range (Table 5.1). To generalize from the irreversible mutation model, we can expect that the process of mutation does influence allele frequencies but that substantial changes to allele frequency caused by mutation alone will take thousands or tens of thousands of generations depending on the mutation rate.

The assumption of irreversible mutation is not biologically realistic. Mutation can usually change the state of all alleles, resulting in both forward ( $A \rightarrow a$ ) and reverse ( $a \rightarrow A$ ) mutation for a diallelic locus. The **bi-directional or reversible mutation** model takes this possibility into account using independent rates of forward mutation ( $\mu$ ) and reverse mutation ( $\nu$ , pronounced “nu”). With mutation pressure in both directions we can again ask how mutation will change allele frequencies in a population over time. Each generation,  $\mu$  of the  $A$  alleles mutate to  $a$  alleles while at the same time  $\nu$  of the  $a$  alleles mutate to  $A$  alleles. The allele frequency after one generation is therefore

$$p_{t+1} = p_t(1 - \mu) + (1 - p_t)\nu \quad (5.21)$$

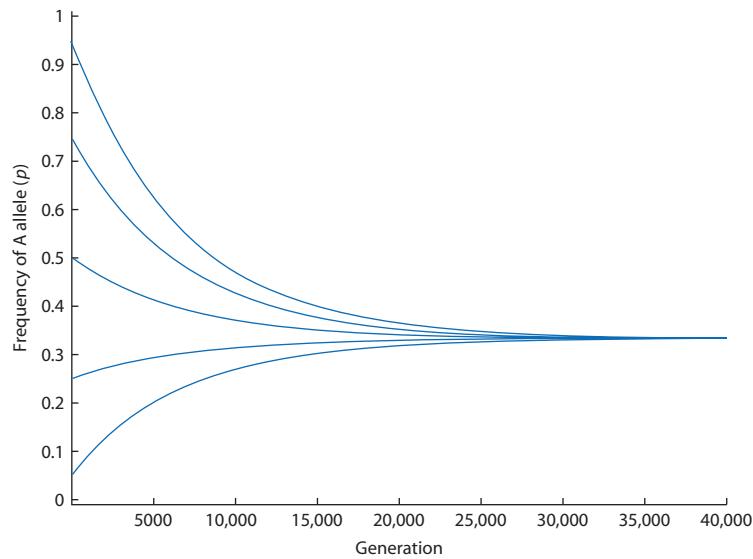
because the frequency of  $A$  alleles will decline due to the proportion of alleles that experience forward mutation ( $p_t(1 - \mu)$ ) but increase due to the proportion of the alleles in the population that experience reverse mutation ( $(1 - p_t)\nu$ ). The general result is that the equilibrium value of the frequency of  $A$  is determined by the net balance of the two rates of mutation:

$$p_{equilibrium} = \frac{\nu}{\mu + \nu} \quad (5.22)$$

**Figure 5.10** Expected change in allele frequency due to reversible or two-way mutation for a diallelic locus for five initial allele frequencies. Here the chance that an  $A$  allele mutates to an  $a$  allele ( $A \rightarrow a$ ) is 0.0001 and the chance that an  $a$  allele mutates to an  $A$  allele ( $a \rightarrow A$ ) is 0.00005. These mutation rates are toward the high end of the range of estimated mutation rate values (see Table 5.1). The expected equilibrium value is  $p = 0.333$  according to equation 5.22, an allele frequency that is reached only after tens of thousands of generations. The time to equilibrium is proportional to the absolute magnitudes of the mutation rates while the equilibrium value depends only on a function of the ratio of the mutation rates.

as derived in Math box 5.1. So whatever the starting frequency of the  $A$  allele, the population will converge to  $p_{equilibrium}$  that is closer to one for the allele produced by the higher of the two mutation rates. Figure 5.10 shows the frequency of the  $A$  allele over time with bi-directional mutation for five different initial allele frequencies. Because the forward and backward mutation rates used for the figure are not equal but are within a factor of five, both alleles have intermediate frequencies at equilibrium. The number of generations required to reach the equilibrium allele frequency is again very long, just as it is with the irreversible mutation model.

It turns out that the process of mutation within a population is exactly analogous to the process of gene flow among several subpopulations. Compare allele frequency in Fig. 5.9 with irreversible mutation and Fig. 4.13 which shows allele frequency under one-way gene flow in the continent-island model. Both processes cause allele frequencies to change toward a state of fixation and loss and the shape of both curves is identical. Then compare allele frequency in Fig. 5.10 with the process of bi-directional mutation and the process of bi-directional gene flow in the two-island model shown in Fig. 4.14. Here too the shape of both curves is identical and both processes result in intermediate frequencies of both alleles at equilibrium. The major differences in the mutation and gene-flow graphs are the time scales. In the absence of other processes, gene flow causes allele frequencies to approach equilibrium values in tens or hundreds of generations whereas mutation requires tens to hundreds of thousands of generations to approach equilibrium allele frequencies. It is



### Math box 5.1 Equilibrium allele frequency with two-way mutation

To determine the equilibrium allele frequency for a diallelic locus with the possibility of both backward and forward mutation, we take the basic equation that predicts allele frequency over one generation:

$$p_{t+1} = p_t(1 - \mu) + (1 - p_t)\nu \quad (5.23)$$

and try to express it as

$$p_{t+1} - a = (p_t - a)b \quad (5.24)$$

where  $a$  and  $b$  are constants that depend only on the forward and backward mutation rates  $\mu$  and  $\nu$ . Expressing the equation in this way allows us to equate  $p_{t+1}$  with  $a$  if the  $(p_t - a)b$  term goes to zero under certain limiting conditions. Equation 5.24 can be rearranged by adding  $a$  to both sides:

$$p_{t+1} = (p_t - a)b + a \quad (5.25)$$

and then multiplying:

$$p_{t+1} = p_t b - ab + a \quad (5.26)$$

and lastly factoring terms containing  $a$  to give

$$p_{t+1} = p_t b + a(1 - b) \quad (5.27)$$

Equation 5.23 containing the mutation rates can be put into this same form by expanding to give

$$p_{t+1} = p_t - p_t\mu + \nu - p_t\nu \quad (5.28)$$

which then can be factored to give

$$p_{t+1} = p_t(1 - \mu - \nu) + \nu \quad (5.29)$$

important to understand that this difference in time scale is just a product of the vastly different rates for the two processes rather than a fundamental difference in the processes themselves. In the figures, the chance that a gamete migrated was one in 10 while the chance of an allele mutated was between one in 1000 and one in 10,000. While these rates are likely to be on the high end of the range of values found in natural

Comparing equations 5.27 and 5.29 we see that

$$b = (1 - \mu - \nu) \quad (5.30)$$

and

$$a(1 - b) = \nu \quad (5.31)$$

Substituting the expression for  $b$  into the equation above gives the solution for  $a$ :

$$a = \frac{\nu}{\mu + \nu} \quad (5.32)$$

We can then substitute these values of  $a$  and  $b$  into equation 5.24 to get a new expression for the change in allele frequency over one generation:

$$p_{t+1} - \frac{\nu}{\mu + \nu} = (p_t - \frac{\nu}{\mu + \nu})(1 - \mu - \nu) \quad (5.33)$$

Since the expression for change in allele frequency over any one generation interval is identical and the mutation rates are constant over time, we can recast the equation above in terms of the initial allele frequency  $p_0$  and the number of generations that have elapsed:

$$p_{t+1} - \frac{\nu}{\mu + \nu} = (p_0 - \frac{\nu}{\mu + \nu})(1 - \mu - \nu)^t \quad (5.34)$$

Notice that as the number of generations grows very large ( $t \rightarrow \infty$ ) the  $(1 - \mu - \nu)^t$  term approaches zero, making the entire right-hand side of the equation zero. Therefore, when many generations have elapsed, the equilibrium allele frequency is expected to be

$$p_{t+1} = \frac{\nu}{\mu + \nu} \quad (5.35)$$

populations, gene flow is expected to occur at much higher rates than mutation as a general rule. The conclusion from this comparison is that gene flow is a much more potent force to change allele frequencies at single loci over the short term compared to mutation. Mutation does have an effect, but it is longer term.

The parallel nature of the processes of gene flow and mutation can be used as an advantage to

### Interact box 5.4 Simulating irreversible and bi-directional mutation

Both the irreversible and two-way mutation models are available in PopGene.<sup>S<sub>2</sub></sup> by clicking on the **Mutation** menu and then selecting **Irreversible** or **Two-way**. In the irreversible model only the forward mutation rate can vary while the reverse mutation rate is always 0.0. As an example, compare how rapidly equilibrium is approached with forward mutation rates of 0.01 and 0.001 from an initial allele frequency of 0.9 and the scale set to 2000 generations. For the two-way model, compare approach to equilibrium over 2000 generations when both backward and forward mutation rates are equal (e.g. both 0.001) and when they are unequal (e.g. 0.0015 and 0.0005) starting at an initial allele frequency of 0.9. Note that these mutation rates serve as an illustration only and that biologically realistic mutation rates are usually much lower.

understand more about the process of mutation. In particular, we can learn more about how mutation will impact autozygosity in finite populations where genetic drift is also operating. Recall from Chapter 3 the expression for the level of autozygosity in a finite population caused by genetic drift:

$$F_t = \frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1} \quad (5.36)$$

Mutation breaks the chain of descent by changing the state of alleles and therefore reduces the probability that a genotype is composed of two alleles identical by descent (autozygous). Genotypes with no alleles, one allele, or two alleles impacted by mutation each generation have frequencies of  $(1 - \mu)^2$ ,  $2\mu(1 - \mu)$ , and  $\mu^2$ , respectively. Only the  $(1 - \mu)^2$  genotypes with no mutated alleles can contribute to the pool of alleles that may become identical by descent due to finite sampling. From the opposite perspective, we note that  $2\mu$  genotypes heterozygous and  $\mu^2$  genotypes homozygous for a new mutation are expected each generation. Together, these two classes of genotypes with mutations reduce the autozygosity by a factor of  $1 - 2\mu - \mu^2 = (1 - \mu)^2$ . (This is

identical to the reasoning used in Chapter 4 for the case of gene flow.)

Mutation will therefore reduce the autozygosity caused by finite sampling in the present generation (chance of  $\frac{1}{2N_e}$ ) by a factor of  $(1 - \mu)^2$ . In addition, mutation will also reduce any autozygosity from past generations ( $F_{t-1}$ ) because some alleles that are identical by descent may change to new states via mutation, leaving the proportion  $(1 - \mu)^2$  of genotypes unaffected by mutation and at the same level of autozygosity. Putting these two separate adjustments for the autozygosity together gives

$$F_t = \frac{1}{2N_e}(1 - \mu)^2 + (1 - \frac{1}{2N_e})(1 - \mu)^2 F_{t-1} \quad (5.37)$$

Assuming that the mutation rate is small and much, much less than the effective population size (see derivation in Math box 4.1 for the case of gene flow), an *approximation* for the expected amount of autozygosity at equilibrium in a finite population experiencing mutation is

$$F_{equilibrium} = \frac{1}{4N_e\mu + 1} \quad (5.38)$$

This result also depends on each mutation giving rise to a new allele that is not present in the population or the infinite alleles model. Since the allozygosity or expected heterozygosity is just one minus the autozygosity,

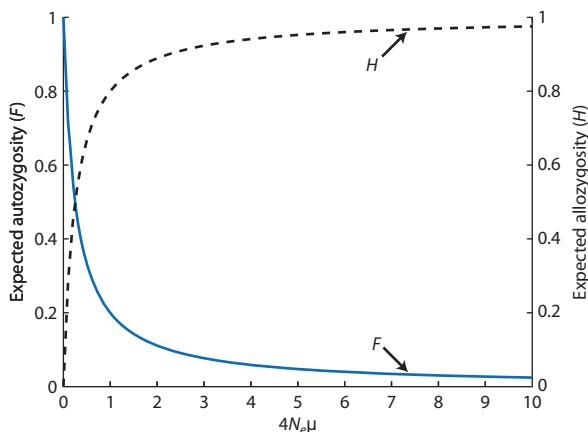
$$\begin{aligned} H_{equilibrium} &= 1 - F_{equilibrium} = \frac{4N_e\mu + 1}{4N_e\mu + 1} - \frac{1}{4N_e\mu + 1} \\ &= \frac{4N_e\mu}{4N_e\mu + 1} \end{aligned} \quad (5.39)$$

This is the expected heterozygosity in a finite population where the “push” on allele frequencies toward fixation and loss by genetic drift and the “push” on allele frequencies away from fixation and loss by mutation has reached a net balance.

The quantity  $4N_e\mu$  has a ready biological interpretation when  $N_e$  is large and  $\mu$  is small. In a population of  $2N_e$  alleles, the expected number of mutated alleles each generation is  $2N_e\mu$ . In a sample of two alleles that compose a diploid genotype, the chance that both alleles have experienced mutation and are therefore not identical by descent is  $4N_e\mu$ . For example, in

a population of  $2N_e = 100$  alleles with a mutation rate of one in 10,000 alleles per generation ( $\mu = 0.0001$ ), the expected number of mutations is 0.01 and the chance that a sample contains two alleles that are not autozygous is 0.02. The quantity  $4N_e\mu$  is frequently symbolized by  $\theta$  (pronounced “theta”). Under the infinite alleles model,  $\theta$  is the probability that two alleles sampled at random from a population at drift–mutation equilibrium will be allozygous. With  $\theta = 0.02$ , the expected heterozygosity at drift–mutation equilibrium is 0.0099. It is important to note that equilibrium heterozygosity will be lower than that predicted by  $\theta$  if the infinite alleles or infinite sites model is not met. This is the case because with a finite number of allelic states not all mutation events will produce a novel allele that forms an allozygous pair, or a heterozygote, when sampled with an existing allele in the population. In fact, mutations that make additional copies of existing alleles actually increase the perceived homozygosity due to homoplasy.

Figure 5.11 shows the expected probability of autozygosity and allozygosity at mutation–genetic drift equilibrium. At small values of  $4N_e\mu$  there will be an intermediate equilibrium level of autozygosity due to the balance of mutation introducing new alleles and genetic drift moving allele frequencies toward fixation or loss. As  $4N_e\mu$  gets large there is either



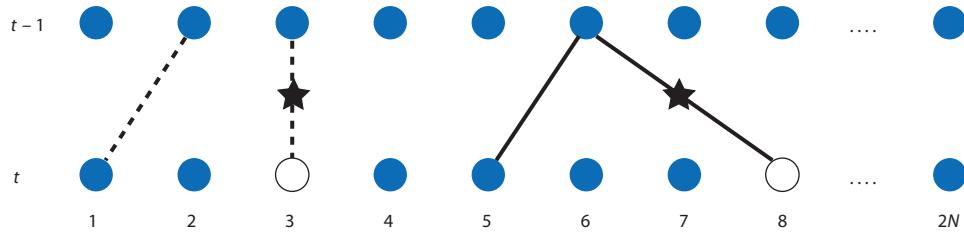
**Figure 5.11** Expected homozygosity ( $F$  or autozygosity, solid line) and heterozygosity ( $H$  or allozygosity, dashed line) at equilibrium in a population where the processes of both genetic drift and mutation are operating. The chance that two alleles sampled randomly from the population are identical in state depends on the net balance of genetic drift working toward fixation of a single allele in the population and mutation changing existing alleles in the population to new states. A critical assumption is the infinite alleles model, which guarantees that each mutation results in a unique allele and thereby maximizes the allozygosity due to mutations.

little drift or lots of mutation so there will be almost complete heterozygosity (no autozygosity). In the other direction,  $4N_e\mu$  near zero indicates very strong genetic drift or very infrequent mutation resulting in high levels of autozygosity and very low heterozygosity. Bear in mind that reaching the expected equilibrium autozygosity or heterozygosity will take many, many generations because mutation rates are low, making mutation a very slow process. If heterozygosity is perturbed from its mutation–drift equilibrium point, a population will take a very long time to return to that equilibrium.

## 5.5 The coalescent model with mutation

- Adding the process of mutation to coalescence.
- Longer genealogical branches experience more mutations.
- Genealogies under the infinite alleles and infinite sites models of mutation.

The genealogical branching model was introduced in Chapter 3 for a single finite population and then extended in Chapter 4 to account for branching patterns expected with population subdivision. The goal of those sections was to predict genealogical branching patterns without reference to the identity of the lineages represented by the branches. Those branching models need to be extended to account for the possibility that mutation occurs. Mutations will alter the genes or DNA sequences that are represented by each lineage or branch in the genealogical tree. Therefore, accounting for mutation will be a critical step in developing a coalescent model that explains differences among a sample of lineages in the present. This section will focus on the action of mutation in the coalescent model along with the state of each lineage in a genealogy. This is accomplished by coupling the process of coalescence and the process of mutation while moving back in time toward the most recent common ancestor. The ultimate goal is to build a genealogical branching model that can be used to predict the numbers and types of alleles that might be expected in a sample of lineages taken from an actual population. For example, one prediction might be the number of alleles expected in a single finite population for a given mutation rate. In this way, the combination of the coalescent process and the mutation process is used to form quantitative expectations about patterns of genetic variation produced by various population genetic processes.



**Figure 5.12** Haplod reproduction in the context of coalescent and mutation events. In a haplod population the probability of coalescence is  $\frac{1}{2N}$  (solid lines) while the probability that the two lineages do not have a common ancestor in the previous generation is  $1 - \frac{1}{2N}$  (dashed lines). The process of mutation can also occur simultaneously (stars), changing the state of alleles (filled circles to unfilled circles). Compare with Figure 3.23.

Building a coalescent model with mutation is as simple as adding another type of possible event that can occur between the present and some time in the past (Fig. 5.12). We will assume that both coalescent and mutation events are rare (or that  $N_e$  is large and the rate of mutation is small), so that when an event does occur going back in time it is *either* coalescence *or* mutation. In other words, we will assume that mutation and coalescence events are mutually exclusive.

Every generation there is the chance a mutation occurs. The rate of mutation,  $\mu$ , can be thought of as the chance that each lineage experiences a mutation each generation. The chance that a lineage does not experience a mutation is therefore  $1 - \mu$  each generation. The chance that  $t$  generations pass before a mutation event occurs is then the product of the chances of  $t - 1$  generations of no mutation followed by a mutation, or

$$P(T_{mutation} = t) = (1 - \mu)^{t-1}\mu \quad (5.40)$$

This equation has an identical form to the chance that a coalescent event occurs after  $t$  generations given in Chapter 3. Like the probability of coalescence, the probability of a mutation through time is a geometric series that can be approximated by the exponential distribution (see Math box 3.2).

To obtain the exponential expression or the exponent of  $e$  that describes the frequency of mutation events, we need to determine the rate at which mutations are expected to occur. When time is measured on a continuous scale with  $t = \frac{j}{2N_e}$ , one unit of time is equivalent to  $2N_e$  generations. If  $2N_e$  generations elapse and  $\mu$  is the rate of mutation per generation, then  $2N_e\mu$  mutations are expected

during one unit of continuous time. If we define  $\theta = 4N_e\mu$  then  $\theta/2$  is equivalent to  $2N_e\mu$ . This leads to the exponential approximation for the chance that a mutation occurs in a single lineage at generation  $t$ :

$$P(T_{mutation} = t) = e^{-\frac{\theta}{2}} \quad (5.41)$$

on a continuous time scale. When there is more than one lineage, each lineage has an independent chance of experiencing a mutation but only one lineage can experience a mutation. When events are independent but mutually exclusive, the probability of each event is added over all possible events to obtain the total chance that an event occurs. Adding the  $e^{-\frac{\theta}{2}}$  chance of a mutation for each lineage over all  $k$  lineages gives the total chance of mutation:

$$P(T_{mutation} = t) = e^{-\frac{\theta}{2}k} \quad (5.42)$$

for  $k$  lineages (compare this with the chances of coalescence at time  $t$  with  $k$  lineages,  $e^{-\frac{k(k-1)}{2}}$ , based on similar logic). The chance that a mutation occurs in one of  $k$  lineages at or before a certain time can then be approximated with the cumulative exponential distribution

$$P(T_{mutation} \leq t) = 1 - e^{-\frac{\theta}{2}k} \quad (5.43)$$

in exactly the same fashion that times to coalescent events are approximated.

When two independent processes are operating, the coalescence model becomes one of following lineages back in time and waiting for an event to happen. The possible events are mutation or

coalescence, so the total chance of *any* event is the sum of the independent probabilities of each type of mutually exclusive event. The total chance of an event occurring while going back in time (increasing  $t$ ) is then

$$P(T_{event} \leq t) = 1 - e^{-t \frac{k(k-1+\theta)}{2}} \quad (5.44)$$

where the exponent of  $e$  is  $-t \left[ k \frac{\theta}{2} + \frac{k(k-1)}{2} \right]$  or the

sum of the intensities of mutation and coalescence. When an event does occur at a time given by this exponential distribution, it is then necessary to decide whether the event is a coalescence or a mutation. The total chance that the event is either a mutation event or a coalescence event is  $\frac{k\theta}{2} + \frac{k(k-1)}{2}$ . Therefore the chance the event is a mutation is

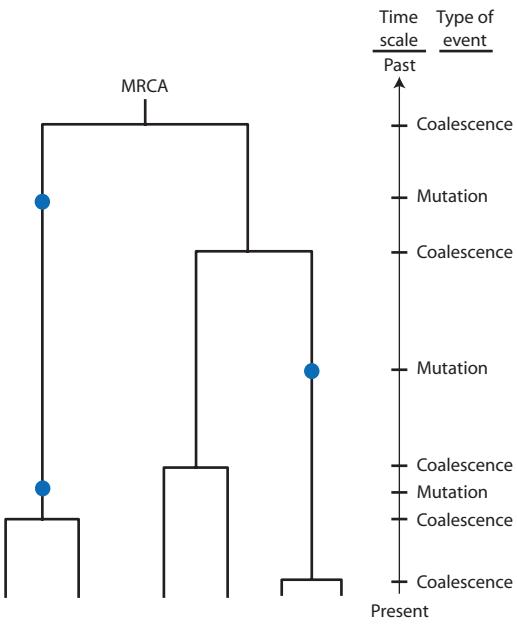
$$\frac{\frac{k\theta}{2}}{\frac{k\theta}{2} + \frac{k(k-1)}{2}} = \frac{\theta}{k-1+\theta} \quad (5.45)$$

while the chance that the event is a coalescence is

$$\frac{\frac{k(k-1)}{2}}{\frac{k\theta}{2} + \frac{k(k-1)}{2}} = \frac{k-1}{k-1+\theta} \quad (5.46)$$

Using the cumulative exponential distribution specified by equation 5.44 and then determining whether each event is a mutation or a coalescence, it is possible to construct a coalescent genealogy that includes the possibility of mutations occurring along each branch (Fig. 5.13).

The pattern of mutations on genealogical trees has some general features. Since the chance of mutation is assumed to be constant through time, the more time that passes, the greater the chance that a mutation occurs. This means that longer branches tend to experience more of the mutations on average in genealogical trees, while shorter branches are less likely to exhibit mutations. Recall the metaphor of branch length as a road that was used in Chapter 3 to describe the total branch length of a genealogy. If mutations are road signs with a constant chance of appearing per distance of roadway, then longer stretches of road are expected to



**Figure 5.13** A genealogy constructed under the simultaneous processes of coalescence in a single finite population and mutation. Working backward in time from the present, both mutation and coalescence events can occur. The blue dots represent mutation events, each assigned at random to a lineage present when the event occurred. Mutation events alter the state of a lineage, causing divergence from the ancestral state of the most recent common ancestor (MRCA) of all the lineages in the present.

have more signs. Applying this logic to genealogies like that in Fig. 5.13 tells us that more mutations are expected during the long average waiting time for coalescence with two lineages ( $k = 2$ ) than are expected to occur when there are six lineages that can coalesce ( $k = 6$ ). Another example would be the pattern of mutations expected for lineages in two demes with different levels of migration (see Fig. 4.17). With very limited migration, multiple mutations are expected on the long branches before the single lineages within each deme coalesce. Alternatively, when migration rates are high then many fewer mutations are expected between migration events. In the former case mutations cause lineages to diverge substantially between the two demes, while in the latter case the lineages in the different demes have less opportunity to accumulate differences.

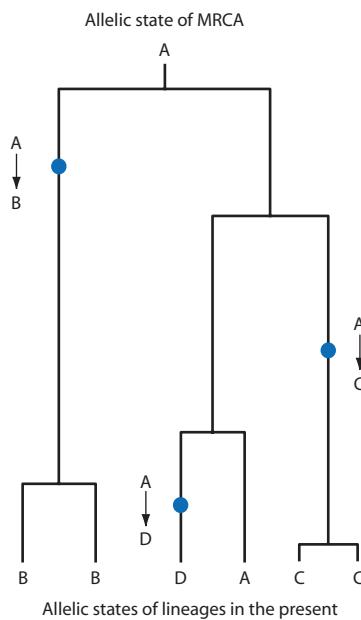
A genealogy with generic mutations like that shown in Fig. 5.13 is an abstraction until it is joined with a mutation model. Figure 5.14 shows the same genealogy under the assumptions that the most recent common ancestor has an allelic state of A and mutational changes follow the infinite alleles

### Interact box 5.5 Build your own coalescent genealogies with mutation

Again building a few coalescent trees can help you to better understand the evolution of genealogies when both the processes of mutation and coalescence are operating. You can use an expanded version of the Microsoft Excel spreadsheet used to build coalescent trees in Chapter 3 that now models waiting times for both mutation and coalescence. The spreadsheet contains the cumulative exponential distributions used to determine the time until a coalescent or mutation event (see equation 5.44) for up to six lineages. To determine the time that an event occurs for a given number of lineages  $k$  and mutation rate, a random number between zero and one is picked and then compared to the cumulative exponential distribution. The time interval on the distribution that matches the random number is taken as the event time. The next step is to determine whether the event was a mutation or a coalescence, again accomplished by comparing a random number to the chances of each type of event (equations 5.45 and 5.46).

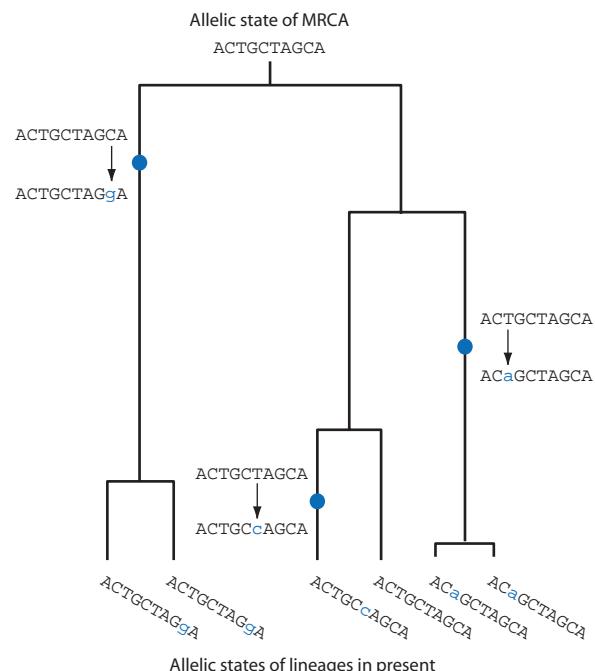
- Step 1 Open the spreadsheet and click on cells to view the formulas used, especially the cumulative probability of coalescence for each  $k$ . This will help you understand how the equations in this section of the chapter are put into practice. You can compare the cumulative probability distributions graphed for  $k = 6$  and  $k = 2$ .
- Step 2 Look at the section of the spreadsheet under the heading "Event times:" on the right side of the sheet. This section gives the waiting times until an event occurs and then determines if the event was a coalescence or a mutation. Press the recalculate key(s) to generate new sets of random numbers (see Excel help if necessary). Watch the times to an event change.
- Step 3 Now draw a genealogical tree with the possibility of mutations (do not recalculate again until Step 6 is complete). Along the bottom of a blank sheet of paper, draw six evenly-spaced dots to represent six lineages.
- Step 4 Start at the first "Decide event time:" panel to determine how much time passes (going backward in time) until either a mutation or a coalescence occurs. Then use the entries under "Decide what type of event:" to determine if the event was a coalescence or a mutation. If the event is a mutation go to Step 5, otherwise go to Step 6.
- Step 5 If the event is a mutation, draw the lines for all lineages back in time by a length equal to the waiting time (e.g. if the time is 0.5, draw lines that are 0.5 cm). Use the random number table to pick one lineage and draw an X on the lineage at the event time to indicate a mutation occurred. If a mutation occurred the number of lineages ( $k$ ) remains the same. Move down to the next "Decide event time:" panel and obtain the next event time for the same value of  $k$ . Repeat Step 5 until the event is a coalescent event.
- Step 6 Using the random number table, pick two lineages that will experience coalescence. Label the two left-most dots with these lineage numbers. Then, using a ruler, draw two parallel vertical lines that start at the end of the last event and extend as long as the time to coalescence in continuous time (e.g. if the time is 0.5, draw lines that are 0.5 cm). Connect these vertical lines with a horizontal line. Assign the lineage number of one of the coalesced lineages to the pair's single ancestor at the horizontal line. Record the other lineage number on a list of lineages no longer present in the population (skip over these numbers if they appear in the random number table). There are now  $k - 1$  lineages.
- Step 7 Return to Step 4 until all lineages have coalesced ( $k = 1$ ).

You should obtain a coalescence tree with mutation events on the branches like that in Fig. 5.13. Your trees will be different because the random coalescence and mutation times vary around their averages, but the overall shape of your trees (e.g. shorter branches when  $k$  is large) and frequencies of mutations (for a given mutation rate) will be similar.



**Figure 5.14** A genealogy constructed under the simultaneous processes of coalescence in a single finite population and mutation. Here the infinite alleles model of mutation is assumed to determine the allelic state of each lineage in the genealogy. Arbitrarily assigning allelic state A to the most recent common ancestor (MRCA), each mutational event then alters the state of the lineage experiencing the mutation. Each mutation changes the allelic state of the lineage to a new allele not present in the population, giving rise to a variety of allelic states among the lineages in the present.

model. Each mutation event is an instance where the current state of a lineage changes to an allelic state not currently present in the population. Because of mutational changes to the ancestral allelic state, the six lineages in the present represent four allelic states. Two of the alleles have a frequency of  $2/6 = 33\%$  while the remaining two alleles have a frequency of  $1/6 = 16.6\%$ . The lineages sharing B or C alleles are identical in state and therefore can be considered identical by descent. Figure 5.15 shows a third version of the genealogy with mutations, this time representing each allele as a DNA sequence and employing the infinite sites mutation model. Each mutational event alters a randomly picked site in the DNA sequence under the constraint that a site can only experience mutation a single time. The result is a set of DNA sequences that differ at three of 10 nucleotide sites. As with the infinite alleles model, DNA sequences in the present that are identical in state are therefore identical by descent.



**Figure 5.15** A genealogy constructed under the simultaneous processes of coalescence in a single finite population and mutation. Here the infinite sites model of mutation for DNA sequences is assumed to determine the allelic state of each lineage in the genealogy. Arbitrarily assigning the DNA sequence ACTGCTAGCA to the most recent common ancestor (MRCA), each mutational event then alters the DNA sequence of the lineage experiencing the mutation. Each mutation occurs at a random site in the DNA sequence that has not previously experienced a mutation (bases in blue lower-case letters), giving rise to differences in the DNA sequences among the lineages in the present. Here each base is equally likely to be produced by a mutation, although there are numerous models to specify the pattern of nucleotide changes expected by mutation. Under the finite sites model of mutation, each site in the DNA sequence could experience mutation repeatedly.

When a genealogy containing mutations is combined with a mutation model, it results in an explicit prediction of the diversity and types of allelic states expected under the processes that influence the branching patterns. Although the two examples shown here both utilize genealogies resulting from genetic drift in a finite population, mutation could also be combined with processes such as population structure, growing or shrinking population sizes, or natural selection that are used to generate a coalescent genealogy. Chapter 7 covers genealogies expected with natural selection and Chapter 8 explains methods to compare the expected patterns of allelic states in genealogies generated by different population genetic processes.

## Chapter 5 review

- The spectrum of relative fitness for genotypes containing mutations expresses the frequencies of mutations with a range of fitness effects. Mutations that are strongly deleterious or lethal will be purged by natural selection, while strongly advantageous mutations will be fixed by natural selection. Mutations with smaller effects may be neutral or nearly neutral and therefore be subject to stochastic fixation or loss due to genetic drift.
- New mutations may be lost simply by Mendelian segregation since there is a  $\left(\frac{1}{2}\right)^k$  chance that a given single allele will not be transmitted to  $k$  progeny. Neutral mutations are eventually lost while the chance of a new selectively favored mutation escaping loss (becoming fixed) is approximately twice its selective advantage, under the assumption that the population size is very large.
- New selectively neutral mutations are initially at a frequency of  $\frac{1}{2N_e}$  so the chance of fixation is  $\frac{1}{2N_e}$  and the chance of loss is  $1 - \frac{1}{2N_e}$ . Most new neutral mutations are expected to be lost from a population very rapidly since their initial frequency is very near zero. An average of  $4N_e$  generations will elapse before a new mutation reaches fixation.
- Fisher's geometric model of mutation shows that mutations with small effects on phenotype are more likely to be fixed by natural selection since these are the mutations with the greatest chance of improving fitness.
- The combination of mutation, genetic drift, and natural selection in genomes where recombination is absent or restricted leads to a growing accumulation of deleterious mutations in a phenomenon known as Muller's Ratchet.
- The infinite alleles model assumes discrete allelic states where each mutation creates an allele not currently present in the population. The infinite sites model assumes alleles are DNA sequences and that each mutation changes a single nucleotide at a site that has never before experienced mutation. Since mutation cannot form the same allele twice in both of these models, identity in state is always equivalent to identity by descent.
- Irreversible mutation will eventually lead to loss of the original allele in a population since

there is no process to restore the original allele. Bi-directional mutation leads to a net balance of two alleles changing state and an intermediate allele frequency that depends on the forward and reverse mutation rates. Both models show that mutation alone will take thousands or tens of thousands of generations to attain equilibrium allele frequency in a population, depending on the mutation rate.

- The process of mutation can be added to coalescent genealogies by modeling the waiting time to any event with an appropriate cumulative exponential distribution. When an event does occur, it can be either a coalescence or a mutation that is then reflected in the genealogy. More mutations are likely to occur on longer branches in genealogies since the chance of mutation is constant through time.
- Each mutation event in a genealogy can be interpreted under a specific mutation model such as infinite alleles or infinite sites. The combination of a coalescent genealogy containing mutations and a mutation model yields a prediction for the number and frequency of alleles expected under the process or processes that produced the genealogy.

## Further reading

For a detailed review of mutation rate estimates and the processes influencing the evolution of mutation rates see:

Drake JW, Charlesworth B, Charlesworth D, and Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148: 1667–86.

For an overview of the numerous evolutionary and population genetic phenomena related to mildly deleterious mutation see:

Charlesworth B and Charlesworth D. 1998. Some evolutionary consequences of deleterious mutation. *Genetica* 102/103: 3–19.

A full explanation of the experimental methods, assumptions, and statistical analysis used in mutation-accumulation studies along with a critical review of past results can be found in:

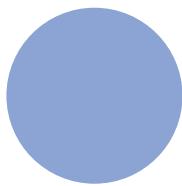
Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, and Willis J. 1999. Perspective: spontaneous deleterious mutation. *Evolution* 53: 645–63.

This paper provides an overview of the role mutation plays as the source of genetic variation in the larger process of adaptation by natural selection:

Orr HA. 2005. The genetic theory of adaptation: a brief history. *Nature Reviews Genetics* 6: 119–27.

For more detail on incorporating mutation into genealogical branching models see:

Hein J, Schierup MH, and Wiuf C. 2005. *Gene Genealogies, Variation and Evolution*. Oxford University Press, New York.



# CHAPTER 6

## Fundamentals of natural selection

### 6.1 Natural selection

- Translating Darwin's ideas into a model.
- Natural selection as differential population growth.
- Natural selection with clonal reproduction.
- Natural selection with sexual reproduction and its assumptions.

Charles Darwin's (1859) statement of the process of natural selection can be summarized as three basic observations about populations:

- all species have more offspring than can possibly survive and reproduce,
- individual organisms vary in phenotypes that influence their ability to survive and reproduce, and
- within each generation, the individuals possessing phenotypes that confer greater survival and reproduction will contribute more offspring to the next generation.

The result is that phenotypes which cause a predictably greater chance of survival and reproduction will increase in frequency over generations to the extent that such traits have a genetic basis. Darwin's observations initially served as a qualitative model, since an accurate model of genetic inheritance was lacking until Mendel's results were recognized. Once particulate inheritance was understood, the unification of genetics with the principle of natural selection took place in what is now called the **modern synthesis** or **neo-Darwinian synthesis** of evolutionary biology. The major challenge in the modern synthesis for population genetics was to develop expectations for the genetic changes that are caused by natural selection. This section of the chapter develops these basic population genetic expectations for natural selection.

#### *Natural selection with clonal reproduction*

At its core natural selection is actually a process of population growth, so let's start off by examining a

simple population growth model. If a population is assumed to have no upper limit in its size, the number of individuals one generation in the future ( $N_{t+1}$ ) is a product of the number of individuals present now ( $N_t$ ) multiplied by the **finite rate of increase** of the population  $\lambda$  (pronounced "lambda") to give the expression:

$$N_{t+1} = \lambda N_t \quad (6.1)$$

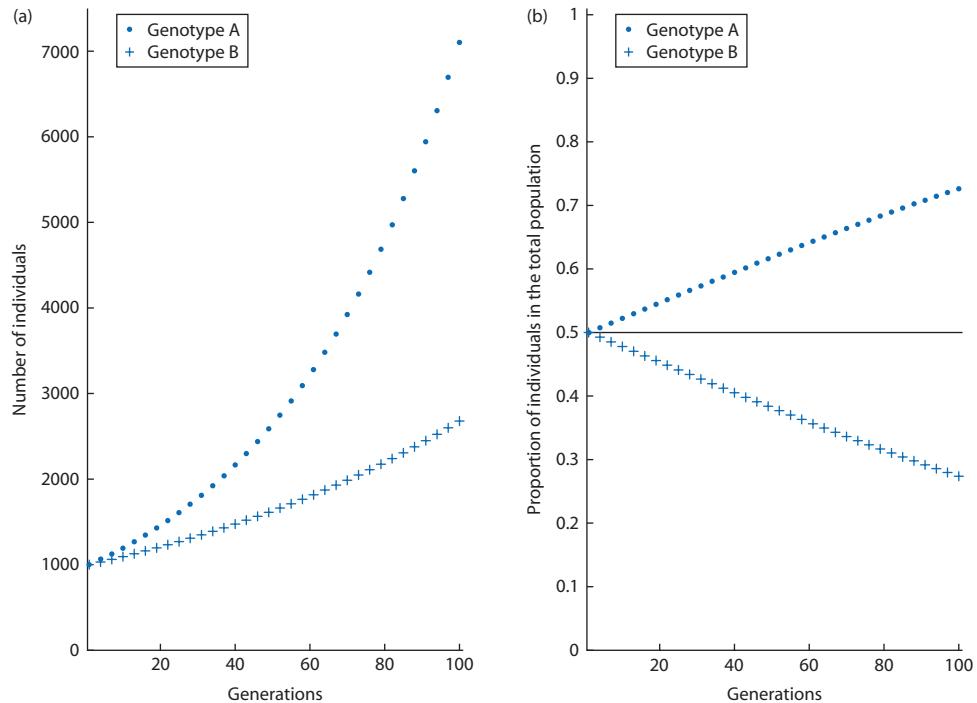
In this equation for unbounded population growth,  $\lambda$  is a multiplier that represents the net difference between the number of individuals lost from the population due to death and the number of new individuals recruited to the population by reproduction each generation. If the number of births and deaths are exactly equal then  $\lambda$  is one and the population does not change in size. If there are more births than deaths then  $\lambda > 1.0$  and the population grows whereas if there are more deaths than births then  $\lambda < 1.0$  and the population shrinks. The population growth rate can be thought of as the chance that an individual contributes one offspring to the next generation.

Natural selection is really just a special case of this basic population growth model where each genotype has its own growth rate. To see how this works, let's consider a population composed of two genotypes of an asexual organism like a bacterial species that reproduces only by clonal division over discrete generations. Call the two genotypes A and B with genotype-specific growth rates or **absolute fitnesses** of  $\lambda_A$  and  $\lambda_B$ . The proportions of each genotype in the total population in any generation are:

$$p = \frac{N_A}{N_A + N_B} \quad (6.2)$$

and

$$q = \frac{N_B}{N_A + N_B} \quad (6.3)$$



**Figure 6.1** Population growth in two genotypes with clonal reproduction, starting out with equal numbers of individuals and therefore equal proportions in the total population. Genotype A grows 3% per generation ( $\lambda = 1.03$ ) and genotype B grows 1% per generation ( $\lambda = 1.01$ ). (a) Individuals of both genotypes increase in number over time. (b) Because the genotypes grow at different rates, their relative proportions in the total population change over time. The solid line shows the initial equal proportions. Eventually, genotype A will approach 100% and genotype B 0% of the total population. Values are plotted for every third generation.

where  $N_A + N_B$  is the total population size. Figure 6.1a shows numbers of individuals over time for an example where genotype A grows faster than genotype B. In absolute numbers of individuals, the population sizes of both genotypes increase over time. However, the proportion of the population made up of individuals with A and B genotypes changes over time in the population (Fig. 6.1b). Since the A genotype grows faster, A individuals represent an increasing proportion of the individuals in the total population. This is equivalent to saying that  $p$  increases over time while  $q$  decreases over time. Thus, Fig. 6.1 shows a case of natural selection favoring the A genotype since it has a higher level of absolute fitness.

An alternative way to represent the changing proportions of the two genotypes in the population is to follow the ratio of the number of A and B individuals,  $N_A/N_B$ , over time. The value of the ratio at any point in time will depend on the initial numbers of A and B individuals (call them  $N_A(0)$  and  $N_B(0)$ ), the growth rates of the two genotypes, and the number of generations that have elapsed. The ratio  $N_B/N_A$  after one generation of population growth is given by

$$\frac{N_B(t=1)}{N_A(t=1)} = \left( \frac{\lambda_B}{\lambda_A} \right) \frac{N_B(0)}{N_A(0)} \quad (6.4)$$

which is akin to dividing a version of equation 6.1 for genotype A by a version of equation 6.1 for genotype B. In general, we can predict the ratio of  $N_A/N_B$  at any time  $t$  using

$$\frac{N_B(t)}{N_A(t)} = \left( \frac{\lambda_B}{\lambda_A} \right)^t \frac{N_B(0)}{N_A(0)} \quad (6.5)$$

by assuming that genotype-specific growth rates ( $\lambda_A$  and  $\lambda_B$ ) remain constant through time.

The ratio of genotype-specific growth rates is called the **relative fitness** and is represented by the symbol  $w$  in models of natural selection with discrete generations. Substituting the relative fitness for the ratio of the genotype-specific growth rates in equation 6.5 gives

$$\frac{N_B(t)}{N_A(t)} = w^t \frac{N_B(0)}{N_A(0)} \quad (6.6)$$

Recalling that  $N_A + N_B$  gives the total population size  $N$  at any point in time and then multiplying both

sides of equation 6.6 by  $\frac{1}{N}$  gives

$$\frac{\frac{N_B(t)}{N}}{\frac{N_A(t)}{N}} = w^t \frac{\frac{N_B(0)}{N}}{\frac{N_A(0)}{N}} \quad (6.7)$$

which can be simplified by utilizing equations 6.2 and 6.3:

$$\frac{q_t}{p_t} = w^t \frac{q_0}{p_0} \quad (6.8)$$

to represent each genotype by its proportion of the total population at any time  $t$ . When  $w = 1.0$  the two genotypes have identical growth rates and the proportion of each genotype remains constant in the population through time. If  $w > 1.0$  then the genotype in the numerator grows faster than the genotype in the denominator and it will represent a larger proportion of the population over time. Conversely, if  $w < 1.0$  then the genotype in the numerator grows less rapidly than the genotype in the denominator and it will represent a shrinking proportion of the population over time. Using the A genotype as the standard of comparison and the absolute fitness values from Fig. 6.1,  $w_A = 1.03/1.03 = 1.0$  and  $w_B = 1.01/1.03 = 0.981$  and so the frequency of the A genotype is expected to increase over time.

The relative fitness can be used to determine the change in frequency of a genotype over time, as shown in Table 6.1. The *change* in genotype frequency is the

**Table 6.1** The expected frequencies of two genotypes after natural selection, for the case of clonal reproduction. The top section of the table gives expressions for the general case. The bottom part of the table uses absolute and relative fitness values identical to Fig. 6.1 to show the change in genotype proportions for the first generation of natural selection. The absolute fitness of the A genotype is highest and is therefore used as the standard of comparison when determining relative fitness.

	Genotype	
	A	B
<b>Generation t</b>		
Initial frequency	$p_t$	$q_t$
Genotype-specific growth rate (absolute fitness)	$\lambda_A$	$\lambda_B$
Relative fitness	$w_A = \frac{\lambda_A}{\lambda_A}$	$w_B = \frac{\lambda_B}{\lambda_A}$
Frequency after natural selection	$p_t w_A$	$q_t w_B$
<b>Generation t + 1</b>		
Initial frequency $p_{t+1}$	$\frac{p_t w_A}{p_t w_A + q_t w_B}$	$\frac{q_t w_B}{p_t w_A + q_t w_B}$
Change in genotype frequency	$\Delta p = p_{t+1} - p_t$	$\Delta q = q_{t+1} - q_t$
<b>Generation t</b>		
Initial frequency	$p_t = 0.5$	$q_t = 0.5$
Genotype-specific growth rate (absolute fitness)	$\lambda_A = 1.03$	$\lambda_B = 1.01$
Relative fitness	$w_A = \frac{\lambda_A}{\lambda_A} = \frac{1.03}{1.03} = 1.0$	$w_B = \frac{\lambda_B}{\lambda_A} = \frac{1.01}{1.03} = 0.981$
Frequency after natural selection	$p_t w_A = (0.5)(1.0) = 0.5$	$q_t w_B = (0.5)(0.981) = 0.4905$
<b>Generation t + 1</b>		
Initial frequency $p_{t+1}$	$\frac{0.5}{0.5 + 0.4905} = 0.5048$	$\frac{0.4905}{0.5 + 0.4905} = 0.4952$
Change in genotype frequency	$0.5048 - 0.5 = 0.0048$	$0.4952 - 0.5 = -0.0048$

difference between frequencies in two generations,  $p_{t+1} - p_t$ . A difference is commonly symbolized with the Greek capital letter delta ( $\Delta$ ), so we could say that the change in the frequency of the A genotype is given by  $\Delta p = p_{t+1} - p_t$ . To generate an expression for  $\Delta p$  we can compare the initial genotype frequency  $p_t$  with its frequency a generation later after natural selection has acted via differential growth. We start with the basic expression for the difference in genotype frequency

$$\Delta p = p_{t+1} - p_t \quad (6.9)$$

If  $\Delta p$  is positive then the A genotype will increase in proportion in the population while it will decrease in proportion if  $\Delta p$  is negative. Substituting the expression for the expected frequency of the A genotype after natural selection (Table 6.1) gives

$$\Delta p = \frac{p_t w_A}{p_t w_A + q_t w_B} - p_t \quad (6.10)$$

The  $p_t w_A + p_t w_B$  term in the denominator on the right side of equation 6.10 is the **average relative fitness** of the population (it is a frequency-weighted average and so depends on the sum of the product of the frequency and relative fitness for each genotype). Positive values of  $\Delta p$  occur when the frequency of the A genotype after natural selection is greater than the average fitness of both genotypes after natural selection. The computations in Table 6.1 show that the frequency of the A genotype multiplied by its relative fitness ( $p_t w_A$ ) is greater than the average fitness so the A genotype will increase in proportion in the population over time. The average fitness will be covered in more detail when considering natural selection in sexual diploid populations.

One advantage of using the relative fitness is that the population growth rates of each genotype do not have to be known to model the proportions of the genotypes over time. Rather, the outcome of the growth process in terms of the relative frequencies of genotypes can be predicted strictly from the ratio of growth rates. This means that equation 6.8 potentially applies to organisms with very high absolute growth rates like bacteria as well as to species with absolute growth rates very near one such as elephants. Equation 6.8 even applies to cases where population sizes are declining through time. If a population is headed to extinction because it is composed of genotypes that all have growth rates less than one, the relative fitness will nonetheless accurately express the

change in the proportion of genotypes in the population over time. In practice, the relative fitness can be estimated in competition experiments where two or more genotypes are placed in the same environment and their proportions estimated at a later point in time (see Problem box 6.1, below).

**Absolute fitness** The genotype-specific rate of increase or population growth that predicts the absolute number of individuals of a given genotype in a population over time. Commonly symbolized as  $W$  or  $\lambda$ .

**Average or mean fitness** ( $\bar{w}$ ) The frequency-weighted sum of the relative fitness values of each genotype in the population.

**Relative fitness** The growth rate of genotypes relative to one genotype picked as the standard of comparison (often the genotype with the highest absolute fitness). Called Darwinian fitness after Charles Darwin and symbolized as  $w$  in models where time is represented in discrete generations (also called Malthusian fitness after Thomas Malthus and symbolized as  $m$  in models where time is continuous).

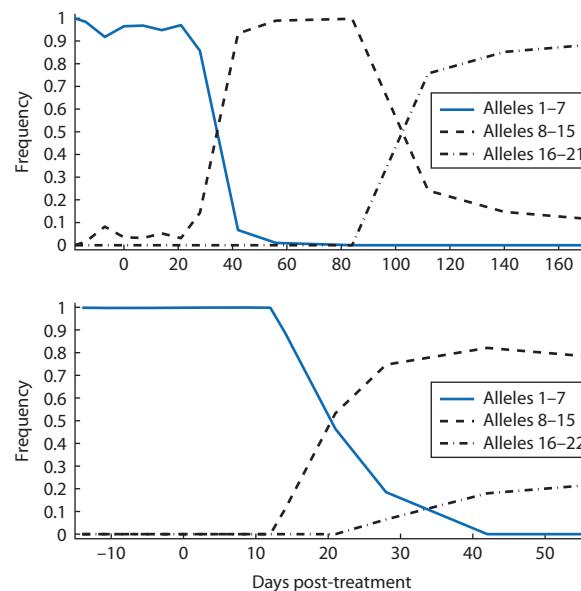
Although simplistic and requiring many assumptions, the model of natural selection among genotypes in organisms with clonal reproduction is nonetheless pertinent to a range of practical situations. One example is the evolution of drug resistance by natural selection in the human immunodeficiency virus (HIV). The genome of HIV (and other retroviruses) is single-stranded RNA. All the proteins inside a virus particle as well as the viral protein envelope itself are encoded by genes in this RNA genome. After infecting a host cell, HIV uses reverse transcriptase produced from its own gene to reverse-transcribe its genome into double-stranded DNA. This DNA version of the retrovirus genome is then integrated into the DNA of the host cell, where it is transcribed by the host cell into many new virus RNA genomes. These new viral RNA genomes are packaged into virus particles released from the host cell through a viral protease. One treatment strategy for HIV has utilized drugs that mimic nucleosides (nucleotides without phosphate groups) that interfere with the virus reverse transcriptase but do not interfere with host cell DNA polymerase. Another treatment uses protease inhibitors that interfere with polyprotein

cleavage necessary to produce new infectious virus particles. Unfortunately, HIV has shown rapid evolution of drug-resistant genotypes via natural selection. Figure 6.2 shows allele frequencies over time in the population of HIV particles infecting two patients who began protease inhibitor treatment at day 0. Individual HIV particles with protease alleles resistant to the drug have higher replication rates than HIV particles with wild-type protease alleles. This differential growth rate of HIV genotypes, or natural selection at the protease locus, rapidly changed the protease gene allele frequencies in the HIV population found within each patient. The combination of short generation time, high mutation rate, and large effective population size make natural selection a rapid process acting to change allele frequencies in HIV populations.

### Problem box 6.1 Relative fitness of HIV genotypes

It is commonly thought that drug-resistant alleles have lower relative fitness than non-resistant alleles in the absence of drug exposure. To test this hypothesis for HIV-1, Goudsmit et al. (1996) monitored the frequency of alleles at codon 215 of the reverse transcriptase gene in an individual newly infected with HIV but who was not undergoing treatment with the nucleoside analog azidothymidine (AZT). Initially, the HIV alleles were all sequences (90% TAC and 10% TCC codons) known to confer AZT resistance. Over time, the non-resistant allele (a TCC codon) increased in frequency to 49% after 20 months.

Use this change in allele frequencies over 601 days and equation 6.8 to estimate the relative fitness of the non-resistant allele in the absence of AZT. Assume that the generation time of HIV is 2.6 days and that generations are discrete, and that the wild-type allele was initially present at a frequency of 1.0% and so was not initially detectable. Note that the exponent in an equation like  $a = y(x^t)$  where  $a$ ,  $y$  and  $x$  are constants can be removed by taking the log of both sides to get  $\log(a) = \log(y) + t \log(x)$ .



**Figure 6.2** Allele frequencies at the protease locus over time in the HIV populations in two patients undergoing protease inhibitor (ritonavir) treatment (Doukhan & Delwart 2001). Alleles found at very low frequencies before drug treatment come to predominate in the HIV population after drug treatment, due to natural selection among HIV genotypes for drug resistance. Alleles are bands observed in denaturing-gradient gel electrophoresis (DGGE), a technique that is capable of discriminating single-base-pair differences among different DNA fragments. DGGE was used to identify the number of different protease locus DNA sequences present in a sample of HIV particles. Protease inhibitor treatment began on day 0. Dr. E. Delwart kindly provided the original data used to draw this figure.

### Natural selection with sexual reproduction

The model of natural selection with clonal reproduction leaves out a critical part of the biology of many organisms, namely sexual reproduction. To build a model of natural selection for sexual reproduction, we can combine the Hardy–Weinberg model of genotype frequencies with genotype-specific growth rates to get a general model of natural selection operating on the three genotypes produced by a single locus with two alleles. The blending of these two models leads to a number of assumptions that are listed in Table 6.2 (compare with the assumptions of Hardy–Weinberg alone given in Chapter 2). For now, let's utilize the assumptions that yield expected genotype frequencies. The consequences of many of the other assumptions are explored throughout the chapter.

Imagine a population of  $N$  diploid individuals formed by random mating among the parents and then random fusion of gametes to produce zygotes.

**Table 6.2** Assumptions of the basic natural selection model with a diallelic locus.

Genetic
<ul style="list-style-type: none"> <li>Diploid individuals</li> <li>One locus with two alleles</li> <li>Obligate sexual reproduction</li> </ul>
Reproduction
<ul style="list-style-type: none"> <li>Generations do not overlap</li> <li>Mating is random</li> </ul>
Natural selection
<ul style="list-style-type: none"> <li>Mechanism of natural selection is genotype-specific differences in survivorship (fitness) that lead to variable genotype-specific growth rates, termed viability selection</li> <li>Fitness values are constants that do not vary with time, over space, or in the two sexes</li> </ul>
Population
<ul style="list-style-type: none"> <li>Infinite population size so there is no genetic drift</li> <li>No population structure</li> <li>No gene flow</li> <li>No mutation</li> </ul>

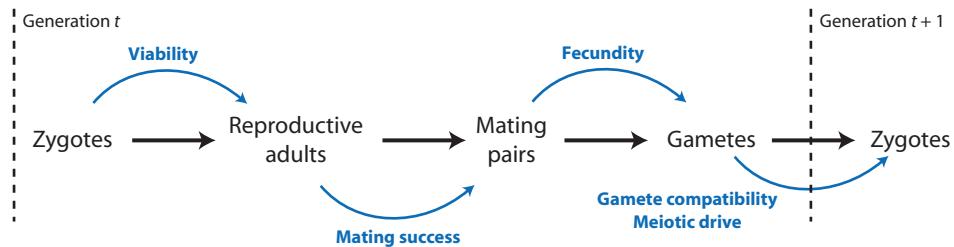
When the  $N$  zygotes have just formed, before any natural selection, the genotypes are in Hardy–Weinberg expected frequencies. If the total population size at this time is  $N_t$ , then the number of zygotes of each genotype is

$$AA : p^2N_t \quad Aa : 2pqN_t \quad aa : q^2N_t \quad (6.11)$$

which defines the initial number of each of the three genotypes analogously to  $N_A(0)$  and  $N_B(0)$  used in the case of clonal reproduction.

After the initial population of zygotes is formed, natural selection will then operate on the three genotypes. The mechanism of natural selection takes a particular form under the assumptions of the one

locus selection model. Each genotype is assumed to experience genotype-specific survival and reproduction during the course of a single generation as diagrammed in Fig. 6.3. This leads to a possible reduction in the number of zygotes of each genotype present in the population at the very beginning of the life cycle of a single generation. For the time being let's assume that any reduction in the numbers of individuals of any genotype comes exclusively from failure to survive to reproductive age but that all adults reproduce equally regardless of genotype. In this situation the fitness values of each genotype specify the probability of survival to reproduction, termed **viability**. Natural selection then takes the form of **viability selection**.



**Figure 6.3** A diagram of the life cycle of organisms showing some points where differential survival and reproduction among genotypes can result in natural selection. Viability is the probability of survival from zygote to adult, mating success encompasses those traits influencing the chances of mating and the number of mates, and fecundity is the number of gametes and progeny zygotes produced by each mating pair. Gametic compatibility is the probability that gametes can successfully fuse to form a zygote whereas meiotic drive is any mechanism that causes bias in the frequency of alleles found in gametes. Most models of natural selection assume a single fitness component such as viability. In reality, all of these components of fitness can influence genotype frequencies simultaneously.

**Viability selection** A form of natural selection where fitness is equivalent to the probability that individuals of given genotype survive to reproductive age but all surviving individuals have equal rates of reproduction.

**Marginal fitness** The frequency-weighted and allele-copy-weighted sum of the relative fitness values of genotypes that contain a specific allele; a special case of the average fitness for only those genotypes that contain a certain allele.

As an analog of  $\lambda$  used for clonal reproduction, let  $\ell$  (the cursive letter l) represent the genotype-specific probability of survival to reproductive age. The numbers of individuals of each genotype after viability selection at the point of reproduction is then

$$AA : \ell_{AA} p^2 N_t \quad Aa : \ell_{Aa} 2pqN_t \quad aa : \ell_{aa} q^2 N_t \quad (6.12)$$

These are the numbers of individuals of each genotype that will engage in random mating to form the next generation. The total number of individuals in the population after selection is then

$$\ell_{AA} p^2 N_t + \ell_{Aa} 2pqN_t + \ell_{aa} q^2 N_t \quad (6.13)$$

This is a quantity that can be used to determine the frequency of a genotype or allele in the population after selection. For example, the proportion of the total population made up of individuals with an AA genotype after selection is

$$\text{Frequency of AA genotype} = \frac{\ell_{AA} p^2 N_t}{\ell_{AA} p^2 N_t + \ell_{Aa} 2pqN_t + \ell_{aa} q^2 N_t} \quad (6.14)$$

Since there are fewer alleles than genotypes, the results of natural selection are often summarized in terms of allele frequencies rather than in terms of genotype frequencies. The allele frequencies in the gametes made by surviving individuals (under the assumptions in Table 6.2) are:

$$\text{Frequency of A allele in gametes} = \frac{\ell_{AA} p^2 N_t + \frac{1}{2}(\ell_{Aa} 2pqN_t)}{\ell_{AA} p^2 N_t + \ell_{Aa} 2pqN_t + \ell_{aa} q^2 N_t} \quad (6.15)$$

and

Frequency of a allele in gametes

$$= \frac{\ell_{aa} q^2 N_t + \frac{1}{2}(\ell_{Aa} 2pqN_t)}{\ell_{AA} p^2 N_t + \ell_{Aa} 2pqN_t + \ell_{aa} q^2 N_t} \quad (6.16)$$

In each of these equations the number of heterozygote individuals after selection is multiplied by one-half since a heterozygote contributes one copy of a given allele to the gamete pool for every two copies of that same allele contributed by a homozygote. These expressions simplify to:

Frequency of A allele in gametes

$$= \frac{\ell_{AA} p^2 + \ell_{Aa} pq}{\ell_{AA} p^2 + \ell_{Aa} 2pq + \ell_{aa} q^2} \quad (6.17)$$

and:

Frequency of a allele in gametes

$$= \frac{\ell_{aa} q^2 + \ell_{Aa} pq}{\ell_{AA} p^2 + \ell_{Aa} 2pq + \ell_{aa} q^2} \quad (6.18)$$

because  $N_t$  can be factored out of each term in the numerator and denominator and then cancelled, and the constants of  $1/2$  and  $2$  cancel in the numerator.

As in the case of clonal reproduction, we can utilize relative fitness values for each genotype rather than absolute values of survivorship to reproductive age. We can then replace the  $\ell_{AA}$ ,  $\ell_{Aa}$ , and  $\ell_{aa}$  values with the relative fitness values  $w_{AA}$ ,  $w_{Aa}$ , and  $w_{aa}$  to give

$$p_{t+1} = \frac{w_{AA} p^2 + w_{Aa} pq}{w_{AA} p^2 + w_{Aa} 2pq + w_{aa} q^2} \quad (6.19)$$

and

$$q_{t+1} = \frac{w_{aa} q^2 + w_{Aa} pq}{w_{AA} p^2 + w_{Aa} 2pq + w_{aa} q^2} \quad (6.20)$$

Notice that the denominator in both expressions is the sum of the fitness-weighted genotype frequencies, or the mean relative fitness  $\bar{w}$ . Making this substitution gives even more compact expressions for the allele frequencies:

$$p_{t+1} = \frac{w_{AA} p^2 + w_{Aa} pq}{\bar{w}} \quad (6.21)$$

**Table 6.3** The expected frequencies of three genotypes after natural selection for a diallelic locus with sexual reproduction and random mating. The absolute fitness of the AA genotype is used as the standard of comparison when determining relative fitness.

	Genotype	
	AA	Aa
	aa	
Generation $t$		
Initial frequency	$p_t^2$	$2p_t q_t$
Genotype-specific survival (absolute fitness)	$\ell_{AA}$	$\ell_{Aa}$
Relative fitness	$w_{AA} = \frac{\ell_{AA}}{\ell_{AA}}$	$w_{Aa} = \frac{\ell_{Aa}}{\ell_{AA}}$
Frequency after natural selection		
Average fitness	$p_t^2 w_{AA} + 2p_t q_t w_{Aa} + p_t^2 w_{aa}$	$2p_t q_t w_{Aa}$
Generation $t+1$		
Genotype frequency	$\frac{p_t^2 w_{AA}}{\bar{w}}$	$\frac{q_t^2 w_{aa}}{\bar{w}}$
Allele frequency	$p_{t+1} = \frac{p_t(p_t w_{AA} + q_t w_{Aa})}{\bar{w}}$	$q_{t+1} = \frac{q_t(q_t w_{aa} + p_t w_{Aa})}{\bar{w}}$
Change in allele frequency	$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}}$	$\Delta q = \frac{pq[q(w_{aa} - w_{Aa}) + p(w_{Aa} - w_{AA})]}{\bar{w}}$

and

$$q_{t+1} = \frac{w_{aa}q^2 + w_{Aa}pq}{\bar{w}} \quad (6.22)$$

These expressions show that the increase or decrease in allele frequencies depends on a comparison of the average fitness of genotypes that contain a certain allele (the quantity in the numerator), called the **marginal fitness**, and the average fitness of all genotypes in the population. A larger marginal fitness occurs when individuals with genotypes that make up the marginal fitness have higher viability for a given allele frequency. Table 6.3 summarizes the key quantities used to construct expected genotype and allele frequencies after one generation of viability selection.

As in the case of clonal reproduction, the change in allele frequency over one generation is given by  $\Delta p = p_{t+1} - p_t$ . For sexual reproduction,

$$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}} \quad (6.23)$$

and

$$\Delta q = \frac{pq[q(w_{aa} - w_{Aa}) - p(w_{AA} - w_{Aa})]}{\bar{w}} \quad (6.24)$$

as derived in Math box 6.1. This equation provides three generalizations that match with our intuitions about natural selection. Allele frequencies do not

change when  $pq = 0$  or when there is no genetic variation since one allele or the other has reached loss ( $p$  or  $q = 0$ ). Allele frequencies do not change when all fitness values are identical – meaning there is no natural selection – so the terms inside the square brackets give a value of zero. Lastly, allele frequencies do not change when the fitness differences weighted by an allele frequency (the  $p(w_{AA} - w_{Aa})$  and  $q(w_{Aa} - w_{aa})$  terms) cancel each other out to yield a zero inside the square brackets.

## 6.2 General results for natural selection on a diallelic locus

- Selection against a recessive phenotype.
- Selection against a dominant phenotype.
- The general effects of dominance.
- Heterozygote disadvantage and advantage.
- The strength of natural selection.

The previous section presented the basic building blocks of a model for natural selection acting through genotype-specific viability on one locus with two alleles. This section will present the general results of natural selection under this very basic model. This task is simpler than it might seem since all the outcomes of the selection model can be represented by five general categories of fitness values for the three genotypes (Table 6.4). Notice that Table 6.4 presents fitness values in terms of **selection coefficients** rather than relative fitness. Selection coefficients are simply the difference between a relative fitness value and one:

**Table 6.4** The general categories of relative fitness values for viability selection at a diallelic locus. The variables  $s$  and  $t$  are used to represent the decrease in viability of a genotype compared to the maximum fitness of 1 ( $1 - w_{xx} = s$ ). The degree of dominance of the A allele is represented by  $h$  with additive gene action (sometime called codominance) when  $h = 1/2$ .

Category	Genotype-specific fitness		
	$w_{AA}$	$w_{Aa}$	$w_{aa}$
Selection against a recessive phenotype	1	1	$1 - s$
Selection against a dominant phenotype	$1 - s$	$1 - s$	1
General dominance (dominance coefficient $0 \leq h \leq 1$ )	1	$1 - hs$	$1 - s$
Heterozygote disadvantage (underdominance for fitness)	1	$1 - s$	1
Heterozygote advantage (overdominance for fitness)	$1 - s$	1	$1 - t$

### Math box 6.1

#### The change in allele frequency each generation under natural selection

To solve the equation for the change in allele frequency over one generation due to natural selection, start with

$$\Delta p = \frac{p^2 w_{AA} + pq w_{Aa}}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}} - p \quad (6.25)$$

where the allele frequencies  $p$  and  $q$  are all taken in the same generation so the generation subscripts are dropped. First, put both of the terms over a common denominator so they can be subtracted:

$$\Delta p = \frac{p^2 w_{AA} + pq w_{Aa}}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}} - \frac{p(p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa})}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}} \quad (6.26)$$

Then notice that a factor of  $p$  can be taken from the numerator of the left-hand term of the difference:

$$\Delta p = \frac{p(pw_{AA} + qw_{Aa})}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}} - \frac{p(p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa})}{p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}} \quad (6.27)$$

which leads to the following (the denominator is hereafter given as  $\bar{w}$  for simplicity):

$$\Delta p = \frac{p[pw_{AA} + qw_{Aa} - p^2 w_{AA} - 2pq w_{Aa} - q^2 w_{aa}]}{\bar{w}} \quad (6.28)$$

A trick comes in at this point utilizing the fact that  $p = 1 - q$  with a diallelic locus, so that  $pq = p(1 - p) = p - p^2$ . The first and third terms inside the square brackets of the numerator of equation 6.28 ( $pw_{AA} - p^2 w_{AA}$ ) can be expressed alternatively as  $pqw_{AA}$ . This then gives

$$\Delta p = \frac{p[pqw_{AA} + qw_{Aa} - 2pq w_{Aa} - q^2 w_{aa}]}{\bar{w}} \quad (6.29)$$

A  $q$  can then be factored out of the terms inside the square brackets of the numerator to give

$$\Delta p = \frac{pq[pw_{AA} + w_{Aa} - 2pw_{Aa} - qw_{aa}]}{\bar{w}} \quad (6.30)$$

Then, notice the middle two terms inside the square brackets ( $w_{Aa} - 2pw_{Aa}$ ). Since  $p + q = 1$ ,  $w_{Aa} - 2pw_{Aa} = (p + q)w_{Aa} - 2pw_{Aa} = qw_{Aa} - pw_{Aa}$ . Making this substitution leads to

$$\Delta p = \frac{pq[pw_{AA} + qw_{Aa} - pw_{Aa} - qw_{aa}]}{\bar{w}} \quad (6.31)$$

which can finally be rearranged to

$$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}} \quad (6.32)$$

The same approach can be taken to obtain the expression for  $\Delta q$ .

$$s_{xx} = 1 - w_{xx} \text{ or } w_{xx} = 1 - s_{xx} \quad (6.33)$$

where the subscript  $xx$  represents a genotype and the maximum relative fitness is one. Selection coefficients therefore represent the difference in viability between a given genotype and the genotype with the highest viability.

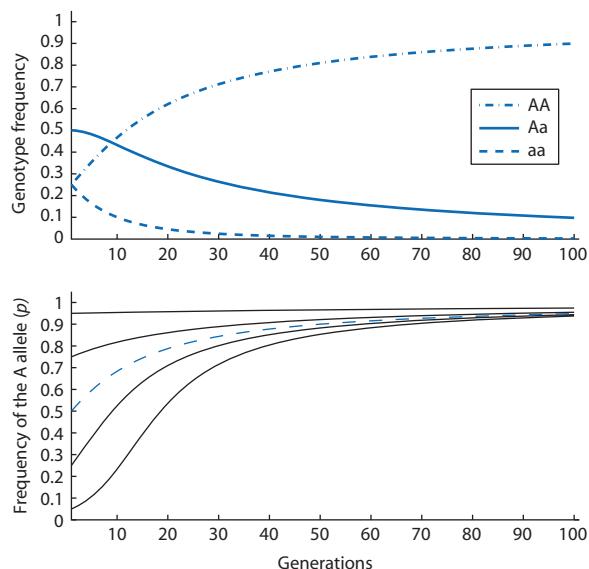
Examining the outcome of selection for each category of fitness values or selection coefficients will illustrate how viability selection is expected to change genotype and allele frequencies in populations. By iterating versions of equation 6.14 for all three genotypes as well as equations 6.21 and 6.22, we can visualize the action of natural selection. The behavior of allele frequencies under natural selection can be understood by examining plots of allele frequencies over time to see the direction and rate of allele frequency change. An important general feature of natural selection is the allele frequency reached when allele frequencies eventually stop changing, or the equilibrium allele frequency. The goal of this section is to understand both how and why genotype and allele frequencies change when acted on by a constant force of natural selection over time.

Although it is common to speak of an *allele* favored by natural selection, any change in allele frequencies is really caused by natural selection on *genotypes* due to their different-viability phenotypes. Alleles themselves do not have phenotypes nor fitness values in most types of natural selection (natural selection on gametes or haploid genomes are exceptions). The changing frequency of genotypes is what causes allele frequencies to change. Although two allele frequencies can be displayed with more economy than three genotype frequencies, it is critical not to forget that natural selection directly causes changes in genotype frequency and that change in allele frequencies is an indirect consequence.

The process of natural selection has the special quality that the genotype frequencies reached at equilibrium are always the same as long as the starting frequencies and relative fitness values are constant. Processes that always lead to the same outcome from a given set of initial conditions are called **deterministic** because the end state is completely determined by the initial state. Similar patterns of genotype frequencies in independent populations are therefore evidence that the process of natural selection is operating. In contrast, the stochastic process of genetic drift would result in random outcomes in each independent population. This also means that there is no need to view replicate outcomes of natural selection for the same set of initial conditions.

### Selection against a recessive phenotype

The results of natural selection acting against a completely recessive homozygous genotype (see Table 6.4) are shown in Fig. 6.4. The top panel shows the frequencies of the three genotypes over time starting from an initial allele frequency of  $p = q = 0.5$ . The frequency of the recessive homozygote (aa) declines because that genotype has lower viability. At the same time, the frequency of the dominant homozygote (AA) increases since it has a higher viability. Even though the heterozygote also has the maximum fitness, its frequency declines from a maximum of 0.5 as A alleles become more frequent and a alleles less frequent over time, reducing the value of  $2pq$ . The bottom panel summarizes the results of natural selection in terms of allele frequencies over time for five initial allele frequencies. (The one allele frequency trajectory that corresponds to the genotype frequencies in the top panel is given as a colored, dashed line.)



**Figure 6.4** The change in genotype and allele frequencies caused by viability selection against the aa genotype exhibiting the recessive phenotype. The top panel shows the change in genotype frequencies over time and the bottom panel shows the frequency of the dominant allele (A) over time. The colored, dashed line in the bottom panel corresponds to the allele frequencies in the top panel. Because of changes in genotype frequency caused by natural selection, the frequency of the dominant allele rapidly approaches fixation from all five initial allele frequencies. In this illustration  $w_{AA} = w_{Aa} = 1.0$  while  $w_{aa} = 0.8$ , meaning that eight individuals with the aa genotype are expected to survive to reproduce for every 10 individuals with the AA or Aa genotype that survive to reproduce each generation. Genotype frequencies assume random mating.

Allele frequencies change more rapidly in the early generations when the initial allele frequency is lower because the selectively favored dominant homozygote and heterozygotes are relatively frequent in the population. Even at an initial dominant allele frequency of 0.05, 9.75% (or  $1 - q^2$ ) of the genotypes are AA and Aa. As the frequency of the recessive allele decreases (the dominant allele approaches higher frequencies), the change in allele frequency from one generation to the next steadily declines. For example, starting from an initial allele frequency of 0.05 the allele frequency changes by 0.1 in just a few generations early on but requires many generations when the frequency of the recessive allele is low. This occurs because there are progressively fewer recessive homozygotes and progressively more of the highest fitness genotypes (the dominant homozygote and heterozygotes) in the population as selection changes the genotype and allele frequencies.

Does the dominant allele go to fixation when there is natural selection against the recessive homozygote? The answer is no, because the heterozygote fitness is equal to the maximum fitness and every generation heterozygotes will produce gametes that can combine to make the recessive homozygote. In essence, the recessive allele is shielded from natural selection in the heterozygote due to dominance. This is true no matter how large the selection coefficient against a recessive homozygote.

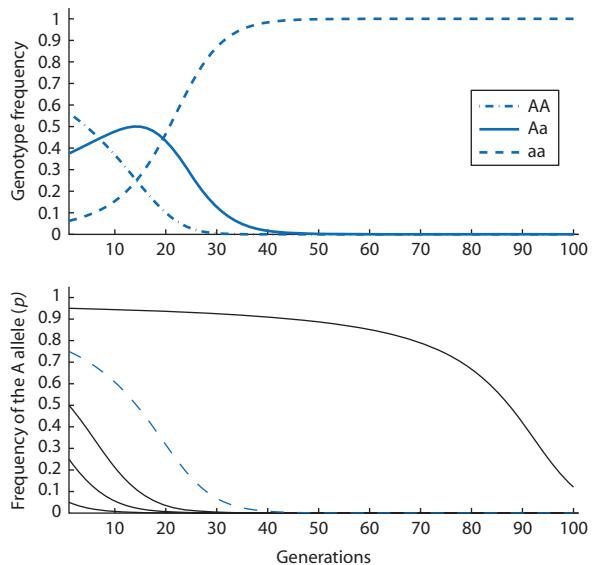
One way to quantify the sheltering effect of heterozygotes is to examine the proportion of recessive alleles present in heterozygotes compared to recessive alleles present in homozygotes

$$\frac{pq}{q^2} = \frac{p}{q} \quad (6.34)$$

where the expected frequency of heterozygotes is weighted by one-half since each contains only one recessive allele. When the frequency of the recessive allele is low,  $q = 0.05$  for example, the proportion of the genotype frequencies is  $0.0475/0.0025 = 19$ . This means that there are 19 recessive alleles protected against natural selection in heterozygotes for each recessive allele impacted by natural selection in a homozygous genotype.

### **Selection against a dominant phenotype**

The results of natural selection acting against a completely dominant phenotype shared by the dominant homozygotes and the heterozygotes (see Table 6.4) are shown in Fig. 6.5. The top panel shows the fre-



**Figure 6.5** The change in the genotype and allele frequency of a completely dominant allele (A) when natural selection acts against the AA and Aa genotypes exhibiting the dominant phenotype. Notice that the frequency of the A allele decreases slowly at first when the A allele is common in the population since the aa genotype is infrequent. The colored, dashed line in the bottom panel corresponds to the allele frequencies in the top panel. In this illustration  $w_{AA} = w_{Aa} = 0.8$  while  $w_{aa} = 1.0$ . Genotype frequencies assume random mating.

quencies of the three genotypes over time starting from an initial allele frequency of  $p = 0.75$ . The frequency of the dominant homozygote (AA) declines due to its lower viability while the frequency of the recessive homozygote (aa) increases due to its higher viability. Even though the heterozygote has a lower relative fitness than the recessive homozygote, its frequency initially increases since the frequency of the two alleles approaches equality. The frequency of the heterozygote temporarily peaks at the maximum value of  $2pq = 0.5$  at the same point that the frequency of the two homozygotes both equal 0.25. The heterozygote frequency then drops again as the frequency of the recessive homozygote continues to increase and the frequency of the dominant homozygote continues to decrease.

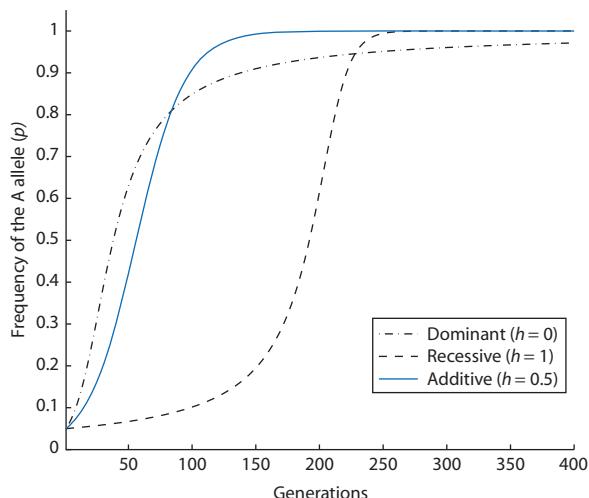
The bottom panel of Fig. 6.5 shows that the frequency of the dominant allele decreases toward zero under this type of natural selection. (Again, the one allele frequency trajectory that corresponds to the genotype frequencies in the top panel is given as a colored, dashed line.) At an initial dominant allele frequency of  $p = 0.95$ , only 0.25% (or  $q^2$ ) of the genotypes are aa. This makes natural selection slow to change allele frequencies until the frequency of the recessive allele increases enough to make the higher

fitness aa genotype more common in the population. The allele frequency trajectories that start at lower initial frequencies for the A allele change more rapidly and bear out this point. Does the recessive allele go to fixation when there is natural selection against the dominant homozygote and heterozygote? In this case yes, since both the dominant homozygote and the heterozygote have a lower fitness than the favored homozygote and therefore the dominant allele is not shielded from natural selection in the heterozygote.

### General dominance

The previous two examples of natural selection against dominant and recessive phenotypes cover the extremes of dominance. The impact of dominant and recessive alleles on the outcome of natural selection on a diallelic locus can be made more general by employing a dominance coefficient, symbolized  $h$ . Complete dominance (the heterozygote and a homozygote having identical phenotypes) for one allele is represented by  $h = 0$  and complete dominance for the other allele is represented by  $h = 1$ . When the heterozygote has a phenotype that is the average of the two homozygotes then  $h = 1/2$ , a situation sometimes called codominance. A dominance coefficient of  $h = 1/2$  is more descriptively referred to as **additive gene action** since the phenotype of the heterozygote is the sum of the phenotypic effects of each allele. For example, if phenotypes are AA = 3 spots, Aa = 2 spots and aa = 1 spot, an A allele contributes 1.5 spots and an a allele contributes 0.5 spots in the heterozygote when gene action is additive. Look at Table 6.4 and verify the fitness of the heterozygote when  $h = 0$ , 1, and  $1/2$ . This method to specify fitness has the advantage that the results of natural selection can be predicted for any degree of dominance. There is also a strong biological motivation, since alleles commonly show a wide range of dominance or gene action in actual populations, ranging between being completely dominant or completely recessive.

The outcome of selection for three cases of gene action are shown in Fig. 6.6. All three cases start at the same initial allele frequency and share the same selection coefficient. However, gene action varies from completely dominant to completely recessive with the additive case in between. The results of natural selection on a completely dominant allele (rapid change in allele frequency initially but never reaching fixation) and on a completely recessive allele (slow initial change in allele frequency then more rapid change and eventual fixation) are identical to the dynamics seen in earlier examples. The allele



**Figure 6.6** Allele frequencies over time for three types of gene action with a low initial allele frequency. In all three cases the equilibrium allele frequency is fixation or near fixation for the A allele. With complete dominance, natural selection initially increases the allele frequency very rapidly. The approach to fixation for the A allele slows as aa homozygotes become rare since heterozygotes harbor a alleles that are concealed from natural selection by dominance. Natural selection initially changes the frequency of a recessive allele very slowly since homozygote recessive genotypes are very rare. As the recessive homozygotes become more common, allele frequency increases more rapidly. With additive gene action the phenotype of the heterozygote is intermediate between the two homozygotes so all genotypes differ in their viability. Additive gene action has the most rapid overall approach to equilibrium allele frequency. The degree of dominance is represented by the dominance coefficient,  $h$ . In this illustration the selection coefficient is  $s = 0.1$ .

frequency trajectory for additive gene action is intermediate. It combines the rapid initial change in allele frequency of complete dominance with the later-stage rapid approach to equilibrium and fixation of the complete recessive. Equilibrium allele frequency (fixation or near fixation) is reached most quickly with additive gene action.

With completely dominant or recessive alleles, natural selection cannot discriminate between two of the three genotypes since their fitness values are identical. How this lack of difference in fitness values affects natural selection depends on genotype frequencies. In the early generations, the recessive case shows slow change because the heterozygote is selected against and the fittest genotype is rare. In the later generations of the dominance case, heterozygotes shelter the recessive allele from natural selection slowing further change in allele frequency as the recessive homozygote becomes very infrequent. In contrast, the fitness values of all three genotypes are distinct and uniformly different with additive gene

action. Additive gene action gives the maximum difference in marginal and average fitness values across the entire range of possible genotype frequencies under random mating.

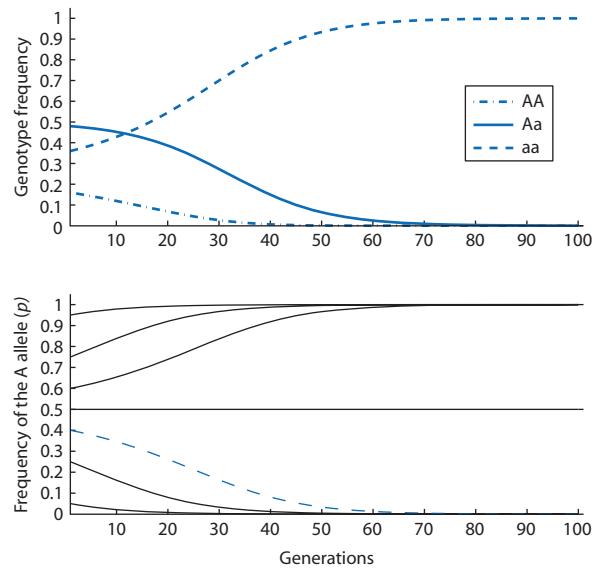
Gene action is an important factor in understanding the fate of new mutations acted on by natural selection. Imagine a new mutation in a population that has a high relative fitness when homozygous. As covered in Chapter 5, the initial frequency of any new mutation will be low ( $\frac{1}{2N_e}$ ). A completely or nearly recessive mutation will take a very long time to increase in frequency under natural selection. In contrast, a completely or nearly dominant mutation with the same fitness as a homozygote and starting at the same frequency will increase in frequency very rapidly. The examples in Fig. 6.6 where the initial frequency of the A allele is 0.05 are equivalent to a new mutation in a population of  $N_e = 10$ .

### Heterozygote disadvantage

The results of natural selection acting against the heterozygote phenotype, a situation known as heterozygote disadvantage, underdominance for fitness, or disruptive selection (see Table 6.4), are shown in Fig. 6.7. Starting from an initial allele frequency of  $p = 0.4$ , the top panel shows how the aa homozygote eventually reaches fixation over time. The bottom panel requires close attention in this case, since the equilibrium allele frequency depends strongly on the initial allele frequency in the population. Initial allele frequencies above  $p = 0.5$  all lead to fixation of the AA homozygote while all initial allele frequencies below  $p = 0.5$  lead to fixation of the aa homozygote. When the initial allele frequency in the population is exactly  $p = 0.5$ , allele frequencies remain constant over time. It turns out that this equilibrium point is not robust to any change in allele frequency, and so is called an **unstable** equilibrium. Any slight change in allele frequency will result in the allele frequencies changing to alternative stable equilibrium points of fixation or loss. Such an unstable equilibrium is very unlikely to persist in a finite population, since even a slight amount of genetic drift would alter allele frequencies in the population toward one of the stable equilibrium points.

### Heterozygote advantage

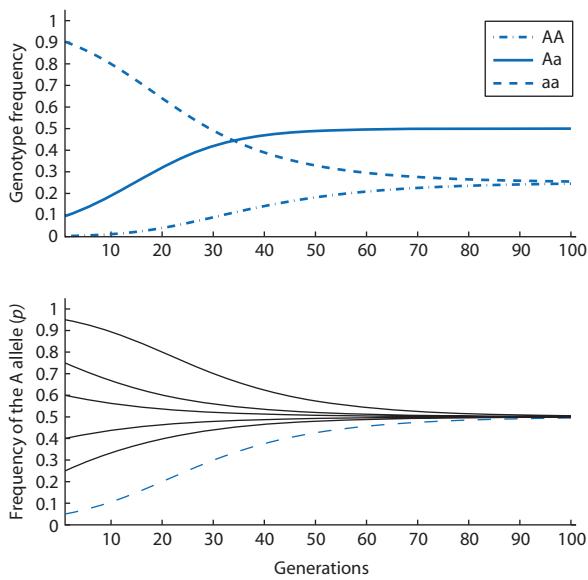
The results of natural selection acting to increase the frequency of the heterozygous genotype, commonly



**Figure 6.7** The change in the genotype and allele frequency when there is underdominance for fitness and natural selection acts against individuals with Aa genotypes. The equilibrium allele frequency depends on the initial allele frequency. Starting below 0.5 populations head toward loss while starting above 0.5 populations go to fixation. There is an unstable equilibrium at an initial allele frequency of exactly 0.5. From any initial allele frequency the population converges on a minimum frequency of heterozygotes. The colored, dashed line in the bottom panel corresponds to the allele frequencies in the top panel. In this illustration  $w_{AA} = w_{aa} = 1.0$  and  $w_{Aa} = 0.9$ . Genotype frequencies assume random mating.

referred to as heterozygote advantage, overdominance for fitness, or balancing selection (see Table 6.4), are shown in Fig. 6.8. The top panel shows the frequencies of the three genotypes over time starting from an initial allele frequency of  $p = 0.05$ . The heterozygous genotype increases in frequency due to its higher relative fitness. At the same time, the aa homozygote (initially 90% of the population) declines due to its lower viability. Although the relative fitness of the AA homozygote is lower than that of the heterozygote, its frequency increases toward 25% as allele frequencies approach  $p = q = 0.5$  due to the increasing frequency of heterozygotes. The bottom panel shows that for all initial allele frequencies, natural selection causes the population to approach  $p = q = 0.5$ .

Overdominance for fitness represents a unique exception for the outcome of natural selection on a diallelic locus. Selection against a dominant phenotype results in fixation of the recessive allele and loss of the dominant allele. Similarly, selection against a recessive phenotype results in near fixation of the



**Figure 6.8** The change in the genotype and allele frequencies when there is overdominance for fitness and natural selection favors individuals with Aa genotypes. From any initial allele frequency the population converges on a maximum frequency of heterozygotes. This corresponds to equal allele frequencies with random mating. The colored, dashed line in the bottom panel corresponds to the allele frequencies in the top panel. In this illustration  $w_{AA} = w_{aa} = 0.9$  and  $w_{Aa} = 1.0$ . Genotype frequencies assume random mating.

dominant allele and near loss of the recessive allele. Selection against a heterozygote also results ultimately in fixation of one allele and loss of the other allele. These three forms of natural selection all produce an equilibrium with little or no genetic variation, known as a **monomorphic equilibrium**. In contrast, when heterozygotes have the highest fitness natural selection maintains both alleles in the population at equilibrium, resulting in a **polymorphic equilibrium**. Thus, overdominance for fitness is one type of natural selection that is consistent with the permanent maintenance of genetic variation in populations.

The allele frequencies expected at equilibrium with overdominance can be obtained from equation 6.23, as shown in Math box 6.2. The equilibrium allele frequencies are

$$p_{equilibrium} = \frac{t}{s+t} \quad (6.35)$$

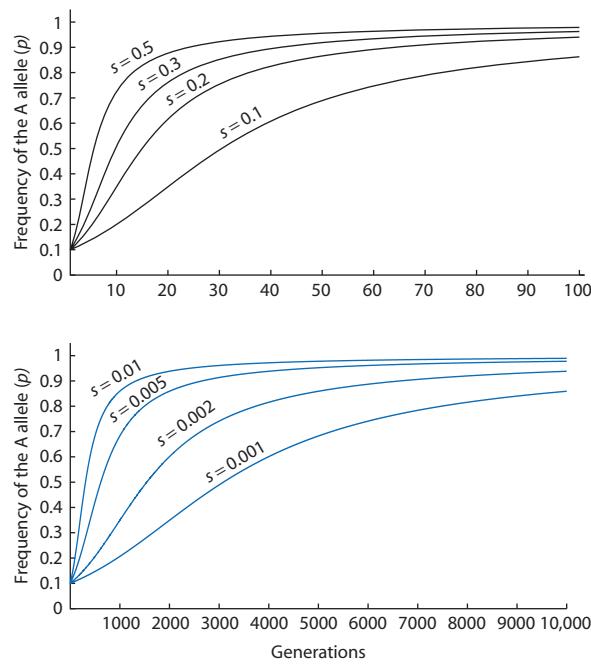
and

$$q_{equilibrium} = \frac{s}{s+t} \quad (6.36)$$

where  $s$  and  $t$  are the selection coefficients against the AA and aa homozygotes, respectively (see Table 6.4). The equilibrium allele frequency is higher for the allele in the homozygous genotype that has the smaller selection coefficient (or higher relative fitness).

### The strength of natural selection

The strength of selection against a genotype can vary from weak, such as a viability 0.1% less than the most fit genotype, to very strong, such as 50% viability or even zero viability (lethality) of a genotype. Allele frequencies over time (starting from the same initial allele frequency) are plotted in Fig. 6.9 for a wide range of selection coefficients in the case of natural selection against a homozygous recessive genotype. Notice that the shapes of the curves in the top and bottom panels of Fig. 6.9 are very similar but that the time scale of each plot is very different. Selection coefficients of 10% or greater bring the dominant allele to high frequencies within 100 generations. In contrast, reaching these same allele frequencies takes 10,000 generations when the



**Figure 6.9** The strength of natural selection influences the rate of change in genotype and allele frequencies. In this illustration, selection acts against the recessive homozygote (aa). The top panel shows strong natural selection where viability of the aa genotype is 10–50% less than that of the other genotypes. The bottom panel shows weak natural selection where viability of the aa genotype is 1–0.1% less than that of the other genotypes. Note the vastly different time scales in the two plots.

### Math box 6.2 Equilibrium allele frequency with overdominance

By definition, equilibrium allele frequencies are reached when allele frequencies stop changing from one generation to the next. This means that  $\Delta p$  as expressed by

$$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}} \quad (6.37)$$

which was first shown as equation 6.23, should be equal to zero.

Two equilibrium points occur when  $p = 0$  or  $q = 0$ , biologically equivalent to situations where there is no genetic variation in a population. When there is genetic variation (both  $p \neq 0$  and  $q \neq 0$ ), the equilibrium point depends on the fitness differences contained in the numerator. Taking the numerator term in square brackets in equation 6.37 and setting it equal to zero,

$$p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa}) = 0 \quad (6.38)$$

and then solving  $p$  or  $q$  in terms of relative fitness values, will give allele frequencies where  $\Delta q$  is zero. The first step is to substitute  $q = 1 - p$ :

$$p(w_{AA} - w_{Aa}) + (1 - p)(w_{Aa} - w_{aa}) = 0 \quad (6.39)$$

and then expand by multiplying the terms:

$$pw_{AA} - pw_{Aa} + w_{Aa} - w_{aa} - pw_{Aa} + pw_{aa} = 0 \quad (6.40)$$

The relative fitness values that are multiplied by  $p$  can be brought together:

$$p(w_{AA} - 2w_{Aa} + w_{aa}) + w_{Aa} - w_{aa} = 0 \quad (6.41)$$

and then subtracted:

$$w_{Aa} - w_{aa} = -p(w_{AA} - 2w_{Aa} + w_{aa}) \quad (6.42)$$

Dividing both sides by  $-(w_{AA} - 2w_{Aa} + w_{aa})$  gives

$$p = \frac{w_{Aa} - w_{aa}}{2w_{Aa} - w_{AA} - w_{aa}} \quad (6.43)$$

which expresses  $p$  as a function of relative fitness values alone. Substituting the relative fitness values of  $w_{AA} = 1 - s$ ,  $w_{Aa} = 1$ , and  $w_{aa} = 1 - t$  as given in Table 6.4:

$$p = \frac{1 - (1 - t)}{2(1) - (1 - s) - (1 - t)} \quad (6.44)$$

and then carrying out the addition and subtraction then gives the equilibrium allele frequency in terms of selection coefficients for the two homozygous genotypes:

$$p = \frac{t}{s + t} \quad (6.45)$$

selection coefficient is between 1.0 and 0.1%. This illustrates the general principle that stronger natural selection (larger selection coefficients or larger fitness differences) results in a more rapid approach to equilibrium allele frequencies. This conclusion applies to all of the situations given in Table 6.4 and to the process of natural selection in general.

### 6.3 How natural selection works to increase average fitness

- Natural selection acts to increase mean fitness.
- The fundamental theorem of natural selection.

In the five fitness situations shown in Table 6.4 for natural selection on a diallelic locus, there are

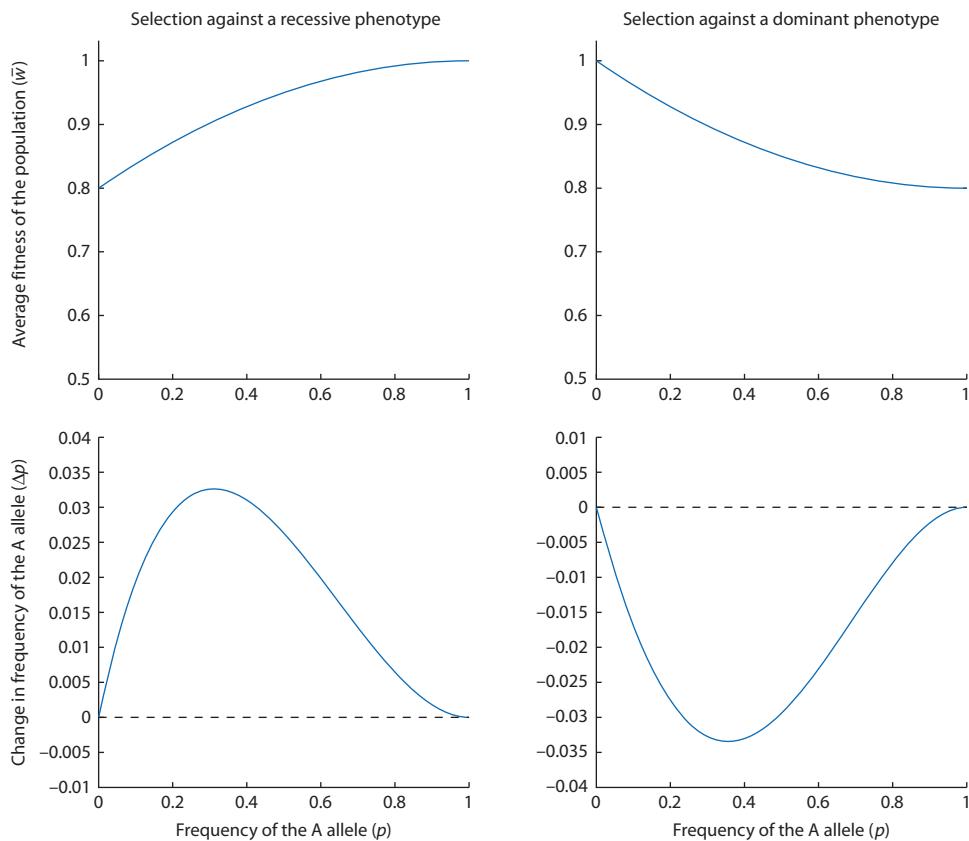
always two general outcomes. Directional selection of any type ends in fixation and loss (selection against a dominant phenotype) or nearly fixation and loss (selection against a recessive phenotype). Underdominance too results in fixation or loss (with one exception unlikely to be realized in finite populations). Overdominance is the exception that maintains both alleles in the population indefinitely. So the two outcomes are either fixation and loss or intermediate frequencies for both alleles (sometimes called a **balanced polymorphism**). The reason why these two general outcomes occur can be understood by examining the average fitness of a population ( $\bar{w}$ ) as well as the rate of change in allele frequency ( $\Delta p$ ) over the entire range of allele frequencies.

### Average fitness and rate of change in allele frequency

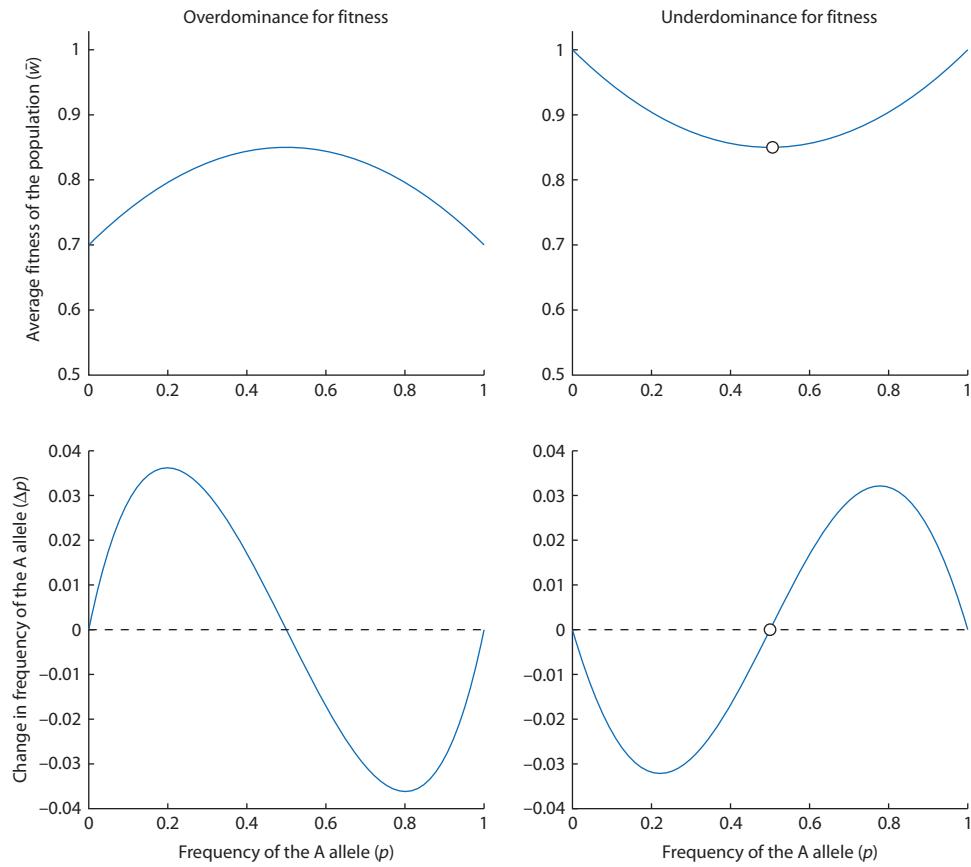
The mean fitness ( $\bar{w}$ ) over all possible allele frequencies is plotted for each case of natural selection on a diallelic locus in Figs 6.10 and 6.11. For the cases of directional selection in Fig. 6.10, notice that the highest mean fitness corresponds exactly to fixation of the A allele for selection against a recessive phenotype and to loss of the A allele for selection against a dominant phenotype. This same pattern is evident in Fig. 6.11 where the highest mean fitness is found at an intermediate allele frequency for overdominance or at fixation or loss for underdominance. These plots of mean fitness by allele frequency show that natural selection acts to increase the mean fitness of the population to its maximum. It is the maximum mean fitness in a population that really defines the equilibrium points for genotype and allele frequencies. The plots of  $\bar{w}$  against  $p$  reveal the

generalization that the process of natural selection acts to increase the population mean fitness every generation if possible and stops when the mean fitness can no longer increase. In this sense, natural selection can be metaphorically likened to a mountain climber who continually works to find the highest point, but who cannot ever go downhill and will only rest at the summit. Keeping with this metaphor, plots of  $\bar{w}$  against  $p$  are called **fitness surfaces**, **adaptive landscapes**, or **adaptive topographies** and represent a topographic map of the mountain at any point where our imaginary mountain climber might venture.

Figures 6.10 and 6.11 also show the change in allele frequency over a single generation ( $\Delta p$ ) over all possible allele frequencies for each case of natural selection. Plots of  $\Delta p$  against  $p$  reveal when allele frequencies are increasing ( $\Delta p$  is positive) or decreasing ( $\Delta p$  is negative) as well as when allele frequencies are changing rapidly (the absolute value of  $\Delta p$  is



**Figure 6.10** Mean fitness in a population ( $\bar{w}$ ) and change in allele frequency over a single generation ( $\Delta p$ ) as a function of allele frequency for directional selection. Directional selection reaches allele frequency equilibrium at either fixation or loss, the point of highest mean fitness. Positive values of  $\Delta p$  (above the dashed line) indicate that allele frequency is increasing under selection while negative values of  $\Delta p$  (below the dashed line) indicate that allele frequency is decreasing under selection. The change in allele frequencies is faster when average fitness changes more rapidly (the slope of  $\bar{w}$  is steeper). Here  $w_{AA} = w_{Aa} = 1.0$  and  $w_{aa} = 0.8$  for selection against a recessive phenotype and  $w_{AA} = w_{Aa} = 0.8$  and  $w_{aa} = 1.0$  for selection against a dominant phenotype.



**Figure 6.11** Mean fitness in a population ( $\bar{w}$ ) and change in allele frequency over a single generation ( $\Delta p$ ) as a function of allele frequency for balancing and disruptive selection. Natural selection changes allele frequencies to increase the average fitness in each generation, eventually reaching an equilibrium when the mean fitness is highest. The change in allele frequencies is faster when average fitness changes more rapidly (the slope of  $\bar{w}$  is steeper). The dashed line in the plots of  $\Delta p$  by  $p$  shows where allele frequencies stop changing ( $\Delta p = 0$ ) and thus are allele frequency equilibrium points. With underdominance for fitness,  $\Delta p$  is zero when  $p = 0.5$  and so defines an equilibrium point marked by the circle. However, this equilibrium point is unstable since  $\Delta p$  on either side of  $p = 0.5$  changes allele frequencies *away* from the equilibrium point (below  $p = 0.5$   $\Delta p$  is negative leading toward loss and above  $p = 0.5$   $\Delta p$  is positive leading toward fixation). In contrast, with overdominance  $\Delta p$  on either side of  $p = 0.5$  changes allele frequencies *toward* the equilibrium point (below  $p = 0.5$   $\Delta p$  is positive and above  $p = 0.5$   $\Delta p$  is negative) and thus  $p = 0.5$  is a stable equilibrium point. Here  $w_{AA} = w_{aa} = 1.0$  and  $w_{Aa} = 0.7$  for underdominance and  $w_{AA} = w_{aa} = 0.7$  and  $w_{Aa} = 1.0$  for overdominance.

large) or slowly (the absolute value of  $\Delta p$  is small). When allele frequencies are not changing at all ( $\Delta p$  is zero), then an equilibrium allele frequency has been reached. Notice that fixation or loss of the A allele corresponds to  $\Delta p = 0$  for directional selection. For overdominance and underdominance,  $\Delta p = 0$  for fixation and loss as well as for the intermediate allele frequency of  $p = 0.5$ . These allele frequencies are therefore equilibrium points because natural selection is not causing any change in allele frequency at these specific allele frequencies.

Also compare each plot of  $\Delta p$  against  $p$  to the corresponding plot of  $\bar{w}$  against  $p$ . There is a striking relationship between  $\Delta p$  and  $\bar{w}$ . Both the magnitude and sign of  $\Delta p$  correspond exactly to the slope of

the  $\bar{w}$  line at any value of  $p$ . The slope of  $\bar{w}$  is always positive, just as  $\Delta p$  is always positive for selection against a recessive, while the slope of  $\bar{w}$  is always negative, just as  $\Delta p$  is always negative for selection against a dominant (Fig. 6.10). This same pattern is seen in Fig. 6.11 for overdominance and underdominance, where the slope of  $\bar{w}$  is zero at fixation and loss as well as at  $p = 0.5$ . The slope of  $\bar{w}$  explains why the polymorphic equilibrium point for overdominance is stable while that for underdominance is unstable. With overdominance, if there is any shift of allele frequencies away from  $p = 0.5$ , say by genetic drift or mutation, natural selection will return the population back to the equilibrium of  $p = 0.5$  ( $\Delta p$  is positive for  $p < 0.5$  and negative for  $p > 0.5$ ). In contrast,

### Problem box 6.2 Mean fitness and change in allele frequency

Using equations 6.35 and 6.36 allows us to predict the allele frequencies at equilibrium for selection with overdominance for fitness. We also need to understand *why* the equilibrium is the point at which genotype frequencies stop changing. Let the fitness values be  $w_{AA} = 0.9$ ,  $w_{Aa} = 1.0$ , and  $w_{aa} = 0.8$ . First, calculate expected frequency of the A allele at equilibrium or  $p_{equilibrium}$ . Then compute  $\Delta p$  and  $\bar{w}$  at  $p_{equilibrium}$ ,  $p = 0.9$ , and  $p = 0.2$ . How do the values of  $\Delta p$  and  $\bar{w}$  at the three allele frequencies compare? Use  $\Delta p$  and  $\bar{w}$  to explain why equilibrium allele frequency is between  $p = 0.9$  and  $p = 0.2$ .

with underdominance, if there is any shift of allele frequencies away from  $p = 0.5$ , natural selection will change allele frequencies to the maximum mean fitness values found at fixation and loss ( $\Delta p$  is negative for  $p < 0.5$  and positive for  $p > 0.5$ ).

### The fundamental theorem of natural selection

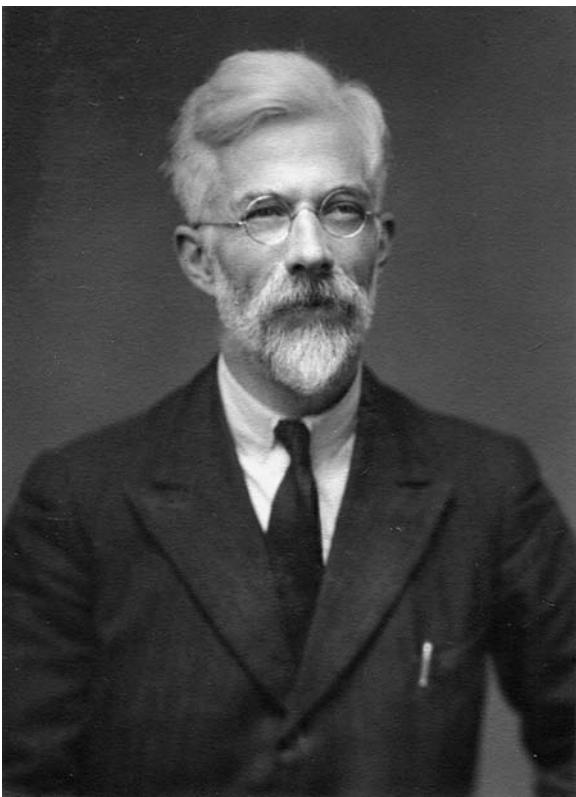
To close this section, let's examine one last generalization about the action of natural selection on a diallelic locus. Sir Ronald Fisher (Fig. 6.12) proposed the impressive sounding **fundamental theorem of natural selection** (Fisher 1999, originally published in 1930) as a way to summarize and generalize the process. In Fisher's words, the fundamental theorem of natural selection was that "the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time." A modern restatement of the theorem is that "the rate of increase in the mean fitness of any organism at any time ascribable to natural selection acting through changes in gene frequencies is exactly equal to its genic variance in fitness at that time" (Edwards 1994). The fundamental theorem has also been interpreted as showing that any change in mean fitness caused by natural selection must always be positive. As Crow (2002) and Edwards (2002) recount, this cryptic yet simultaneously insightful statement about natural selection has lead to a great deal of controversy, misunderstanding, and just plain confusion over many years.

### Interact box 6.1 Natural selection on one locus with two alleles

Launch Populus. In the **Model** menu, choose **Natural Selection** and then **Selection on a Diallelic Autosomal Locus**. In the options dialog box set view to  $p$  vs.  $t$  and use fitness coefficients. Set the fitness values as given below. Click on the button for **One Initial Frequency** and set the initial allele frequency to 0.5. Press the **View** button to see the simulation results. Clicking on the different radio buttons for **Plot Options** switches the view among the four different modes of display ( $p$  vs.  $t$ , Genotypic frequency vs.  $t$ ,  $\Delta p$  vs.  $p$ , and  $\bar{w}$  vs.  $p$ ). When selection is weak, set the generations to view in the range of 500 to 1000, but when selection is strong 50 to 100 generations will be more appropriate. Using the **Six Initial Frequencies** option shows the outcome of natural selection for six different initial allele frequencies like Figs 6.4–6.8, but the Genotypic frequency vs.  $t$  plot will not be available.

Here are some fitness values to simulate.

- Weak selection against recessive:  $w_{AA} = 1$ ;  $w_{Aa} = 1$ ;  $w_{aa} = 0.9$  ( $h = 0$ ,  $s = 0.1$ ). Compare with selection against recessive lethal:  $w_{AA} = 1$ ;  $w_{Aa} = 1$ ;  $w_{aa} = 0.0$  ( $h = 0$ ,  $s = 1.0$ )
- Weak selection with additive gene action:  $w_{AA} = 1$ ;  $w_{Aa} = 0.95$ ;  $w_{aa} = 0.9$  ( $h = 0.5$ ,  $s = 0.1$ ). Compare with strong selection with additive gene action:  $w_{AA} = 1$ ;  $w_{Aa} = 0.7$ ;  $w_{aa} = 0.4$  ( $h = 0.5$ ,  $s = 0.6$ )
- Weak selection with overdominance:  $w_{AA} = 0.98$ ;  $w_{Aa} = 1$ ;  $w_{aa} = 0.95$ . Compare with strong selection with overdominance:  $w_{AA} = 0.2$ ;  $w_{Aa} = 1$ ;  $w_{aa} = 0.4$ .
- Selection against the heterozygote:  $w_{AA} = 1$ ;  $w_{Aa} = 0.8$ ;  $w_{aa} = 1$ . For this case be sure to examine the Genotypic frequency vs.  $t$  plot for several different initial allele frequencies such as 0.2, 0.5 and 0.8.



**Figure 6.12** Sir Ronald A. Fisher (1890–1963) photographed in 1943, was a pioneer in the theory and practice of statistics. He invented the techniques of analysis of variance and maximum likelihood as well as numerous other statistical tests and methods of experimental design. Fisher's 1930 book *The Genetical Theory of Natural Selection* established a rigorous mathematical framework that coupled Mendelian inheritance and Darwin's qualitative model of natural selection and is one of the foundation works of modern population genetics. Much of Fisher's work stressed the effectiveness of natural selection in changing gene frequencies in infinite, panmictic populations. Photograph courtesy of the Master and Fellows of Gonville and Caius College, Cambridge.

One way to illustrate the idea behind the fundamental theorem is to examine natural selection and the change in the average fitness of a population over time. For simplicity, assume that the organisms are entirely haploid and reproduce asexually or clonally and that generations are discrete (these assumptions are not required by the fundamental theorem itself but make the math much simpler). In the haploid case, the average fitness is fitness of each haplotype weighted by its frequency summed over all haplotypes in the population (recall equation 6.10). In an equation the mean fitness is

$$\bar{w} = \sum_{i=1}^k (p_i w_i) \quad (6.46)$$

where  $k$  is the total number of haplotypes in the population. Extending the results in Table 6.1 to an arbitrary number of alleles, the frequency of any single haplotype, call it haplotype  $i$ , after natural selection is

$$p'_i = \frac{p_i w_i}{\bar{w}} \quad (6.47)$$

where the prime symbol is used to represent quantities after one generation of natural selection. Based on this haplotype frequency after selection, the average fitness after one generation of selection is then

$$\bar{w}' = \sum_{i=1}^k (p'_i w_i) \quad (6.48)$$

which when substituting in the expression for  $p'_i$  gives

$$\bar{w}' = \frac{1}{\bar{w}} \sum_{i=1}^k p_i w_i^2 \quad (6.49)$$

The change in fitness from one generation to the next standardized by the mean fitness in the initial generation is

$$\Delta \bar{w} = \frac{\bar{w}' - \bar{w}}{\bar{w}} \quad (6.50)$$

which when substituting in the expression for  $\bar{w}'$  from equation 6.48 gives

$$\Delta \bar{w} = \frac{\frac{1}{\bar{w}} \sum_{i=1}^k p_i w_i^2 - \bar{w}}{\bar{w}} \quad (6.51)$$

an equation that can be rearranged by multiplying by  $\frac{1}{\bar{w}}$  rather than dividing by  $\bar{w}$  to give

$$\Delta \bar{w} = \frac{1}{\bar{w}} \left[ \frac{1}{\bar{w}} \sum_{i=1}^k p_i w_i^2 - \bar{w} \right] = \frac{1}{\bar{w}^2} \sum_{i=1}^k p_i w_i^2 - \bar{w} \quad (6.52)$$

It turns out that the term  $\sum_{i=1}^k p_i w_i^2 - \bar{w}$  is the variance in fitness (the variance is  $\sum (p_i w_i - \bar{w})^2$  which is equivalent to  $\sum p_i w_i^2 - \bar{w}^2$ ) and when both terms are multiplied by the constant  $\frac{1}{\bar{w}}$  it gives the term in

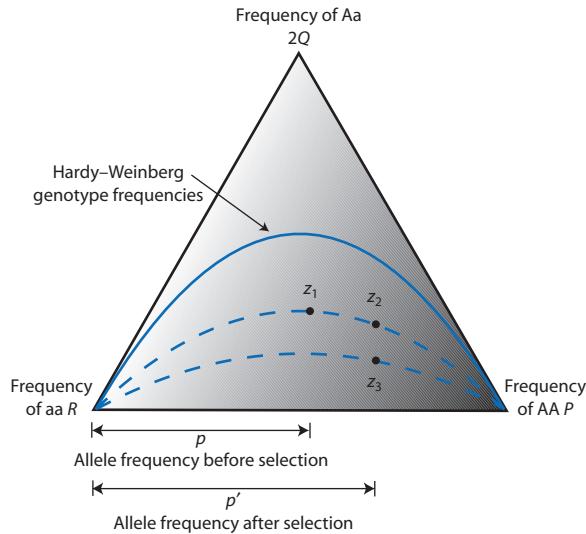
brackets in equation 6.52). In addition, the relative fitness values of all the haplotypes can be scaled so that  $\bar{w} = 1$ . This then leads to

$$\Delta\bar{w} = \text{var}(w) \quad (6.53)$$

and the conclusion that the change in mean fitness of the population after one generation of natural selection is equal to the variation in fitness. This variation in fitness is really genetic variation in the case of haploids, due to the frequencies of the different haplotypes in the population as well as to the different fitness values of each haplotype. Therefore, the change in fitness under natural selection is equal to the genetic variation in fitness. Further, since a variance can never be negative, the change in mean fitness by natural selection must then be greater than or equal to zero.

The point of Fisher's fundamental theorem can also be shown graphically for a diploid diallelic locus using a De Finetti diagram (introduced in Chapter 2) that also displays the mean fitness of the population (Fig. 6.13). To see this, let  $2Q$ ,  $P$ , and  $R$  represent the frequencies of the genotypes Aa, AA, and aa, respectively. The ratio of the genotype frequencies can be expressed as the square of half the heterozygote frequency divided by the product of the homozygote frequencies or  $\lambda = Q^2/PR$ , where  $\lambda$  is a measure of departure from Hardy–Weinberg genotype frequencies akin to the fixation index  $F$ . When genotype frequencies are in Hardy–Weinberg proportions, genotype frequencies are then  $2Q = 2pq$ ,  $P = p^2$ ,  $R = q^2$ , and  $\lambda = 1$ . The two dashed lines in Fig. 6.13 have values of  $\lambda$  less than one. Each point of the De Finetti diagram in Fig. 6.13 will also represent a value of the mean fitness of the population, depending on the specific values of the relative fitness values of the genotypes. Mean fitness on the De Finetti diagram is represented by the grayscale gradient with darker tones representing higher mean fitness.

The change in mean fitness under natural selection can be thought of as a two-step process on the De Finetti diagram (Edwards 2002). In the first step, allele frequencies change from their current value to some new value while keeping the ratio of genotype frequencies constant. This is equivalent to moving from point  $z_1$  to point  $z_2$  while remaining on the line that defines a constant value of  $\lambda$ . In the second step, the population moves from point  $z_2$  to point  $z_3$  by changing its genotype frequencies but not altering its allele frequencies. The first part of the change in mean fitness caused by selection is due to the change



**Figure 6.13** A graphical illustration of R.A. Fisher's fundamental theorem of natural selection. The curved lines represent the product of the homozygote frequencies ( $P = p^2$  and  $R = q^2$ ) as a constant proportion of the square of the product of the allele frequencies ( $Q = pq$ ) or  $\lambda = Q^2/PR$ . Hardy–Weinberg genotype frequencies produced by random mating represent the special case of  $\lambda = 1$  (solid colored line). Mean fitness is represented by the grayscale gradient with darker tones representing higher mean fitness. In this illustration, genotype frequencies start out at  $z_1$ . Suppose that natural selection over one generation changes genotype frequencies to point  $z_3$  (under the conditions that genotype AA has the highest fitness and additive gene action, for example). This change in genotype frequencies can be decomposed into two distinct parts. One part is the change from  $z_1$  to  $z_2$  moving along the curve where  $\lambda$  is constant but allele frequencies change from  $p$  to  $p'$ . The other part is the change in the genotype frequencies (changing the value of  $\lambda$ ) that occurs by moving vertically on the De Finetti diagram from  $z_2$  to  $z_3$  but keeping allele frequencies constant. The fundamental theorem says that the change in the mean fitness by natural selection is proportional to the change in allele frequency alone. Processes other than natural selection, such as mating system, dictate the change in genotype frequencies. When natural selection moves the genotype frequencies along a curve of constant  $\lambda$ , then the total change in mean fitness is completely due to changes in allele frequency and genetic variation in fitness is completely additive. Modified from Edwards (2002).

in allele frequencies alone with everything else held constant. This *partial* change in the mean fitness due exclusively to the change in allele frequencies is exactly the same as the **genic variance** or the **additive genetic variance** that is present in the population at point  $z_1$ . The second part of the change in mean fitness is due to changes in genotype frequency and is therefore caused by factors such as mating patterns or physical linkage resulting in gametic disequilibrium that will change the value

of  $\lambda$  as allele frequencies change. The fundamental theorem says that natural selection will change mean fitness by an amount proportional to the additive genetic variance alone. If  $\lambda$  is constant, the total change in mean fitness is just the change due to the variation in allele frequencies. When  $\lambda$  is not constant, changes in genotype frequencies can either increase or decrease mean fitness and can be thought of as causing an average change of zero.

Genetic variation in phenotype due to the substitution of alleles (additive genetic variation) and due to the effects of genotypes is examined from a completely distinct perspective in Chapters 9 and 10 on quantitative genetics. Those chapters also demonstrate the distinction that is made in the fundamental theorem between genetic variation due to changes in allele frequencies and changes in genotype frequencies. Both approaches give the same result that additive genetic variation is the basis of changes in mean phenotype due to natural selection.

- The outcomes of natural selection on viability for a diallelic locus can be generalized into directional selection (a homozygote most fit) that results in fixation and loss (or very nearly fixation and loss), balancing selection (heterozygote advantage) that maintains both alleles forever, and disruptive selection (heterozygote disadvantage) that results in fixation or loss depending on initial genotype frequencies.
- The fundamental theorem of natural selection shows us that the change in mean fitness by natural selection is proportional to the additive genetic variation in fitness.
- The degree of dominance and recessivity for viability phenotypes impacts the rate of change of genotype frequencies under natural selection because there is not a perfect relationship between genotype and phenotype. Natural selection changes genotype frequencies fastest when gene action is additive.

### Chapter 6 review

- For haploid organisms, natural selection is a population growth process where different genotypes vary in genotype-specific population growth rates. The ratio of genotype-specific growth rates is the relative fitness and it predicts the genotype that will approach fixation in an infinitely expanding population over time.
- Natural selection in diploid organisms also relies on the relative fitness to express genotype-specific growth rates with the addition of sexual reproduction such that pairs of parents can produce a predictable frequency of genotypes in their progeny under random mating.

### Further reading

Arguably the first comprehensive treatment of natural selection that came out of the modern synthesis and still a stimulating read today is:

Fisher RA. 1999. *The Genetical Theory of Natural Selection: a Complete Variorum Edition*. Oxford University Press, Oxford (originally published in 1930).

Another early classic of the modern synthesis that established the mathematical connections between Mendelian genetics and natural selection is:

Haldane JBS. 1990. *The Causes of Evolution*. Princeton University Press, Princeton, NJ (originally published in 1932).

### Problem box answers

#### Problem box 6.1 answer

To solve for relative fitness given initial and final allele frequencies and time elapsed, we need to rearrange equation 6.8 by taking the logarithm of both sides:

$$\log\left(\frac{q_t}{p_t}\right) = t \log(w) + \log\left(\frac{q_0}{p_0}\right)$$

to remove the exponent. We will let  $p$  represent the frequency of the wild-type allele and  $q$  the combined frequency of the drug-resistant alleles. Based on 601 days between allele frequency estimates,  $t = 231$  generations elapsed. Substituting these values gives

$$\log\left(\frac{0.51}{0.49}\right) = (231) \log(w) + \log\left(\frac{0.99}{0.01}\right)$$

$$\begin{aligned}
 0.01737 &= (231) \log(w) + 1.9956 \\
 -1.9782 &= (231) \log(w) \\
 -1.9782/231 &= \log(w) \\
 -0.008564 &= \log(w) \\
 10^{-0.008564} &= w \\
 w &= 0.9805
 \end{aligned}$$

The relative fitness of the drug-resistant alleles is 98% of the wild type allele and therefore the wild-type allele increases in frequency over time when AZT is not present.

### Problem box 6.2 answer

$$p_{equilibrium} = t/(s+t) = 0.2/(0.1+0.2) = 2/3$$

For equilibrium allele frequencies:

$$\begin{aligned}
 \bar{w} &= 0.9(0.667)^2 + (1)2(0.667)(0.333) \\
 &\quad + 0.8(0.333)^2 \\
 &= 0.9333 \\
 \Delta p &= \frac{(0.667)(0.333)[0.667(0.9-1) + 0.333(1-0.8)]}{0.9333} \\
 &= 0
 \end{aligned}$$

or calculated using the marginal fitness

$$\begin{aligned}
 p_{t+1} &= \frac{0.9(0.667)^2 + 1(0.667)(0.333)}{0.9333} = 0.667 \\
 \Delta p &= 0.667 - 0.667 = 0
 \end{aligned}$$

At  $p = 0.9$  ( $p > p_{equilibrium}$ )

$$\begin{aligned}
 \bar{w} &= 0.9(0.9)^2 + (1)2(0.9)(0.1) + 0.8(0.1)^2 \\
 &= 0.917
 \end{aligned}$$

$$\begin{aligned}
 \Delta p &= \frac{(0.9)(0.0)[0.9(0.9-1) + 0.1(1-0.8)]}{0.917} \\
 &= -0.0069
 \end{aligned}$$

or calculated using the marginal fitness

$$\begin{aligned}
 p_{t+1} &= \frac{0.9(0.9)^2 + 1(0.9)(0.1)}{0.917} = 0.8931 \\
 \Delta p &= 0.8931 - 0.9 = -0.0069
 \end{aligned}$$

At  $p = 0.2$  ( $p < p_{equilibrium}$ )

$$\begin{aligned}
 \bar{w} &= 0.9(0.2)^2 + (1)2(0.2)(0.8) + 0.8(0.8)^2 \\
 &= 0.868 \\
 \Delta p &= \frac{(0.2)(0.8)[0.2(0.9-1) + 0.8(1-0.8)]}{0.868} \\
 &= 0.0258
 \end{aligned}$$

or calculated using the marginal fitness

$$\begin{aligned}
 p_{t+1} &= \frac{0.9(0.2)^2 + 1(0.2)(0.8)}{0.868} = 0.2258 \\
 \Delta p &= 0.2258 - 0.2 = 0.0258
 \end{aligned}$$

At  $p = 0.9$ ,  $\bar{w}$  is lower than at  $p_{equilibrium}$ . Therefore,  $\Delta p$  is negative meaning that natural selection is causing allele frequencies to decrease. At  $p = 0.2$ ,  $\bar{w}$  is also lower than at  $p_{equilibrium}$ . Therefore, increasing allele frequencies (positive  $\Delta p$ ) by natural selection causes an increase in mean fitness. At  $p_{equilibrium}$ ,  $\bar{w}$  is at its maximum for these relative fitness values and so  $\Delta p$  is zero because selection will no longer change the allele frequencies.

# CHAPTER 7

## Further models of natural selection

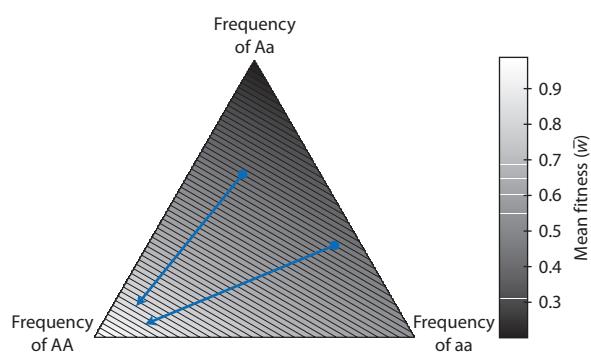
### 7.1 Viability selection with three alleles or two loci

- Mean fitness surfaces.
- Natural selection on one locus with three alleles.
- Natural selection on two diallelic loci.

Chapter 6 established a series of general predictions about the action of natural selection when fitness is equivalent to genotype-specific viability determined by a single locus with two alleles. The conditions required for the basic diallelic locus model of natural selection are quite restrictive and are probably not met often in biological populations. The goal of this chapter is to extend our understanding of the model of natural selection to increasingly complex and general genetic situations. In a sense then, this chapter explores the process of natural selection under assumptions that might better approximate conditions found in some natural populations. In the first section, we will retain the viability natural selection model and its assumptions but modify the numbers of alleles at a locus and the number of loci. The goal is to examine the outcomes of viability selection when fitness is determined by either a single locus with three alleles or two loci each with two alleles.

A useful tool that we will employ to understand the dynamics of genotype frequencies, allele frequencies, and mean fitness under natural selection is called a fitness surface. A fitness surface is a graph that shows genotype frequencies of a population on some axes along with the mean fitness of the population at each possible point in the range of genotype frequencies. For one locus with two or three alleles, a De Finetti diagram can be used as a fitness surface, as shown in Fig. 7.1. The three axes represent genotype frequencies of a population on the plot. Each point inside the triangle defines three genotype frequencies that are then used to compute the mean fitness of the population. The mean fitness is represented by shading as well as contour lines that connect points

of equal mean fitness. Since contour plots of mean fitness are interpreted exactly like topographic maps where contour lines are used to represent elevation, they are also called fitness landscapes or adaptive landscapes. The highest point on a fitness surface represents equilibrium genotype frequencies under natural selection. A fitness surface also shows how natural selection will change genotype frequencies over time if the process of natural selection operates like a hiker who can only travel uphill. For any point on a fitness surface, natural selection will act to increase mean fitness of the population and shift genotype frequencies in a direction that increases the mean fitness. Once the population is at a point where mean fitness cannot increase, natural selection has reached an equilibrium and stops changing genotype frequencies. In Fig. 7.1, the entire surface is a tilted plane that is has its highest point at the



**Figure 7.1** A fitness surface made by including mean fitness on a De Finetti plot of the three genotype frequencies for a diallelic locus. The colored lines indicate the possible trajectories of genotype frequencies as natural selection increases the mean fitness of the population. The fitness values are  $w_{AA} = 1.0$ ,  $w_{Aa} = 0.6$ , and  $w_{aa} = 0.2$  so the highest mean fitness is found in the lower left apex when the population is fixed for the AA genotype. This highest fitness point can be reached by continually increasing mean fitness from any initial point on the surface. Gene action is additive because alleles have a constant impact fitness regardless of the allele they are paired with in a genotype. An A allele always contributes 0.5 and an a allele 0.1 toward the fitness of a genotype.

left vertex or where the AA genotype is fixed in the population. Therefore, natural selection will change genotype frequencies such that the population climbs in mean fitness until reaching fixation for AA.

### Natural selection on one locus with three alleles

With an understanding of fitness surfaces, let's now turn to the classic case of natural selection on three alleles at the human hemoglobin  $\beta$  gene (see Allison 1956; Modiano et al. 2001). The hemoglobin protein is found in red blood cells and is responsible for binding and then carrying oxygen from the lungs to the entire body. Adult hemoglobin is formed from four separate proteins, two  $\alpha$  (or "alpha") proteins and two  $\beta$  (or "beta") proteins. The hemoglobin  $\beta$  gene encodes the  $\beta$  protein, which is often referred to as  $\beta$ -globin or  $Hb$ . The  $Hb$  A allele is the most common allele in human populations. Although several hundred  $Hb$  alleles have been identified in human populations, the  $Hb$  S allele is a common low-frequency allele. The S allele is characterized by a nucleotide change that results in substitution of the hydrophobic amino acid valine in place of the hydrophilic glutamic acid at the sixth amino acid position of the  $\beta$ -globin protein. Individuals homozygous for the S allele exhibit changes in red blood cell morphology ("sickling") and impaired oxygen transport that leads to chronic anemia (Ashley-Koch et al. 2000). The  $Hb$  C allele is also present at low frequencies in West African and southeast Asian populations. Individuals who are CC homozygotes have mild to moderate anemia and enlargement of the spleen that is often asymptomatic (e.g. Fairhurst & Casella 2004).

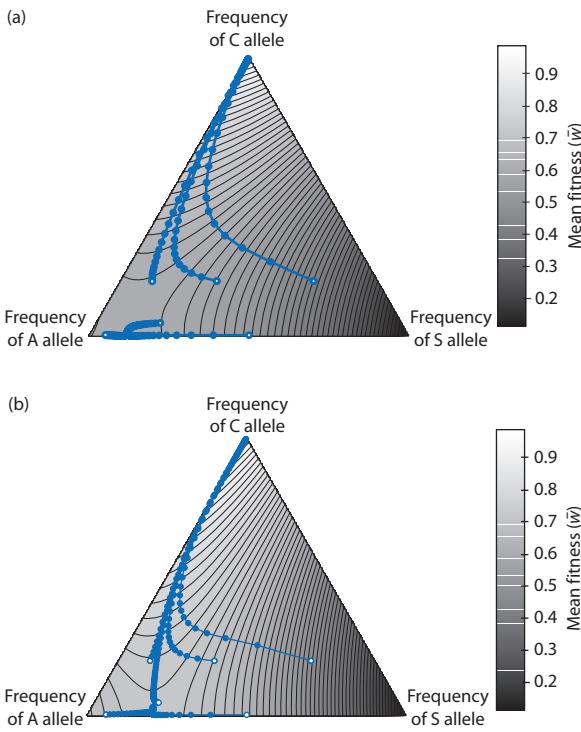
The fitness of  $Hb$  genotypes depends on the environment where people live. In areas of the world without the malarial parasite *Plasmodium falciparum*, genotypes that result in anemia and related conditions have lower fitness. However, in regions where malarial infection is common, certain  $Hb$  genotypes confer resistance to infection by *P. falciparum* that may partly or completely compensate for any disadvantage due to anemia. Two estimates of the relative fitnesses of the six  $Hb$  genotypes in Western Africa where malaria is common are shown in Table 7.1.

A seemingly obvious prediction from Table 7.1 is that natural selection in populations where malaria is common would increase the frequency of the CC genotype and eventually fix the C allele. But is this really what will happen? The answer comes from examining fitness surfaces for the six  $Hb$  genotypes. With three alleles there are six genotype frequencies, which is too many to represent in a De Finetti plot like Fig. 7.1. But since the allele frequencies must sum to one, we can represent the fitness surface on a ternary graph where each axis represents one of the three allele frequencies. Fitness surfaces drawn in this way are shown Fig. 7.2 for the two sets of fitness values given in Table 7.1. These three allele fitness surfaces are now rippled or hilly compared to the fitness surface in Fig. 7.1.

Understanding how genotype frequencies will change on a fitness surface requires calculating the change in allele frequencies due to selection for a series of points on the surface. The sign and magnitude of the change in allele frequency will be a function of the slope of the fitness surface at any point we examine. To do this for the fitness surfaces

**Table 7.1** Relative fitness estimates for the six genotypes of the hemoglobin  $\beta$  gene estimated in Western Africa where malaria is common. Values from Cavallo-Sforza and Bodmer (1971) are based by deviation from Hardy-Weinberg expected genotype frequencies. Values from Hedrick (2004) are estimated from relative risk of mortality for individuals with AA, AC, AS, and CC genotypes and assume 20% overall mortality from malaria.

Genotype . . .	Relative fitness ( $w$ )					
	AA	AS	SS	AC	SC	CC
<b>From Cavallo-Sforza and Bodmer (1971)</b>						
Relative to $w_{CC}$	0.679	0.763	0.153	0.679	0.534	1.0
Relative to $w_{AS}$	0.89	1.0	0.20	0.89	0.70	1.31
<b>From Hedrick (2004)</b>						
Relative to $w_{CC}$	0.730	0.954	0.109	0.865	0.498	1.0
Relative to $w_{AS}$	0.765	1.0	0.114	0.906	0.522	1.048



**Figure 7.2** Fitness surfaces for the A, S, and C alleles at the human hemoglobin  $\beta$  gene when malaria is common. The surface in (a) corresponds to the top set of fitness values in Table 7.1 and (b) shows the surface for the bottom set of values. The tracks of circles represent generation-by-generation allele frequency trajectories due to natural selection over 50 generations calculated with equation 7.3. In (a), when the initial frequency of the C allele is relatively high, the equilibrium of natural selection is the fixation of the CC genotype. In contrast, when the C allele is initially rare (a frequency of less than about 7%) selection reaches an equilibrium with only the A and S alleles segregating and the C allele going to loss. In (b), selection will eventually fix the CC genotype from any initial frequency of the C allele. However, when the C allele is at low frequencies, the increase in the C allele each generation is extremely small so that selection would take hundreds of generations to fix the CC genotype. The six initial allele frequency points, shown as open circles, are identical for the two surfaces.

in Fig. 7.2, we need to extend the viability model of natural selection to three alleles at one locus. We can compute the mean fitness of the population:

$$\bar{w} = w_{AA}p^2 + w_{BB}q^2 + w_{CC}r^2 + w_{AB}2pq + w_{AC}2pr + w_{BC}2qr \quad (7.1)$$

where  $p$ ,  $q$ , and  $r$  represent the frequencies of the three alleles A, B, and C. We can also use the marginal fitness of the genotypes that contain each of the alleles to compute whether an allele will increase or decrease in frequency due to the average fitness of all the geno-

types that carry the allele. When there is random mating, the marginal fitness for the A allele is

$$\begin{aligned}\bar{w}_A &= \frac{w_{AA}p^2 + w_{AB}pq + w_{AC}pr}{p} \\ &= w_{AA}p + w_{AB}q + w_{AC}r\end{aligned}\quad (7.2)$$

where the frequencies of the heterozygous genotypes are multiplied by  $1/2$  since they carry one copy of the A allele. The marginal fitness is a way to compare the ratio of  $p$  in the current generation with frequency of  $p$  in the next generation that will result from natural selection changing genotype frequencies. Allele frequencies change each generation due to differences between the marginal fitness of each allele and the average fitness of the entire population. The change in the frequency of the A allele is

$$\Delta p = p \frac{(\bar{w}_A - \bar{w})}{\bar{w}} \quad (7.3)$$

Allele frequency after one generation of selection is then simply  $p_{t+1} = p + \Delta p$ . Similar expressions are obtained easily for the B and C alleles. Also note that this approach can be extended to an arbitrary number of alleles at one locus as long as genotypes are in Hardy–Weinberg frequencies at the start of each generation before the action of selection.

Returning to the fitness surfaces, Fig. 7.2a is an interesting case because it has two stable equilibrium points. One of the equilibrium points matches our intuition after inspecting Table 7.1 that the CC genotype should be fixed by selection. When the initial frequency of the C allele is relatively high, all three trajectories of allele frequencies over 10 generations of selection calculated with equation 7.3 are clearly headed for fixation of CC. In contrast, when the initial frequency of the C allele is low the trajectories of allele frequencies show that the C allele will be lost from the population. This is counterintuitive given that the CC genotype has the highest relative fitness. This result is a consequence of the fitness surface. When C is at low frequency its marginal fitness is actually less than the mean fitness. In other words, the fitness surface is going down in elevation toward higher frequencies of the C allele. Since natural selection only works to increase mean fitness, the C allele is reduced in frequency to loss.

To see one possible consequence of this fitness surface, imagine that the A and S alleles are older in human populations than the C allele and the A and S alleles have reached equilibrium frequencies.

Using equation 6.35 and Table 7.1, the equilibrium frequency of the A allele would be  $t/(s+t) = 0.8/(0.11+0.8) = 0.88$  and therefore the equilibrium frequency of S would be  $1 - 0.88 = 0.11$ . Next imagine that the C allele occurs in the population at a later time due to mutation. Since mutation rates are low, the resulting frequency of the C allele will also be low and most C alleles would occur in AC and SC heterozygotes. All heterozygotes have overdominance (AS) or underdominance (SC and AC) for fitness. In particular, the SC heterozygote has a lower relative fitness than AA and AS genotypes and so its marginal fitness would be negative when C is at a low frequency. Thus, to get from a mean fitness state where C is infrequent we would have to go through a dip of lower mean fitness as C initially increases. At higher initial frequencies of C, however, mean fitness increases steadily until CC fixes. So if A and S alleles were ancestral, natural selection alone would drive a newly introduced C allele to loss despite the high relative fitness of the CC homozygote.

For the fitness surface in Fig. 7.2b, natural selection will eventually fix the CC genotype from any initial frequency of the C allele. However, when the C allele is at low frequency selection increases the frequency of the C allele very slowly. This is because the marginal fitness of the C allele is only very slightly greater than the mean fitness below a frequency of about 15% when the frequency of the A allele is also high. This can be seen on the fitness surface by noting the

### Problem box 7.1 Marginal fitness and $\Delta p$ for the Hb C allele

Compute the mean fitness, marginal fitness of the C allele, and the change in the C allele using the two sets of initial allele frequencies given below and relative fitness values from the top of Table 7.1. Use  $\Delta p$  along with Fig. 7.2a to predict the equilibrium that will be reached by natural selection for both initial allele frequencies.

Initial allele frequencies set 1:  $p = 0.75$ ,  
 $q = 0.20$ ,  $r = 0.05$

Initial allele frequencies set 2:  $p = 0.70$ ,  
 $q = 0.20$ ,  $r = 0.10$

wide spacing between contour lines toward the left vertex. Widely spaced contour lines indicate areas with little slope. These are areas where the mean fitness of the population is either constant or nearly constant for a range of genotype frequencies. Such flat areas on fitness surfaces can be stable or unstable equilibrium points and are regions where selection is a weak process because the marginal fitnesses are very close in value to the mean fitness.

Determining which of the different hemoglobin  $\beta$  genotype fitness values best describe actual

### Interact box 7.1 Natural selection on one locus with three or more alleles

Direct simulation of selection on one locus with three alleles is an easy way to see that equilibrium points depend strongly on over- and underdominance for fitness. Populus has the ability to simulate selection on a locus with three or more alleles. Launch Populus and, in the **Model** menu, choose **Natural Selection** and then **Selection on a Multi-allelic Locus**. Click on each of the radio buttons for the display options to see how the results can be displayed. Note that when using the default fitness values the  $P_3$  allele goes to fixation. The options dialog box can be made larger by dragging the tab at the bottom right, making it easier to see the parameter fields.

Then try some different fitness values:

Additivity:  $w_{11} = 0.6$ ,  $w_{12} = 0.7$ ,  $w_{13} = 0.8$ ;  $w_{21} = 0.7$ ,  $w_{22} = 0.8$ ,  $w_{23} = 0.9$ ;  $w_{31} = 0.8$ ,  $w_{32} = 0.9$ ,  $w_{33} = 1.0$

Overdominance:  $w_{11} = w_{22} = w_{33} = 0.3$ ;  $w_{12} = w_{13} = w_{21} = w_{23} = w_{13} = w_{31} = 1.0$

Underdominance:  $w_{11} = w_{22} = w_{33} = 1.0$ ;  $w_{12} = w_{13} = w_{21} = w_{23} = w_{13} = w_{31} = 0.3$

Also be sure to vary the allele frequencies for each set of fitness values. You might try frequencies of all alleles equal at  $1/3$ , and then one allele more common, with 0.67, 0.12, and 0.21 (the default values).

populations is not the main point of this example. Rather, the hemoglobin  $\beta$  gene serves to illustrate that dominance for fitness, the order of appearance of alleles in a population and the relative fitness values may all interact to determine the outcome of natural selection with three alleles.

### Natural selection on two diallelic loci

Since phenotypes, and therefore fitness, may be caused by more than one locus, a logical step is to extend the model of natural selection to two loci. Biologically there is strong motivation to consider selection at more than one locus since many phenotypes are known to show variation caused by multiple loci (see Chapter 9). The fate of two mutations could also be considered as two-locus selection. Natural selection on two loci is inherently more complicated than at a single locus because of gametic disequilibrium. As covered in Chapter 2, both linkage and natural selection itself produce gametic disequilibrium that must be accounted for in a two-locus model of natural selection. Because natural selection on two loci is considerably more complex than on just one locus, the goal of this section is to provide a general introduction to two-locus models. It is important to recognize at the outset that there is no easily summarized set of equilibria for two-locus selection as there are for selection on a diallelic locus. The outcome of two-locus selection depends on the balance between natural selection and recombination between loci as well as the initial genotype frequencies in the population.

Two-locus natural selection is commonly approached from the perspective of gametes because gametic disequilibrium is expressed in terms of gamete frequencies. With two diallelic loci there are 16 possible genotypes that result from the union of four possible gametes. Let the frequencies of the gametes AB, Ab, aB, and ab be  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$ . Table 7.2 shows the relative fitness values for all possible combinations of four gametes. There are only 10 unique fitness values if the same gamete inherited from either parent has the same fitness in a progeny genotype. For example, if an Ab gamete from either a male or female parent has the same fitness in an AB/Ab progeny genotype, then  $w_{12} = w_{21}$  in the fitness matrix. Table 7.3 then gives the expected frequencies of each progeny genotype under the assumption of random mating (compare with Table 2.12). The frequencies of each gamete in the next generation can be obtained by summing each of the columns in Table 7.3 while also weighting each expected frequency by the relative fitness of each genotype. For example, the

**Table 7.2** Matrix of fitness values for all combinations of the four gametes formed at two diallelic loci (top). If the same gamete inherited from either parent has the same fitness in a progeny genotype (e.g.  $w_{12} = w_{21}$ ), then there are 10 gamete fitness values shown outside the shaded triangle. These 10 fitness values can be summarized by a genotype fitness matrix (bottom) under the assumption that double heterozygotes have equal fitness ( $w_{14} = w_{23}$ ) and representing their fitness value by  $w_H$ . The double heterozygote genotypes are of special interest since they can produce recombinant gametes.

	<b>AB</b>	<b>Ab</b>	<b>aB</b>	<b>ab</b>
<b>AB</b>	$w_{11}$	$w_{12}$	$w_{13}$	$w_{14}$
<b>Ab</b>	$w_{21}$	$w_{22}$	$w_{23}$	$w_{24}$
<b>aB</b>	$w_{31}$	$w_{32}$	$w_{33}$	$w_{34}$
<b>ab</b>	$w_{41}$	$w_{42}$	$w_{43}$	$w_{44}$

	<b>BB</b>	<b>Bb</b>	<b>bb</b>
<b>AA</b>	$w_{11}$	$w_{12}$	$w_{22}$
<b>Aa</b>	$w_{13}$	$w_H$	$w_{24}$
<b>aa</b>	$w_{33}$	$w_{34}$	$w_{44}$

expected frequency of AB gametes after one generation of natural selection and recombination is

$$x_{1(t+1)} = \frac{w_{11}x_1^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + (1-r)w_{14}x_1x_4 + rw_{23}x_2x_3}{\bar{w}} \quad (7.4)$$

which is directly comparable with equation 6.21 for one allele at a diallelic locus. This can be simplified by expanding the  $(1-r)w_{14}x_1x_4$  term:

$$x_{1(t+1)} = \frac{w_{11}x_1^2 + w_{12}x_1x_2 + w_{13}x_1x_3 + w_{14}x_1x_4 - rw_{14}x_1x_4 + rw_{23}x_2x_3}{\bar{w}} \quad (7.5)$$

and then factoring an  $x_1$  out of the first four terms and an  $r$  out of the last two terms

$$x_{1(t+1)} = \frac{x_1(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4 - r(w_{14}x_1x_4 + w_{23}x_2x_3))}{\bar{w}} \quad (7.6)$$

An additional simplification is possible if we assume that the fitness of genotypes with the same number

**Table 7.3** Expected frequencies of gametes under viability selection for two diallelic loci in a randomly mating population with a recombination rate of  $r$  between the loci. The expected gamete frequencies assume that the same gamete coming from either parent will have the same fitness in a progeny genotype (e.g.  $w_{12} = w_{21}$ ). Eight genotypes have non-recombinant and recombinant gametes that are identical and so do not require a term for the recombination rate. Two genotypes produce novel recombinant gametes, requiring inclusion of the recombination rate to predict gamete frequencies. Summing down each column of the table gives the total frequency of each gamete in the next generation due to mating and recombination.

Genotype	Fitness	Total frequency	Frequency of gametes in next generation			
			AB	Ab	aB	ab
AB/AB	$w_{11}$	$x_1^2$	$x_1^2$			
AB/Ab	$w_{12}$	$2x_1x_2$	$x_1x_2$	$x_1x_2$		
AB/aB	$w_{13}$	$2x_1x_3$	$x_1x_3$		$x_1x_3$	
AB/ab	$w_{14}$	$2x_1x_4$	$(1 - r)x_1x_4$	$(r)x_1x_4$	$(r)x_1x_4$	$(1 - r)x_1x_4$
Ab/Ab	$w_{22}$	$x_2^2$		$x_2^2$		
Ab/aB	$w_{23}$	$2x_2x_3$	$(r)x_2x_3$	$(1 - r)x_2x_3$	$(1 - r)x_2x_3$	$(r)x_2x_3$
Ab/ab	$w_{24}$	$2x_2x_4$		$x_2x_3$	$x_2x_3$	
aB/aB	$w_{33}$	$x_3^2$			$x_3^2$	
aB/ab	$w_{34}$	$2x_3x_4$			$x_3x_4$	$x_3x_4$
ab/ab	$w_{44}$	$x_4^2$				$x_4^2$

of A and B alleles is equal. For example, the double heterozygotes AB/ab and Ab/aB have the same number of A and B alleles and so we can reasonably assume they have equal fitness values (Table 7.2). This assumption allows us to equate the fitnesses of those double heterozygotes where recombination plays a role in the gametes that are produced. Applying this assumption to equation 7.6, we can set  $w_{14} = w_{23}$  and then the  $r(w_{14}x_1x_4 + w_{23}x_2x_3)$  term becomes  $rw_{14}(x_1x_4 + x_2x_3)$  to give

$$x_{1(t+1)} = \frac{x_1(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4) - rw_{14}(x_1x_4 + x_2x_3)}{\bar{w}} \quad (7.7)$$

This helps because the gametic disequilibrium parameter  $D$  is the difference between the product of the coupling gametes and the product of the repulsion gametes (see equation 2.24). In the notation of this section  $D = x_1x_4 - x_2x_3$ . We can then substitute  $-D$  for  $x_1x_4 + x_2x_3$  in equation 7.7 to give

$$x_{1(t+1)} = \frac{x_1(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4) - rw_{14}D}{\bar{w}} \quad (7.8)$$

Equation 7.8 shows that the frequency of AB gametes after one generation of natural selection is a function of three things. First, the viabilities of the three

genotypes that produce AB gametes can change genotype frequencies and thereby impact the frequency of AB gametes in the next generation (recombination does *not* alter the frequency of AB gametes produced by the AB/AB, AB/Ab, and AB/aB genotypes). Then, an additional part of the frequency of AB gametes is determined by the combination of recombination, fitness values of the double heterozygotes, and initial gametic disequilibrium in the population. Double heterozygotes could be more or less frequent than expected by random mating as measured by  $D$ . Also the frequency of recombination and the relative fitness of the genotypes will determine how many AB gametes are produced. If  $D$  and  $r$  could be ignored, the frequency of AB gametes would be analogous to the frequency of one of the four possible gametes for a single locus with four alleles.

Expanding on the idea that the four gamete frequencies can be treated like the frequencies of four alleles at one locus, we can utilize some of the expressions developed earlier in the chapter for a single locus. The marginal fitness for each of the two-locus gametes ( $\bar{w}_i$ ) is obtained by summing the frequency-weighted fitness value of each of the gametes that a given gamete could pair with to make a genotype:

$$\bar{w}_i = \sum_{j=1}^4 x_j w_{ij} \quad (7.9)$$

Similarly, the average fitness of the population is the frequency-weighted average of the fitness values for all of the possible gamete combinations:

$$\bar{w} = \sum_{i=1}^4 \sum_{j=i}^4 x_i x_j w_{ij} \quad (7.10)$$

The marginal fitness and mean fitness can be combined with equation 7.8 to give an expression for the change in a gamete frequency under selection and recombination.

To continue with the AB gamete as an example, notice that the marginal fitness  $\bar{w}_1$  is equal to  $x_1 w_{11} + x_2 w_{12} + x_3 w_{13} + x_4 w_{14}$ . Making this substitution in equation 7.8, dividing by the mean fitness, and substituting  $w_H$  for  $w_{14}$  or  $w_{23}$  gives the change in the AB gamete frequency over one generation of natural selection:

$$\Delta x_1 = \frac{x_1 \bar{w}_1 - r w_H D}{\bar{w}} \quad (7.11)$$

This is exactly like the expression for  $\Delta p$  for a diallelic locus (compare with equation 6.23). Using analogous steps for the other three gametes gives the recursion equations for change in gamete frequency after one generation of natural selection and recombination:

$$\Delta x_2 = \frac{x_2 \bar{w}_2 + r w_H D}{\bar{w}} \quad (7.12)$$

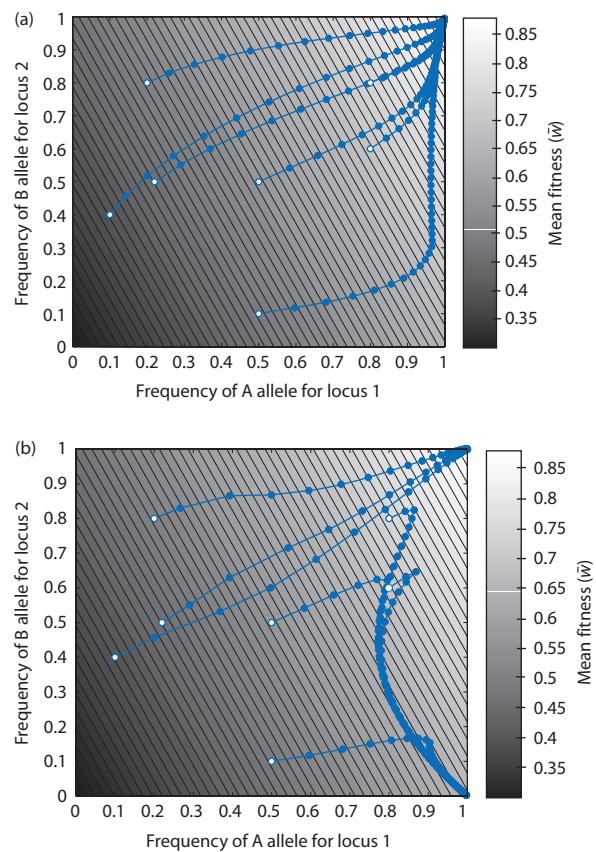
$$\Delta x_3 = \frac{x_3 \bar{w}_3 + r w_H D}{\bar{w}} \quad (7.13)$$

$$\Delta x_4 = \frac{x_4 \bar{w}_4 - r w_H D}{\bar{w}} \quad (7.14)$$

Equations 7.11–7.14 show that the change in gamete frequency under natural selection is due to both fitness values and recombination. If there is no recombination ( $r = 0$ ), then each gamete is analogous to a single allele. The outcome of selection is then like four alleles at a single locus as dictated by the gamete fitness values. The process of recombination may either reinforce or oppose the changes in gamete frequencies due to natural selection. For example, if gametes Ab and aB have the highest fitness values and there is no recombination then  $\Delta x_2$  and  $\Delta x_3$  would be positive while  $\Delta x_1$  and  $\Delta x_4$  would be negative (when not at equilibrium). The gamete frequency changes caused by recombination would amplify the effect of

natural selection on gamete frequencies since the  $r w_H D$  term would increase  $\Delta x_2$  and  $\Delta x_3$  but decrease  $\Delta x_1$  and  $\Delta x_4$ . In contrast, if gametes AB and ab have the highest fitness and there is recombination, then the  $r w_H D$  term would decrease  $\Delta x_1$  and  $\Delta x_4$  but increase  $\Delta x_2$  and  $\Delta x_3$  in opposition to natural selection.

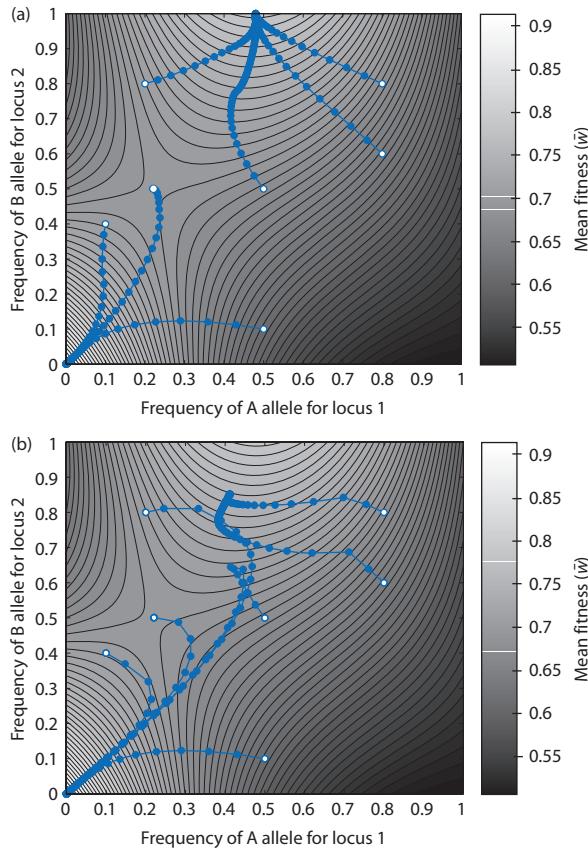
Examining natural selection on two loci with and without recombination demonstrates that selection and recombination working in opposition can produce counterintuitive equilibrium gamete frequencies. Figure 7.3 shows a fitness surface where



**Figure 7.3** A fitness surface for two loci that each have two alleles where gene action is additive. The blue dots show generation-by-generation allele frequencies based on equations 7.11–7.14 for seven different initial sets of four gamete frequencies. When recombination is a weak force ( $r = 0.05$ ), equilibrium allele frequencies are dictated by natural selection and all initial gamete frequencies eventually reach the highest mean fitness point (a). In contrast, when recombination is a strong force ( $r = 0.5$ ) then equilibrium allele frequencies depend on initial gamete frequencies (b). When recombination is strong, equilibrium allele frequencies may not correspond to the highest mean fitness. Relative fitness values are  $w_{AABB} = 0.9$ ,  $w_{AAbb} = 0.8$ ,  $w_{AAbb} = 0.7$ ,  $w_{AaBB} = 0.7$ ,  $w_{AaBb} = 0.6$ ,  $w_{Aabb} = 0.5$ ,  $w_{aaBB} = 0.5$ ,  $w_{aaBb} = 0.4$ , and  $w_{aabb} = 0.3$ . The seven initial allele frequency points, shown as open circles, are identical for the two surfaces.

gene action is completely additive. Since the fitness surface is a tilted plane, our earlier experience with one locus selection suggests that the equilibrium under natural selection should be the highest fitness point. When recombination is a weak force relative to selection (Fig. 7.3a), then the change in gamete frequencies follows the slope of the fitness surface and the equilibrium point reached from all initial gamete frequencies is the highest mean fitness. However, when recombination is strong relative to selection (Fig. 7.3b), then equilibrium gamete frequencies depend strongly on initial gamete frequencies. Figure 7.4 shows another example of two-locus selection on a fitness surface with two peaks shaped like a saddle due to dominance and epistasis at the two loci. When recombination is weak (Fig. 7.4a), then the equilibrium points depend on the slope of the fitness surface at the initial gamete frequencies since populations move uphill due to the strong force of selection. However, when recombination is strong relative to selection (Fig. 7.4b) then gamete frequencies will change in opposition to selection and change in directions that decrease mean fitness. When recombination is strong in Figs 7.3 and 7.4, the gamete frequency trajectories take sharp turns and move downhill on the fitness surfaces due to the force of recombination. This happens because recombination works toward gametic equilibrium ( $D = 0$ ) whereas selection works toward the highest mean fitness. When one process is much stronger than it will win out over the other process to determine the equilibrium. When the two processes are of approximately equal strength then the result is a compromise that may produce an equilibrium that is neither gametic equilibrium nor maximum mean fitness.

The fitness surface in Fig. 7.4 demonstrates an additional point about the action of natural selection on two loci. Gene action is a key variable in determining the equilibrium reached by natural selection. With additive gene action for two loci, the genotype at one locus has the same fitness value regardless of the genotype at the other locus. This means continual small changes in genotype frequencies that each increase mean fitness will eventually reach the highest mean fitness. In contrast, with non-additive gene action (dominance and epistasis) those same small, generation-by-generation changes in allele frequencies may lead to local maxima because the fitness surface is not a plane. Such peaks and valleys of mean fitness occur when the genotype at one locus has an impact on fitness values at another locus. Fitness surfaces therefore have increasingly com-



**Figure 7.4** A fitness surface for two loci that each have two alleles where gene action exhibits epistasis. When recombination is a weak force ( $r = 0.05$ ), equilibrium allele frequencies are dictated by natural selection. Equilibrium allele frequencies depend on initial gamete frequencies since the two highest mean fitness points are separated by a fitness valley (a). When recombination is strong ( $r = 0.5$ ), allele frequencies change such that mean fitness actually decreases for a time before increasing again to eventually reach the lower of the two mean fitness peaks (b). The two initial gamete frequencies in the upper right of the surface reach an equilibrium point where fitness is not maximized and there is gametic disequilibrium ( $D = 0.041$ ). Relative fitness values are  $w_{AABB} = 0.61$ ,  $w_{AABb} = 0.58$ ,  $w_{AAbb} = 0.50$ ,  $w_{AaBB} = 1.0$ ,  $w_{AaBb} = 0.77$ ,  $w_{Aabb} = 0.50$ ,  $w_{aaBB} = 0.64$ ,  $w_{aaBb} = 0.62$ , and  $w_{aabb} = 0.92$ . The seven initial allele frequency points, shown as open circles, are identical for the two surfaces.

plex topography as dominance and epistasis increase and additive gene action decreases. For this reason, natural selection is sometimes described as short-sighted or myopic because it operates based on mean fitness each generation rather than on some plan that accounts for the entire mean fitness surface. The result is that the equilibrium produced by natural selection when mean fitness surfaces possess multiple maxima depends strongly on initial genotype frequencies.

Although there is no general set of equilibrium gamete frequencies for two-locus selection with an arbitrary set of fitness values, many special cases have been examined that have produced some general conclusions (see Hastings 1981, 1986; reviewed by Ewens 2004). By itself, low frequencies of recombination (small  $r$ ) make it more likely that selection will result in gametic disequilibrium at equilibrium gamete frequencies even with random mating. The combination of non-additive gene action and infrequent recombination also make gametic disequilibrium at equilibrium gamete frequencies more likely. High rates of self-fertilization (strong departure from random mating) adds an additional force on gamete frequencies that can either compliment or act in opposition to selection and recombination (Hastings 1985; Holsinger & Feldman 1985). Since mean fitness may decrease with selection and recombination, Fisher's fundamental theorem does not hold for two-locus selection (see Turner 1981; Hastings 1987). A critical conclusion from examining two-locus natural selection is that generalizing from the results of one-locus selection models to multiple loci may be biologically misleading except in limiting cases such as when there is very little recombination and there is no epistasis.

## 7.2 Alternative models of natural selection

- Moving beyond the assumptions of fitness as constant viability in an infinitely growing population.
- Natural selection via different levels of fecundity.
- Natural selection with frequency-dependent fitness.
- Natural selection with density-dependent fitness.

The model of natural selection considered thus far equates fitness with the viability of genotypes. This is equivalent to assuming that while individuals of different genotypes vary in survival to adulthood, all genotypes are equal in terms of any other phenotypes that may impact numbers of progeny an individual contributes to the next generation. Looking again at Fig. 6.3, you can see that there are numerous points in the reproductive life cycle where genotypes may have differential success or performance. Genotypes may differ in phenotypes such as the production and survival of gametes, mating success, gamete genetic compatibility with other gametes, and parental care. It is even possible that some alleles at a locus have an advantage during segregation of homologous chromosomes and are more likely to be found in gametes, a phenomenon called **meiotic drive** (see the historical background on this process in Birchler

et al. 2003). Each of these points in the life cycle is a situation where genotypes will potentially have different levels of performance, eventually leading to different frequencies of genotypes in the progeny. The basic viability model of natural selection also assumes that fitness values are constant through time and space. Instead, it may be that fitness actually changes in response to the conditions found in different populations or in response to the changes in genotype frequency brought on by natural selection. In order to accommodate these potential biological situations, modifications to the model of natural selection are required. This section is devoted to extending the basic viability model of natural selection in a variety of ways to predict how natural selection works for different components of fitness and for changing fitness values. It is not possible to cover all possible models of natural selection exhaustively since there are many. Instead, each of the three models detailed in this section gives some insight into the dynamics of natural selection when one of the major assumptions of the one-locus, two-allele viability model is changed.

### **Natural selection via different levels of fecundity**

Natural selection due to differences in genotype viability is sometimes called **hard selection** since genotype frequency changes come about from the death of individuals and their complete failure to reproduce. In contrast, natural selection due to differences in the fecundity (production of offspring) of individuals with different genotypes causes changes in the frequency of genotypes within the progeny of each generation. Fecundity selection is called **soft selection** because all individuals in the parental generation reproduce, although by differing amounts.

A fecundity model for natural selection on a diallelic locus requires a different approach than was taken for viability selection. A major difference is that fitness depends on the *pair* of genotypes that mate. This means that there are nine different fitness values in a fecundity selection model, as shown in Table 7.4. Another difference is that predicting the genotype frequencies of the progeny is going to be slightly more complicated than for simple random mating. Variation in fecundity may alter the number of progeny produced by each mating pair from the frequency expected by random mating alone. This requires accounting for the expected progeny genotype frequencies that arise from each mating pair weighted by the fecundity of that mating pair as

**Table 7.4** Fitness values based on the fecundities of mating pairs of male and female genotypes for a diallelic locus along with the expected genotype frequencies in the progeny of each possible male and female mating pair weighted by the fecundity of each mating pair. The frequencies of the AA, Aa, and aa genotypes are represented by X, Y, and Z respectively.

Male genotype	Female genotype . . .	Fitness value			Expected progeny genotype frequency
		AA	Aa	aa	
AA		$f_{11}$	$f_{12}$	$f_{13}$	
Aa		$f_{21}$	$f_{23}$	$f_{23}$	
aa		$f_{31}$	$f_{32}$	$f_{33}$	
Parental mating	Fecundity	Total frequency	AA	Aa	aa
AA × AA	$f_{11}$	$X^2$	$X^2$	0	0
AA × Aa	$f_{12}$	$XY$	$1/2XY$	$1/2XY$	0
AA × aa	$f_{13}$	$XZ$	0	$XZ$	0
Aa × AA	$f_{21}$	$YX$	$1/2YX$	$1/2YX$	0
Aa × Aa	$f_{22}$	$Y^2$	$Y^2/4$	$(2Y^2)/4$	$Y^2/4$
Aa × aa	$f_{23}$	$YZ$	0	$1/2YZ$	$1/2YZ$
aa × AA	$f_{31}$	$ZX$	0	$ZX$	0
aa × Aa	$f_{32}$	$ZY$	0	$1/2ZY$	$1/2ZY$
aa × aa	$f_{33}$	$Z^2$	0	0	$Z^2$

shown in Table 7.4. X, Y, and Z are the frequencies of the genotypes AA, Aa, and aa, respectively. This is the same notation used for the proof of Hardy–Weinberg in Chapter 2.

This fecundity selection model is more complex than a viability model because the equations used to solve for genotype frequencies after one generation are functions of genotype frequencies in the parental generation. The mean number of offspring of each genotype after one generation of fecundity selection are found by summing the offspring frequencies (columns in Table 7.4) each weighted by its fecundity. For the progeny with the AA genotype the average fecundity, or  $\bar{f}$ , is

$$\bar{f}X_{t+1} = f_{11}X^2 + f_{12}^{1/2}XY + f_{21}^{1/2}YX + f_{22}\frac{Y^2}{4} \quad (7.15)$$

which simplifies to

$$\bar{f}X_{t+1} = f_{11}X^2 + (f_{12} + f_{21})^{1/2}YX + f_{22}^{1/4}Y^2 \quad (7.16)$$

Using similar steps, the equations for the average fecundities of the Aa and aa genotypes are

$$\begin{aligned} \bar{f}Y_{t+1} &= (f_{12} + f_{21})^{1/2}XY + (f_{13} + f_{31})XZ \\ &\quad + \frac{1}{2}f_{22}Y^2 + (f_{23} + f_{32})^{1/2}YZ \end{aligned} \quad (7.17)$$

and

$$\bar{f}Z_{t+1} = f_{33}Z^2 + (f_{32} + f_{23})^{1/2}YZ + f_{22}^{1/4}Y^2 \quad (7.18)$$

The total average fecundity ( $\bar{f}$ ) is the sum of the average fecundity for each genotype, so that  $\bar{f}X_{t+1}$ ,  $\bar{f}Y_{t+1}$ , and  $\bar{f}Z_{t+1}$  give the proportion of the total number of offspring composed of any genotype after one bout of reproduction. Compare these equations for the average fecundity as functions of genotype frequencies with equations 6.21 and 6.22 that are in terms of allele frequencies. Since random mating does not occur by definition when there is fecundity selection, general equilibrium points cannot be found for arbitrary sets of the nine fecundity values shown in Table 7.4. Rather, the change in genotype frequencies caused by fecundity selection model must be understood by considering special cases of fecundity values.

One special case of fecundity selection occurs when the total fecundity of a mating is always the sum of the fecundity value of each genotype for each sex,

a situation called additive fecundities, analogous to additive gene action (Penrose 1949). Let the fecundity values of females be  $f_{AA}$ ,  $f_{Aa}$ , and  $f_{aa}$  and the fecundity values of males be  $m_{AA}$ ,  $m_{Aa}$ , and  $m_{aa}$ . With an additive fecundity model, the fecundity values given in Table 7.4 would be  $f_{11} = f_{AA} + m_{AA}$  and  $f_{12} = f_{AA} + m_{Aa}$  as two examples. With additive fecundities, higher fecundities for heterozygotes result in both alleles being maintained in a population at equilibrium, as is true for overdominance in the viability selection model. A second special case is when fecundities are multiplicative (Bodmer 1965). For example,  $f_{11} = f_{AA}m_{AA}$  and  $f_{12} = f_{AA}m_{Aa}$ . Depending on fecundity values for the three genotypes, there can be equilibrium points where both alleles are maintained in the population. A third special case that has been examined extensively is when there are four fecundity parameters that correspond to the degree of heterozygosity of each mating pair (Hadeler & Liberman 1975; Feldman et al. 1983). In these cases, depending on the specific fecundity values used, it is also possible that fecundity selection can maintain both alleles in the population, since equilibrium points are reached when all three genotypes have non-zero frequencies. Nevertheless, the fecundity selection model does not result in the maintenance of genetic variation for arbitrary fitness values more often than the basic viability model of selection (Clark & Feldman 1986). This means that fecundity models predict that natural selection frequently results in fixation or loss of alleles at equilibrium, just as the viability model does for directional selection.

Pollak (1978) has shown that mean fecundity does not necessarily increase with fecundity selection. This means that the mean fecundity is not necessarily maximized at equilibrium genotype frequencies for fecundity selection, in contrast to the way natural selection maximizes mean fitness in the viability model for one locus with two alleles.

Hybridization between genetically modified and wild sunflowers provides an example of how a simple fecundity selection model can be used to understand changes in allele frequencies. Using transgenic biotechnology, it is now routine practice to permanently incorporate foreign genes into crop plants. There is the possibility that such transgenes can escape into the wild through the hybridization that occurs between some crop plants and wild relatives that are often weeds (reviewed by Snow & Palma 1997). In the case of sunflowers, mating between pure crop genotypes and wild plants produces hybrids with seed production that is only 2% of that shown by wild plants but hybrids and wild plants have identical survival rates.

Cummings et al. (2002) established three experimental populations with half crop–wild-plant F1 hybrids and half wild plants. In these populations the initial frequency of crop-specific alleles was 25%. The frequencies of the crop-specific alleles at three allozyme loci dropped to about 5% in the next generation. The crop-specific allele frequency in the next generation best matched the allele frequency predicted by a fecundity selection model with additive fecundities.

### **Natural selection with frequency-dependent fitness**

In the basic viability model we considered fitness values as invariant properties of the genotypes. Another way to say this is that a fitness value  $w_{xx}$  is constant regardless of conditions or the frequencies of any of the genotypes. It seems intuitive to expect that the fitness of a genotype may depend on its frequency in a population and there is direct evidence for frequency-dependent fitness in natural populations. For example, mating success of males with different chromosomal inversion genotypes in *Drosophila* depends on chromosome inversion frequencies in the population (see Álvarez-Castro & Alvarez 2005). In plants, the frequency of different flower colors in a population may impact the frequency of visits by pollinators and thereby cause frequency-dependent mating success (e.g. Gigord et al. 2001; Jones & Reithel 2001). A whimsical example of frequency-dependent fitness values comes from the left and right curvature of the mouths of Lake Tanganyikan cichlid fish *Perissodus microlepis* that pluck and eat scales from other fish. There appears to be an advantage to the rarer phenotype, presumably since the fish that are attacked anticipate approach of the cichlid from the side of the more frequent mouth phenotype (Hori 1993).

We can also construct a simple selection model where fitness (as genotype-specific viability) depends on genotype frequency and therefore changes as genotype frequencies change. The key concept in frequency-dependent selection models is creating a measure of fitness that changes. Suppose that the fitness of a genotype decreases as that genotype becomes more common in the population, called **negative frequency dependence**. The relative fitness values are:

$$\begin{aligned} w_{AA} &= 1 - s_{AA}p^2 \\ w_{Aa} &= 1 - s_{Aa}2pq \\ w_{aa} &= 1 - s_{aa}q^2 \end{aligned} \quad (7.19)$$

where  $s_{xx}$  represents the genotype-specific selection coefficient. Genotypes have higher fitness when they are rare since relative fitness decreases as the product of the selection coefficient and the genotype's frequency increases. Note that the selection coefficient itself is a constant and can be thought of as a *per-capita* decrease in relative fitness.

As with other models of selection, the equilibrium points in this model of natural selection can be found by determining when the change in allele frequency ( $\Delta p$ ) is equal to zero. The expression for change in allele frequency over one generation of fecundity selection is

$$\Delta p = \frac{pq s(q-p)(p^2 - pq + q^2)}{\bar{w}} \quad (7.20)$$

for the special case of the selection coefficient being equal for all genotypes, as derived in Math box 7.1. Two equilibrium points occur at fixation and loss

( $p = 1.0$  and  $p = 0.0$ ) since the  $pq$  term in the numerator is zero. There is also an equilibrium point at  $p = 1/2$  since the  $q - p$  term is zero.

Figure 7.5 shows the relative fitness values in equation 7.19 when all selection coefficients are equal to one. It is interesting to note that at  $p = 1/2$  the fitness of the heterozygote is less than that of the two homozygotes, so this model of natural selection does not require overdominance for fitness to maintain genetic variation at equilibrium. However, with independent selection coefficients for each genotype there is not a stable polymorphism in general. With arbitrary fitness values for the three genotypes, there are many possible outcomes, many without stable polymorphism.

#### Natural selection with density-dependent fitness

An assumption in most models of natural selection is that populations are able to grow without any

### Math box 7.1 The change in allele frequency with frequency-dependent selection

Start with the expression for change in allele frequency under viability selection given in equation 6.23:

$$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}} \quad (7.21)$$

and then substitute the definitions of the frequency-dependent fitness values given in equation 7.19. If the selection coefficients for all genotypes are equal so that they can be represented by  $s$  without any subscripts, this gives

$$\Delta p = \frac{pq[p(1 - sp^2) - (1 - s2pq)] + q(1 - s2pq - (1 - sq^2))}{\bar{w}} \quad (7.22)$$

Expanding the terms inside the square brackets gives

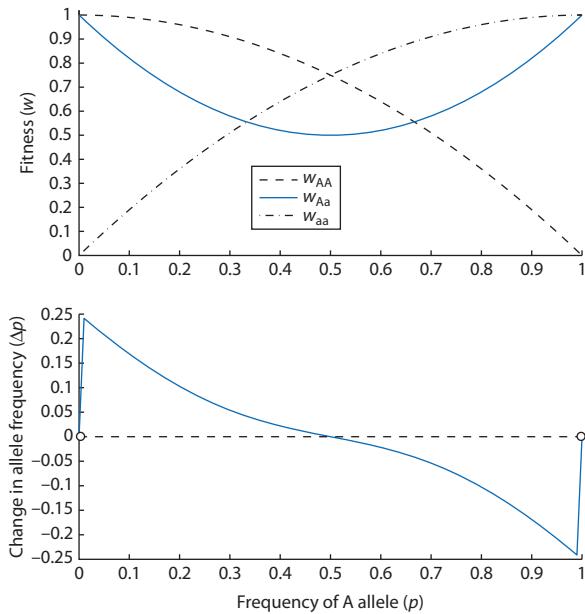
$$\Delta p = \frac{pq[p - sp^3 - p + s2p^2q + q - s2pq^2 - q + sq^3]}{\bar{w}} \quad (7.23)$$

which simplifies by canceling the positive and negative  $p$  and  $q$  terms and factoring out an  $s$  to give

$$\Delta p = \frac{pq s[-p^3 + 2p^2q - 2pq^2 + q^3]}{\bar{w}} \quad (7.24)$$

The term in square brackets can then be factored to give

$$\Delta p = \frac{pq s(q-p)(p^2 - pq + q^2)}{\bar{w}} \quad (7.25)$$



**Figure 7.5** The relative fitness of each genotype ( $w_{xx}$ ) and the change in allele frequency ( $\Delta p$ ) across all frequencies of the A allele under frequency-dependent natural selection. There is a stable equilibrium point at  $p = 0.5$  in this particular case, even though the heterozygote has the lowest fitness. Two unstable equilibria at fixation and loss are marked with open circles. Here the relative fitness values are  $w_{AA} = 1 - s_{AA}p^2$ ,  $w_{Aa} = 1 - s_{Aa}2pq$ , and  $w_{aa} = 1 - s_{aa}q^2$  with  $s_{AA} = s_{Aa} = s_{aa}$ .

limits. The first section of this chapter developed a natural selection model where the size of the population of any genotype one generation in the future was its population size currently multiplied by a constant (refer back to equation 6.1). This model is obviously unrealistic because no organism can grow without some eventual limits on population size. Organisms are limited by the space and resources available to them, limitations that lead to changes in the rate of growth as the density of individuals changes over time. To incorporate such limits in a model of natural selection, we can alter our basic genotype-specific population growth equations to incorporate an upper bound on the population size as well as a rate of population growth that changes with population size.

A simple model where population growth has an upper bound is called **logistic growth** and the upper limit is called the **carrying capacity** (symbolized by  $K$ ). Logistic growth depends on feedback between the growth rate and the size of the population according to

$$\lambda = 1 + r - \frac{r}{K}N \quad (7.26)$$

### Interact box 7.2 Frequency-dependent natural selection

Launch PopGeneS<sup>2</sup> and in the **Selection** menu, choose **Frequency dependent selection**. In the model dialog you can set the frequency-sensitive relative fitness values for the three genotypes produced by one locus with two alleles. The  $s_1$ ,  $s_2$ , and  $s_3$  values correspond to  $s_{AA}$ ,  $s_{Aa}$ , and  $s_{aa}$  in equation 7.19. First try values of  $s_1 = 0.3$ ,  $s_2 = 1.0$ , and  $s_3 = 0.3$ . Explain why the graph of genotype-specific fitness by allele frequency (the middle panel) looks the way it does.

Next enter selection coefficients of  $s_1 = 0.7$ ,  $s_2 = 1.0$ , and  $s_3 = 0.2$ . Then compare the  $p$  by  $\Delta p$  graph in the top panel with the  $p$  by mean fitness graph in the bottom panel to see the mean fitness at equilibrium allele frequency. Does natural selection always cause allele frequencies to reach an equilibrium that corresponds to the maximum value of the mean fitness? Through educated guesses, try to find selection coefficients where equilibrium allele frequencies do and do not correspond to the maximum mean fitness value.

where  $N$  is the population size and  $r$  is the rate of increase ( $\lambda$  used as the growth rate earlier in Chapter 6 can be equated to  $r$  by  $\lambda = 1 + r$ ). Biologically,  $r$  represents the rate of growth in excess of the replacement rate of one. When  $N = 0$  then  $\frac{r}{K}N$  is 0 and the growth rate is at its maximum. But when  $N = K$  then  $\frac{r}{K}N$  equals  $r$  and the population replaces itself but does not change in size.

Logistic population growth can be applied to the three genotypes at a diallelic locus by defining genotype-specific carrying capacities and rates of increase to obtain absolute fitness values for each genotype:

$$\lambda_{AA} = 1 + r_{AA} - \frac{r_{AA}}{K_{AA}}N \quad (7.27)$$

$$\lambda_{Aa} = 1 + r_{Aa} - \frac{r_{Aa}}{K_{Aa}}N \quad (7.28)$$

and

$$\lambda_{aa} = 1 + r_{aa} - \frac{r_{aa}}{K_{aa}}N \quad (7.29)$$

Here the sum of the numbers of individuals with each genotype is equal to the total size of the population ( $N_{AA} + N_{Aa} + N_{aa} = N$ ). The average absolute fitness in the population is the average of the  $r$  and  $r/K$  values for each genotype:

$$\bar{\lambda} = 1 + \bar{r} - \frac{\bar{r}}{K}N \quad (7.30)$$

where  $\bar{r} = p_t^2 r_{AA} + 2p_t q_t r_{Aa} + q_t^2 r_{aa}$  and  $\frac{\bar{r}}{K} = \frac{p_t^2 r_{AA}}{K_{AA}} + \frac{2p_t q_t r_{Aa}}{K_{Aa}} + \frac{q_t^2 r_{aa}}{K_{aa}}$ .

The final step is to use these results to modify our previous results for unbounded population growth to take logistic growth into account. First we can modify the growth in the total size of the population (equation 6.1 for unbounded growth) so that

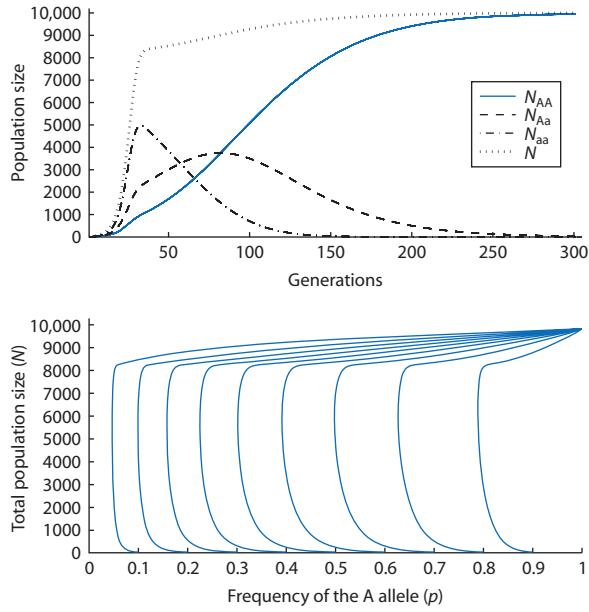
$$N_{t+1} + \bar{\lambda}N_t \quad (7.31)$$

This equation predicts that the total population size cannot exceed the largest carrying capacity. We can also follow allele frequencies over time by modifying the expression for allele frequency after one generation of selection:

$$p_{t+1} = \frac{\lambda_{AA} p_t^2 + \lambda_{Aa} p_t q_t}{\bar{\lambda}} \quad (7.32)$$

which is the logistic growth version of equation 6.19.

Both the numbers of individuals of each genotype and the allele frequency under density-dependent natural selection can be seen in Fig. 7.6 when the AA genotype has the highest carrying capacity. Starting out with a very small  $N$ , the numbers of individuals of all genotypes increase over time. However, once  $N$  approaches the lowest carrying capacity,  $K_{aa}$  in this illustration, the number of aa individuals peaks and then declines. This happens because the absolute fitness of aa approaches one in the fewest generations while the other two genotypes continue to add individuals to their populations because their carrying capacities are higher. The same phenomenon also occurs to the Aa genotype because it has the next



**Figure 7.6** The results of density-dependent natural selection on the numbers of individuals of different genotypes ( $N_{AA}$ ,  $N_{Aa}$ , and  $N_{aa}$ ) and allele frequencies in a population of total size  $N$ . At the upper limit of  $N$ , the equilibrium allele and genotype frequencies are determined by the genotype with the highest carrying capacity ( $K$ ). In contrast, the genotype with the highest growth rate ( $r$ ) has the greatest impact on allele frequency when the population is small. In this example  $K_{AA} = 10,000$ ,  $K_{Aa} = 9000$ , and  $K_{aa} = 8000$  with  $r_{AA} = 0.2$ ,  $r_{Aa} = 0.25$ , and  $r_{aa} = 0.3$ .

lowest carrying capacity. The AA genotype has the highest carrying capacity and the number of individuals of that genotype eventually grows to the point where it makes up the entire population.

The general result for density-dependent selection is that the carrying capacities of the three genotypes will determine the eventual genotype and allele frequencies when populations approach their carrying capacities. When  $K_{AA}$  is the highest then the A allele goes to fixation and when  $K_{aa}$  is the largest then the a allele goes to fixation. Alternatively, when  $K_{Aa}$  is the largest then there is an equilibrium with both alleles segregating and when  $K_{Aa}$  is the smallest then either A or a will reach fixation depending on initial allele frequencies. These results are qualitatively identical to those for unbounded growth.

In contrast, the results of density-dependent and density-independent natural selection do not agree when population size is restricted to low numbers. The total population size  $N$  may be much less than the carrying capacity in highly disturbed or inhospitable environments where individuals have low reproductive output or high turnover. To see this, consider equations 7.27–7.29 for genotype absolute

### Interact box 7.3 Density-dependent natural selection

Launch Populus. In the **Model** menu, choose **Natural Selection** and then **Density-Dependent Selection w/ Genetic Variation**. In the options dialog you can set the genotype-specific carrying capacity and growth rates. Click on the radio button for **Nine-Frequency** to display results for nine initial allele frequencies (the **Single Frequency** button shows the results in terms of the total population size  $N$ ). The  $N$  text box sets the initial population size. Press the **View** button to see the simulation results.

Parameter values to simulate:

$K_{AA} = 8000$ ,  $K_{Aa} = 8000$ , and  $K_{aa} = 10,000$ ;  $r_{AA} = 0.4$ ,  $r_{Aa} = 0.4$ , and  $r_{aa} = 0.3$ ; generations = 100  
 $K_{AA} = 8000$ ,  $K_{Aa} = 10,000$ , and  $K_{aa} = 8000$ ;  $r_{AA} = 0.4$ ,  $r_{Aa} = 0.3$ , and  $r_{aa} = 0.35$ ; generations = 100  
 $K_{AA} = 8000$ ,  $K_{Aa} = 6000$ , and  $K_{aa} = 9000$ ;  $r_{AA} = 0.5$ ,  $r_{Aa} = 0.3$ , and  $r_{aa} = 0.4$ ; generations = 100

fitnesses. As  $N$  for each genotype approaches zero the  $\frac{r}{K}N$  term will also approach zero, leaving each absolute fitness increasingly determined by the genotype-specific growth rate. The genotype with the highest growth rate should then increase fastest and dictate genotype and allele frequencies at equilibrium (if  $r_{AA}$  is highest A fixes, if  $r_{aa}$  is highest a fixes, if  $r_{Aa}$  is highest both alleles segregate, and if  $r_{Aa}$  is lowest then initial allele frequencies dictate fixation of either A or a). This effect can be seen in Fig. 7.6 where allele frequencies change toward a lower frequency of  $p$  when the population is small. Despite the fact that the population is expected to fix for  $p$  at carrying capacity, the growth rate of the aa genotype is the greatest and so it has the greatest impact on allele frequencies when the population is small.

### 7.3 Combining natural selection with other processes

- Natural selection and genetic drift acting simultaneously.
- The balance between natural selection and mutation.

Natural selection takes place at the same time that other processes are also operating and having an impact on allele frequencies. These other processes may work in concert with natural selection and work toward the same equilibrium allele frequencies, or they may work against natural selection toward alternative equilibrium allele frequencies. Since many population genetic processes are likely to be acting simultaneously in actual biological populations, it is

important to put natural selection into the context of other processes that impact allele frequencies. This section first considers allele frequencies when natural selection and genetic drift are in opposition and then natural selection and mutation acting in opposition. When natural selection and other processes act in concert, this simply shortens the number of generations required to reach equilibrium and does not alter equilibrium allele frequencies.

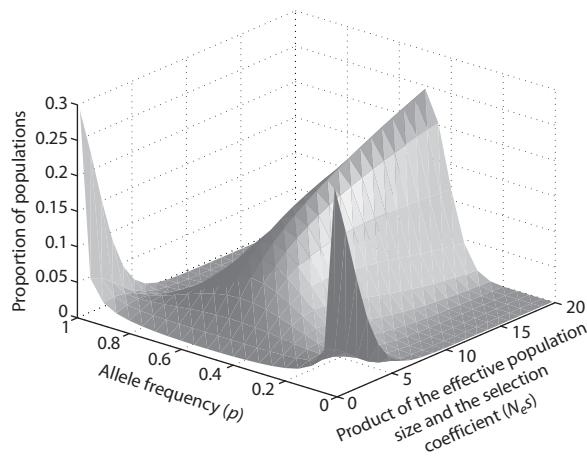
#### **Natural selection and genetic drift acting simultaneously**

Wright (1931) showed the probability that a population has a given allele frequency when exposed to the simultaneous processes of natural selection, genetic drift, and mutation as given by

$$\phi(p) = Cp^{(4N_e\mu-1)}q^{(4N_e\nu-1)}e^{(4N_e spq)} \quad (7.33)$$

where  $\phi$  (pronounced “phi”) means a probability density,  $p$  and  $q$  are the allele frequencies,  $N_e$  is the effective population size,  $\mu$  and  $\nu$  are the forward and backward mutation rates,  $s$  is the selection coefficient, and  $C$  is a constant used to adjust the total probability across all allele frequencies to sum to 1.0 for each value of  $N_e s$ . This equation is a probability density function for an ensemble population like those discussed in Chapter 3 for genetic drift. It describes the chance that one of many replicate populations will reach any allele frequency between zero and one at equilibrium given values of the effective population size and the selection coefficient.

This equation is most easily understood by examining its predictions in graphical form (Fig. 7.7). When  $N_e s$  is near zero, either genetic drift is very



**Figure 7.7** The expected distribution of allele frequencies for a very large number of replicate finite populations under natural selection where there is overdominance for fitness ( $w_{AA} = w_{aa} = 1 - s$  and  $w_{Aa} = 1$ ). In an infinite population the expected allele frequency at equilibrium is 0.5. However, in finite populations the equilibrium allele frequency will depend on the balance of natural selection and genetic drift. This balance is determined by the product of the effective population size and the selection coefficient ( $N_e s$ ). Low values of  $N_e s$  mean that selection is very weak compared to drift and each population reaches fixation or loss. High values of  $N_e s$  mean that selection is strong compared to drift and most populations reach an equilibrium allele frequency near 0.5. Here forward and backward mutation rates are equal ( $\mu = v = 0.00001$ ).

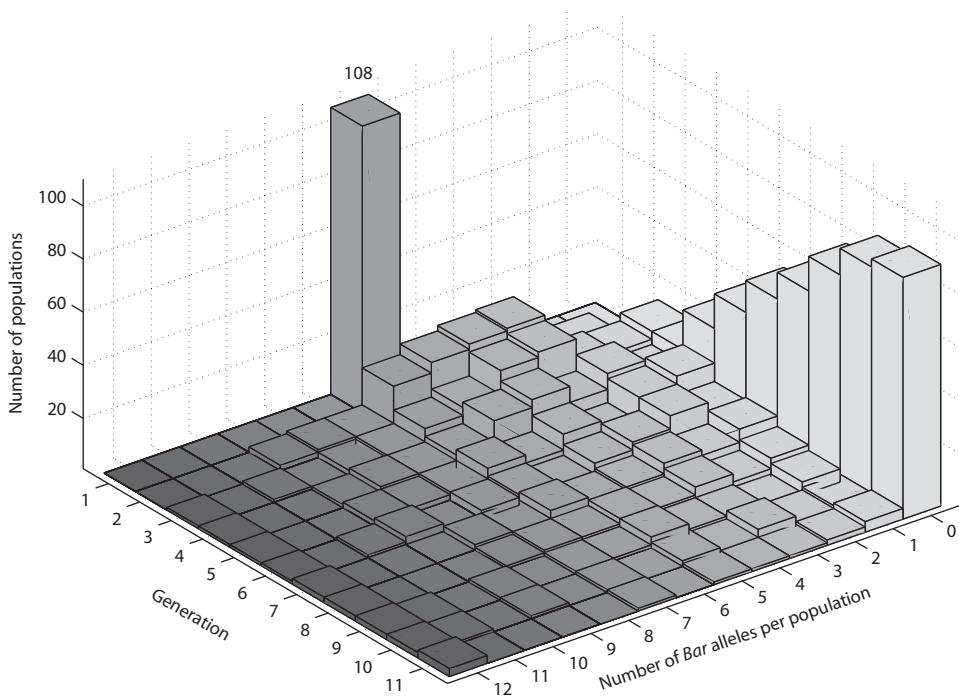
strong because the population is tiny or the selection coefficient is extremely small and so the populations are evolving in a neutral fashion under drift alone. In either of these cases genetic drift is the dominant process and will eventually result in either fixation or loss in all populations. At the lowest values of  $N_e s$  in Fig. 7.7, populations are most likely to have allele frequencies near zero or near one as expected under genetic drift alone. Alternatively,  $N_e s$  can take on large values in two general situations. One is when genetic drift is very weak because the effective population size is very large and there is some natural selection favoring heterozygotes ( $s$  can occupy a wide range as long as it is not extremely small). The other is when the selection coefficient is large and the effective population size is at least 10 or so individuals and therefore genetic drift is not extreme. At the largest values of  $N_e s$  in Fig. 7.7, the probability is greatest that the allele frequency in a population will be near 0.5. Equation 6.35 shows that 0.5 is the expected equilibrium allele frequency under balancing selection when  $w_{AA} = w_{aa}$  and a population is infinite. Therefore, when  $N_e s$  is large, selection is stronger than drift and the equilibrium allele frequency is deter-

mined mostly by natural selection. At intermediate values of  $N_e s$ , the equilibrium allele frequency for many populations is between the equilibrium allele frequencies expected under genetic drift acting alone or natural selection acting alone.

To summarize the balance of natural selection and genetic drift, Motoo Kimura suggested a simple rule of thumb for a diploid locus (Kimura 1983a, 1983b). If four times the product of the effective population size and the selection coefficient is much less than one ( $4N_e s \ll 1$ ) then selection is weak relative to sampling, and genetic drift will dictate allele frequencies. Alternatively, if four times the product of the effective population size and the selection coefficient is much greater than one ( $4N_e s \gg 1$ ) then selection is strong relative to sampling, and natural selection will dictate allele frequencies. When four times the product of the effective population size and the selection coefficient is approximately one ( $4N_e s \approx 1$ ) then allele frequencies are unpredictable.

Figure 7.8 shows an example of the balance between genetic drift and natural selection in replicated laboratory populations of the fruit fly *Drosophila melanogaster* (Wright & Kerr 1954). The plot shows allele frequencies at the *Bar* locus for 108 replicate populations that were each founded from four males and four females every generation (since *Bar* is hemizygous in males the effective population size is equivalent to six rather than eight diploid individuals). Although the vast majority of populations fix for the wild-type allele due to strong natural selection favoring the recessive phenotype of wild-type homozygotes, three populations fix for the *Bar* allele by the end of the experiment. The selection coefficient against the *Bar* homozygote was estimated at 0.63, giving an upper bound estimate of  $4N_e s \approx 15$  in the experiment ( $N_e$  may have been less than six in actuality). Even with this value of  $4N_e s$ , natural selection is not sufficiently strong to dictate equilibrium allele frequency in all populations.

It is also possible to gain biological insight into  $N_e s$  by recognizing that it is analogous to the quantity  $N_e m$  that dictates the balance between genetic drift and gene flow discussed in Chapter 4. Both  $N_e s$  and  $N_e m$  represent the net balance of the pressure on allele frequencies toward eventual fixation or loss due to genetic drift and the countervailing force driving allele frequencies toward a specific allele frequency caused by either natural selection or by gene flow. In the case of natural selection the specific allele frequency is dictated by the relative fitness values of genotypes while in the case of gene flow



**Figure 7.8** Frequency of the *Bar* allele in 108 replicate *D. melanogaster* populations over 10 generations (Wright & Kerr 1954). Each population was founded from four males and four females. The *Bar* locus is found on the X chromosome and so is hemizygous in males, making the effective population size equivalent to about six individuals. The eyes of *D. melanogaster* individuals homozygous for the wild-type allele are oval, but heterozygotes and homozygotes for the partially dominant *Bar* allele have bar-shaped eyes with a reduced number of facets. Females homozygous for the *Bar* allele produced 37% of the progeny compared to females homozygous or heterozygous for the wild-type allele. Despite this strong natural selection against *Bar*, three populations fixed for *Bar* by the end of the experiment. Compare with the similar example in Figure 3.11 where the locus is selectively neutral.

#### Interact box 7.4 The balance of natural selection and genetic drift at a diallelic locus

You can use PopGene.S<sup>2</sup> to simulate the simultaneous action of genetic drift and natural selection in many identical finite populations. From the **Drift** menu select the **Drift + Selection + Mutation**. In the module dialog, run the model a few times using the default parameters to understand the output. The graph on the left shows the allele frequency trajectories of each replicate population over time while the histogram on the right shows the distribution of allele frequencies for all populations (analogous to a two-dimensional slice through Fig. 7.7 for a single value of  $N_e s$ ). Note that mutation will have almost no effect on the outcome of allele frequencies as long as backward and forward mutation rates are equal and very small (you can test this by setting one very high mutation rate such as 0.1 to see the impact).

Try two sets of simulations, one keeping the selection coefficient constant and varying  $N_e$ , and another for comparison keeping  $N_e$  constant but varying the selection coefficient. Before each of these runs you should compute the value of  $4N_e s$  and make a prediction about the distribution of allele frequencies among the populations. First set  $w_{AA} = w_{aa} = 0.9$  and  $w_{Aa} = 1.0$  (or  $s = 0.1$ ) and then run separate simulations for  $N_e = 2, 20$ , and  $200$ . Then set  $N_e = 10$  and use selection coefficients of 0.05, 0.5, and 0.95.

the specific allele frequency is the average allele frequency for all demes.

### The balance between natural selection and mutation

Natural selection takes place at the same time that mutation is working to alter allele frequencies and reintroduce alleles that selection may be driving to loss. Therefore, the process of natural selection may be counteracted to some degree by mutation. If a completely recessive allele is both deleterious when homozygous and also produced by spontaneous mutation, there are contrasting forces acting on its frequency. Mutation pressure will continually reintroduce the allele into a population while natural selection will continually work to drive the allele to loss. What equilibrium allele frequency is expected when the opposing processes of mutation and natural selection balance out?

Let's assume that there are two alleles at one locus and that the *a* allele is completely recessive and has a frequency of  $q$ . Also assume the case of selection against a recessive as given in Table 6.4. An expression for the change in allele frequency per generation for the specific case of natural selection against a recessive homozygote can be obtained by substituting the fitness values of  $w_{AA} = w_{Aa} = 1$  and  $w_{aa} = 1 - s$  into equation 6.24:

$$\Delta q_{selection} = \frac{pq[q((1-s)-1) + p(1-1)]}{(1)p^2 + (1)2pq + (1-s)q^2} \quad (7.34)$$

which then rearranges to

$$\Delta q_{selection} = \frac{-spq^2}{1-sq^2} \quad (7.35)$$

Let's further assume that mutation is irreversible and that the probability that an *A* allele mutates to an *a* allele is  $\mu$ . The change in the frequency of the allele due to mutation each generation is then

$$\Delta q_{mutation} = \mu p \quad (7.36)$$

At equilibrium, the action of natural selection pushing the allele toward fixation and the pressure of mutation increasing the allele frequency exactly balance so that allele frequency does not change. This means that, at equilibrium,

$$\Delta q_{mutation} + \Delta q_{selection} = 0 \quad (7.37)$$

Substituting in the expressions for  $\Delta q_{mutation}$  and  $\Delta q_{selection}$  into this equation yields

$$\mu p = \frac{spq^2}{1-sq^2} \quad (7.38)$$

If we assume that the frequency  $q$  of the *a* allele is low, then  $q^2$  is very small and the quantity  $1 - sq^2$  is approximately one. This approximation leads to

$$\mu p = spq^2 \quad (7.39)$$

which can be solved in terms of genotype frequency:

$$q^2 = \frac{\mu}{s} \quad (7.40)$$

or in terms of allele frequency:

$$q_{equilibrium} = \left( \frac{\mu}{s} \right)^{1/2} \quad (7.41)$$

Thus, the expected frequency of a deleterious recessive allele at an equilibrium between natural selection and mutation depends on the ratio of the mutation rate and the selection coefficient. Equation 7.41 shows that even if a recessive homozygous genotype is lethal ( $s = 1$ ), the expected frequency of the allele is  $\mu^{0.5}$  and the expected frequency of the lethal genotype is  $\mu$  due to recurrent mutation.

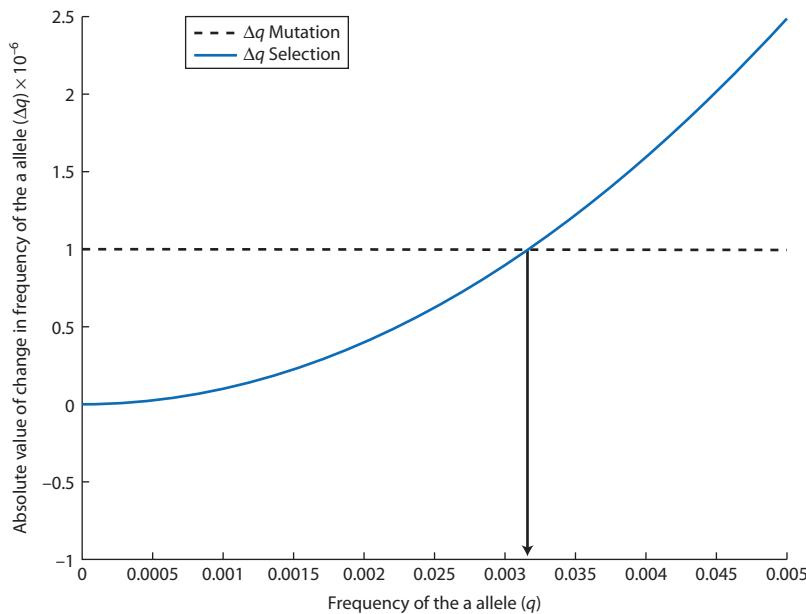
The balance between selection and mutation is illustrated in Fig. 7.9. The  $\Delta q$  due to mutation is always positive while the  $\Delta q$  due to selection in this case is negative. Figure 7.9 uses the absolute value of each  $\Delta q$  to show where the change in allele frequency for the two processes intersects. This intersection is the equilibrium point. The expected equilibrium allele

frequency is  $q_{equilibrium} = \left( \frac{1 \times 10^{-6}}{0.1} \right)^{1/2} = 0.0032$ , in

agreement with the figure.

Consanguineous mating results in an excess of homozygosity and a deficit of heterozygosity compared to random mating. If  $f$  is the degree of departure from Hardy-Weinberg expectations (see equation 2.20) then the equilibrium allele frequency from equation 7.40 can be restated as

$$q^2 + fpq = \frac{\mu}{s} \quad (7.42)$$



**Figure 7.9** The absolute value of the change in allele frequency due to mutation ( $\Delta q_{\text{mutation}}$ ) and due to natural selection ( $\Delta q_{\text{selection}}$ ) when there is selection against a recessive homozygote. Mutation continually makes new copies of the recessive allele while selection continually works toward loss of the recessive allele. The equilibrium allele frequency occurs when the processes of mutation and selection exactly counteract each other. Here  $s = 0.1$  and  $\mu = 1 \times 10^{-6}$  so the expected equilibrium is  $q_{\text{equilibrium}} = 0.0032$  as shown by the vertical arrow.

Under the assumption that  $q$  is small compared to  $f$ , the approximate equilibrium allele frequency when mutation and selection reach a balance with consanguineous mating is

$$q_{\text{equilibrium}} = \frac{\mu}{fs} \quad (7.43)$$

### Interact box 7.5 Natural selection and mutation

You can use PopGene.S<sup>2</sup> to simulate the simultaneous action of natural selection and mutation. From the **Drift** menu select the **Drift + Selection + Mutation**. In the module dialog set  $N_e = 20$ ,  $w_{AA} = w_{Aa} = 1.0$ , and  $w_{aa} = 0.9$  (or  $s = 0.1$  with complete dominance for the A allele) and  $\mu = 1 \times 10^{-3}$ . Compute the expected equilibrium for these parameter values using equation 7.41 and then run the simulation.

Genetic drift will have a minimal effect on the outcome of allele frequencies as long as the effective population size is large. You can test this prediction by trying simulations with  $N_e = 10, 100, 500$ , and 1000 and comparing the equilibrium allele frequencies for each case.

(see Haldane 1940; Morton 1971). Since recessive deleterious mutations are only perceived by natural selection when homozygous, consanguineous mating increases the effectiveness of selection by increasing the proportion of homozygous genotypes in the population. This means that selection is more effective at eliminating the recessive homozygote (there are fewer heterozygotes that shelter the allele) and the equilibrium allele frequency for mutation–selection balance occurs at a lower allele frequency. It is counterintuitive that populations which cease consanguineous mating and engage in random mating may temporarily experience an increase in deleterious allele frequencies and a decrease in average fitness due to less effective natural selection.

### 7.4 Natural selection in genealogical branching models

- The problem with selection in genealogical branching models.
- Directional selection and the ancestral selection graph.
- Genealogies and balancing selection.

The final topic in this chapter is the process of natural selection in the context of genealogical branching models. Representing natural selection in genealogical branching models will require a change in perspective about how selection works and also an expansion of the ways that events are represented on genealogies. The major goal of this section is to

introduce ways of modeling selection on genealogies to understand how the operation of natural selection might change the height and total branch length of genealogical trees compared with the case of coalescence patterns due to genetic drift alone.

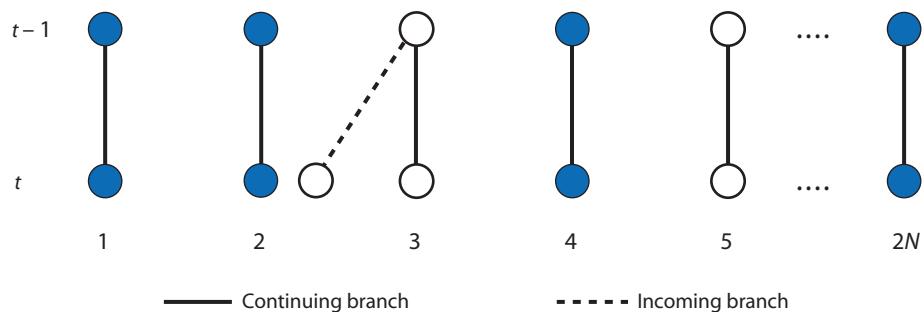
Adding natural selection to the genealogical branching model introduces a serious complication to the Wright–Fisher model of sampling that the basic genealogical branching model is built on. Recall from Chapter 3 that the basic coalescent model assumes that when going one generation back in time, the chance that any two lineages coalesce is  $\frac{1}{2N}$ . This probability results from the assumption that all lineages in any given generation have an equal chance of being chosen as a common ancestor working back in time from the present. In the basic coalescent model where alleles are selectively neutral, each lineage within a generation has an *equal and constant* probability of becoming an ancestral lineage when working back in time to find the most recent common ancestor (MRCA). From the time-forward perspective, with neutral evolution each lineage has an equal probability of being sampled and represented in the next generation. In general, with selective neutrality the haplotype of a lineage does not influence its sampling properties.

Natural selection violates the basic assumption that all lineages have equal and constant probabilities of coalescence. When natural selection operates, some lineages tend to increase in frequency over time whereas other lineages tend to decrease in frequency over time due to fitness differences among the lineages caused by differences in haplotype relative fitnesses. These changes in the frequencies of lineage

copies translate into probabilities of coalescence that change over time as well. A lineage bearing a haplotype favored by selection will increase over time and therefore have a *decreasing* probability of coalescence working back in time from the present. Similarly, a lineage bearing a lower-fitness haplotype will decrease in frequency over time and therefore have an *increasing* probability of coalescence moving back in time. Thus, natural selection presents a fundamental contradiction to the sampling process built into the genealogical branching model.

### **Directional selection and the ancestral selection graph**

Fortunately, there is a clever and relatively simple way to modify the genealogical branching model to accommodate directional natural selection (Neuhauser & Krone 1997; Neuhauser 1999). This modification relies on treating coalescence and natural selection as distinct processes that can both possibly occur working back in time from the present toward the MRCA in the past exactly as was done to combine coalescence with migration or mutation. The first step in including natural selection in the genealogical branching model is to alter slightly our view of the sampling process. Figure 7.10 shows five lineages of  $2N$  total across the span of one generation. If there is selective neutrality, then moving forward in time each lineage is sampled once to found the next generation. This is equivalent to the absence of a coalescence event moving backward in time. Alternatively, if there is natural selection operating then lineages with one haplotype are favored and will increase in frequency over time.



**Figure 7.10** Haplod reproduction with the possibility of coalescence and natural selection events. Each haploid lineage replicates itself and is included in the next generation if there are no coalescence events (solid lines). A lineage makes an extra copy of itself (dashed line) and has the potential to displace one copy of another lineage. If the lineage making the extra copy of itself (open circle) has a higher-fitness haplotype than a randomly chosen lineage (blue circle), then it will displace the lineage of the lower fitness haplotype. Therefore, the outcome of a lineage-duplication event that may result in natural selection depends on the haplotype states of the specific lineages involved. The solid lines are continuing branches and the dashed line is an incoming branch. Compare with Figures 3.23 and 5.12.

In Fig. 7.10, the possible action of natural selection is shown by the dotted line. If the lineage represented by the open circles has a higher-fitness haplotype, then it will displace the lineage of the lower-fitness haplotype (closed circles). This displacement event is analogous to growth in the population size of the fitter haplotype given that the total population size is constant.

We can treat the dual continuing/incoming branching process as two independent parts of the overall coalescence process. When two independent processes are operating, the coalescence model is based on waiting for *any* event to occur and then deciding which type of event happened. When events are independent but mutually exclusive, the probability of each event is added over all possible events to obtain the total chance that an event occurs. As was done for both migration and mutation, we assume that coalescent and natural selection events are rare or that  $N_e$  is large and the selection coefficient is small. This assumption makes sure that natural selection and coalescence events are mutually exclusive and that when an event does occur going back in time it is *either* coalescence *or* natural selection.

Natural selection depends on the chance that the fitter haplotype makes an incoming branch that displaces a lineage bearing a less-fit haplotype. Twice the rate of natural selection events is

$$\sigma = 4N_e s \quad (7.44)$$

( $\sigma$  is pronounced “sigma”) where  $2N_e s$  is the expected number of natural selection events that will occur for a single lineage during one unit of continuous time (see section 5.5 for a fuller explanation of such a rate in the context of mutation rates) and the fitter haplotype has a relative fitness of  $1 + s$  compared to a relative fitness of 1 for the less-fit haplotype. When  $s$  is of the order of  $\frac{1}{2N}$  then the rate of natural selection events would be comparable to the rate of coalescent events due to neutral sampling. The exponential approximation for the chance that a natural selection event occurs at generation  $t$  is then

$$P(T_{\text{incoming branch}} = t) = e^{-t\frac{\sigma}{2}} \quad (7.45)$$

for a single lineage and

$$P(T_{\text{incoming branch}} = t) = e^{-t\frac{\sigma_k}{2}} \quad (7.46)$$

for  $k$  lineages on a continuous time scale. The chance that an incoming branch due to natural selection displaces one of  $k$  lineages at or before a certain time can then be approximated with the cumulative exponential distribution,

$$P(T_{\text{incoming branch}} \leq t) = 1 - e^{-t\frac{\sigma_k}{2}} \quad (7.47)$$

in exactly the same fashion that times to coalescent events are approximated.

Combining natural selection with coalescence and mutation into a three-process waiting-time distribution is now simple. Since the three types of events are mutually exclusive, we add the chance that a natural selection event occurs to the chance that a coalescence or mutation event occurs to get the expected waiting time until an event:

$$P(T_{\text{event}} \leq t) = 1 - e^{-t\left(\frac{k(k-1)}{2} + \frac{\sigma}{2}k + \frac{\theta}{2}\right)} \quad (7.48)$$

When an event does occur according to the waiting time in equation 7.48, it is then necessary to determine the type of event. Since the total chance that an

event occurs is  $\frac{k(k-1)}{2} + \frac{k\sigma}{2} + \frac{k\theta}{2}$ , the chance that an event is a coalescence is

$$\frac{\frac{k(k-1)}{2}}{\frac{k\sigma}{2} + \frac{k(k-1)}{2} + \frac{k\theta}{2}} = \frac{k-1}{k-1+\sigma+\theta} \quad (7.49)$$

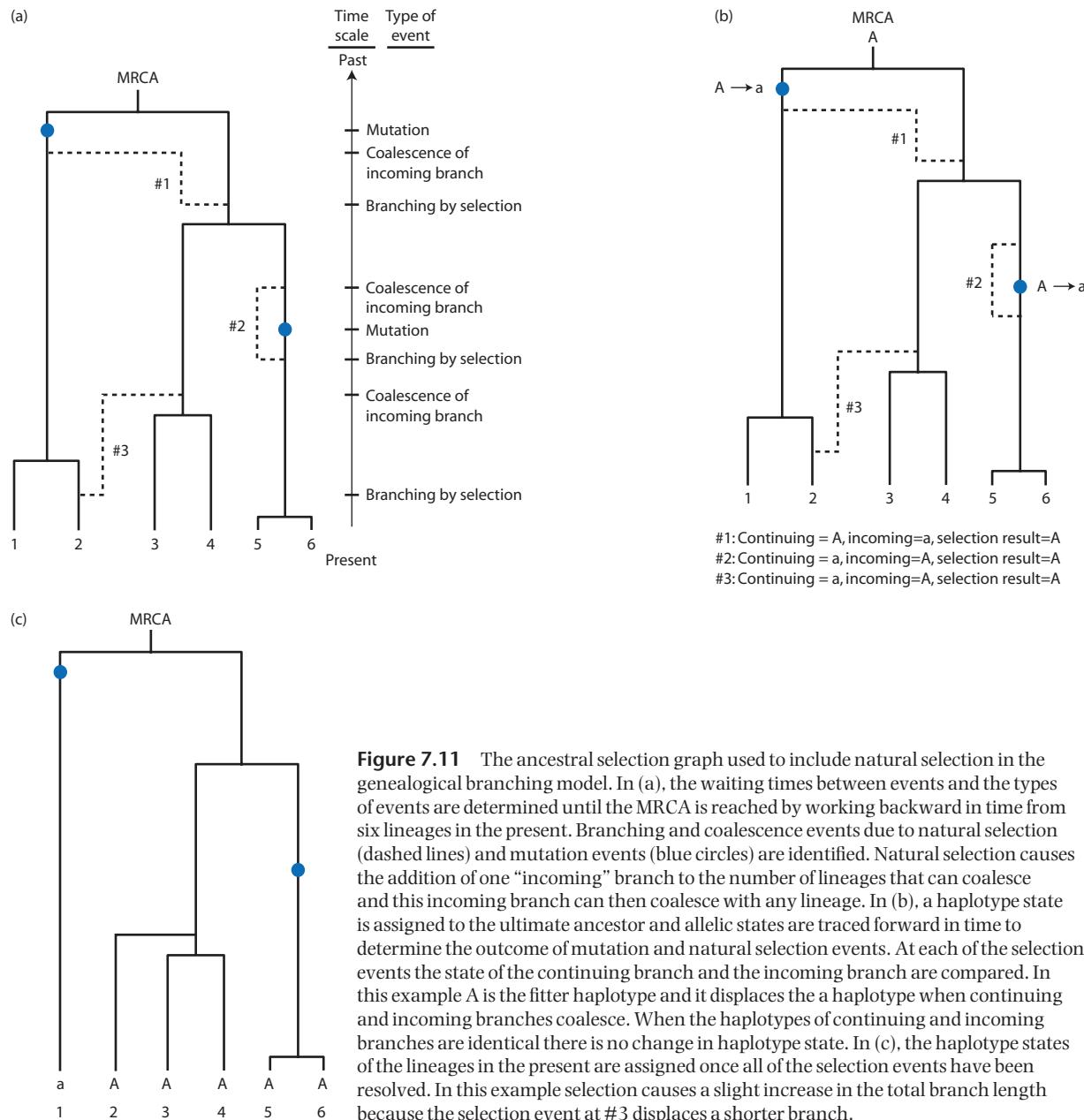
while the chance that an event is due to natural selection is

$$\frac{\frac{k\sigma}{2}}{\frac{k\sigma}{2} + \frac{k(k-1)}{2} + \frac{k\theta}{2}} = \frac{\sigma}{k-1+\sigma+\theta} \quad (7.50)$$

and the chance that the event is a mutation is

$$\frac{\frac{k\theta}{2}}{\frac{k\sigma}{2} + \frac{k(k-1)}{2} + \frac{k\theta}{2}} = \frac{\theta}{k-1+\sigma+\theta} \quad (7.51)$$

Using equation 7.48 and then determining whether each event is selection, mutation, or coalescence, it is possible to construct what is known as an **ancestral selection graph** (Fig. 7.11). The term ancestral



selection graph is used to describe the outcome of the three processes since it explicitly shows possible natural selection events. Natural selection events result in an *addition* of branches and thereby serve to visualize selection events that are not apparent on a genealogy alone. When branching occurs due to a natural selection event (going back in time), the resulting branch is called the **incoming branch** to represent a possible lineage displacement. The lineage that the incoming branch splits off from is called the **continuing branch**. The incoming branch coalesces with a randomly chosen lineage at a later

time determined by the waiting time distribution and assumes the state of the branch where it coalesces. Refer again to Fig. 7.10 to see the distinction between incoming and continuing branches. Even though natural selection events make more branches, the coalescence process is faster and will eventually result in coalescence to the MRCA (the coalescence rate is proportional to  $k^2$  whereas the selection rate is proportional to  $k$ ).

Figure 7.11 shows one outcome of the coalescence–natural-selection–mutation process. Figure 7.11a shows the events that occurred working back in time

from six lineages in the present. The first potential natural selection event (labeled #3 because it is the last event when working forward in time) occurred on lineage 2, causing a branching event. The incoming branch eventually coalesced with the lineage that is ancestral to lineages 3 and 4. The second potential natural selection event caused a branching event on the lineage that is ancestral to lineages 5 and 6. That incoming branch coalesced again with the same lineage after a short time. The final potential natural selection event caused a branch from one of the two internal lineages near the MRCA that coalesced with the other lineage present near the MRCA.

The actual outcome of these three selection events can only be determined once the state of the MRCA and the fitness of the two haplotypes have been assigned. In Fig. 7.11b, the ancestor is assigned a state of A which is also assumed to be the higher fitness haplotype. Then moving forward in time on the ancestral selection graph, the outcome of each instance of natural selection can be determined. For selection event #1, the incoming branch has a haplotype of a due to the mutation while the continuing branch has the ancestral A haplotype. Since A is more fit, the state of the continuing branch is kept and the state of the incoming branch discarded. (This is exactly like the open circles being less fit and the closed circles more fit in Fig. 7.10.) At the second selection event (#2), the incoming branch has a state of A and the continuing branch has a state of a due to a mutation. Here the incoming branch displaces the continuing branch and the lineage has a state of A thereafter. The incoming branch has a state of A and the continuing branch a state of a at the final selection event (#3). This results in displacement of the continuing branch.

The genealogy that results from after resolving the potential natural selection events is shown in Fig. 7.11c. Given the ancestral state and high-fitness haplotype, selection events #1 and #2 had no impact on the branching pattern of the tree. In contrast, selection event #3 caused a change in the branching pattern that moved the coalescence point of lineage 2 from the continuing branch to the coalescence point of the incoming branch. This reflects the fact that after natural selection acted, lineage 2 was identical by descent to a different lineage. This change in the branching pattern causes the total branch length of the tree to be slightly longer than it was without natural selection. In this case, the height of the tree is not changed by natural selection.

### Problem box 7.2 Resolving possible selection events on an ancestral selection graph

Use Fig. 7.11b and trace the lineage states forward in time, assuming that the state of the MRCA is an a haplotype and that A is the fitter haplotype. Are the resulting lineage states in the present the same as originally given in the figure? Has the height of the genealogy changed?

As another exercise to test your knowledge of the ancestral selection graph, resolve the genealogy in Fig. 7.11b using the a allele as the state of the MRCA, alternatively assuming that the A and then the a allele is the fitter haplotype.

The conclusions to be drawn from the ancestral selection graph with two alleles are straightforward. Weak to moderate directional natural selection tends to have only a minor impact on average times to coalescence. Stated another way, the action of directional natural selection does not greatly alter the average times to coalescence compared with a strictly neutral genealogy with the same number of lineages. When the selection coefficient and the mutation rate are approximately equal, the mean time to the MRCA is shortened slightly (Neuhauser & Krone 1997; see also Przeworski et al. 1999). However, the difference in the average coalescence times with directional selection and with strict neutrality is slight given the wide variation in coalescence times due to finite sampling. Strong natural selection for advantageous alleles or selection against deleterious mutations is expected to reduce the total height of genealogical trees because of lineages bearing states that are strongly disadvantageous (see Charlesworth et al. 1993, 1995).

#### ***Genealogies and balancing selection***

Natural selection where heterozygotes have the highest fitness, also called **balancing selection**, can also be incorporated in genealogical branching models (Hudson & Kaplan 1988; Kaplan et al. 1988; Nordborg 1997; see Hudson 1990). Earlier in Chapter 6, we saw that balancing selection is expected to maintain both alleles at a diallelic locus segregating

### Interact box 7.6 Coalescent genealogies with directional selection

The coalescent process can be simulated for directional selection via an ancestral selection graph. The simulation displays a genealogy based on parameter values for the number of lineages ( $n:$ ), the strength of natural selection ( $S:$ ), and the mutation rate ( $U:$ ). The strength of selection parameter corresponds to the rate at which incoming branches are formed on the genealogy. Mutation events will change allelic states of branches in the genealogy. The blue dots in the genealogy (labeled **Phony?** in the legend) are incoming branches. The button at the top right of the genealogy animation labeled **Show mutations** can be used to show mutations by coloring the branches. Clicking this button again when it is labeled **Show types** displays the allelic states of each node in the genealogy.

Click the **Recalc** button to begin a new simulation. The resulting genealogy can be viewed as an animation going back in time (click **Animation** tab at top left and use playback controls at bottom) or as a genealogy (click **Trees** tab at top left).

Set the selection parameter to 1.0 and view a few resulting genealogies. Then set the selection parameter to 5.0 and view numerous genealogies. How do the genealogies compare under weak or strong selection? Rerun the simulation a few times while trying out a range of values for the selection strength. Be sure to view multiple replicate genealogies for each set of parameter values.

in the population at equilibrium. The two allele frequencies at equilibrium will depend on the selection coefficients against the two homozygous genotypes. Since the haploid genealogical model does not have diploid genotypes nor sexual reproduction, we will need to take an alternative approach rather than specifying multiple genotype fitness values.

Balancing selection is a special case of natural selection because it works counter to the fixation and loss due to genetic drift. In a genealogical branching model genetic drift is represented by the process of coalescence. So to approximate the overall effect of balancing selection we need a process that will delay coalescence to the same degree that selection favors heterozygotes in the diploid selection model. This same overall result can be obtained by modeling balancing selection along the lines of population structure with two demes. Although it sounds like an odd approach, population structure and balancing selection have similar effects for different reasons. In structured populations, two lineages cannot coalesce unless they are in different demes (refer back to Fig. 4.16). Gene-flow events that move lineages into different demes therefore tend to delay coalescence events. Using this same logic, we can model balancing selection in a single panmictic population as a process where there are two lineage types. A switching process akin to gene flow (or mutation) changes lineage types at random while the coalescence process operates at the same time. If two lineages must be of the same type to coalesce, then the switching

process will prevent coalescence among the lineages that are of different types.

Let the two lineage types be A and B and their respective frequencies in the population be  $p$  and  $q$  so that  $p + q = 1$  (see Fig. 7.12). Every generation, lineages of one type may switch to the other type with rate  $\mu$ . Twice the expected number of the  $2N$  total lineages in the population that switch types each generation is then  $v = 4N\mu$ . The expected number of lineages switching each generation serves as a surrogate for the strength of balancing selection since frequency switching will let lineages escape coalescence. Using this switching rate, the expected waiting time until an A lineage switches to a B lineage is

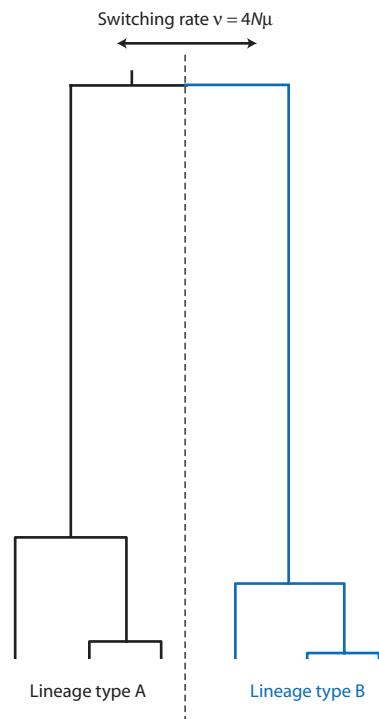
$$P(T_{A \rightarrow B} \leq t) = 1 - e^{-k_A \frac{v}{2} \left( \frac{q}{p} \right)} \quad (7.52)$$

and the expected waiting time until a B lineage switches to an A lineage is

$$P(T_{B \rightarrow A} \leq t) = 1 - e^{-k_B \frac{v}{2} \left( \frac{p}{q} \right)} \quad (7.53)$$

The ratio of the frequencies ( $q/p$  and  $p/q$ ) serves to adjust the exponent for the relative frequencies of the two lineage types.

Next we need to express the waiting time until a coalescence event, keeping in mind that lineages can only coalesce if they are of the same type. Given that there are  $2Np$  lineages of type A and  $2Nq$  lineages of



**Figure 7.12** A genealogy where balancing natural selection is modeled by type switching. Every generation, lineages of one type (here A and B) may switch to the other type with rate  $\mu$ . Twice the expected number of the  $2N$  total lineages in the population that switch types each generation is then  $v = 4N\mu$ . Since lineages can only coalesce when they are of the same type, type switching increases the average time to coalescence. This is analogous to natural selection favoring heterozygotes because overdominance also extends the segregation times of alleles. Genealogical trees that result from balancing selection modeled as type switching tend to have longer branches compared to genealogies that result from genetic drift or directional natural selection.

type B and coalescence events are mutually exclusive, the expected waiting time until a coalescence event is

$$P(T_{coalescence} \leq t) = 1 - e^{-\frac{k_A(k_A-1)}{2}\left(\frac{1}{p}\right) + \frac{k_B(k_B-1)}{2}\left(\frac{1}{q}\right)} \quad (7.54)$$

If switching events and coalescence events are all mutually exclusive, the individual exponents can be added together to obtain the total waiting time to any event:

$$\begin{aligned} P(T_{coalescence} \leq t) \\ = 1 - e^{-\frac{k_A(k_A-1)}{2}\left(\frac{1}{p}\right) + \frac{k_B(k_B-1)}{2}\left(\frac{1}{q}\right) + k_A \frac{v}{2} \left(\frac{q}{p}\right) + k_B \frac{v}{2} \left(\frac{p}{q}\right)} \end{aligned} \quad (7.55)$$

Given that an event has occurred with a known waiting time, the type of event can be determined by

drawing a random number between zero and one and comparing it with the cumulative total of the chance of each event divided by the total probability of all events.

Genealogical trees that result from balancing selection modeled as type switching tend to have longer branches compared with genealogies that result from genetic drift alone (Fig. 7.12). This is due to the increase in average waiting times between coalescence events caused by lineage type switching. The final two lineages in particular are expected to take a long time to coalesce since they must switch to the same type. The results of two allele balancing selection are qualitatively similar to genealogies with long waiting times for the last two lineages expected with subdivided populations. If mutation is also operating along with balancing selection, then genealogies with longer branches would also accumulate more mutations since the number of mutation events is proportional to the total branch length of a genealogy. Lineages of the two different types, in particular, are expected to have more mutational changes between them than would be expected in a genealogy under the basic neutral coalescent model.

An additional model of balancing selection exists for populations with more than two alleles and equivalent (overdominant) fitness values for the possible heterozygous genotypes (Vekemans & Slatkin 1994; Uyenoyama 1997). Such multi-allelic balancing selection is distinct from balancing selection with only two alleles, resulting in genealogies that have long coalescence times for the lineages near the present (or shorter times to coalescence further back in time nearer the MRCA) compared to neutral genealogies. Classic examples of multi-allelic balancing selection are the many alleles found at single self-incompatibility loci in some plants (e.g. Schierup et al. 1998).

Chapter 8 on molecular evolution further considers the consequences of natural selection on the height and shape of genealogies in the context of comparisons designed to test whether or not a genealogy is different than expected by the processes of drift and mutation alone.

### Chapter 7 review

- Mean fitness in a population can be viewed as a graph of average fitness by genotype or allele frequencies called a fitness surface. Natural selection acts as an uphill climber on a fitness surface,

- moving genotype frequencies uphill based on the slope at the current genotype frequencies.
- Fitness surfaces will have multiple peaks and valleys if there is dominance or epistasis.
  - When fitness depends on three alleles at one locus, the results of natural selection depend on initial genotype frequencies in the population if there is strong over- and underdominance.
  - The net balance of recombination and natural selection may result in equilibria that do not correspond to mean fitness maxima. Recombination works toward gametic equilibrium irrespective of mean fitness and may act in opposition to natural selection.
  - Although the viability model of fitness is used as a standard, changes in genotype frequencies and their equilibria are often distinct when fitness is defined as differential fecundity or carrying capacity, or when fitness values vary in time and space.
  - When both natural selection and genetic drift are acting, selection is strong relative to genetic drift when  $4N_e s$  is much greater than one, selection is weak relative to genetic drift when  $4N_e s$  is much less than one, and when  $4N_e s$  is approximately one then selection and drift are about the same strength.
  - When both natural selection and mutation are acting, deleterious alleles will be maintained in a population at a level that increases with the mutation rate but decreases with consanguineous mating and the selection coefficient against the allele when homozygous.
  - In genealogical branching models, directional selection can be modeled as an ancestral selection graph. Weak directional selection does not greatly alter the total branch length nor the total height of genealogical trees on average.
  - In genealogical branching models, balancing selection can be modeled as a lineage type switching process similar to gene flow or mutation. With two alleles, balancing selection lengthens

the average time to coalescence for the final two lineages since they have to switch to the same type to coalesce. With three or more alleles, balancing selection will tend to increase coalescence times and lengthen terminal branches in a genealogical tree.

### Further reading

A classic and approachable treatment of two-locus selection that features fitness surfaces can be found in:

Lewontin RC and White MJD. 1960. Interaction between inversion polymorphisms of two chromosome pairs in the grasshopper, *Moraba scurra*. *Evolution* 14: 116–29.

For case studies, perspective, and basic theory relating to genotype interactions at two or more loci that influence fitness see chapters in:

Wolf JB, Brodie ED III, and Wade MJ (eds). 2000. *Epistasis and the Evolutionary Process*. Oxford University Press, Oxford.

Extensions of the basic viability natural selection model that account for biological variations such as spatial and temporal variation in fitness, fitness trade-offs, competition, and predation can be found in:

Roff DA. 2001. *Life History Evolution*. Sinauer Associates, Sunderland, MA

Roughgarden J. 1996. *Theory of Population Genetics and Evolutionary Ecology: an Introduction*. Prentice Hall, Upper Saddle River, NJ.

A genealogical model of directional natural selection where haplotypes are neutral when rare and become favored after reaching a threshold frequency can be found in:

Przeworski M, Coop G, and Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59: 2312–23.

### Problem box answers

#### Problem box 7.1 answer

$$\begin{aligned} w_{AA} &= 0.679 & w_{AS} &= 0.763 & w_{SS} &= 0.153 \\ w_{AC} &= 0.679 & w_{SC} &= 0.534 & w_{CC} &= 1.0 \end{aligned}$$

Initial allele frequencies set 1:

$$p = 0.75, q = 0.20, r = 0.05$$

$$\begin{aligned} \bar{w} &= 0.679(0.75)^2 + 0.153(0.2)^2 \\ &\quad + 1(0.05)^2 + (0.763)2(0.75)(0.2) \\ &\quad + (0.679)2(0.75)(0.05) \\ &\quad + (0.534)2(0.2)(0.05) \end{aligned}$$

$$\begin{aligned} \bar{w} &= 0.382 + 0.00612 + 0.0025 + 0.2289 \\ &\quad + 0.051 + 0.0107 = 0.6812 \end{aligned}$$

$$\begin{aligned} \bar{w}_C &= 1(0.05) + (0.679)(0.75) + (0.534)(0.2) \\ &= 0.6661 \end{aligned}$$

$$\Delta p = \frac{0.05(0.6661 - 0.6812)}{0.6812} = -0.0011$$

$$p_{t+1} = 0.05 - 0.0011 = 0.0489$$

The C allele goes to loss because its marginal fitness is lower than the mean fitness.

Initial allele frequencies set 2:

$$p = 0.7, q = 0.20, r = 0.1$$

$$\begin{aligned} \bar{w} &= 0.679(0.7)^2 + 0.153(0.2)^2 + 1(0.1)^2 \\ &\quad + (0.763)2(0.7)(0.2) + (0.679)2(0.7)(0.1) \\ &\quad + (0.534)2(0.2)(0.1) \end{aligned}$$

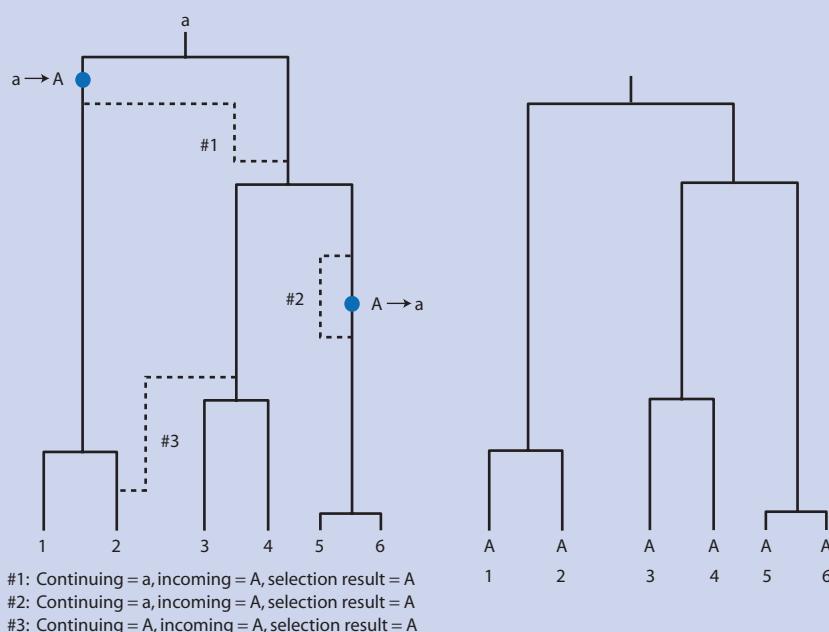
$$\begin{aligned} \bar{w} &= 0.333 + 0.0061 + 0.01 + 0.2136 \\ &\quad + 0.0951 + 0.0214 = 0.6792 \\ \bar{w}_C &= 1(0.1) + (0.679)(0.7) + (0.534)(0.2) \\ &= 0.6821 \\ \Delta p &= \frac{0.1(0.6821 - 0.6792)}{0.6792} = 0.05 \\ p_{t+1} &= 0.1 + 0.05 = 0.15 \end{aligned}$$

The C allele goes to fixation because its marginal fitness is greater than the mean fitness, a condition that will hold until the C allele is fixed in the population.

#### Problem box 7.2 answer

At selection events #1 and #2, the incoming branch displaces the continuing branch.

The incoming and continuing branches have identical states at selection event #3. The result is that all lineages have the state A in the present. In this case the height of the genealogy is shorter because of the outcome of selection event #1. See Figure 7.13.



**Figure 7.13** An example ancestral selection graph showing how natural selection operates in the genealogical branching model. The genealogy on the left shows incoming and continuing branches and their haplotype states. The genealogy on the right shows the final pattern of branch states once all natural selection events have been resolved and states reassigned as necessary. In this illustration, the A haplotype is at a selective advantage over the a haplotype. Haplotype A increases in frequency when the continuing and incoming branches are resolved. See Problem box 7.2.

# CHAPTER 8

## Molecular evolution

### 8.1 The neutral theory

- The neutral theory and its predictions for levels of polymorphism and rates of divergence.
- The nearly neutral theory.

The field of molecular evolution involves the study of DNA, RNA, and protein sequences with the goal of elucidating the processes that cause both change and constancy among sequences over time. One approach to molecular evolution is to focus on a specific gene, seeking to test hypotheses about what parts of that specific sequence are most likely involved in some function or in the regulation of transcription. Another type of inquiry in molecular evolution involves testing hypotheses about the population genetic processes that have operated on sequences in the past using DNA sequence data. This latter type of research often seeks to distinguish whether a pattern of variation in a sample of DNA sequences is consistent with genetic drift or with certain forms of natural selection. The common feature of all hypothesis tests in studies of molecular evolution is the use of null and alternative hypotheses for the patterns and rates of sequence change. This chapter will introduce the conceptual foundations behind many of the most commonly used null and alternative hypotheses in molecular evolution. Although this chapter focuses exclusively on DNA sequences, the concepts presented are sometimes applicable to protein sequences as well.

The **neutral theory** now forms the basis of the most widely employed null model in molecular evolution. The neutral theory adopts the perspective that most mutations have little or no fitness advantage or disadvantage and are therefore **selectively neutral**. Genetic drift is therefore the primary evolutionary process that dictates the fate (fixation or loss) of newly occurring mutations. When it was originally proposed, the neutral theory was a major departure from orthodox population genetic theory of the time.

In the 1950s and 1960s, it was widely thought that most mutations would have substantial fitness differences and therefore the fate of most mutations was dictated by natural selection. Motoo Kimura (shown in Fig. 8.1) argued instead that the interplay of mutation and genetic drift could explain many of the patterns of genetic variation and the evolution of protein and DNA sequences seen in biological populations (Kimura 1968, reviewed in Kimura 1983a). King and Jukes (1969) also proposed a similar idea at around the same time. (The controversy generated by the neutral theory as well as some of the logic behind Kimura's proposal of the neutral theory is



**Figure 8.1** Motoo Kimura (on left) and James Crow in 1986 on the occasion of Kimura being awarded an honorary doctoral degree at the University of Wisconsin, Madison, WI, USA. Kimura pioneered the use of diffusion equations to determine quantities such as the average time until fixation or until loss for neutral mutations. Based on these foundations, he proposed the neutral theory of molecular evolution in 1968. Kimura and Crow collaborated to develop some of the basic expectations for neutral genetic variation. Crow mentored and encouraged many people who became influential contributors to population genetics, including Kimura. Photograph kindly provided by J. Crow.

covered in Chapter 11.) The neutral theory null model makes two major predictions under the assumption that genetic drift alone determines the fate of new mutations. One prediction is the amount of **polymorphism** for sequences sampled within a population of one species. The other prediction is the degree and rate of **divergence** among sequences sampled from separate species.

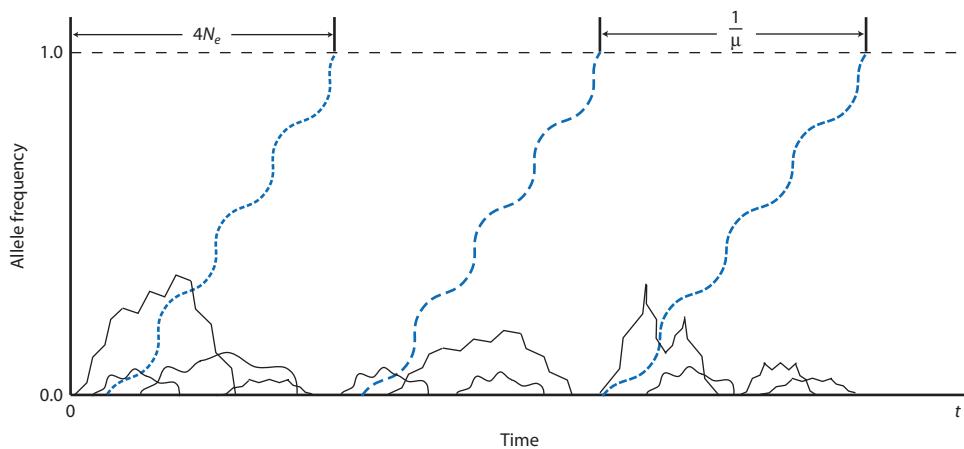
**Divergence** Fixed genetic differences that accumulate between two completely isolated lineages that were originally identical when they diverged from a common ancestor.  
**Polymorphism** The existence in a population of two or more alleles at one locus.  
 Populations with genetic polymorphisms have heterozygosity, gene diversity, or nucleotide diversity measures that are greater than zero.

### Polymorphism

The balance of genetic drift and mutation that determines polymorphism in the neutral theory is diagrammed in Fig. 8.2. Each line indicates the change in frequency of an allele over time. New mutations enter the population (lines at the bottom edge) and

their frequency in the population then changes due to genetic drift. The frequency of each allele is therefore a random walk between fixation and loss. To see how this random walk results in polymorphism, hold a straight edge such as a ruler to form a vertical line at any single time point. If the vertical line intersects any allele frequency lines, then the population has genetic polymorphism at that time point since there are multiple alleles segregating in the population. More alleles segregating in the population indicate more polymorphism. Segregating alleles, and therefore polymorphism, result from the random walk in frequency that each mutation takes under genetic drift. Most mutations segregate for short periods of time and then are lost from the population. However, since their frequency is dictated by random sampling, some alleles may reach high frequencies before eventually being lost. A small proportion of mutations will eventually be fixed in the population after a random walk in allele frequency. Under neutral theory, polymorphism results from the transient dynamics of allele frequencies before they reach fixation or loss end points. The process that underlies Fig. 8.2 can be approximately simulated in Interact box 5.1.

The neutral theory's prediction for levels of polymorphism in a population follows directly from the predicted dynamics of allele frequency under genetic drift (see Chapters 3 and 5). Chapter 3 showed that the initial frequency of an allele is also its chance



**Figure 8.2** The fate of selectively neutral mutations in a population. New mutations enter the population at rate  $\mu$  and an initial frequency of  $\frac{1}{2N}$ . Allele frequency is a random walk determined by genetic drift. The time that a new mutation segregates in the population, or the dwell time of a mutation, depends on the effective population size. However, the chance that a new mutation goes to fixation (equal to its initial frequency) is also directly related to the effective population size. These two effects of the effective population size cancel each other out for neutral alleles. The neutral theory then predicts that the rate of fixation is  $\mu$  and therefore the expected time between fixations is  $1/\mu$  generations. For that subset of mutations that eventually fix, the expected time from introduction to fixation is  $4N_e$  generations. After Figure 3.1 in Kimura (1983a).

of eventual fixation. For new mutations that start out as one copy in the entire population of  $2N$  allele copies, the chance of eventual fixation is  $\frac{1}{2N}$  while the chance of eventual loss is  $1 - \frac{1}{2N}$ . The diffusion

approximation of genetic drift in Chapter 3 also showed that the average time to fixation of a new mutation approaches  $4N$  generations while the average time to loss approaches just  $2\left(\frac{N_e}{N}\right)\ln(2N)$  generations as  $N$  gets large. (To see these results set  $p = \frac{1}{2N}$  in equation 3.40 and evaluate the expressions as  $N$  increases toward infinity.) The average time to fixation also has a large variance, so the standard deviation of the number of generations to fixation is expected to be about half of the mean or  $2.15N_e$  generations (Kimura & Ohta 1969b; Narain 1970; Kimura 1970, 1983a).

For example, in a population of  $N = 1000$ , the average time to loss is about 15 generations (assuming  $\frac{N_e}{N} = 1$ ) but the average time to fixation for those alleles that eventually fix is 4000 generations. In addition, the time to fixation is highly variable since genetic drift is a stochastic process or a random walk. The standard deviation of the time to fixation is large, at approximately 2150 generations, consistent with a broad range in times to fixation. Relatively few mutations are expected to fix, but those mutations that do go to fixation segregate for a much longer time on average than the large proportion of mutations that go to loss. While mutations are segregating before their eventual end point of fixation or loss, there is polymorphism in the population.

An additional way to understand the expected polymorphism of neutral alleles in a population is to examine the equilibrium balance between genetic drift causing alleles to go to fixation and the input of new alleles in a population by mutation. Chapter 5 showed that for the infinite alleles model of mutation, the combined processes of mutation and genetic drift produce equilibrium heterozygosity:

$$H_{equilibrium} = \frac{4N_e\mu}{4N_e\mu + 1} \quad (8.1)$$

that depends on the effective population size  $N_e$  and the mutation rate  $\mu$  (see equation 5.39). In this view

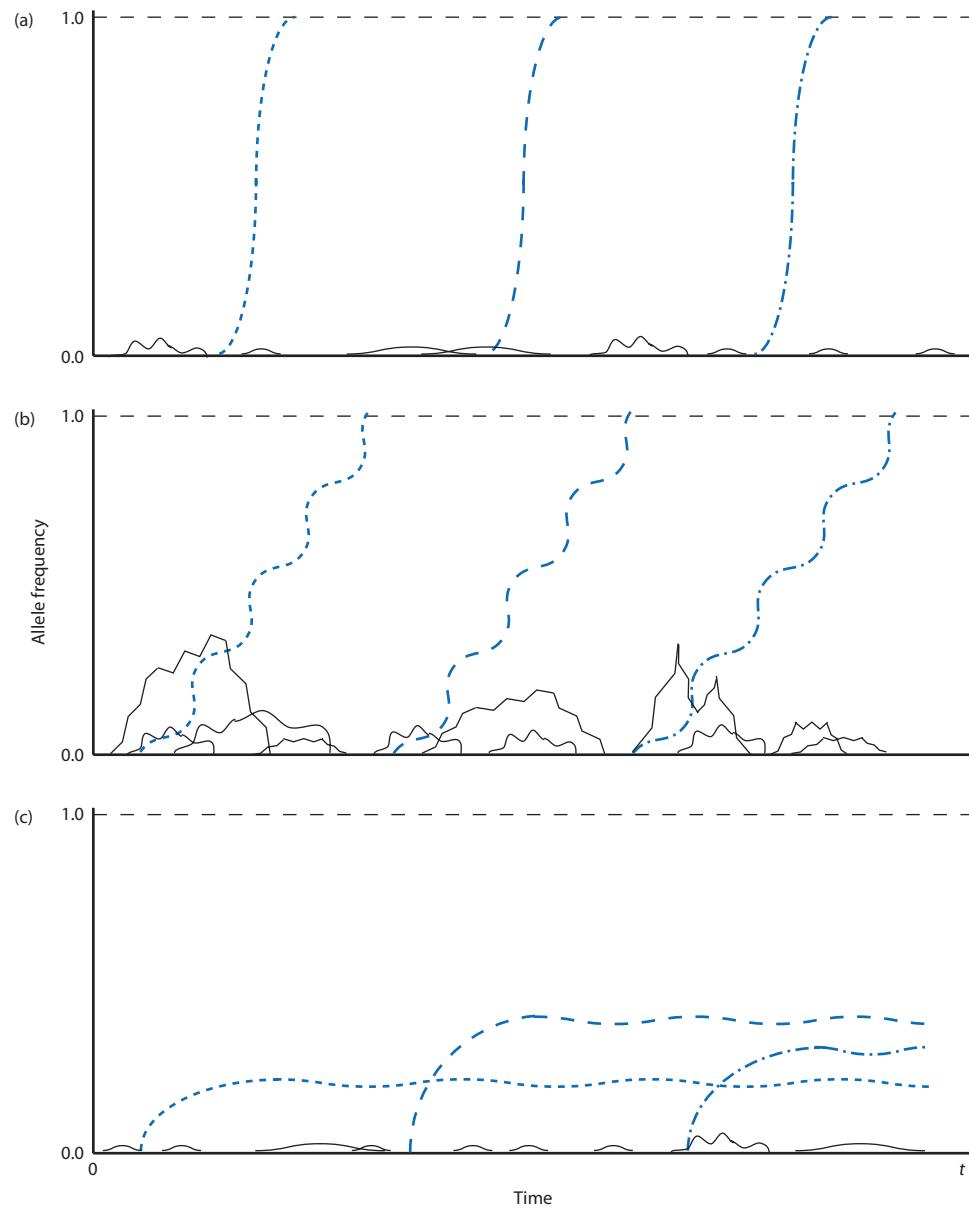
of neutral mutations, polymorphism results from either a high rate of input of mutations even if drift is strong, a long dwell time for each mutation due to a large effective population size even if mutations are infrequent, or intermediate levels of mutation and genetic drift.

The neutral theory prediction for polymorphism can be readily compared with polymorphism expected under positive (higher than average genotype fitness) and negative (lower than average genotype fitness) natural selection (Fig. 8.3). New mutations that are deleterious will go to loss faster than neutral mutations since natural selection will deterministically reduce their frequencies and there will be little or no random walk in allele frequency. In contrast, new mutations that are advantageous will increase in frequency to fixation, again deterministically under natural selection without a random walk in frequency. So a locus with new mutations that are influenced by directional natural selection should show less polymorphism than a locus with neutral mutations. The other possibility is that some advantageous mutations are influenced by balancing selection due to overdominance for fitness. In that case two or more alleles will have very long times of segregation since natural selection will maintain several alleles at intermediate frequencies between fixation and loss with the result of increased levels of polymorphism in the population.

As a whimsical metaphor, compare the average times to fixation under directional selection, neutral evolution, and balancing selection with the time it takes a population genetics professor to go from his or her office (initial mutation) to the coffee shop and back (fixation) at different career stages (different processes). A new, over-worked professor goes directly to get a cup of coffee and returns immediately without stopping to talk to anyone, so the trip is short and direct like directional selection. In mid-career, a professor has a bit more free time and will pause more often to greet friends, like a random walk. Late in their career, a professor takes a roundabout path to the shop and stops to talk frequently such that the coffee break takes a very long time, like balancing selection.

### Divergence

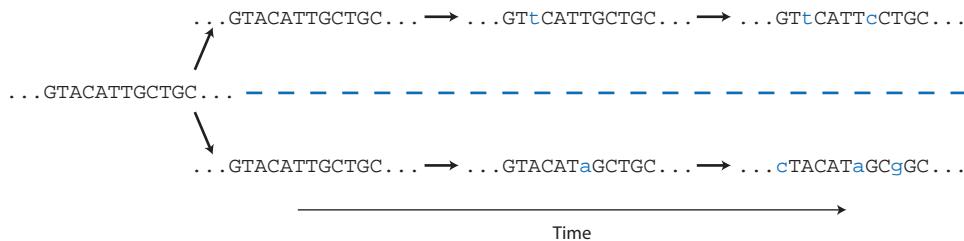
The neutral theory also predicts the rate of divergence between sequences. Genetic divergence occurs by **substitutions** that accumulate in two DNA sequences over time. Think of two DNA sequences that are



**Figure 8.3** The dwell time for new mutations is different if fixation and loss is due to genetic drift or natural selection. With neutral mutations (b), most mutations go to loss fairly rapidly and a few mutations eventually go to fixation. For eventual fixation or loss of neutral mutations the path to that outcome is a random walk, implying that the time to fixation or loss has a high variance. For mutations that fix because they are advantageous (a), directional selection fixes them rapidly in the population. Therefore under directional selection alleles segregate for a shorter time and there is less polymorphism than with neutrality. For mutations that show overdominance for fitness, natural selection favoring heterozygote genotypes maintains several alleles in the population indefinitely. Therefore balancing selection greatly increases the segregation time of alleles and increases polymorphism compared to neutrality. Both cases of natural selection (a and c) are drawn to show negative selection acting against most new mutations. If new mutations are deleterious then the time to loss is very short and there is very little random walk in allele frequency since selection is nearly deterministic.

copies of the same ancestral sequence (Fig. 8.4). The two sequences were originally identical before any substitutions occurred. Over time, mutations occurred in each population and some were fixed by chance due to genetic drift (see Fig. 8.2). Each fixed mutation causes a change in the base pairs at random

nucleotide positions in the sequence, causing each sequence to diverge slightly from its ancestor as well as from its sibling sequence. A biological example is two species that recently diverged from an ancestral species with no genetic variation. Both of the new species would be founded with identical DNA



**Figure 8.4** The process of divergence for two DNA sequences that descended as identical copies of an ancestral sequence. Each sequence experiences neutral mutations, some of which are eventually fixed by genetic drift. These fixed mutations replace all other alleles and are therefore substitutions (indicated by blue lower-case letters). As substitutions accumulate, the two sequences diverge from the ancestral sequence as well as from each other. In this example, the two sequences are eventually divergent at five of 12 nucleotide sites due to substitutions. The dashed line indicates complete isolation of the two populations containing the derived sequences.

sequences and thereafter be reproductively isolated from each other. DNA sequences compared between the two species would each experience DNA sequence divergence due to mutations occurring over time.

**Substitution** The complete replacement of one allele previously most frequent in the population with another allele that originally arose by mutation.

The neutral theory predicts the rate at which allelic substitutions occur and thereby the rate at which divergence occurs. Predicting the substitution rate for neutral alleles requires knowing the probability that an allele becomes fixed in a population and the number of new mutations that occur each generation. A new mutation in a population of diploid individuals is initially present as just a single copy out of a total of  $2N$  copies of the locus. Therefore, the initial frequency of a new mutation is  $\frac{1}{2N}$ . Under genetic drift, the chance of fixation of any neutral allele is simply its initial frequency (see Chapter 3). Each generation, the chance that an allele copy mutates is  $\mu$  and there are a total of  $2N$  allele copies. Therefore, the expected number of new mutations in a population each generation is  $2N\mu$ . Multiplying the probability of fixation by the expected number of mutations per generation,

$$k = (2N\mu) \frac{1}{2N} \quad (8.2)$$

gives the rate at which alleles that originally entered the population as mutations go to fixation per gen-

eration, symbolized by the substitution rate  $k$ . Notice that this equation simplifies to

$$k = \mu \quad (8.3)$$

The necessary assumption is that the substitution process is viewed on a time scale that is long relative to the average time to fixation for an individual mutation. If more than  $4N_e$  generations have elapsed, then it is likely that all the alleles in a population will have descended from one allele due to genetic drift. The probability that the lucky allele that is fixed in the population was a new mutation is  $\mu$ .

This result is remarkable because it says that the probability that a neutral mutation goes to fixation each generation, or the rate of substitution, is simply equal to the mutation rate. Notice that the predicted substitution rate does not depend on the effective population size. This is because a mutation in a smaller population has a greater chance of fixation but there are fewer new mutations each generation, while a mutation in a larger population has a smaller chance of reaching fixation but there are more mutations introduced each generation. The rate of input of new mutations in a population and the chance of fixation due to genetic drift exactly balance out when  $N$  changes. Note that this same result holds in the case of haploid loci since there are a total of  $N$  alleles and the probability of fixation of a new mutation is  $\frac{1}{N}$ .

Based on the rate of substitution, the neutral theory also predicts that the substitutions that ultimately cause divergence should occur at a regular average rate. For waiting time processes, the time between events is the reciprocal of the rate of events. Using a clock that chimes on the hour as an example, the rate of chiming is 24 per day (or 24/day). Therefore, the expected time between chiming events is

$1/24$  of a day or 1 hour. Since the rate of neutral substitution is  $\mu$ , the *expected* time between neutral substitutions is  $1/\mu$  generations (see Fig. 8.2). For example, if the mutation rate of a locus is  $1 \times 10^{-6}$  (one nucleotide change per  $10^6$  gametes per generation) then the expected time between neutral substitutions is  $10^6$  generations *on average*. This offers one explanation of why different loci diverge at different rates: the different loci simply have distinct mutation rates that lead to variable neutral substitution rates.

### Nearly neutral theory

The **nearly neutral theory** considers the fate of new mutations if some portion of new mutations are acted on by natural selection of different strengths (Ohta & Kimura 1971, Ohta 1972, reviewed in Ohta 1992 and Gillespie 1995). The nearly neutral theory recognizes three categories of new mutations: neutral mutations, mutations acted on strongly by either positive or negative natural selection, and mutations acted on weakly by natural selection relative to the strength of genetic drift. This last category contains mutations that are nearly neutral since neither natural selection nor genetic drift will determine their fate exclusively.

For a new mutation in a finite population that experiences natural selection, the forces of directional selection and genetic drift oppose each other. Recall from Chapter 3 that genetic drift causes heterozygosity to decrease at a rate of  $\frac{1}{2N_e}$  per generation (see equation 3.51). Thus,  $\frac{1}{2N_e}$  quantifies the “push” on a new mutation toward fixation caused by genetic drift. The selection coefficient ( $s$ ) on a genotype describes the “push” on alleles toward fixation or loss due to natural selection. The force of selection on a new mutation can be quantified using the result from section 5.2 that the chance of fixation is approximately  $2s$ . Setting these forces equal to each other,

$$2s = \frac{1}{2N_e} \quad (8.4)$$

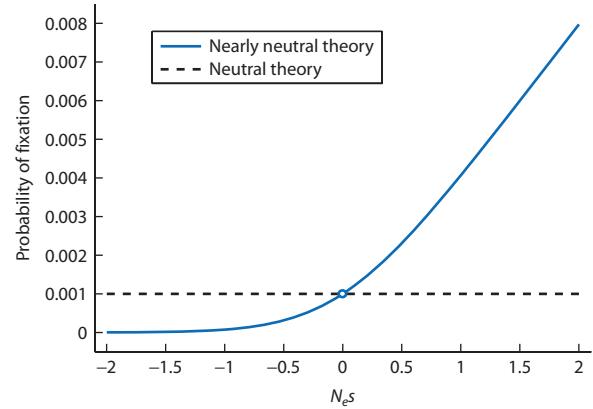
gives the conditions where genetic drift and natural selection have approximately equal influence on the fate of allele frequencies. When  $2s$  is within an order of magnitude of the reciprocal of effective population size, an allele can be described as **net neutral** or **nearly neutral** since natural selection and genetic

drift are approximately equal forces dictating the probability of fixation of an allele. Next, notice that multiplying both sides of equation 8.4 by  $2N_e$  gives  $4N_e s = 1$  as the condition where the processes of genetic drift and natural selection are equal. When  $4N_e s$  is much greater than one natural selection is the stronger process whereas when  $4N_e s$  is much less than one genetic drift is the stronger process.

Using more sophisticated mathematical techniques, Kimura (1962) showed that the probability of fixation for a new mutation in a finite population is

$$P_{\text{fixation}} = \frac{1 - e^{-4N_e sp}}{1 - e^{-4N_e s}} \quad (8.5)$$

where  $p$  is the allele frequency ( $p = \frac{1}{2N_e}$  for a single mutation but in general  $p$  is assumed to be much less than one),  $N_e$  is the effective population size, and  $s$  is the selection coefficient assuming codominance. This equation is plotted in Fig. 8.5 along with the constant probability of fixation for a new mutation expected under neutral theory.



**Figure 8.5** The probability of eventual fixation for a new mutation under the neutral and nearly neutral theories. Under the nearly neutral theory the probability of fixation depends on the balance between natural selection and genetic drift, expressed in the product of the effective population size and the selection coefficient ( $N_e s$ ). When negative selection operates against a deleterious allele, the selection coefficient and  $N_e s$  are negative. Values of  $N_e s$  near 0 yield a fixation probability close to that predicted by neutral theory. Only when the absolute value of  $N_e s$  is large does natural selection exclusively determine the probability of fixation. Neutral theory assumes that neutral mutations are not influenced by selection and have a constant probability of fixation dictated by the effective population size. In this example the initial allele frequency is 0.001, or the frequency of a new mutation at a diploid locus in a population of 500. After Ohta (1992).

### Interact box 8.1 The relative strengths of genetic drift and natural selection

Examine the fitness values given below. Based on your knowledge of natural selection, predict for each case the equilibrium frequency of the A allele under *only* the process of natural selection. The answer is given below.

	1	2	3	4
$w_{AA}$	0.975	0.8	0.9	0.4
$w_{Aa}$	0.9875	0.9	1.0	1.0
$w_{aa}$	1.0	1.0	0.9	0.4

Now, launch Populus and use the **Model** menu to select **Mendelian Genetics** and then **Drift and Selection**. In the options dialog box set population size to 20, initial allele frequency to 0.5, and generations to 200. Enter each set of genotype fitness values and then track the fate of the a allele over time. Run the simulation 10 times for each set of fitness values, tabulating your results as you go (numbers in each row should total to 10).

Fitness set	No. fixed for A	No. lost for A	No. segregating for A and a
1			
2			
3			
4			

Explain your simulation results by calculating  $4N_e s$  for each set of fitness values.

The interplay of genetic drift and natural selection can also be simulated in PopGene.S<sup>2</sup> by selecting the **Drift** menu and then the **Drift + Selection + Mutation** model. First try parameters of 20 populations, population size of 20, initial allele frequency of 0.5, 100 generations, relative fitness values of 1, 0.95, and 0.90, and mutation rates of 0.005. Modify these parameter values to get outcomes you would expect when genetic drift is the stronger process as well as when natural selection is the stronger process.

Answer: In sets 1 and 2 natural selection will fix the A allele and drive the a allele to loss, as expected for directional selection with additive gene action. In sets 3 and 4 the heterozygotes have the highest fitness, so balancing selection should maintain both alleles indefinitely.

The nearly neutral theory predicts that the rate of substitution will depend on the effective population size for the proportion of mutations in a population that are nearly neutral ( $4N_e s \approx 1$ ). The nearly neutral theory therefore predicts that the amount of polymorphism in populations depends for some mutations on the effective population size. A consequence is that subdivided populations and different species can exhibit different levels of polymorphism based on their effective population size. Similarly, rates of divergence can also vary between species, due to differences in  $N_e$ . This is in contrast to the neutral theory, which predicts that the rate of substitution is independent of the effective population size.

### 8.2 Measures of divergence and polymorphism

- Measuring divergence of DNA sequences.
- Nucleotide substitution models correct divergence estimates for saturation.
- DNA polymorphism measured by number of segregating sites and nucleotide diversity.

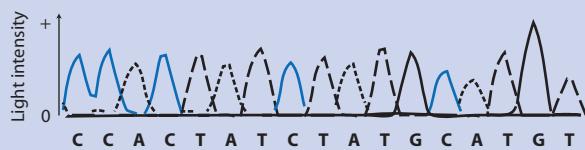
Most natural and laboratory populations contain at least some, and often a large amount, of genetic variation represented by different alleles found at the many loci in the genome. The smallest possible unit of the genome is a homologous **nucleotide site**, or single base-pair position in the exact same genome

### Box 8.1 DNA sequencing

Figure 8.6 shows a DNA sequence **electropherogram** from a capillary electrophoresis DNA sequencer. In the chain-termination method (also called the Sanger method after its inventor Frederick Sanger), a DNA polymerase and oligonucleotide primer are used to copy a DNA template sequence in the presence of a low concentration of dideoxynucleotides. The dideoxynucleotides lack a 3' hydroxyl (-OH) group necessary for continued 5'-to-3' DNA synthesis and therefore the replicating copy ends whenever a dideoxynucleotide is added by the polymerase. The dideoxynucleotides also carry a radioactive or molecular dye marker. When the population of DNA copies

is separated by size using electrophoresis, the molecular label is used to identify the type of dideoxynucleotide base (an A, T, C, or G) that terminated the replicating copy. Modern DNA sequencers use laser light to detect one of four molecular dyes attached to each dideoxynucleotide during electrophoresis.

In Fig. 8.6 the y axis represents the intensity of molecular dye signal measured as DNA fragments pass by a laser reader. Each line style represents the intensity of a different color light detected by the laser. Each of the four light colors represents a different dideoxynucleotide. See the text web page for a full-color version of this figure.



**Figure 8.6** A DNA sequence electropherogram. See the text web page for a full-color version of this figure.

location, that could be compared among individuals. Genetic variation at such nucleotide sites is characterized by the existence of DNA sequences that have different nucleotides (e.g. some individuals have an A and other individuals have a T at the 37th base pair from the start codon in the same gene) and is called **nucleotide polymorphism**. Nucleotide polymorphisms are sometimes referred to as **single nucleotide polymorphisms** or **SNPs** (pronounced “snips”). This section of the chapter covers commonly used measures of divergence and polymorphism estimated from DNA sequence data.

#### DNA divergence between species

The most fundamental method to quantify molecular evolution is by comparing two DNA sequences. This comparison is a two-step procedure. First, the two DNA sequences must be **aligned** such that homologous nucleotide sites for each sequence are all lined up in the same columns. For example, if two coding genes were sequenced then one way to align them would be to match up the first three nucleotides that make up the start codon. (Methods of sequence alignment are beyond the scope of this text, but readers can refer to text such as Page & Holmes (1998)

for more details.) The second step is to determine the number of sites that have different nucleotides. The number of nucleotide sites that differ between two DNA sequences divided by the total number of nucleotide sites compared gives the proportion of nucleotide sites that differ, often called a **p distance** as a shorthand for proportion-distance. This is a basic measure of the evolutionary events that have occurred since two DNA sequences descended from a common ancestor, when they were each identical copies of the same sequence.

**p Distance** The number of nucleotide sites that differ between two DNA sequences divided by the total number of nucleotide sites, a shorthand for proportion-distance. Sometimes symbolized as  $d$  for distance.

An example of divergence between a pair of sequences is shown in Fig. 8.4. Consider the two sequences at the far right in the present time after some divergence that has introduced substitutions. There are five nucleotide sites that have different nucleotides out of a total of 16 nucleotide sites. Therefore, the  $p$

distance is  $5/12 = 0.3125$ , or 31.25% of the nucleotide sites have diverged.

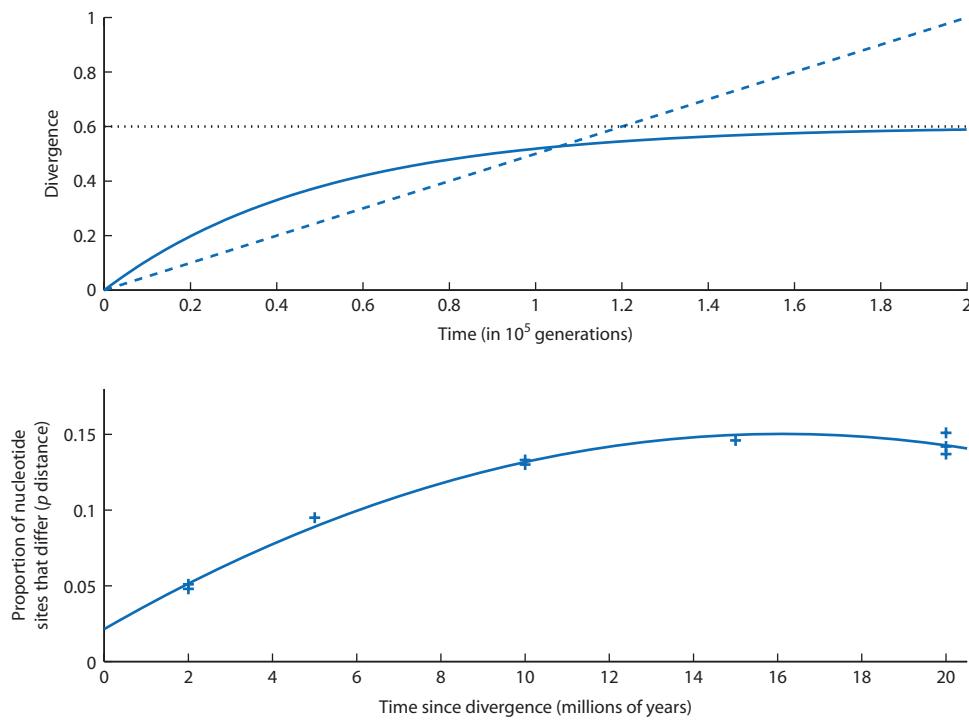
The  $p$  distance between two DNA sequences sampled from completely independent populations should increase over time as substitutions within each population replace the nucleotide that was originally shared at each site due to identity by descent. If the two DNA sequences represent two distinct species or completely isolated populations, then the  $p$  distance is a measure of divergence between the two species.

### DNA sequence divergence and saturation

**Saturation** is the phenomenon where DNA sequence divergence appears to slow and eventually reaches a plateau even as time since divergence continues to increase. Saturation in nucleotide changes over time is caused by substitution occurring multiple times at the same nucleotide site, a phenomenon called **multiple hit** substitution (see the related topic of multiple hit mutation in Chapter 5). Substitutions that

occur repeatedly at the same site have the effect of covering up information about past substitutions, since only the most recent substitution can be observed and measured as divergence between two DNA sequences. Computing the  $p$  distance between two sequences leads to under-estimates of number of substitutions that have occurred and therefore an under-estimate of the degree of divergence. The top panel of Fig. 8.7 shows divergence that increases linearly with time since divergence (dashed line) and that exhibits saturation (solid line). Actual DNA sequence data routinely exhibit some degree of saturation, as shown in the bottom panel of Fig. 8.7 for the mitochondrial cytochrome *c* oxidase subunit II gene sequenced for several bovine species (ungulates including domestic cattle, bison, water buffalo, and yak) that diverged between 2 and 20 million years ago.

Saturation can be understood by imagining the process of assembling a DNA sequence at random and comparing it with another existing DNA sequence. Think of drawing individual nucleotides out of a bucket containing equal numbers of A, C, G, and T base



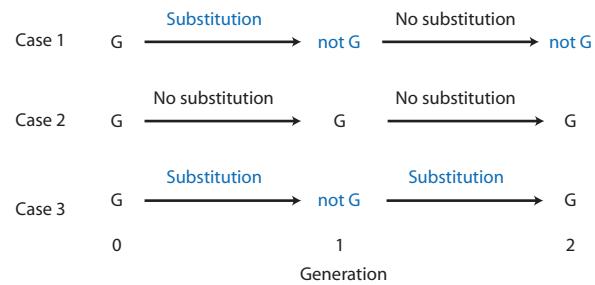
**Figure 8.7** Substitutions that occur repeatedly at the same nucleotide site lead to saturation of nucleotide changes as time since divergence from a common ancestor increases. The rate of substitutions does not change and the total number of substitutions continues to increase over time, as shown by the dashed line in the top panel representing the true number of substitutions. In contrast, multiple substitutions at the same sites leads to a slowing and leveling off in the estimate of divergence (solid line, top panel). Therefore, the amount of divergence leads to the perception that the rate of divergence decreases over time. The bottom panel shows divergence and saturation at the mitochondrial cytochrome *c* oxidase subunit II gene among bovine species (ungulates including domestic cattle, bison, water buffalo, and yak) that diverged between 2 and 20 million years ago. In the top panel  $\alpha = 1 \times 10^{-6}$  ( $\alpha$  is explained overleaf). The bottom panel data are from Janecek et al. (1996) and the line is a quadratic regression fit.

pairs. If a nucleotide site in the existing sequence is an A, there is a 25% chance that a randomly drawn nucleotide will be an A and the sites will match. On the other hand, there is a 75% chance it will not be an A (it will be a T, C, or G) and the two sites will be diverged. Thus, a DNA sequence assembled from random draws of nucleotides at equal frequency should be 75% divergent or 25% identical to another sequence on average. The consequence is that two sequences which originated from a common ancestor do not continue to get increasingly divergent over time. Eventually, the maximum divergence will plateau at 75% as continued mutation essentially randomizes the shared sites between the two sequences.

There are a wide variety methods to correct the perceived divergence between two DNA sequences to obtain a better estimate of the true divergence after accounting for multiple hits. These correction methods are called **nucleotide substitution models** and use parameters for DNA base frequencies and substitution rates to obtain a modified estimate of the divergence between two DNA sequences. The simplest of these is the Jukes and Cantor (1969) nucleotide-substitution model, named for its authors. Working through the derivation of the Jukes–Cantor model is worthwhile to gain some insight into how nucleotide-substitution models operate.

The Jukes–Cantor model starts out by assuming that any nucleotide in a DNA sequence is equally likely to be substituted with any of the other three nucleotides. For example, if a site currently has a C then substitution of an A, T, or G all have the same chance of occurring. Figure 8.8 shows three possible events for a nucleotide site. The site may (1) experience one and only one substitution, (2) not experience any substitutions over time, and (3) experience a substitution that changes the nucleotide at the site and then another independent substitution that restores the original nucleotide at the site. In the first situation perceived divergence and actual divergence are the same and no correction is required. The perceived divergence in the second and third cases is the same but very different events have occurred. In the third case, substitutions have occurred for some portion of nucleotide sites that appear to have no divergence. Nucleotide substitution models such as Jukes–Cantor serve to estimate the frequency of nucleotide sites that appear to have not diverged but actually have diverged.

In the Jukes–Cantor model, the probability of a nucleotide substitution is customarily represented by  $\alpha$  (pronounced “alpha”). Since there are three nucleotides that can each be substituted for a



**Figure 8.8** The three types of event that a single nucleotide site may experience over two generations. A nucleotide site may initially have a G, for example. In case 1 a single substitution event in generation one changes the G to an A, C, or T nucleotide, the nucleotide also present at generation two. A  $p$  distance measure of divergence will accurately count the number of substitutions in case 1. In cases 2 and 3 the nucleotide site still retains the same nucleotide it had initially, giving the impression that there have been no substitutions. In case 2 this impression is accurate. However, in case 3 there have been two substitution events that are not accounted for in a simple  $p$  distance measure of divergence.

nucleotide currently at a site and all three are equally likely to be substituted, the probability of any substitution is  $3\alpha$ . So if the nucleotide is initially a G at generation zero, the probability that it is also a G one generation later is

$$P_{G(t=1)} = 1 - 3\alpha \quad (8.6)$$

Since the chance of substitution is independent in each generation, the probability of no substitutions over two generations is

$$P_{G(t=2)} = (1 - 3\alpha)^2 \quad (8.7)$$

This gives the probability that a nucleotide does not change over two generations as shown in case 2 of Fig. 8.8.

We also need to determine the probability that a nucleotide changes twice, as shown in case 3 of Fig. 8.8. From generation zero to generation one, the probability of a substitution is  $3\alpha$ . This probability can also be written as one minus the probability of no substitution, or  $1 - P_{G(t=1)}$ . From generation one to generation two, there is only one base that can be substituted to make the site match its initial state. The chance that this occurs is the probability of a substitution or  $\alpha$ . Bringing these two independent probabilities together gives the probability that a multiple hit nucleotide substitution occurs which restores the nucleotide initially present:

$$P_{G(t=2)} = \alpha(1 - P_{G(t=1)}) \quad (8.8)$$

This is the probability that two substitutions occur, neither of which would be detectable by comparing two DNA sequences at time two.

Combining these two results gives the probability that a nucleotide site has the same base pair after two generations:

$$P_{G(t=2)} = (1 - 3\alpha)P_{G(t=1)} + \alpha(1 - P_{G(t=1)}) \quad (8.9)$$

regardless of the number of substitutions that have occurred over two generations. Since the probabilities of substitution and no substitution are independent each generation, this equation can be written in a more general form as:

$$P_{G(t+1)} = (1 - 3\alpha)P_{G(t)} + \alpha(1 - P_{G(t)}) \quad (8.10)$$

or as a recurrence equation that applies to any two time periods one generation apart. This recurrence equation can also be expressed as the change in the probability that an initial nucleotide site remains unchanged over time. Since the change in a quantity is the difference between what it is now and what it was one time step  $t$  in the past,  $\Delta P_{G(t)} = P_{G(t+1)} - P_{G(t)}$ . The change in the probability that a given nucleotide is found at a site over one generation is then

$$\Delta P_{G(t)} = (1 - 3\alpha)P_{G(t)} + \alpha(1 - P_{G(t)}) - P_{G(t)} \quad (8.11)$$

Expanding the terms on the right side of this equation gives

$$\Delta P_{G(t)} = P_{G(t)} - 3\alpha P_{G(t)} + \alpha - \alpha P_{G(t)} - P_{G(t)} \quad (8.12)$$

which then simplifies to

$$\Delta P_{G(t)} = \alpha - 4\alpha P_{G(t)} \quad (8.13)$$

The model we have considered to this point treats time as discrete steps, as shown in Fig. 8.8. If we consider the rate of change at any time  $t$ , then the change in the probability that a nucleotide site appears the same with changes in time is a differential equation,  $\frac{dP_{G(t)}}{dt} = \alpha - 4\alpha P_{G(t)}$ . The solution to this equation is:

$$P_{G(t)} = \frac{1}{4} + (P_{G(t=0)} - \frac{1}{4})e^{-4\alpha t} \quad (8.14)$$

which is analogous to an exponential growth equation for a population with a carrying capacity. As

$t$  gets large, the  $e^{-4\alpha t}$  term approaches zero so that  $P_{G(t)}$  approaches  $1/4$ .

Using the continuous time equation, the probability that a nucleotide site is a G over time depending on its initial state and the rate of substitution is shown in Fig. 8.9. If the nucleotide at a site is initially a G, then  $P_{G(t)} = 1$  and the probability the site remains a G over time is

$$P_{G(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \quad (8.15)$$

This is a probability that declines exponentially toward  $1/4$  as  $t$  increases. In contrast, if the nucleotide at a site is not initially a G then  $P_{G(t)} = 0$ . The probability the site remains a G over time is

$$P_{G(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \quad (8.16)$$

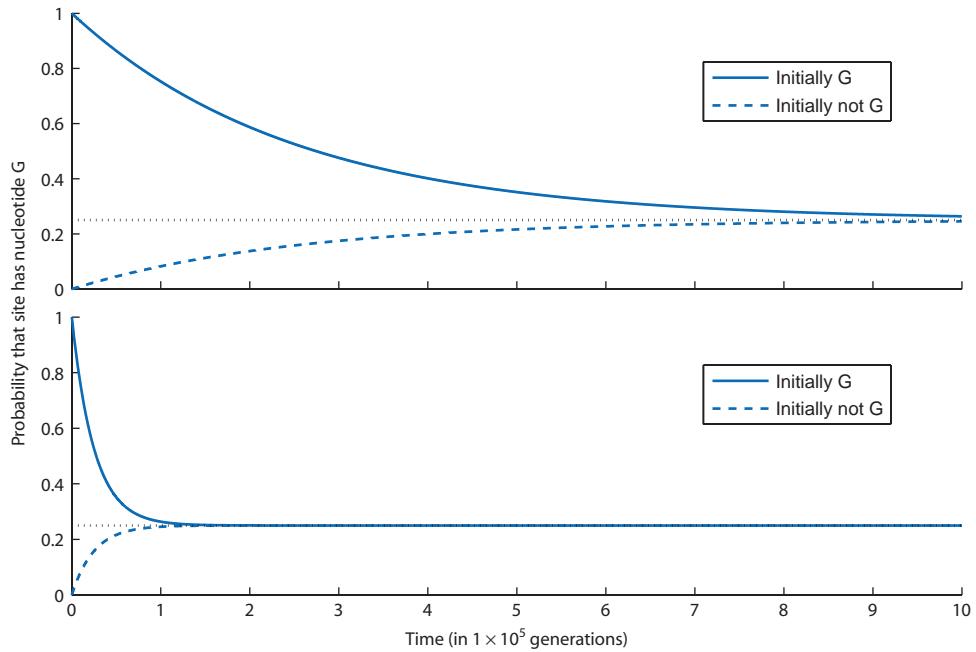
which increases from zero to  $1/4$  as time increases. Recall that  $1/4$  is the probability that a site in a sequence assembled from randomly drawn nucleotides (that are equally frequent) matches the same site in an existing sequence. Also notice that the approach to  $1/4$  will be faster as the substitution rate  $\alpha$  increases due to the  $-4\alpha t$  term in the exponent.

Let's not forget that the original goal was to correct observed divergence between sequences or  $p$  distances for multiple hit mutations. The model of sequence change we have so far is the foundation of a correction, but we need to do some more work to obtain an actual correction method. If we think of two DNA sequences originally identical by descent at every nucleotide site at time 0, at some later time  $t$  the probability that any site will possess the same nucleotide is

$$P_{I(t)} = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t} \quad (8.17)$$

where  $P_{I(t)}$  indicates the probability that the nucleotides at a given site are identical. The exponential term is now  $e^{-8\alpha t}$  because there are two DNA sequences that can change independently so that the chance of the nucleotide remaining the same decreases twice as fast with time. The probability that two sites are different or divergent – call it  $d$  – over time is one minus the probability that the sites are identical, so that

$$d = \frac{3}{4}(1 - e^{-8\alpha t}) \quad (8.18)$$



**Figure 8.9** The probability that a nucleotide site retains its original base pair under the Jukes–Cantor model of nucleotide substitution. If a nucleotide site originally has a G base, for example, the probability of the same base being present declines steadily over time. If a nucleotide site was initially not a G (it was an A, C, or T), the probability that a G is present at the site increases over time. The probability that a given base is present always converges to 25% because that is the probability of sampling a given base at random if the probability of substitution to each nucleotide is equal. In the top panel  $\alpha = 1 \times 10^{-6}$  whereas in the bottom panel  $\alpha = 1 \times 10^{-5}$ .

The exponent can be removed from this equation by taking the natural logarithm of the right side and the equation rearranged to give

$$8\alpha t = -\ln\left(1 - \frac{4d}{3}\right) \quad (8.19)$$

This equation states that eight times the substitution rate multiplied by time is related to the amount of divergence we expect to see between two DNA sequences. For two DNA sequences that were originally identical by descent, we expect that each site has a  $3\alpha t$  chance of substitution. Since there are two sequences, there is a  $6\alpha t$  chance of a site being divergent between the two sequences. If we set expected divergence  $K = 6\alpha t$ , then we notice  $K$  is close to the  $8\alpha t$  above. In fact,  $K$  is  $3/4$  of the expression for  $8\alpha t$ , so that

$$K = -\frac{3}{4} \ln\left(1 - \frac{4d}{3}\right) \quad (8.20)$$

where  $d$  is the observed proportion of sites that differ between two DNA sequences, or the  $p$  distance.  $K$  is then the estimate of the actual number of sites

that have experienced divergence events corrected for multiple hits with the Jukes–Cantor nucleotide-substitution model.

A few examples will help show how the Jukes–Cantor model correction works in practice. Imagine two DNA sequences that differ at 1 site in 10 so the  $p$  distance is 10% or  $d = 0.10$ . This level of observed divergence is an under-estimate because it does not account for multiple hits. To adjust for multiple hits we compute corrected divergence as

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3}(0.10)\right) = 0.1073 \quad (8.21)$$

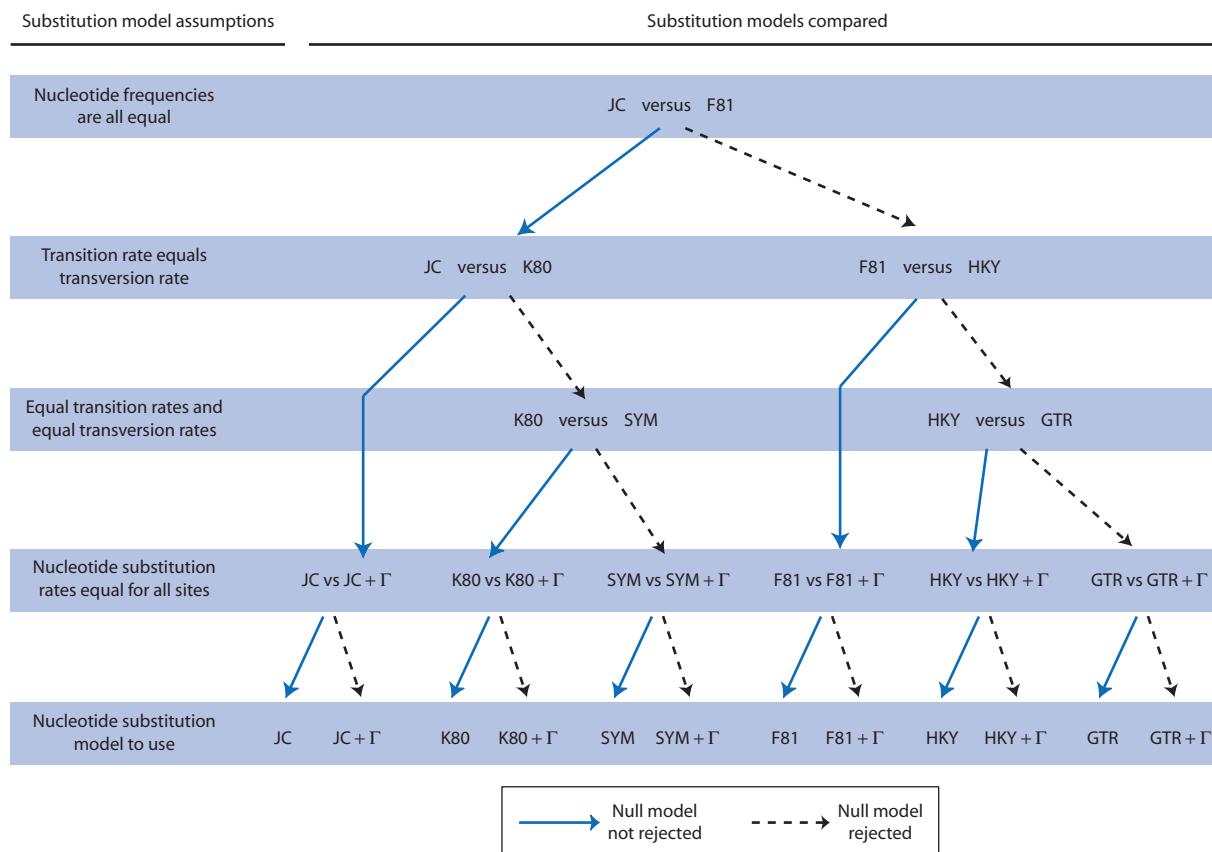
which shows that at the low apparent divergence of 10% there are expected to be 0.7% of sites that had experienced multiple hits. The true divergence is then estimated as 10.73% or slightly greater than the apparent divergence. If apparent divergence was greater, say  $d = 0.40$ , then we should expect a larger correction. In that case

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3}(0.40)\right) = 0.5813 \quad (8.22)$$

so the correction for multiple hits is much larger at a bit over 18% for a total corrected divergence of 58.13% of nucleotide sites. A frequently used convention is that a capital  $K$  refers to a saturation-corrected estimate of divergence while lower case  $k$  or  $d$  is an uncorrected estimate of divergence.

The Jukes–Cantor is the simplest possible nucleotide substitution model because it assumes that all nucleotides are equally frequent in DNA sequences and that all sites experience the same substitution rate. Many DNA sequences, however, exhibit variation in these parameters, which is not accounted for in the Jukes–Cantor model. There are numerous models of nucleotide substitution of increasing complexity that take these factors into account by using an increasing number of parameters to represent the different types of substitution rates (Posada &

Crandall 2001). Figure 8.10 illustrates a hierarchy of some of the nucleotide-substitution models available. These different models can be distinguished by examining DNA sequence data to test each model assumption. For example, the Jukes–Cantor model assumes that all nucleotides have equal frequencies. If a sample of DNA sequences shows base frequencies that deviate significantly from equal frequencies of 25%, then the F81 model is a better choice because it assumes arbitrary base frequencies. Both the JC and F81 nucleotide-substitution models assume that transition and transversion rates are equal and that substitution rates are constant among sites. It is now common practice to estimate the substitution model that best approximates the patterns of nucleotide change in a DNA sequence data set (Posada & Crandall 1998).



**Figure 8.10** The hierarchy of nucleotide-substitution models that can be used to correct apparent divergence between DNA sequences to better estimate the actual number of substitutions that have occurred. The Jukes–Cantor (JC) model is the simplest and assumes that there is just one rate of substitution that applies to all nucleotide changes and is constant among nucleotide sites. Other nucleotide substitution models include an increasing number of parameters to represent more features of DNA sequence evolution, in particular variable rates of substitution among various categories of nucleotides. If nucleotide-substitution rates are variable among different sites, this variation can be modeled by a gamma distribution indicated by the Greek letter  $\Gamma$ . Nucleotide-substitution models: JC, Jukes–Cantor (Jukes & Cantor 1969); F81, Felsenstein 81 (Felsenstein 1981); K80, Kimura 80 (Kimura 1980); HKY, Hasegawa–Kishino–Yano (Hasegawa et al. 1985); SYM, symmetrical model (Zharkikh 1994); GTR, general time reversible (Rodriguez et al. 1990). Figure after Posada and Crandall (1998).

### DNA polymorphism

Variable DNA sequences at one locus within a species represent different alleles that are present in the population. Since DNA sequences are composed of many nucleotide sites, defining alleles is somewhat more complex than if alleles are discrete (i.e. A or a). Imagine obtaining a sample of  $n$  individuals from a population and determining the DNA sequence of  $L$  nucleotides for one gene or genomic region for each individual (see Tajima 1993b). For simplicity consider each individual as haploid or homozygous. The first step would be to construct a **multiple sequence alignment** so that the homologous nucleotide sites for each sequence are all lined up in the same columns (Fig. 8.11). With such a multiple sequence alignment there are two commonly used measures that characterize the pattern of DNA polymorphism in a sample of DNA sequences from a single species.

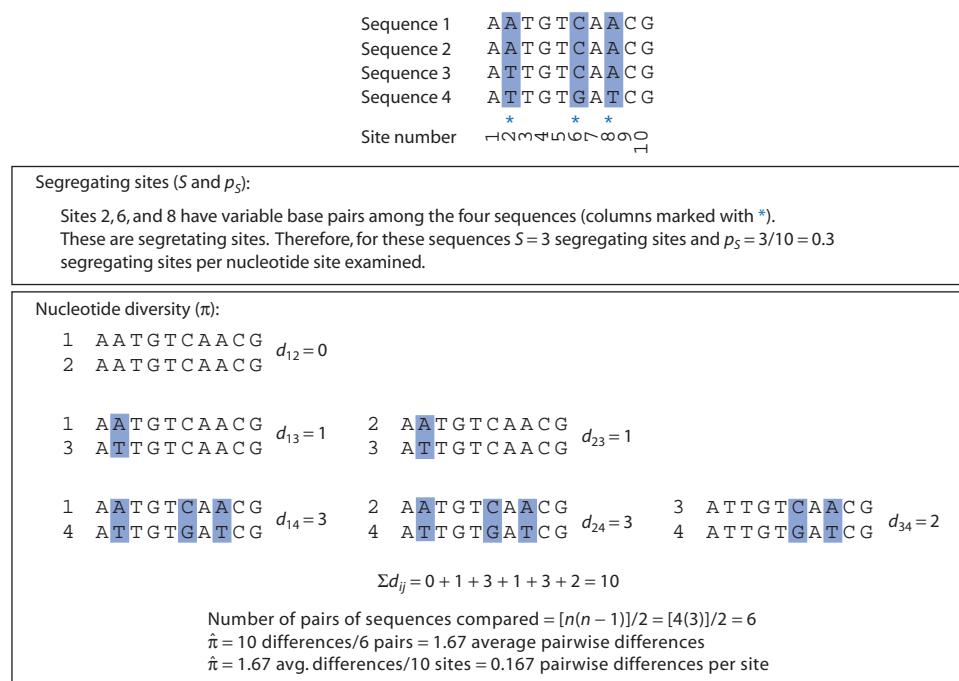
One measure of DNA polymorphism is the **number of segregating sites**,  $S$ . A segregating site is any of the  $L$  nucleotide sites that maintains two or more nucleotides within the population, such as sites 2, 6, and 8 in Fig. 8.11. The total number of segregating sites is  $S$  and can be expressed as the number of segregating sites per nucleotide site,  $p_S$ ,

by dividing the number of segregating sites by the total number of sites:

$$p_S = \frac{S}{L} \quad (8.23)$$

The frequency of DNA sequences with a given nucleotide at a site does not influence  $S$  (compare sites 2 and 6 or 8 in Fig. 8.11), but  $S$  will increase as the number of individuals sampled increases since DNA sequences with additional polymorphisms will be added to the sample.

The number of segregating sites ( $S$ ) under neutrality is a function of the scaled mutation rate  $4N_e\mu$ . Watterson (1975) first developed a way to estimate  $\theta$  from the number of segregating sites observed in a sample of DNA sequences. The expected number of segregating sites at drift–mutation equilibrium can more easily be determined using the logic of the coalescent model (Watterson used a different approach). Under the infinite sites model of mutation, each mutation that occurs increases the number of segregating sites by one. The expected number of segregating sites is therefore just the expected number of mutations for a given genealogy. If each lineage has the probability  $\mu$  of mutating each generation and there are  $k$



**Figure 8.11** A hypothetical sample of four DNA sequences that are each 10 nucleotides long. There are a total of three segregating sites ( $S = 3$ ) or three-tenths of a segregating site per nucleotide ( $p_S = 0.3$ ). The nucleotide diversity is calculated by summing the nucleotide sites that differ between each unique pair of DNA sequences. In this example there are 1.67 average pairwise nucleotide differences or 0.167 average pairwise nucleotide differences per nucleotide site.

lineages, then the expected number of mutations in one generation is  $k\mu$ . If the expected time to coalescence for  $k$  lineages is  $T_k$ , then  $k\mu T_k$  mutations are expected for each value of  $k$ . The expected number of mutations ( $E$  indicates an expectation or average) is obtained by summing over all  $k$  between the present and the most recent common ancestor (MRCA):

$$E[S] = E\left[\sum_{k=2}^n \mu k T_k\right] = \mu \sum_{k=2}^n k E[T_k] \quad (8.24)$$

where  $n$  is the total number of lineages in the present. To see an illustration of this equation, refer to Fig. 3.26 where  $n = 6$  in the summation of equation 8.24 and imagine summing up the probability of a mutation in each time interval between coalescent events.

A fundamental result of the coalescent model is that the probability of  $k$  lineages coalescing is  $\frac{k(k-1)}{2} \left(\frac{1}{2N_e}\right)$ . Therefore, the expected time to coalescence is the inverse of the probability of coalescence or  $\frac{2(2N_e)}{k(k-1)}$ . This expected time to coalescence can then be substituted into equation 8.24 to give

$$E[S] = \mu \sum_{k=2}^n k \frac{2(2N_e)}{k(k-1)} \quad (8.25)$$

This equation simplifies by canceling each  $k$ , taking the constant  $4N_e$  outside the summation, and adjusting the range of the summation to remove the  $-1$  after  $k$  in the denominator:

$$E[S] = 4N_e \mu \sum_{k=1}^{n-1} \frac{1}{k} \quad (8.26)$$

to give the expected number of segregating sites in a sample of  $n$  DNA sequences. Once the expected number of segregating sites  $E[S]$  is known, it can be solved for  $\theta$ . Notice that  $\theta = 4N_e \mu$  can be substituted in equation 8.26 to give

$$E[S] = \theta \sum_{k=1}^{n-1} \frac{1}{k} \quad (8.27)$$

and then rearranging,

$$\theta = \frac{E[S]}{\sum_{k=1}^{n-1} \frac{1}{k}} \quad (8.28)$$

to solve for  $\theta$  in terms of the number of segregating sites divided by the total branch length of the genealogy. An estimate of the scaled mutation rate determined from the number of segregating sites in a sample of DNA sequences is symbolized as  $\hat{\theta}_W$  ( $W$  for Watterson) or  $\hat{\theta}_S$  ( $S$  for segregating sites). If we define

a new variable,  $a_1 = \sum_{k=1}^{n-1} \frac{1}{k}$ , then

$$\hat{\theta}_S = \frac{E[S]}{a_1} \quad (8.29a)$$

using the absolute number of segregating sites, or

$$\hat{\theta}_S = \frac{p_S}{a_1} \quad (8.29b)$$

using the number of segregating sites per nucleotide site sampled. The importance of these two final equations is that  $4N_e \mu$  can be estimated from the number of segregating sites and the number of sequences in a sample.

A second measure of DNA polymorphism is the **nucleotide diversity** in a sample of DNA sequences, symbolized by  $\pi$  (pronounced “pie”), and also known as the **average pairwise differences** in a sample of DNA sequences (Nei & Li, 1979; Nei & Kumar 2000). The nucleotide diversity is equivalent to the heterozygosity measured using alleles represented by DNA sequences (assuming random mating and the infinite sites model of mutation). The nucleotide diversity summarizes nucleotide polymorphism by averaging the number of nucleotide site differences found when each unique pair of DNA sequences in a sample is compared. In contrast with the proportion of segregating sites, the nucleotide diversity is sensitive to the frequency of each DNA sequence allele in a sample, since more frequent sequences appear in more of the pairwise comparisons. The nucleotide diversity is the sum of the number of nucleotide differences seen for each pair of DNA sequences:

$$\hat{\pi} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j>i}^n d_{ij} \quad (8.30)$$

where  $i$  and  $j$  are indices that refer to individual DNA sequences,  $d_{ij}$  is the number of nucleotide sites that differ between sequences  $i$  and  $j$ , and  $n$  is the total number of DNA sequences in the sample. The number of unique pairwise comparisons in a sample

**Table 8.1** Nucleotide diversity ( $\pi$ ) estimates reported from comparative studies of DNA sequence polymorphism from a variety of organisms and loci. All estimates are the average pairwise nucleotide differences per nucleotide site. For example, a value of  $\pi = 0.02$  means that two in 100 sites vary between all pairs of DNA sequences in a sample.

Species	Locus	$\pi$	Reference
<i>Drosophila melanogaster</i>	<i>anon1A3</i>	0.0044	Andolfatto 2001
	<i>Boss</i>	0.0170	
	<i>transformer</i>	0.0051	
<i>Drosophila simulans</i>	<i>anon1A3</i>	0.0062	Graustein et al. 2002
	<i>Boss</i>	0.0510	
	<i>transformer</i>	0.0252	
<i>Caenorhabditis elegans</i> <sup>a</sup>	<i>tra-2</i>	0.0	Graustein et al. 2002
	<i>glp-1</i>	0.0009	
	<i>COII</i>	0.0102	
<i>Caenorhabditis remanei</i> <sup>b</sup>	<i>tra-2</i>	0.0112	
	<i>glp-1</i>	0.0188	
	<i>COII</i>	0.0228	
<i>Arabidopsis thaliana</i> <sup>a</sup>	<i>CAUL</i>	0.0042	Wright et al. 2003
	<i>ETR1</i>	0.0192	
	<i>RbcL</i>	0.0012	
<i>Arabidopsis lyrata</i> ssp. <i>Petraea</i> <sup>b</sup>	<i>CAUL</i>	0.0135	
	<i>ETR1</i>	0.0276	
	<i>RbcL</i>	0.0013	

<sup>a</sup>Mates by self-fertilization.

<sup>b</sup>Mates by outcrossing.

of  $n$  sequences is  $(n(n - 1))/2$  and so dividing the sum of  $d_{ij}$  by this number gives the average number of differences per pair of sequences. The average number of pairwise differences can also be divided by the number of nucleotide sites examined ( $L$ ) to express  $\hat{\pi}$  per nucleotide site. Figure 8.11 shows an example computation of  $\hat{\pi}$  for a hypothetical sample of four DNA sequences.

In larger samples that may include multiple identical DNA sequences, the nucleotide diversity can be estimated by

$$\hat{\pi} = \frac{k}{k-1} \sum_{i=1}^k \sum_{j>i} p_i p_j d_{ij} \quad (8.31)$$

where  $p_i$  and  $p_j$  are the frequencies of alleles  $i$  and  $j$ , respectively, in a sample of  $k$  different sequences that each represent one allele. This version of the formula just provides an average of  $d_{ij}$  that is weighted by the frequency of each type of DNA sequence found in a sample. The nucleotide diversity can be underestimated if there are rare sequence polymorphisms

in a population that are unlikely to be sampled (see Renwick et al. 2003). Information on the sampling variance of  $\pi$  can be found in Nei and Kumar (2000).

Some values of  $\pi$  from different organisms and loci are shown in Table 8.1. Estimates of nucleotide diversity are useful because  $\pi$  is a measure of heterozygosity for DNA sequences. As such, the value of  $\pi$  is a function of  $4N_e\mu$  under an equilibrium between genetic drift and mutation. With an estimate of  $\pi$  and the mutation rate at a locus ( $\mu$ ), it is then possible to estimate the effective population size. Because  $\pi$  is an estimator of the scaled mutation rate  $\theta$ , it is sometimes referred to as  $\hat{\theta}_\pi$ .

### 8.3 DNA sequence divergence and the molecular clock

- The molecular clock hypothesis for DNA divergence.
- Dating divergence events with a molecular clock.

One key result of the neutral theory is the prediction that the rate of substitution is equal to the mutation

### Interact box 8.2 Estimating $\pi$ and $S$ from DNA sequence data

The number of segregating sites ( $S$ ) and the nucleotide diversity ( $\pi$ ) can be estimated in PopGene.S<sup>2</sup>. First, you will need to download a file of DNA sequence data from the GenBank web page onto a computer that also has a copy of PopGene.S<sup>2</sup>. The text web page gives step-by-step instructions to obtain DNA sequences for the mitochondrial cytochrome *b* gene in a sample of 30 African sable antelope (Pitra et al. 2002).

Once you have downloaded the DNA sequence data file from Genbank, open PopGene.S<sup>2</sup> and select **Molecular Population Genetics** in the main menu. In the dialog window that appears, click on the **Open File** button and then use the file dialog to find and open the downloaded data file. The sequences will be displayed in the top window. Verify that PopGene.S<sup>2</sup> found 30 sequences in the file and that the longest sequence was 343 base pairs. The first step is to create a multiple sequence alignment for the 30 DNA sequences by pressing the **Align Sequences** button (pressing the button will cause a window to open temporarily). Once the sequences are aligned, the number of segregating sites, the number of gaps, and the nucleotide diversity will then be estimated. Click on the **Pairwise Nucleotide Diversity** button to see the nucleotide diversity for all possible pairs of DNA sequences. The **Nucleotide site distribution** button gives a list of the frequencies of each base pair found at each of the sites in the multiple sequence alignment. Use the **Save Aligned Sequences** button to save a file with the multiple sequence alignment. Then open this file in a text editor such as Notepad to view the aligned sequences.

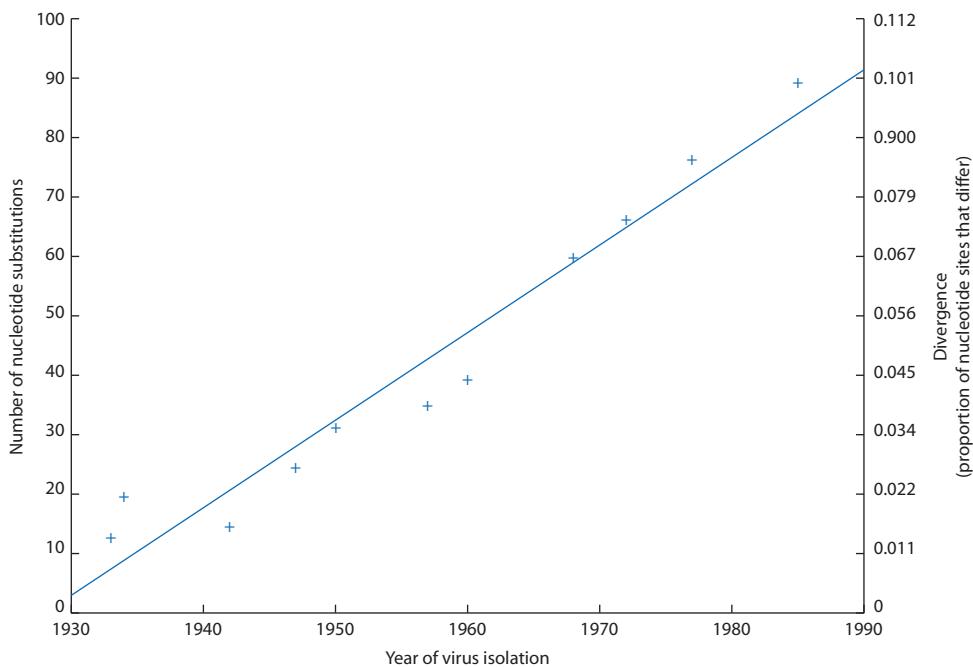
rate. A corollary of this prediction is that the expected number of generations between substitutions is the reciprocal of the mutation rate. For example, if the mutation rate is  $1 \times 10^{-5}$  base pairs replicated in error per generation then we expect to wait an average of  $10^5$  generations to see one mutation in a single gene copy. Thus, the neutral theory provides a null model for the rate of divergence of homologous genes or genome regions between isolated populations or species called the molecular clock hypothesis. This section will first present data that demonstrate the molecular clock and then show how the molecular clock hypothesis can be used to date evolutionary events based on DNA divergence. The section will conclude by showing why divergence may appear to decrease over time as many substitutions accumulate and how divergence estimates can be corrected using models of the mutation process.

As the clock metaphor suggests, the molecular clock hypothesis predicts that divergence accumulates with uniform regularity over time, just like the ticking of a clock. This means that the divergence between two species should increase as the time since they shared a common ancestor recedes further into the past. Such a pattern was originally observed for hemoglobin proteins by Zuckerkandl and Pauling (1962, 1965), who first hypothesized a molecular clock. A classic example of the molecular clock is the increase of divergence with increasing time

seen in the NS gene of the human influenza A virus (Fig. 8.12). Buonagurio et al. (1986) used influenza virus isolated from samples originally taken between 1933 and 1986. They then estimated the number of nucleotide substitutions, or the *p*-distance, between each sequence and the inferred ancestral sequence. The linear increase in divergence with time is the pattern expected by the molecular clock hypothesis.

**Molecular clock hypothesis** The neutral theory prediction that divergence should occur at a constant rate over time so that the degree of molecular divergence between species is proportional to their time of separation. Synonymous with rate constancy or rate homogeneity.

Another important early advance for the molecular clock came when Richard Dickerson (1971) compared rates of substitution in proteins from cytochrome *c*, hemoglobin, and fibrinopeptide genes and observed that the average rates of change were very different for the three proteins (Fig. 8.13). Based on knowledge of the function of the proteins at the time, Dickerson argued that the rate of molecular evolution was faster when fewer sites were subject to functional constraints on amino acid changes. That



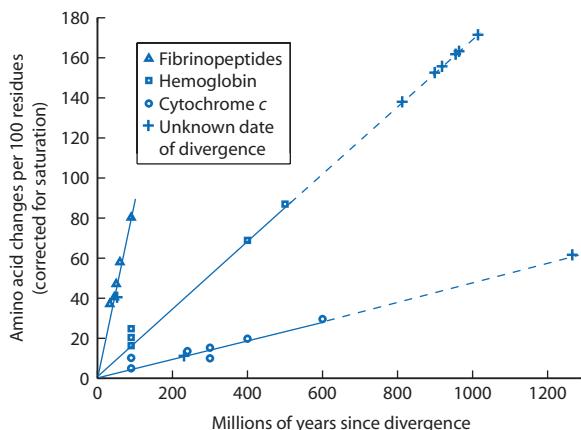
**Figure 8.12** Rates of nucleotide change in the NS gene that codes for “nonstructural” proteins based on 11 human influenza A virus samples isolated between 1933 and 1985. The number of years since isolation and DNA sequence divergence from an inferred common ancestor are positively correlated. The pattern of increasing substitutions over time since divergence is expected under the molecular clock hypothesis. The observed rate of substitution was approximately  $1.9 \times 10^{-3}$  substitutions per nucleotide site per year, a very high rate compared to most genes in eukaryotes. The line is a least-squares fit. Data from Buonagurio et al. (1986).

is, faster molecular evolution occurred when more sites were neutral and free to evolve by genetic drift. Slower molecular evolution occurred when a larger proportion of amino acid changes were eliminated by natural selection because they decreased or eliminated protein function. Thus, those sites that have not diverged in sequences compared among species may be constant due to selective constraint for function. Under this view, novel at the time, the portions of protein or DNA sequences that are invariant over time and shared among species indicate regions of

functional importance. The neutral theory served as a key concept to explain why different loci might have molecular clocks that tick at different rates.

#### Dating events with the molecular clock

A useful application of the molecular clock is to date divergence events between species. For some organisms, the fossil record and geological context provides a means to date when species originated, went extinct, or exhibited evolutionary transitions.



**Figure 8.13** (left) Rates of protein evolution as amino acid changes per 100 residues in fibrinopeptides, hemoglobin, and cytochrome *c* over very long periods of time. Rates of divergence are linear over time for each protein, as expected for a molecular clock. Different proteins have different clock rates due to different mutation rates and degrees of functional constraint imposed by natural selection. Amino acid changes between pairs of taxa with unknown divergence times are plotted on dashed lines with the same slope as lines through points for taxa with estimated divergence times. The six points with unknown divergence times for hemoglobin represent divergence of ancestral globins into hemoglobins and myoglobins in the earliest animals, events that the molecular clock estimates to have happened between 1.1 billion and 800 million years ago. Data from Dickerson (1971).

However, many types of organisms have no fossil record and not all phenotypes fossilize, presenting a problem for dating biological events. If the amount of DNA sequence divergence between two species is known, then this information can be utilized to date their divergence. The molecular clock hypothesis asserts that for neutral alleles the rate of substitution is simply the mutation rate, or  $k = \mu$ . If an absolute substitution rate expressed in fixations per time interval is available, multiplying that rate by a time gives an expected number of substitutions. The number of diverged nucleotide sites between two species also increases at twice the rate of substitutions since each lineage will experience substitutions independently. Bringing these two observations together gives the expected amount of divergence

$$k = 2T\mu \quad (8.32)$$

between two species that diverged  $T$  time units ago. If divergence between two species as well as the rate of divergence is known, this relationship can be rearranged to solve for the unknown of time instead:

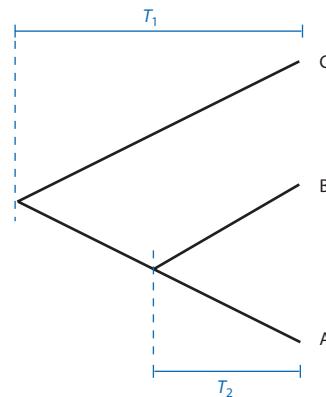
$$T = \frac{k}{2\mu} \quad (8.33)$$

**Absolute substitution rate** A rate of molecular change estimated in substitutions per year based on the combination of a sequence divergence estimate from two taxa and an estimate of the time that has elapsed since those taxa diverged.

Figure 8.14 illustrates a situation with three taxa and two divergence times. Dating events with the molecular clock requires that nucleotide divergences (adjusted for saturation) for each pair of species are estimated ( $K_{AB}$ ,  $K_{AC}$ , and  $K_{BC}$ ) and one divergence time is known. Imagine that our goal is to determine the time of divergence for species A and B ( $T_2$ ) given that divergence time  $T_1$  is known. The absolute rate of substitution that occurred over the known divergence time can be estimated:

$$\mu = \frac{1}{2} \left( \frac{K_{AC}}{2T_1} + \frac{K_{BC}}{2T_1} \right) \quad (8.34)$$

based on the average of the divergences observed between species pair A and C ( $K_{AC}$ ) and species pair B



**Figure 8.14** A schematic phylogenetic tree that can be used to date divergence events under the assumption of a constant rate of divergence over time or a molecular clock.  $T_1$  is the time in the past when species C and the ancestor of species A and B diverged.  $T_2$  is the time in the past when species A and B diverged. If either  $T_1$  or  $T_2$  are known, the rate of molecular evolution per unit of time can be estimated from observed sequence divergences. This rate of divergence can then be used to estimate the unknown amount of time that elapsed during other divergences.

and C ( $K_{BC}$ ) and the time of divergence  $T_1$ . This rate can then be used to solve for the unknown divergence time  $T_2$ :

$$T_2 = \frac{K_{AB}}{2\mu} \quad (8.35)$$

By substituting the definition for  $\mu$  in equation 8.34, equation 8.35 becomes

$$T_2 = \frac{2T_1 K_{AB}}{K_{AC} + K_{BC}} \quad (8.36)$$

As an example, imagine that DNA sequence divergences are estimated as  $K_{AB} = 0.10$ ,  $K_{AC} = 0.31$ , and  $K_{BC} = 0.36$  substitutions per site and the divergence time  $T_1$  is estimated with fossil and geologic data to be 10 million years. The interval  $T_2$  is then estimated as

$$\begin{aligned} T_2 &= \frac{2(10 \text{ million years})(0.10 \text{ substitutions per site})}{(0.31 \text{ substitutions per site})} \\ &\quad + (0.36 \text{ substitutions per site}) \\ &= 2.985 \text{ million years} \end{aligned} \quad (8.37)$$

The answer makes intuitive sense. The DNA divergence between species A and B is about one-third of the average DNA sequence divergence between species pairs A–B and A–C. Since A, B, and C diverged from

### Problem box 8.1 Estimating divergence times with the molecular clock

In the present day, dicotyledonous plants represent the majority of land plants. The divergence of ancestral seed plants into monocotyledonous and dicotyledonous plants was therefore a major evolutionary transition. Based on DNA divergence data for synonymous sites at nine mitochondrial genes in a range of plants (Laroche et al. 1995), a molecular clock can be used to date this event.

Table 8.2 gives DNA divergence data for comparisons of maize and wheat, both monocotyledons, with an estimated divergence time of approximately 60 million years ago. First, use the maize–wheat DNA divergence data to calibrate the absolute rate of substitution per million years for each locus. Then use this rate of change to estimate the time when monocots and dicots split given their degree of DNA sequence divergence. In terms of Fig. 8.14, the maize–wheat split is  $T_2$  and the monocot–dicot split is  $T_1$ .

**Table 8.2** DNA divergence data for comparisons of maize and wheat, both monocotyledons. See Problem box 8.1.

Locus	Nucleotide sites	Synonymous sites	Substitutions per site	
			Wheat–maize	Monocot–dicot
<i>coxl</i>	1461	495	0.0504	0.2060
<i>atp9</i>	195	67	0.1374	0.4439
<i>nad4</i>	1272	456	0.0381	0.1101

a common ancestor 10 million years ago, then  $T_2$  when A and B diverged is about one-third of that time. Assumptions in these time estimates are that rates of substitution are constant over time, among lineages, and over loci. These assumptions will be explored critically later in the chapter.

The molecular clock has been widely used to date major evolutionary transitions, establish times when the ancestors of many different organisms first evolved, and test hypotheses related to divergence times. One example is testing the hypothesis that early mammal evolution was facilitated by ecological niches that opened up when the dinosaurs went extinct. The molecular clock suggests that the earliest mammal lineages had appeared well before the extinction of the dinosaurs was complete (Bromham et al. 1999; Bininda-Emonds et al. 2007). Thus, the estimated divergence time rejects the hypothesis that mammals first originated in habitats left empty as dinosaurs disappeared.

A second illustration is the classic question of when humans and their close ancestors diverged. Using calibration times of 13 million years for the divergence between orangutans and humans and 90 million years between artiodactyls (hoofed mammals with an even number of digits, such as cattle, deer, and

pigs) and primates, as well as numerous loci, Glazko and Nei (2003) estimated the divergence of humans and chimpanzees occurred 5–7 million years ago.

A third example is employment of the molecular clock to date the origin of the human immunodeficiency virus or HIV. The date that HIV was transmitted from primates to humans is a critical question. Identifying related viruses in primates and the genetic features that facilitated transmission and virulence in humans would facilitate the development of treatments for HIV. One highly controversial hypothesis for the origin of HIV was that it was first spread to humans through contaminated human polio vaccine made from cultured chimpanzee cells that was administered in the former Belgian Congo between 1957 and 1960. This hypothesis is supported by indirect evidence, such as the timing and location of the earliest known cases of HIV/AIDS and the fact that the chimpanzee simian virus 40 (SV40) has been transmitted to humans. A molecular clock suggests that HIV was introduced to humans around 1920–30 and that HIV and the simian immunodeficiency virus (SIV) may have diverged as much as 300 years ago (Leitner & Albert 1999; Salemi et al. 2001). However, application of the molecular clock to HIV faces substantial challenges and interpretation

remains controversial due to issues such as rate variation and recombination (see Korber et al. 1998; Schierup & Hein 2000).

The use of the molecular clock to estimate times of divergence is complicated by numerous issues in practice (see review by Arbogast et al. 2002), contributing to both statistical uncertainty as well as controversy over interpretation of date estimates. First, calibration times usually have considerable ranges, leading to uncertainty in any divergence time estimated from the molecular clock. Then, corrections to divergence estimates are required for multiple substitutions occurring at the same nucleotide site. In addition, the rate of substitution is assumed to be constant over time. However, variation in rates of substitution over time and among different loci is now considered the rule rather than the exception, complicating the methods needed to estimate divergence times (see Culter 2000b; Glazko & Nei 2003). Testing for and explaining variation in substitution rates is explored in the next section.

#### 8.4 Testing the molecular clock hypothesis and explanations for rate variation in molecular evolution

- Rate heterogeneity in the molecular clock.
- The Poisson process model of the molecular clock.
- Ancestral polymorphism and the molecular clock.
- Relative rate tests of the molecular clock.
- Possible causes of rate heterogeneity.

The molecular clock predicts that selectively neutral homologous sequences (meaning sequences that were once identical by descent) with equal mutation rates should experience a similar number of substitutions per unit time as divergence increases. Therefore, the molecular clock hypothesis provides a null model to examine the processes that operate during molecular evolution. It is possible to directly test the molecular clock hypothesis and thereby test this null model. Rejecting the molecular clock hypothesis suggests that the sequences compared evolve at unequal rates, a situation referred to as **rate heterogeneity**. Rejecting the molecular clock hypothesis is a way to identify processes that influence the chance of substitution such that rates of fixation are either higher or lower than expected by genetic drift alone. For example, a previous section showed how natural selection changes the probability of fixation and therefore the rate of substitution. So, for example, one sequence taken from a population

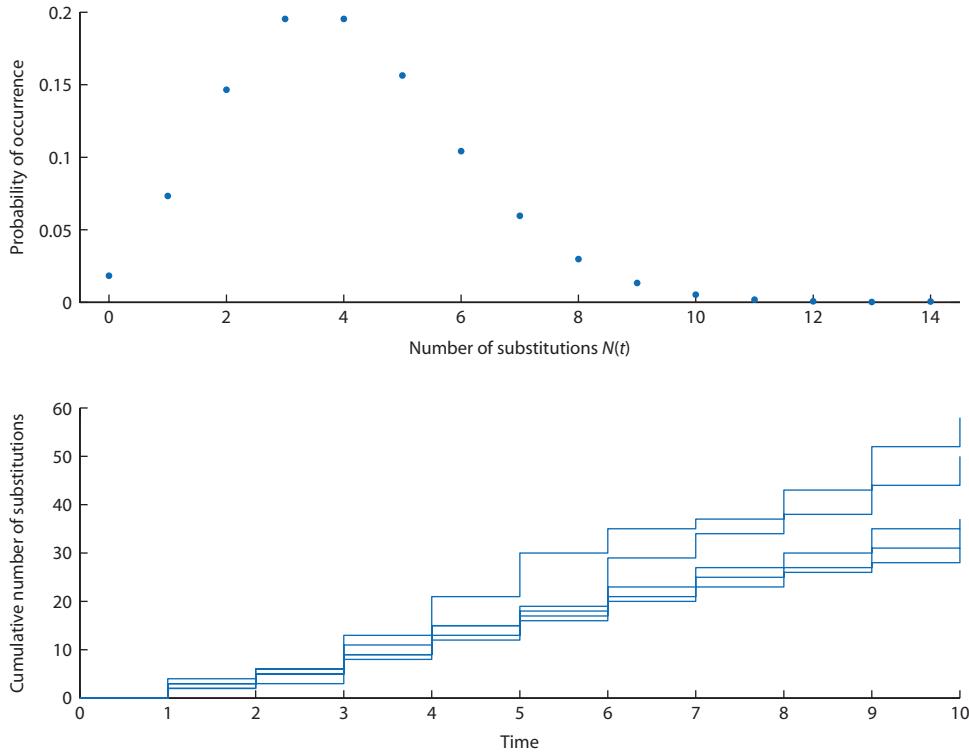
where most mutations are deleterious and selected against and another sequence taken from a population where most mutations are neutral would have different rates of substitution and show different numbers of substitutions over a fixed time interval (see Fig. 8.3). Thus, testing for equal rates of substitution, or **rate homogeneity**, is a useful step in identifying the processes that may be operating in molecular evolution.

**Rate heterogeneity** Variation in the rate of substitution over time or among different lineages for homologous genome regions.

#### *The molecular clock and rate variation*

Since the neutral theory leads to the molecular clock hypothesis, evidence for rate heterogeneity would appear to be evidence that genetic drift is not the main process leading to the ultimate substitution of most mutations. Rejecting the hypothesis of rate homogeneity would suggest that natural selection is operating on mutations such that their rates of substitution are either sped up or slowed down relative to substitution rates under genetic drift. The probability that a new mutation is fixed by natural selection depends on the selection coefficient,  $s$ , and the effective population size rather than just  $\frac{1}{2N_e}$  as it does for genetic drift. Natural selection is therefore very unlikely to produce a molecular clock because  $s$ ,  $N_e$ , and  $\mu$  are not likely to be constant through time or among different lineages. Before reaching the conclusion that natural selection explains all rate heterogeneity, however, it is necessary to dig deeper into the molecular clock hypothesis. The molecular clock is potentially more complex than was revealed at the beginning of the chapter. Understanding these complications is a necessary prerequisite to understanding the range of alternative hypotheses that may explain heterogeneity in rates of molecular evolution.

The molecular clock was originally proposed by Zuckerkandl and Pauling (1962, 1965; Zuckerkandl 1987) to model amino acid substitutions. It was based on a simple statistical method used to describe events that happen at random times given some rate at which events occur (such models are called point processes). The simplest point process for a



**Figure 8.15** Substitution patterns under a Poisson process. The top panel shows the probability distribution for the number of substitutions that might occur during one time interval.  $N(t)$  between 0 and 9 all have probabilities of greater than 0.01. The bottom panel shows the cumulative number of substitutions under a Poisson process for five independent trials. Each trial is akin to an independent lineage experiencing substitutions. The average number of substitutions is approximately 40 (four multiplied by the number of time intervals) but there is variation among the lineages. In both panels the rate of substitution is the same at  $\lambda = 4$ .

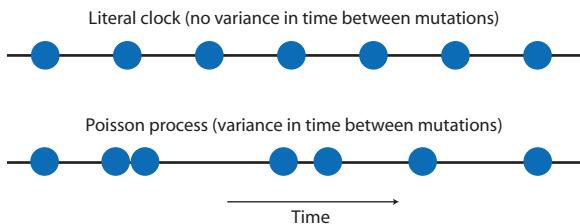
molecular clock is a Poisson process, a stochastic process which is defined in terms of the count of events,  $N(t)$ , since time was equal to zero. In a Poisson process the expected number of events between two times follows a Poisson distribution. Assuming that all substitutions are independent events, the probability of a substitution is very small, and the number of time intervals is very large, a Poisson clock gives the probability of observing some number of substitutions after a time period has elapsed as

$$\text{Probability } (N(t) \text{ substitutions at time } t) = \frac{e^{-\lambda t} (\lambda t)^{N(t)}}{N(t)!} \quad (8.38)$$

where  $N(t)$  is the total number of substitutions (an integer),  $t$  is time in years, and  $\lambda$  is the rate of substitutions per year. Under this model, the expected number of substitutions at time  $t$  is  $\lambda t$  or the product of the substitution rate and the number of time steps that have elapsed. A critical thing to notice in this model is that the rate of substitutions,  $\lambda$ , is constant and does not change with time nor with the total number of substitutions,  $N(t)$ . The Poisson

molecular clock is illustrated in Fig. 8.15. The top panel shows the probability that  $N(t)$  is between zero and 14 for one time step when the rate of substitution is  $\lambda = 4$ . The bottom panel of Fig. 8.15 shows variation in the number of substitutions among five replicate sequences that all evolved for the same period of time at the same constant rate of substitution ( $\lambda = 4$ ).

The Poisson model for a molecular clock implies that the time intervals between substitutions are random in length (Fig. 8.16). Thus, substitutions that follow a Poisson molecular clock will be separated by variable lengths of time. This stands in contrast to our everyday notion of a clock or watch, which has uniform lengths of time separating each event (events are seconds, minutes, and hours). Therefore, a molecular clock that is based on a random process has inherent variation in the number of substitutions that occur over a given time interval even though the rate of substitution remains constant. This means that independent lineages that each diverged from a common ancestor at the same time can display variation in the number of substitutions that have



**Figure 8.16** Two representations of rate at which substitution events (circles) occur over time. Mutations might occur with metronome-like regularity, showing little variation in the time that elapses between each mutation event. If substitution is a stochastic process, an alternative view is that the time that elapses between substitutions is a random variable. The Poisson distribution is a commonly used distribution to model the number of events that occur in a given time interval, so the bottom view is often called the Poisson molecular clock. Note that in both cases the number of substitutions and time elapsed is the same so that the average substitution rate is identical.

occurred. In other words, if substitution is a random process then we expect some variation in the number of substitutions among lineages and loci even if divergence time and the substitution rate are constant. This explanation for variation in the numbers of substitutions is often referred to by saying that rates of molecular evolution follow a **Poisson clock**.

The Poisson process model of the molecular clock leads to a specific prediction about the variation in numbers of substitutions that should be observed if the rate of molecular evolution follows a Poisson process. The Poisson distribution has the special property that the mean is equal to the variance. Therefore, the mean number of substitutions and the variance in the number of substitutions should be equal for independent DNA sequences evolving at the same rate according to a Poisson process. A ratio to compare the mean and variance in the number of substitutions is called the **index of dispersion**, defined as

$$R(t) = \frac{\text{variance } N(t)}{E(N(t))} \quad (8.39)$$

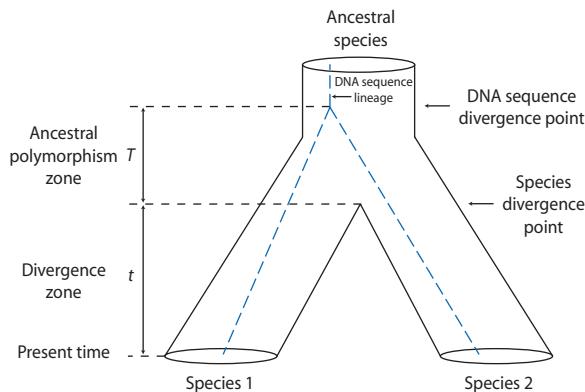
where  $E$  indicates an expected or mean value. The index of dispersion defines the degree of spread among divergence estimates that should be seen under a Poisson process, just as a Markov chain defines the spread in allele frequencies that is expected in an ensemble of finite populations. If the numbers of substitutions in a sample of pairwise sequence divergences follow a Poisson molecular clock then the variance and mean number of substitutions should

be equal and therefore  $R(t)$  should equal one. If the variance is larger than the mean then  $R(t)$  is greater than one, a situation referred to as an **overdispersed molecular clock** since substitution rates have a wider range of values than predicted by the Poisson process model.

**Overdispersed molecular clock** Absolute divergence rates from many independent pairs of species that overall exhibit more variance in divergence rate than expected by the Poisson process molecular clock model; a dispersion index value that is greater than one.

#### Ancestral polymorphism and Poisson process molecular clock

The molecular clock modeled as a Poisson process assumes it is possible to compare pairs of DNA sequences that were derived from a single DNA sequence in the past and then diverged instantly into two completely isolated species. Actual DNA sequences usually have a more complex history that involves processes that operated in the ancestral species followed by the process of divergence in two separate species (Fig. 8.17). In the ancestral species the number and frequency of neutral alleles per locus in the population were caused by population processes such as genetic drift and mutation (assuming the ancestral species was panmictic). This zone of **ancestral polymorphism** is the period of time when genetic variation in the ancestral species was dictated by drift–mutation equilibrium. Within this ancestral population, two lineages split at some point and eventually became lineages within the two separate species (Fig. 8.17). DNA sequences from these two lineages were sampled in the present to estimate substitution rates. Recognizing this more complex history of diverged DNA sequences shows two things. First, it points out that lineages and species may have diverged at different points in time, with lineages often diverging earlier than species. Second, it shows that two distinct processes can contribute to the nucleotide differences between sequences seen as substitutions when observed in the present. Referring to Fig. 8.17, during the time period  $T$ , polymorphisms among sequences were caused by the population processes dictating polymorphism in the ancestral species. Later, during the time period  $t$ ,



**Figure 8.17** An illustration of the history of two DNA sequences that might be sampled from two species in the present time to estimate the rate of substitutions. The history is like a water pipe in an upside-down Y shape. The tube at the top contains the total population of lineages in the ancestral species, eventually splitting into populations of lineages that compose two species. The time when two lineages diverged from a common ancestor is not necessarily the same as the time of speciation. Therefore, a population process governing polymorphism operates for  $T$  generations in the ancestral species while a divergence process operates for  $t$  generations in the diverged species. The polymorphism process initially dictates the number of nucleotide changes between two sequences. Later, the divergence process dictates the number of nucleotide changes between two sequences. In two DNA sequences sampled in the present it is impossible to distinguish which process has caused the nucleotide changes observed.

substitutions were the product of the divergence process between species.

The existence of both ancestral polymorphism and divergence processes complicates testing for overdispersion of the molecular clock (Gillespie 1989, 1994). To make this point, Gillespie articulated the distinction between **origination processes** and **fixation processes**. An origination process describes the times at which the subset of new mutations that will ultimately fix first enter the population. A fixation process, in contrast, describes the times at which the subset of new mutations that will ultimately fix reach a frequency of 1 in the population. At a conceptual level, it is clear that the two processes are not identical. The distribution of origination times is a product of the causes of mutation. Times until fixation depend on both the causes of mutation (the origination process) as well as on the causes of fixation, such as genetic drift in a finite population of neutral alleles or natural selection. Measuring times until fixation of new mutations would require that we are able to follow the populations of diverging species over time and watch as new mutations

segregate and eventually go to fixation and loss, recording the times for those that fix and then calling these the substitution times. In Fig. 8.2 originations are the events at the bottom of the  $y$  axis and fixations are events at the top of the  $y$  axis. In practice, we have only the accumulated amino acid or DNA differences between pairs of species observed at one point in time. Such sequence differences are a product of the origination process because they are a sample of mutations that came into the population some time ago and have fixed by the time we observe them. This is not the same thing as having observed the “tick” of fixations over a long period of time.

Gillespie and Langley (1979) showed that a molecular clock combining polymorphism and divergence does not necessarily comprise a Poisson process where the index of dispersion is expected to equal one. To see this it is necessary to develop expectations for the mean and variance in the number of nucleotide differences between two sequences when both polymorphism and divergence processes are operating over the history of two DNA sequences.

Assuming that the ancestral species is panmictic, there is no recombination, and there is selective neutrality, Watterson (1975) showed that the expected number of segregating sites ( $S$ ) for a sample of two DNA sequences is

$$E(S) = \theta \quad (8.40)$$

and that the variance in the number of segregating sites is

$$\text{Variance}(S) \approx \theta + \theta^2 \quad (8.41)$$

under the infinite sites model of mutation where  $\theta = 4N_e\mu$ . (The relationship between  $\theta$  and the number of segregating sites is derived in Section 8.2. Notice that the factor of  $a_1$  is not shown in equation 8.40 because  $a_1 = 1$  for a sample of two sequences.) This is a prediction for the amount of polymorphism expected under neutrality in a finite population. This result tells us that the mean and variance of the number of nucleotide sites are expected to be different in a sample of two DNA sequences from the ancestral polymorphism zone in Fig. 8.17.

Now shift focus to the divergence zone of Fig. 8.17. For one DNA sequence from species 1 and another from species 2, the divergence time is  $2t$  because each species has diverged independently for  $t$  generations. Based on the Poisson process in equation 8.38, both the expected number of diverged sites and the

**Table 8.3** Mean and variance in the number of substitutions at a neutral locus for the cases of divergence between two species and polymorphism within a single panmictic population. The rate of divergence is modeled as a Poisson process so the mean is identical to the variance. The mutation rate is  $\mu$  and the  $\theta = 4N_e\mu$ . Refer to Fig. 8.17 for an illustration of divergence and ancestral polymorphism.

	Expected value or mean	Variance
Ancestral polymorphism	$\theta$	$\theta + \theta^2$
Divergence	$2t\mu$	$2t\mu$
Sum	$2t\mu + \theta$	$2t\mu + \theta + \theta^2$

variance in the number of diverged sites are  $2\mu t$ . The means and variances in the number of nucleotide sites between two sequences are shown in Table 8.3 as they apply to polymorphism and divergence in Fig. 8.17.

Given the means and variances of the number of changes between two DNA sequences for both polymorphism and divergence processes, we can then combine these expectations into a new expression for the index of dispersion. The index of dispersion

for the number of differences between a sequence from species 1 and a sequence from species 2 is then

$$R(t) = \frac{\text{variance } N(t)}{E[N(t)]} = \frac{2\mu t + \theta + \theta^2}{2\mu t + \theta} \quad (8.42)$$

where  $\theta = 4N_e\mu$ . (This requires the assumptions that the time of lineage divergence  $T$  is a random variable with a geometric distribution as in a genealogical branching model, and that  $N_e$  is large and  $\mu$  is small.) As shown in Math box 8.1, this new version of the dispersion index can be re-written as

$$R(t) = 1 + \frac{\theta^2}{E[N(t)]} \quad (8.43)$$

If we assume that there is no ancestral polymorphism or that  $T = 0$  in Fig. 8.17, then  $\theta$  is zero and  $R(t)$  is identical to what is expected under the Poisson process molecular clock dictated by divergence only.

The major conclusion is that the two-process version of  $R(t)$  is expected to be greater than one when there is any ancestral polymorphism even when DNA changes follow a constant molecular clock. Stated another way, ancestral polymorphism increases the variance in the number of substitutions seen for pairs of sequences evolving under a constant rate compared with a pure divergence process. Unfortunately, the index of dispersion in equation 8.43 seems impossible

### Math box 8.1

#### The dispersion index with ancestral polymorphism and divergence

Define a new variable  $\alpha = \frac{t}{2N_e}$  to scale time by  $2N_e$  generations. Then notice that

$$\theta\alpha = 4N_e\mu \frac{t}{2N_e} = 2\mu t \quad (8.44)$$

The numerator and denominator of

$$R(t) = \frac{2\mu t + \theta + \theta^2}{2\mu t + \theta} \quad (8.45)$$

can then be re-written as functions of  $\alpha$ ,

$$R(t) = \frac{\theta(1 + \alpha) + \theta^2}{\theta(1 + \alpha)} \quad (8.46)$$

and rearranged to

$$R(t) = \frac{\theta(1 + \alpha)}{\theta(1 + \alpha)} + \frac{\theta^2}{\theta(1 + \alpha)} \quad (8.47)$$

The expected number of DNA changes at time  $t$  is  $2\mu t + \theta$  (see Table 8.3), which is equivalent to  $\theta(1 + \alpha)$ . The index of dispersion can then be written

$$R(t) = 1 + \frac{\theta^2}{\theta(1 + \alpha)} = 1 + \frac{\theta^2}{E[N(t)]} \quad (8.48)$$

to estimate in practice because  $\theta$  cannot be estimated in the ancestral species as it does not exist any longer. However, the main point of this model is not to provide a practical test of the molecular clock. Instead, the model shows how  $R(t) > 1$  is not necessarily strong evidence to reject a constant rate of substitution. One cause of  $R(t) > 1$  is that the Poisson process accurately describes the substitution process but that substitution rates are not constant. Alternatively, the Poisson process of model of divergence itself may not be accurate even though the rate of DNA change is constant. The latter possibility suggests that the index of dispersion may be a poor way to test the neutral molecular clock hypothesis.

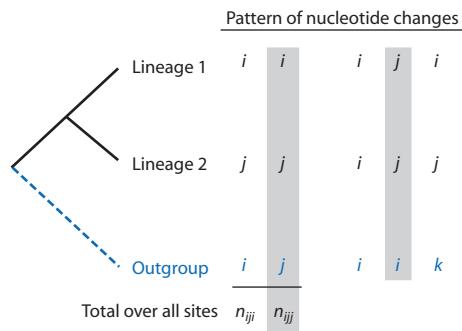
Ancestral polymorphism also presents difficulties for dating divergences using the molecular clock (Maddison 1997; Arbogast et al. 2002). The problem arises because sequence lineage history (genealogy) and species divergence history (species phylogeny) are not identical. Two sequences sampled in the present from two different species have been accumulating substitutions since the most recent common ancestor of the two sequences gave rise to the lineages (Fig. 8.17). The total sequence divergence between two species that would be used to date a speciation event has occurred during two distinct time intervals. One time interval  $T$  is the period when the two lineages accumulated changes in the ancestral species. The second time interval  $t$  is the period when substitutions accumulated after the current species split. Estimates of time since divergence estimate the total elapsed time since the divergence of the two lineages rather than just the time since divergence of the two species. Thus, the use of the molecular clock to date divergence time yields over-estimates of the species divergence time. As the divergence time  $t$  increases relative to the polymorphism time  $T$  the degree of over-estimation shrinks. However, it is usually impossible to determine  $t$  relative to  $T$  in practice and so the degree of over-estimation of the species divergence time is usually unknown.

#### **Relative rate tests of the molecular clock**

One method to circumvent some of the limitations inherent in comparing absolute rates of divergence is to compare relative rates instead. The **relative rate test** compares the number of nucleotide or amino acid changes since divergence from an ancestor represented by a DNA sequence from closely related species (Sarich & Wilson 1967; Fitch 1976). Rates of nucleotide substitution in two different species can

be estimated by comparing the number of DNA or amino acid changes that have occurred independently in each of two species using a third outgroup species to assign sequence changes to each lineage. If rates of substitution are equal in the two species, then the number of sequence changes should be equal in the two species within a statistical confidence interval. Unequal numbers of sequence changes lead to rejection of the null hypothesis that the two species have an equal rate of substitution. Relative rate tests avoid the need for a date of divergence that is often imprecise and also do not rely on the dispersion index and its underlying assumption that the molecular clock is a simple Poisson process.

Tajima's (1993a) 1D test of the molecular clock is a relative rate test that uses the number of nucleotide substitutions that occurred along two lineages being compared as well as an outgroup lineage. The basis of the test is shown in Fig. 8.18. In the figure, the letters  $i$ ,  $j$ , and  $k$  are used to represent the identity of the nucleotide found at the same nucleotide site in each of the three sequences. The outgroup is used to identify the point in time that nucleotide changes took place since lineages 1 and 2 should share the same base pair as the outgroup due to identity by descent if no substitution has occurred. Only changes that can be assigned unambiguously to a lineage are useful when comparing rates between lineages 1 and 2. Nucleotide substitutions of the pattern  $iji$



**Figure 8.18** Patterns of nucleotide changes that are possible when comparing DNA (or amino acid) sequences from two lineages and an outgroup. The letters  $i$ ,  $j$ , and  $k$  are used to represent the identity of the nucleotide found at the same nucleotide site in each of the three sequences. For example,  $iji$  indicates that the first two lineages have an identical base pair and the third lineage has a different base pair. Tajima's 1D relative rate test utilizes substitutions that can be unambiguously assigned to one lineage ( $iji$  and  $ijj$ ). If rates of substitution are identical for lineage 1 and 2,  $E(n_{iji}) = E(n_{ijj})$ . The lineages where the substitution took place are ambiguous for the patterns  $jjj$  and  $ijk$ . The pattern  $iii$  indicates identical nucleotides in all three sequences and therefore no substitution events.

indicate the change occurred on lineage 2 whereas pattern  $ijj$  indicates the change occurred on lineage 1. These two instances allow unambiguous assignment of a substitution to a lineage to estimate the numbers of substitutions. The other three possible nucleotide patterns cannot be used to estimate rates of substitution for one lineage. Nucleotide sites with the pattern  $iii$  are not useful because no substitution occurred and there is no information available to estimate the rate of change. For nucleotide sites with the pattern  $jji$ , the substitution to  $j$  could have occurred in the ancestor to lineages 1 and 2 or both lineages 1 and 2 could have experienced a substitution but it is not clear which event occurred. The pattern  $ijk$  for a nucleotide site indicates that no two lineages share a nucleotide, so again it is unclear at what point in the past these substitutions occurred and they cannot be used to estimate the rates of substitution for lineages 1 and 2.

Under the molecular clock hypothesis, the number of substitutions that occurred on lineage 1 should be identical to the number of substitutions that occurred on lineage 2. Since the divergence time is identical for lineages 1 and 2, identical substitution rates for the two lineages would give the same number of substitutions observed on each lineage. Therefore, the number of substitutions observed for sequence 1 that occurred on lineage 1 ( $ijj$ ) should be equal to the number of substitutions observed for sequence 2 that occurred on lineage 2 ( $iji$ ):

$$E(n_{ijj}) = E(n_{iji}) \quad (8.49)$$

where  $E$  means expected or average value,  $n_{ijj}$  is the total number of nucleotide substitutions that occurred on lineage 1, and  $n_{iji}$  is the total number of nucleotide substitutions that occurred on lineage 2. This expectation can be tested with the Chi-squared statistic

$$\chi^2 = \frac{(n_{ijj} + n_{iji})^2}{n_{ijj} + n_{iji}} \quad (8.50)$$

where there is one degree of freedom. A Chi-squared value greater than 3.84 indicates that it is unlikely that the difference in the number of substitutions between the two lineages is due to chance. In other words, a large Chi-squared value is evidence to reject the molecular clock hypothesis that substitution rates are equal for the two lineages and is evidence of rate heterogeneity. The Chi-square approximation is accurate as long as  $n_{ijj}$  and  $n_{iji}$  are both 6 or greater.

Tajima's 1D test for equal divergence rates in two taxa is simple to employ because it does not require an explicit nucleotide substitution model. Hamilton et al. (2003) took advantage of this aspect of the 1D test when they compared rates of divergence using both nucleotide and insertion/deletion (indel) variation between species of Brazil nut trees. (Because a range of molecular mechanisms leads to the formation of indels, there are no generally employed models of sequence change by indels and many relative rate tests cannot be used with indel variation.) Comparing substitution rates among eight species with the 1D test, they found that two tree species consistently failed to support a molecular clock for both nucleotide and indel changes. One species (*Lecythis zabucajo*) had an accelerated rate of substitution whereas the other species (*Eschweilera romeu cardosoi*) had a slowed rate of substitution.

Relative rate tests provide no information about rates of molecular evolution in the outgroup taxon nor any information about absolute rates of DNA sequence change. The outcome of relative rate tests depends critically on the outgroup used (Bromham et al. 2000). As the time since divergence of the common ancestor of both taxa and the outgroup increases, so does the time over which the evolutionary rates are averaged. If rate heterogeneity is a short-term or recent phenomenon then averaging from a distant outgroup may obscure it. Conversely, if rate heterogeneity is only apparent over long time periods, the rate of substitution may appear homogeneous if a recently diverged outgroup is employed. Finally, since natural selection depends on population-specific fitness values, it is considered unlikely that selection acting simultaneously on both lineages subject to a relative rate test would result in rate homogeneity.

Three-taxon relative rate tests that incorporate nucleotide-substitution models and use a maximum likelihood framework are described in Gu and Li (1992) and Muse and Weir (1992). A variety of relative rate tests that utilize phylogenetic trees are also available that test the molecular clock hypothesis using sequences from many taxa simultaneously (see Nei & Kumar 2000; Page & Holmes 1998).

### **Patterns and causes of rate heterogeneity**

Ohta and Kimura (1971) were the first to carry out a test of the Poisson process molecular clock with rigorous statistical comparisons. They used protein sequences from three loci ( $\beta$  globin,  $\alpha$  globin, and cytochrome  $c$ ) sampled from a range of species.

Based on the observed divergences between pairs of sequences and estimates of the time that has elapsed since those species diverged, they estimated a series of absolute rates of divergences. These absolute rates varied widely (the dispersion index for their data falls between 1.37 and 2.05), leading them to reject the hypothesis of a constant molecular clock (see Gillespie 1991). A few years later, Langley and Fitch (1974) published a larger analysis of absolute substitution rates for the same three loci as well as fibrinopeptide A and used phylogenies to better estimate the number of substitutions for each species. They too found that the dispersion index was greater than one for all loci. These papers attracted a great deal of attention because the variation in rates of sequence change required explanation. Since these early results, a great deal of data on both absolute and relative rates of molecular evolution show clearly that rates of molecular evolution are commonly more variable than expected by a Poisson process model. In fact, rate heterogeneity may now be considered the norm and a constant rate of molecular evolution the exception. This section focuses on hypotheses to explain variation in rates of molecular evolution.

Under neutrality, variation in the rate of divergence at different loci can be explained by differences in rates of mutation. Similarly, variable rates of divergence at the same locus in different species can be explained by different mutation rates among species. Such variation in rates of molecular evolution for the same locus in different species is called a **lineage effect** on the molecular clock (Gillespie 1989). There may be rate heterogeneity evident at a locus even after accounting for variation among lineages, called **residual effects** (reviewed in Gillespie 1991). Residual effects are the variation in the rate of divergence or unevenness in the tick rate of the molecular clock *within* lineages over time (see Fig. 8.16). Residual effects are sometimes described as a pattern where substitutions occur in bursts or clusters with periods of no change in between. The cause of residual effects must be a process that is changing over time within a lineage. Under the neutral theory, the mutation rate within a lineage must change over time to explain residual effects. While temporally variable mutation rates are possible, it is considered more plausible that mutation rates themselves are constant over time but that substitution rates are variable over time. For example, the presence and absence of natural selection would cause changes in the probability of substitution of the mutations that appear constant over time.

**Lineage effect** Variation in the rate of divergence among multiple species that could be explained by the different lineages having variable neutral mutation rates.

**Replication-independent causes of mutation** Causes of mutation that can occur at any time and are therefore independent of the rate of cell division. Examples include environmental mutagens such as ultraviolet radiation,  $\gamma$  particles, and chemicals.

**Residual effect** Variation or unevenness in the rate of divergence within a lineage that cannot be explained by rate heterogeneity among lineages or loci.

Kimura (1983a) argued that mutation rates in different species are roughly constant per year. This could be true if the processes that caused mutations were constant over time units like years. Examples are **replication-independent** causes of mutation such as exposure to ultraviolet radiation,  $\gamma$  particles, or chemical mutagens. The free radical ions constantly produced within cells are another example of a replication-independent cause of mutation. It seems likely that exposure to these extrinsic causes of mutation is constant over calendar time and so a portion of mutations due to replication-independent causes have a rate that is also set in calendar time.

Returning to the basis of the neutral theory shows why different species might not experience substitutions at the same rates. As shown at the beginning of the chapter, neutral theory predicts that the substitution rate is equal to the mutation rate. But since the mutation rate is measured in nucleotide changes per generation, then the substitution rate is also expressed in per generation terms. This leads to the difficulty that a constant molecular clock might not exist if species differ in their generation times. As an example, imagine two species with identical mutation rates of  $\mu = 1 \times 10^{-5}$  errors per base pair per generation. Now imagine the species have generation times of 10 and 100 years. The species with the shorter generation time has

$$\begin{aligned}\mu &= \frac{1 \times 10^{-5} \text{ mutations generation}^{-1}}{10 \text{ years generation}^{-1}} \\ &= 1 \times 10^{-6} \text{ mutations per year}\end{aligned}\quad (8.51)$$

whereas the species with the longer generation time has

$$\mu = \frac{1 \times 10^{-5} \text{ mutations generation}^{-1}}{100 \text{ years generation}^{-1}} \\ = 1 \times 10^{-7} \text{ mutations per year} \quad (8.52)$$

Thus, the constant molecular clock per generation predicted by neutral theory can produce variable rates of substitution per year when comparing species with different generation times.

The observation that neutral mutation rates that are constant per generation may simultaneously be variable per year leads to the **generation-time hypothesis**, a neutral theory explanation for variation in rates of substitution as caused by differences in generation times of species that have constant rates of substitution per generation. Numerous studies have shown evidence for a generation time effect in rates of substitution (Li et al. 1987, 1996; Ohta 1993, 1995). Substitution rates observed over many nuclear genes in different groups of mammals are shown in Table 8.4. Rodents have shorter generation times than primates and artiodactyls. Substitution rates are also negatively correlated with generation times. In contrast, comparisons within these groups, such as comparing rates of substitution between mice and rats, shows nearly equal substitution rates. The rate's speeding up in rodents compared to primates and artiodactyls is a classic example of the generation time effect and is consistent with a neutral explanation for heterogeneity in the rate of molecular evolution.

**Table 8.4** Number of substitutions per nucleotide site observed over 49 nuclear genes for different orders of mammals. Divergences are divided into those observed at synonymous and nonsynonymous sites. Primates and artiodactyls (hoofed mammals such as cattle, deer, and pigs with an even number of digits) have longer generation times than do rodents. There were a total of 16,747 synonymous sites and 40,212 nonsynonymous sites. Data from Ohta (1995).

Mammalian group	Synonymous sites	Nonsynonymous sites
Primates	0.137	0.037
Artiodactyls	0.184	0.047
Rodents	0.355	0.062

**Generation-time hypothesis** The hypothesis that variation in rates of substitution is due to differences in generation times among species that have constant rates of substitution per generation. This explanation for rate heterogeneity is consistent with neutral molecular evolution.

**Replication-dependent causes of mutation** Causes of mutation that occur during replication of DNA, such as replication errors, so that the rate of mutation depends on the rate of cell division.

A generation-time effect can be explained by **replication-dependent** causes of mutation. If mutations occur mostly during the process of cell division when chromosomes are replicated, then more cell replications per generation leads to a higher rate of neutral divergence per generation. In animals, variation in replication-dependent mutation rates per generation may be explained by the fixed number of cell divisions leading to germ-line cells (cells that produce gametes). This explains the observation that mutations occur more frequently in male gametes than in female gametes since more germ-line cell divisions occur in males than in females. The generation-time effect in animals could then be explained if generation times are correlated with the number of germ-line cell divisions (e.g. animals with longer generation times have more germ-line cell divisions). Yet plants with shorter time intervals to first flowering have been shown to have higher rates of substitution (Gaut 1998; Kay et al. 2006). Variation in rates of molecular evolution in plants suggests that germ-line cell divisions are not the only explanation for rate heterogeneity because plants do not have separate germ and somatic cell lines.

The **metabolic rate hypothesis** proposed by Martin and Palumbi (1993; reviewed by Rand 1994) was based on the observation that sharks have rates of synonymous substitution five to seven times lower than those observed in primates and artiodactyls despite the fact that all taxa examined have relatively similar generation times. Mutation rates may be correlated with metabolic rate of organisms for several reasons. Organisms with high metabolic rates have rapidly operating cellular functions and one of these cellular functions is DNA replication. Therefore, high rates of metabolism cause high rates of DNA replication and high rates of replication-dependent

mutation. Alternatively, aerobic respiration within cells produces free oxygen radicals that cause oxidative damage to DNA. Therefore, high metabolism increases the rate of exposure to mutagenic agents and increases replication-independent rates of mutation. These two mechanisms that couple metabolic rate and mutation rates are not mutually exclusive and both can occur at the same time. Gillooly et al. (2005) proposed a model of the substitution rate that explicitly includes effects of body size and temperature and suggested that the molecular clock may indeed be constant after accounting for rate variation from these causes.

Heterogeneity in the rate of divergence can also be explained by the nearly neutral theory (reviewed by Ohta 1992). To see this, let  $f_0$  stand for the proportion of mutations that are selectively neutral because the pressure of negative selection acting on them is weak relative to the effective population size. (All mutations are assumed to be deleterious and advantageous mutations so rare they can be ignored, an assumption of the nearly neutral theory that is problematic (see Gillespie 1995).) The remaining  $(1 - f_0)$  mutations have a large enough deleterious effect that they are acted on by negative selection and are not neutral. The rate of substitution of neutral mutations under the nearly neutral theory is then

$$k = f_0 \mu \quad (8.53)$$

analogous to equation 8.3 for neutral theory. This equation says that the rate of divergence will be higher when more mutations are effectively neutral ( $f_0$  is larger) and lower when fewer mutations are effectively neutral ( $f_0$  is smaller).

Because the proportion of mutations that are effectively neutral depends on the effective population size, the rate of divergence also varies with the effective population size under the nearly neutral theory. In nearly neutral theory all substitutions are the result of genetic drift, as in the neutral theory. But a larger effective population size leads to fewer mutations that are effectively neutral and therefore a smaller pool of neutral mutations that can ultimately reach fixation. In contrast, a smaller effective population size leads to more mutations being effectively neutral and therefore a larger pool of neutral mutations that can ultimately experience substitution. Thus, nearly neutral theory predicts that the rate of divergence is negatively correlated with the effective population size because changes in  $N_e$  result in changes in  $f_0$ . Under the nearly neutral theory, rate

heterogeneity can then be explained by different effective population sizes among lineages or loci that cause  $f_0$  to vary.

Generation time may also influence perceived variation in substitution rates among species under the nearly neutral theory. In the nearly neutral theory, both mutation and substitution rates are expressed in per-generation terms as they are in the neutral theory. This should lead to generation-time effects on the substitution rate just as in the neutral theory. However, generation time effects may be cancelled out under the nearly neutral theory because of a negative correlation between generation time and effective population size. In the nearly neutral theory, the proportion of mutations that are effectively neutral depends on the effective population size. Independently, longer generation times lead to fewer substitutions per year whereas shorter generation times result in more substitutions per year if the mutation rate is constant per generation. The effective population size and the generation time should act independently on substitution rates. However, it turns out that generation time and effective population size are not generally independent (Chao & Carr 1993). For example, mice have short generation times and a large effective population size whereas elephants have long generation times and a small effective population size. Therefore, the impacts of generation time and effective population size tend to cancel each other out, resulting in a nearly neutral theory prediction that substitution rates do not show a generation time effect.

It is possible to test the nearly neutral theory prediction that the impacts of generation time and effective population size on rates of molecular evolution tend to cancel each other out. Ohta (1995) carried out such a test by comparing rates of substitution at synonymous and nonsynonymous sites for 49 genes in primates, artiodactyls, and rodents (recall from earlier in this section that divergence rates for these same animals support the generation time hypothesis). Ohta divided the DNA sequence data into divergence observed at synonymous or nonsynonymous sites within exons. Mutations at nonsynonymous sites are exposed to natural selection since they alter the amino acid sequence of a protein and therefore have a phenotypic effect. In contrast, synonymous site mutations are not perceived by natural selection (or selection is much weaker) since they do not alter the amino acid sequence. The nearly neutral theory predicts that nonsynonymous substitution rates should be lower than synonymous

substitution rates because of negative selection on the pool of nonsynonymous mutations (nonsynonymous  $f_0$  is smaller). In addition, divergence rates at nonsynonymous sites should not exhibit a generation time effect because of the negative correlation between generation time and effective population size for mutations that are nearly neutral. Table 8.4 shows Ohta's divergence data for primates, artiodactyls, and rodents at synonymous and nonsynonymous sites. Synonymous substitution rates are an order of magnitude greater than nonsynonymous rates, as expected if nonsynonymous sites experience frequent negative selection against mutations. The synonymous rate is 2.59 times faster for rodents than for primates. In contrast, the nonsynonymous rate is 1.68 times faster for rodents than for primates. Therefore, divergence at nonsynonymous sites shows less of a generation-time effect, also consistent with nearly neutral theory.

### 8.5 Testing the neutral theory null model of DNA sequence evolution

- The Hudson–Kreitman–Aguadé (HKA) test.
- The McDonald–Kreitman (MK) test.
- Tajima's  $D$  statistic.
- Mismatch distributions.

This section provides the opportunity to apply the conceptual results of the neutral theory developed earlier in the chapter to test the neutral null model for the causes of molecular evolution. Some tests take advantage of the neutral theory predictions for levels of polymorphism and divergence while others rely on coalescent model results that were developed in earlier chapters. These tests have been widely employed in empirical studies of DNA sequences sampled from a wide array of loci, genomes, and species (see review by Ford 2002). The tests described in this section have contributed much to our knowledge of how natural selection has acted on DNA sequences as well as our understanding of how multiple population genetic processes (mating, gene flow, genetic drift, mutation, changes in  $N_e$ , and natural selection) interact in natural populations.

#### **HKA test of neutral theory expectations for DNA sequence evolution**

The HKA test, so named after its authors Hudson, Kreitman, and Aguade (Hudson et al. 1987), is a test that compares neutral theory predictions for

DNA sequence evolution with empirically estimated polymorphism and divergence. The test utilizes the expectation that under neutrality both polymorphism within species and divergence between species are a product of the mutation rate. In fact, under neutral evolution levels of polymorphism and divergence at a locus should be correlated because they are both products of the very same mutations. If a locus has a high mutation rate, for example, then the population should be highly polymorphic (see equation 8.1). At the same time, divergence at that locus when compared with another species should also be substantial since the rate of substitutions that contribute to divergence is also equal to the mutation rate (see equations 8.2 and 8.3). Alternatively, a combination of both low polymorphism and low levels of divergence should be apparent if a neutral locus has a low mutation rate. In this way, expected levels of polymorphism and divergence are not independent under neutrality. Evidence that divergence and polymorphism are not correlated would be at odds with neutral expectations and therefore evidence to reject the neutral null model for the locus under study.

The HKA test requires DNA sequence data from two loci. One locus is chosen because it is selectively neutral and serves as a reference or control locus. Examples of a neutral reference locus include non-coding regions of the genome or duplicate copies of genes that are not functional (pseudo-genes), both of which are expected to be relatively free of functional constraints on nucleotide substitutions. The other locus used is the focus of the test and the locus for which the neutral null model of evolution is being tested.

The HKA test also requires that DNA sequence data for two loci be collected in a particular manner. First, DNA sequences for two loci must be obtained from two species to estimate divergence between the species for both the neutral reference and the test loci. In addition, DNA sequences from multiple individuals within one of the species need to be obtained to estimate levels of polymorphism present at both loci. Polymorphism is measured by nucleotide diversity ( $\pi$ ) for each locus. Divergence is estimated by comparing the DNA sequences for both loci between an individual of each species, employing a nucleotide substitution model to correct for homoplasy.

Once the estimates of polymorphism and divergence are made from DNA sequence data, they can be compared in a format like that shown in Table 8.5. Panel a in Table 8.5 shows the neutral theory expectations for polymorphism and divergence at the two loci.

**Table 8.5** Estimates of polymorphism and divergence for two loci sampled from two species that form the basis of the HKA test. (a) The correlation of polymorphism and divergence under neutrality results in a constant ratio of divergence and polymorphism between loci independent of their mutation rate as well as a constant ratio of polymorphism or divergence between loci. (b) An illustration of ideal polymorphism and divergence estimates that would be consistent with the neutral null model. (c) Data for the *Adh* gene and flanking region (Hudson et al. 1987) is not consistent with the neutral model of sequence evolution because there is more *Adh* polymorphism within *Drosophila melanogaster* than expected relative to flanking region divergence between *D. melanogaster* and *D. sechellia*.

(a) Neutral case expectations			
	Test locus	Neutral reference locus	Ratio (test/reference)
Focal species polymorphism ( $\pi$ )	$4N_e\mu_T$	$4N_e\mu_R$	$\frac{4N_e\mu_T}{4N_e\mu_R} = \frac{\mu_T}{\mu_R}$
Divergence between species ( $K$ )	$2T\mu_T$	$2T\mu_R$	$\frac{2T\mu_T}{2T\mu_R} = \frac{\mu_T}{\mu_R}$
Ratio ( $\pi/K$ )	$\frac{4N_e\mu_T}{2T\mu_T} = \frac{4N_e}{2T}$	$\frac{4N_e\mu_R}{2T\mu_R} = \frac{4N_e}{2T}$	
(b) Neutral case illustration			
	Test locus	Neutral reference locus	Ratio (test/reference)
Focal species polymorphism ( $\pi$ )	0.10	0.25	0.40
Divergence between species ( $K$ )	0.05	0.125	0.40
Ratio ( $\pi/K$ )	2.0	2.0	
(c) Empirical data from <i>D. melanogaster</i> and <i>D. sechellia</i>			
	<i>Adh</i>	5' <i>Adh</i> flanking region	Ratio (Adh/flank)
<i>D. melanogaster</i> polymorphism ( $\pi$ )	0.101	0.022	4.59
Between species divergence ( $K$ )	0.056	0.052	1.08
Ratio ( $\pi/K$ )	1.80	0.42	

Under neutrality and the infinite sites model, DNA sequence polymorphism is expected to be  $\theta = 4N_e\mu$  and divergence is expected to be  $k = 2T\mu$ . The test and reference locus may have different mutation rates. But note that the effective population size is constant when polymorphism is estimated for the two loci since the loci are sampled from the same species. The divergence times are also equal for the two loci since they are estimated from the same species pair. The ratio of the two divergence estimates at the test and reference loci is expected to equal the ratio of the test locus mutation rate over the reference locus mutation rate  $\left(\frac{\mu_T}{\mu_R}\right)$  since the factor of  $4N_e$  cancels out. The ratio of the two divergence estimates at the test and reference loci is also expected to equal  $\frac{\mu_T}{\mu_R}$ . Therefore, under neutrality the ratio of polymorphism estimates at the two loci as well as the ratio of the divergence estimates at the two loci should be equal since they both represent ratios of the mutation rates at the two loci. Similarly, the ratios of polymorphism

over divergence for each locus are expected to be equal under neutrality. The ratios can be tested for equality using a Chi-square test.

Table 8.5b shows an idealized illustration of polymorphism and divergence estimates that would be consistent with the neutral null model of DNA sequence evolution. In this idealization, the two loci do have different mutation rates that lead to different amounts of polymorphism and divergence. However, the ratios are equal as expected if the fate of mutations is due only to genetic drift.

Table 8.5c shows a classic example of divergence and polymorphism estimated in fruit flies to carry out the HKA test (Hudson et al. 1987). The locus tested for neutral evolution is the gene for alcohol dehydrogenase (*Adh*) and the reference locus is sequence upstream (5') to the coding region that does not possess an open reading frame. Polymorphism was estimated for the two genes from a sample of *Drosophila melanogaster* individuals and divergence for the two loci was determined with sequences from *Drosophila sechellia*. If the 5' flanking region is truly neutral, then the *Adh* data show too much

polymorphism within *D. melanogaster*. An excess of *Adh* polymorphism is also indicated by the large ratio of polymorphism for the two loci within *D. melanogaster* compared with the ratio of divergences for the two loci. It is now widely accepted that the *Adh* locus in *D. melanogaster* exhibits an excess of polymorphism consistent with balancing selection.

Although the HKA test is ingenious, it does have some limitations and assumptions. One difficulty in practice is the ability to identify an unambiguously neutral reference locus. For example, the 5' flanking region used by Hudson et al. (1987) as a neutral reference locus very likely contains promoter sequences that are functionally constrained by natural selection. Innan (2006) described a modification to the HKA test to use the average of multiple reference loci that should help avoid misleading results caused by reference loci that do not fit test assumptions.

Implicit in the HKA test is the assumption that each of the two species used are panmictic. Population subdivision has the potential to alter levels and patterns of nucleotide polymorphism and divergence (see review by Charlesworth et al. 2003) depending on how individuals are sampled. Consider levels of polymorphism in a subdivided species where  $F_{ST}$  is greater than zero. Population subdivision causes lower polymorphism for individuals sampled within subpopulations due both to reduced effective population size that increases drift within demes and increased autozygosity resulting from a higher probability of mating within demes. In contrast, there will be larger genetic differences for individuals sampled from two different demes due to differentiation among demes that would result in high perceived levels of polymorphism. If the HKA test is carried out for a species with population structure, sampling needs to be conducted to avoid taking sequences from only one or a few demes that could lead to an erroneous conclusion of too little polymorphism compared to the neutral expectation. Ingvarsson (2004) showed how the HKA test can lead to incorrect rejection of the neutral null hypothesis when there is population subdivision. He also gives an example of a population-structure-corrected HKA test applied to organelle DNA sequence data from the plant species *Silene vulgaris* and *Silene latifolia* which both exhibit strong population structure.

#### **MK test**

The MK test is a test of the neutral model of DNA sequence divergence between two species (McDonald

& Kreitman 1991). Like the HKA test it is named for its authors, McDonald and Kreitman. The MK test is also conceptually similar to the HKA test because it too establishes expected ratios of two classes of DNA changes at a single locus under neutrality. The MK test requires DNA sequence data from a single coding gene. The sample of DNA sequences is taken from multiple individuals of a focal species to estimate polymorphism. The test also requires a DNA sequence at the same locus from another species to estimate divergence.

The neutral expectations for the MK test are given in Table 8.6. The two classes of DNA change used in the MK test are **synonymous** and **nonsynonymous** (or replacement) changes. Nonsynonymous mutations within coding regions may alter the amino acid specified by a codon. Due to the redundancy of the genetic code, some mutations within coding regions will not change the amino acid specified by the codon and are therefore synonymous changes.

If genetic drift is the only process influencing the fate of a new mutation, levels of polymorphism and divergence within each category of DNA change should be correlated because they are both determined in part by the mutation rate. Fixed differences between species are caused by mutations that have gone to fixation, with expected divergence under neutral theory of  $2T\mu$ . Nucleotide sites that have two or more nucleotides within the focal species exhibit polymorphism, with an expected level of  $4N_e\mu$  under neutral theory. Since synonymous and nonsynonymous mutations may occur at different rates, we can assign each category of DNA change a different rate ( $\mu_N$  and  $\mu_S$ ). Both the ratio of nonsynonymous and synonymous fixed differences and the ratio of nonsynonymous and synonymous polymorphic sites are expected to be equal to  $\mu_N/\mu_S$  under neutral theory. The MK test therefore compares these two ratios for equality as a test of neutral theory. The neutral case illustration in Table 8.6b gives an example where  $\mu_N < \mu_S$  and where there is a higher level of polymorphism than divergence. Nonetheless, the ratios of the number of nonsynonymous over synonymous changes are constant for fixed differences and polymorphic sites as expected if both classes of mutations are neutral.

An MK test based on numbers of synonymous and nonsynonymous changes at the *Adh* locus that were fixed between *Drosophila* species or polymorphic within *D. melanogaster* (McDonald & Kreitman 1991) is given in Table 8.6c. Using fixed sequence differences as a reference point, fewer substitutions between

**Table 8.6** Estimates of polymorphism and divergence (fixed sites) for nonsynonymous and synonymous sites at a coding locus form the basis of the MK test. (a) Under neutrality, the number of nonsynonymous sites divided by the number of synonymous sites is equal to the ratio of the nonsynonymous and synonymous mutation rates. This ratio should be constant both for nucleotide sites with fixed differences between species and polymorphic sites within the species of interest. (b) An illustration of ideal nonsynonymous and synonymous site changes that would be consistent with the neutral null model. (c) Data for the *Adh* locus in *D. melanogaster* (McDonald & Kreitman 1991) show an excess of *Adh* nonsynonymous polymorphism compared with that expected based on divergence. (d) Data for the *Hla-B* locus for humans show an excess of polymorphism and more nonsynonymous than synonymous changes, consistent with balancing selection (Garrigan & Hedrick 2003).

	Fixed differences	Polymorphic sites
(a) Neutral case expectations		
Nonsynonymous sites ( $N$ )	$N_F = 2T\mu_N$	$N_p = 4N_e\mu_N$
Synonymous sites ( $S$ )	$S_F = 2T\mu_S$	$S_p = 4N_e\mu_S$
Ratio ( $N/S$ )	$\frac{N_F}{S_F} = \frac{2T\mu_N}{2T\mu_S} = \frac{\mu_N}{\mu_S}$	$\frac{N_p}{S_p} = \frac{4N_e\mu_N}{4N_e\mu_S} = \frac{\mu_N}{\mu_S}$
(b) Neutral case illustration		
Nonsynonymous changes	4	15
Synonymous changes	12	45
Ratio	0.33	0.33
(c) Empirical data from <i>Adh</i> locus for <i>D. melanogaster</i> (McDonald & Kreitman 1991)		
Nonsynonymous changes	2	7
Synonymous changes	42	17
Ratio	0.045	0.412
(d) Empirical data for the <i>Hla-B</i> locus for humans (Garrigan & Hedrick 2003)		
Nonsynonymous changes	0	76
Synonymous changes	0	49
Ratio	—	1.61

species are nonsynonymous than synonymous. The rate of nonsynonymous substitutions is 41.2% of the substitution rate for synonymous substitutions. Under neutrality, we expect about 41% of polymorphic sites within *D. melanogaster* to be at nonsynonymous sites. In contrast, the observed data shows that only about 4.5% of polymorphic sites are nonsynonymous. Thus, polymorphic sites have too many synonymous changes or too few nonsynonymous changes to be consistent with neutral levels of polymorphism. (Note that if using polymorphic sites as the frame of reference, then in this case there is an elevated rate of divergence at nonsynonymous sites compared to neutral expectations.)

A common observation in studies of coding DNA sequences is that numbers of nonsynonymous and synonymous DNA changes are not equal. A neutral explanation for this pattern is that these two types of DNA change have different underlying mutation rates. It is expected that nonsynonymous changes

will be more frequent than synonymous changes if mutations occur at random nucleotide sites. In fact, 96% of nucleotide changes in the first nucleotide position of a codon, all changes in the second position and 30% of changes at the third position are nonsynonymous. Overall, if mutation occurs at random within coding sequences 75.3% of all mutations will be nonsynonymous and 24.7% will be synonymous.

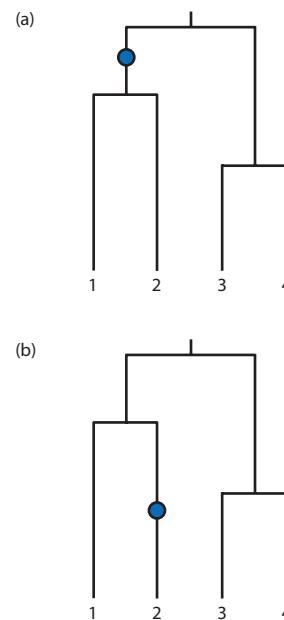
An alternative explanation is that rates of synonymous and nonsynonymous mutation are roughly equal, but that nonsynonymous mutations commonly alter proteins in ways that impair their function. Nonsynonymous mutations that disrupt function also reduce fitness and are therefore acted against by natural selection. This form of natural selection is sometimes called **purifying selection** because selection acts to purify the pool of mutations by removing low fitness sequence changes. It is also possible that some nonsynonymous mutations result

in enhancement of function and are fixed rapidly by positive selection. A third, non-neutral alternative is that nonsynonymous mutations are maintained at intermediate frequencies in the population by balancing selection. An example of strong balancing selection is the human leukocyte antigen (*Hla*) B gene in humans as detected with an MK test by Garrigan and Hedrick (2003) using divergence between humans and chimpanzees (Table 8.6d). There are no fixed DNA differences between humans and chimpanzees for this locus, suggesting low rates of mutation since divergence of these two species. In contrast, the human populations show high levels of polymorphism and 1.6 nonsynonymous changes for every synonymous change, inconsistent with the neutral hypothesis that polymorphism and divergence are correlated. *Hla* genes form the major histocompatibility complex (MHC) region that encodes cell-surface antigen-presenting proteins important in immune system function. Heterozygotes for these loci have higher fitness since they present more diverse cell-surface antigens.

### Tajima's D

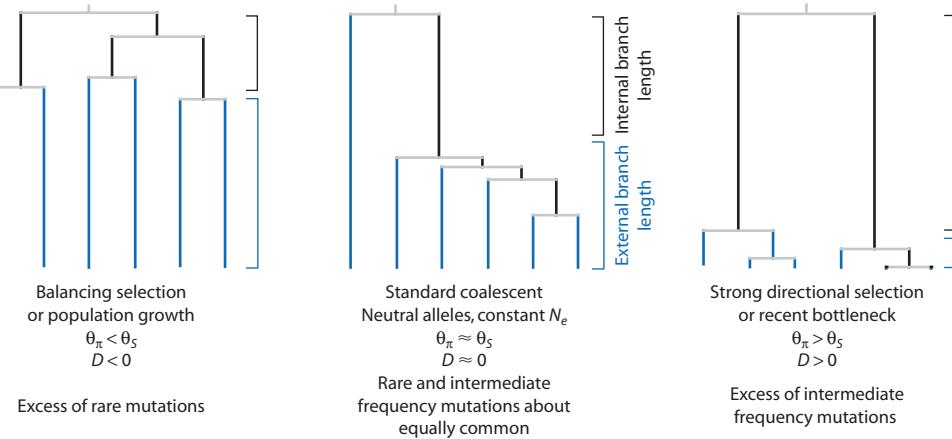
Tajima's *D* is a test of the standard coalescent model (neutral alleles in a population of constant size) that is commonly applied to DNA polymorphism data sampled from a single species (Tajima 1989a, 1989b). The test uses the nucleotide diversity and the number of segregating sites observed in a sample of DNA sequences to make two estimates of the scaled mutation rate  $\theta = 4N_e\mu$ . This section will refer to an estimate of  $\theta$  based on the nucleotide diversity as  $\hat{\theta}_\pi$  and an estimate of  $\theta$  based on the number of segregating sites as  $\hat{\theta}_S$ . Tajima's *D* test relies on the fact that  $\hat{\theta}_\pi$  and  $\hat{\theta}_S$  are expected to be approximately equal under the standard coalescence model where all mutations are selectively neutral and the population remains a constant size through time. The null hypothesis of the test is that the sample of DNA sequences was taken from a population with constant effective population size and selective neutrality of all mutations. Natural selection operating on DNA sequences as well as changes in effective population size through time lead to rejection of this null hypothesis.

Tajima's *D* takes advantage of the fact that mutations that occurred further back in time in a genealogy are counted more times when computing the nucleotide diversity ( $\pi$ ) from all unique pairs of sequences. In contrast, the position of a mutation on a genealogy does not influence the number of segregating sites



**Figure 8.19** Estimates of the scaled mutation rate  $\theta$  are estimated differently using nucleotide diversity ( $\hat{\theta}_\pi$ ) and the number of segregating sites ( $\hat{\theta}_S$ ) depending on the location of mutations in a genealogy. Each mutation makes a single segregating site under the assumptions of the infinite alleles model no matter where it occurs. However, mutations on internal branches will appear in multiple pairwise comparisons and cause  $\pi$  to be larger (a). In contrast, mutations that occur on external branches (b) that cause a nucleotide change in only a single lineage contribute less to  $\pi$ . Each mutation is counted four times ( $d_{13}, d_{23}, d_{14}$ , and  $d_{24}$ ) in (a) but three times ( $d_{12}, d_{23}$ , and  $d_{24}$ ) in (b) when computing  $\pi$ .

( $S$ ) since any number of sequences bearing a given nucleotide always represent just one segregating site (Fig. 8.19). The coalescent process with neutral alleles and constant effective population size results in approximately the same total length along interior and exterior branches in a genealogy. In contrast, processes that alter the probability of coalescence also change the ratio of interior and exterior branch length. Both sustained increases in the effective population size and balancing selection lead to decreasing probabilities of coalescence toward the present time and longer external branches (Fig. 8.20). Longer external branches in a genealogy can also be caused by population structure if the DNA sequences compared are sampled from different demes. Shrinking effective population size or population bottlenecks as well as strong directional selection lead to increasing probabilities of coalescence toward the present time and shorter external branches (Fig. 8.20). Substantial changes to the genealogical branching pattern lead to differences in  $\pi$  and  $S$  that cause Tajima's *D* to differ from zero.



**Figure 8.20** Differences in the shape of genealogies are the basis of Tajima's  $D$  test. In the standard coalescent model of genealogical branching the probability of coalescence is constant per lineage over time. The standard coalescent therefore gives expected branch lengths when all alleles are selectively neutral and the effective population size is constant (center). Changes in the effective population size over time (population growth, population bottlenecks) change the probability of coalescence over time as well. Natural selection also alters the probability of coalescence based on the fitness of alleles each lineage bears. Changes in the effective population size and natural selection alter the expected time to coalescence and therefore the expected branch lengths in a genealogical tree. If the chance of coalescence is greater in the present than in the past (right), most coalescent events occur near the present and internal branches are long in comparison with external branches. If the chance of coalescence is smaller in the present than in the past (left), most coalescent events occurred in the past and external branches are long in comparison with internal branches. Since the chance of a mutation is constant over time, lineages with longer branches are expected to experience more mutations.

An alternative way to think about how Tajima's  $D$  works is to consider the frequency distribution of mutations under different types of natural selection or population histories. Mutations that happen to occur on internal branches in a genealogy have an intermediate frequency because they are inherited by lineages that arise later in time. In contrast, mutations that happen to occur on external branches have a low frequency since they are unique to a single lineage. Since total internal and total external branch length are expected to be about equal under the standard coalescent model, intermediate and rare alleles are also expected to be about equal in frequency. Both population growth and multi-allelic balancing natural selection can lead to an excess of rare mutations since these processes increase the external branch length. In contrast, strong purifying selection, shrinking population size, or a population bottleneck can lead to an excess of intermediate frequency mutations because these processes increase the amount of internal branch length.

Tajima's  $D$  statistic is computed from the difference between  $\hat{\theta}_\pi$  and  $\hat{\theta}_S$  divided by the standard deviation of  $\hat{\theta}_\pi - \hat{\theta}_S$ :

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{\text{var}(\hat{\theta}_\pi - \hat{\theta}_S)}} = \frac{\hat{\theta}_\pi - \frac{p_S}{a_1}}{\sqrt{e_1 p_S + e_2 p_S(p_S - 1)}} \quad (8.54)$$

where  $p_S$  is the number of segregating sites per nucleotide site. Recall that the standard deviation is the square root of the variance, so that dividing by the standard deviation puts  $D$  in units of standard deviations away from the mean of zero expected for standard coalescent genealogies. Only when the observed result is about two standard deviations away from the mean do we reject the null hypothesis of  $D = 0$  and thereby reject the null model of a neutral genealogy with constant effective population size (see confidence limits in Table 2 of Tajima 1989a).

The quantities used to compute the variance are

$$e_1 = \frac{n+1}{3a_1(n-1)} - \frac{1}{a_1^2} \quad (8.55)$$

and

$$e_2 = \frac{c}{a_1^2 + a_2} \quad (8.56)$$

where

$$a_1 = \sum_{k=1}^{n-1} \frac{1}{k} \quad (8.57)$$

$$a_2 = \sum_{k=1}^{n-1} \frac{1}{k^2} \quad (8.58)$$

### Problem box 8.2 Computing Tajima's *D* from DNA sequence data

To study the population history of *Drosophila simulans*, Baudry et al. (2006) sampled flies from multiple populations in Africa, Europe, and the Antilles. From these flies they sequenced four genes located on the X chromosome. Using part of their DNA sequence data, test the hypothesis that *D. simulans* meets the assumptions of the standard neutral coalescent model via Tajima's *D*.

DNA sequences from the *runt* locus for flies sampled in Europe and Mayotte (an overseas collectivity of France composed of several islands in the Indian Ocean, between northern Madagascar and northern Mozambique) exhibited the following patterns

Population	<i>n</i> sequences	Nucleotide sites	<i>S</i>	$\pi$
Europe	15	556	17	0.012436
Mayotte	15	538	34	0.013525

Use the number of segregating sites (*S*) to calculate the number of segregating sites per nucleotide site ( $p_S$ ) and then estimate  $\hat{\theta}_S$  per site according to equation 8.29. Then compute Tajima's *D* according to equation 8.54.

What do you conclude about the history of these two *D. simulans* populations? Note that your estimates of Tajima's *D* will differ from those in Baudry et al. (2006) because they used only synonymous site polymorphisms whereas you have used polymorphisms at all sites. Why did Baudry et al. (2006) use only synonymous site polymorphisms? The DNA sequence data files are available on the text website.

and

$$c = \frac{2(n^2 + n + 3)}{9n(n - 1)} - \frac{n + 2}{a_1 n} + \frac{a_2}{a_1^2} \quad (8.59)$$

where *n* is the number of sequences sampled and assuming that there is no recombination.

Although the variance of *D* is a complex expression, it can still be understood intuitively. The formula quantifies both **sampling variance** and **evolutionary variance**. Sampling variance comes from taking a sample of DNA sequences and using them to estimate  $\pi$  and *S*. As with any finite sample of data from a larger underlying population, repeating the sampling procedure would result in a slightly different estimate of the parameters of interest because the sample is not perfectly representative of the full population. Sampling variance decreases as sample sizes increase since estimates are based on a larger and larger proportion of the underlying population. In contrast, evolutionary variance is caused by the variable outcomes of the random evolutionary processes of genetic drift. Evolutionary variance can only be estimated by sampling multiple independent realizations of the same random process, for example taking samples of DNA sequences from multiple populations that independently experienced genetic drift after being

isolated from the same ancestral population. The coalescence process itself has a great deal of evolutionary variance. For example, under the standard coalescent model there is a wide range of coalescence times for *k* lineages around an average with variance in coalescence time that is largest for two lineages (see equation 3.76). Both sources of variance are taken into account when determining the standard error of *D*.

The value of Tajima's *D* is influenced by changes over time in the size of populations, population structure, and the action of natural selection. Therefore, Tajima's *D* is not a simple test for the action of natural selection alone as sometimes assumed. The null model is based on a constant mutation rate through time (a molecular clock), the infinite sites model of mutation, the Wright–Fisher model with non-overlapping generations, and a panmictic population at drift–mutation equilibrium (see Tajima 1996 on the first two points). Although a large value of *D* serves to reject the standard coalescent model for a given set of DNA polymorphism data, distinguishing among changes in effective size through time, population structure, and natural selection requires more than just an estimate of *D* at a single locus. DNA polymorphism in humans, for example, often shows negative values of Tajima's *D*. These results are considered

by many to be caused by a low level of population structure as well as a history of very rapid population growth in the recent past that characterizes human populations rather than widespread balancing selection operating on human loci (Ptak & Przeworski 2002; Tishkoff & Verrelli 2003).

### **Mismatch distributions**

The previous section explored how a sample of DNA sequences from a population can be used to test neutral expectations by comparing estimates of  $\theta$  based on polymorphism measured with nucleotide diversity ( $\pi$ ) and the number of segregating sites ( $S$ ). Both  $\pi$  and  $S$  summarize patterns of sequence variation into a single number. Specifically, the nucleotide diversity  $\pi$  is really an average of the differences between all pairs of sequences in a sample. Instead of using an average to measure polymorphism, we can directly examine the distribution of all individual pairwise sequence comparisons. This is commonly called the **mismatch distribution** and it is the frequency distribution of the number of nucleotide sites that differ between all unique pairs of DNA sequences in a sample. The mismatch distribution is a tool that can be used to infer the history of the population that gave rise to a sample of DNA sequences. It can be used to infer past changes in the effective size of a population using selectively neutral DNA sequences. Alternatively, in populations that have maintained a constant size over time, these distributions can be used to identify the action of natural selection.

**Mismatch distribution** The frequency distribution of the number of nucleotide sites that differ between all unique pairs of DNA sequences in a sample from a single species. It is also known as the distribution of pairwise differences.

**Haplotype frequency distribution** The distribution of the frequency of each sequence haplotype in a population assuming that individuals are haploid or homozygous. It is also known as the site frequency spectrum.

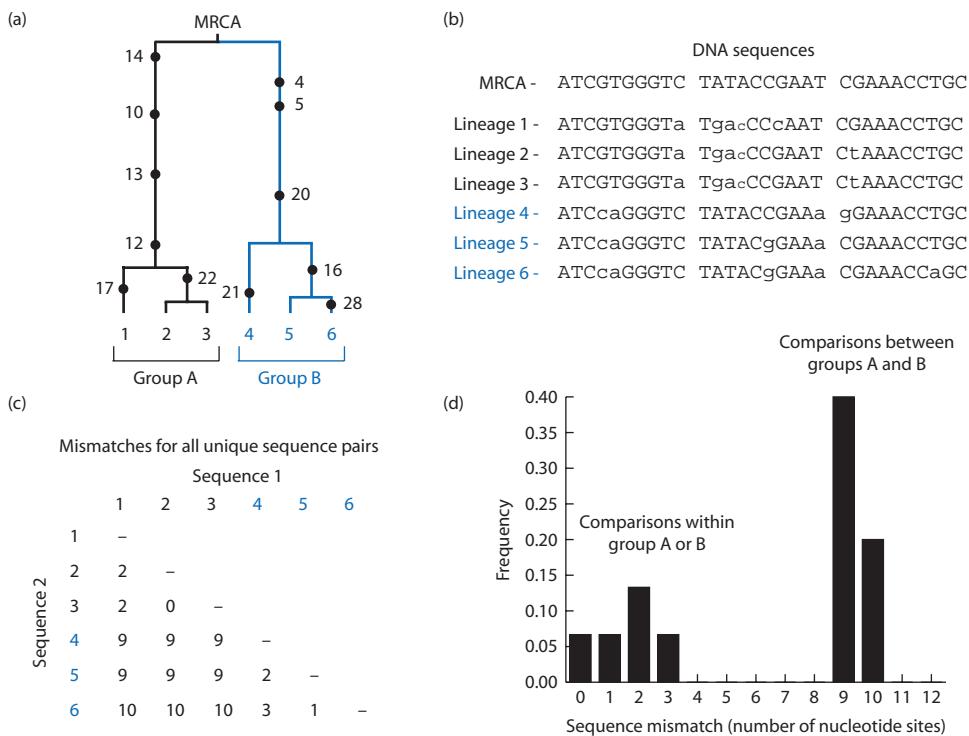
Let's assume complete neutrality of mutations to focus on using mismatch distributions to develop expectations for patterns of DNA sequence differences in stable, growing, and shrinking populations. The properties of the mismatch distribution arise

directly from expected patterns that characterize neutral genealogies. Chapter 3 shows that the last pair of lineages ( $k = 2$ ) takes the longest average time to coalesce in standard neutral genealogies for populations with constant  $N_e$ . When there is mutation, the two oldest lineages in the population also differ by the largest number of mutations since the expected number of mutations is proportional to the length of time a lineage exists. In populations that maintain constant  $N_e$ , these oldest two lineages experience numerous mutations and therefore have a high degree of mismatch. This pattern of long lineages having multiple mutations can be seen in the genealogy in Fig. 8.21.

Working from the past to the present, the two oldest lineages in any genealogy give rise to additional lineages. The younger progeny lineages inherit all the mutations that have occurred on the progenitor lineages and may also experience additional new mutations. Since the lineages closer to the present tend to have shorter times to coalescence (the probability of coalescence increases with larger  $k$ ), they also tend to accumulate fewer mutations. Looking at Fig. 8.21, the three lineages within group A would each inherit the four mutations that occurred on the internal branch that was their ancestor. Because lineages 1, 2, and 3 within group A share the mutations of their ancestral lineage, they also tend to have fewer nucleotide sites that mismatch. For example, lineages 1 and 2 differ by only the two mutations that occurred near the present (mutations at nucleotide sites 17 and 22). Lineages 4, 5, and 6 within group B also have low levels of mismatch by the same logic.

In contrast, the level of sequence mismatch is high when lineages are compared between groups A and B in Fig. 8.21. For example, lineages 1 and 4 differ by nine mutations. This high level of mismatch occurs because sequences from distantly related lineages are separated by much more time since they shared a common ancestral lineage, leading to many more mutational changes that independently altered each DNA sequence. Another way to think of the situation is that closely related lineages differ only by a few young mutations while distantly related lineages differ by more mutations, many of which are old and have been resident in the population for a long time.

The mismatch distribution has distinct patterns depending on the demographic history of the population (Slatkin & Hudson 1991; Rogers & Harpending 1992). Mismatch distributions from populations that have experienced a constant  $N_e$  over time tend to have two clusters of values in the mismatch distribution.



**Figure 8.21** The basis of the mismatch distribution. (a) A neutral genealogy that bears multiple mutation events. Each mutation event is represented by a circle and the number of the random nucleotide site that mutated assuming the infinite sites mutation model. The six lineages in the present can be separated into two groups (called A and B) based on their ancestral lineage when there were only two lineages in the population. (b) The DNA sequences for each lineage are shown based on the 30 base-pair sequence assigned to the most recent common ancestor (MRCA) with mutations shown in lower-case letters. (c) The number of nucleotide sites that are different or mismatched between pairs of DNA sequences. (d) The mismatch distribution shown is a histogram of the mismatches for the 15 pairs of DNA sequences compared. Neutral genealogies from populations with constant  $N_e$  through time tend to show bimodal mismatch distributions. The cluster of observations with few mismatches results from sequence comparisons between recently related lineages (comparisons within group A or group B). In contrast, sequences from distantly related lineages that do not share the same ancestor when  $k = 2$  (comparisons between groups A and B) tend to have more mismatches.

Such a bimodal distribution is the characteristic signature of genealogies in populations with a relatively constant  $N_e$  in the past. The bimodal pattern is caused by roughly equal times to coalescence of all internal and external branches. In contrast, populations that had rapidly growing or shrinking  $N_e$  in the past tend to have distinct mismatch distributions. In populations that have rapidly growing  $N_e$ , most coalescence events happen early in the genealogy near the MRCA since the probability of coalescence decreases toward the present (see the left-hand genealogy in Fig. 8.20). This leads to long external branches that each experience many unique mutations. The mismatch distribution then has a high frequency of sequence pairs with a high degree of mismatch and few sequence pairs with a low degree of mismatch. Alternatively, populations that experienced continual declines in  $N_e$  have genealogies where most coalescence events happen near the present because

the probability of coalescence increases toward the present (see the right-hand genealogy in Fig. 8.20). In a shrinking population, the mismatch distribution tends to have a high frequency of sequence pairs with low mismatch counts.

A related way to view polymorphism is by examining the distribution of haplotype frequencies in a sample of sequences. Such **haplotype frequency distributions** show the proportion of sequences in a population that represent each of the observed sequence alleles (assuming that individuals are haploid or homozygous). Under neutrality and constant effective population size (see Fig. 8.20) a range of haplotype frequencies are expected from very frequent to rare. When populations are growing rapidly or there is balancing selection, there is expected to be an excess of rare haplotypes produced by the excess length of external branches in the genealogy. When populations are shrinking rapidly or there is

### Interact box 8.3 Mismatch distributions for neutral genealogies in stable, growing, or shrinking populations

TreeToy is a java applet that simulates genealogies for neutral alleles in populations that are growing in size through time. The lineages experience mutations with a rate that depends on  $\theta = 4N_e\mu$  (called **Theta0** in the simulation for  $\theta$  at time zero since  $N_e$  changes over time). The simulation displays the genealogy along with the mismatch distribution and the frequency histogram of each of the haplotypes in the population. The simulation requires values for the number of lineages in the genealogy (**Sample Size**), the growth rate of the population (**Growth Factor**) and a scaling factor for the time to coalescence of two genes in units of  $1/(2\mu)$  generations (**Tau**).

Set the **Sample Size** to 30, **Theta0** to 20, **Growth Factor** to 1, leave **Tau** at the default value of 8, and then click on **Draw Tree**. The resulting genealogy has waiting times similar to those expected under the standard coalescent model with constant population size through time. It should have a somewhat bimodal mismatch distribution (the graph at bottom left) and a wide range of haplotype frequencies (graph at bottom right labeled **Freq. Spectrum**). Press the **Draw Tree** button several times to see that there is substantial variation in these distributions even for the same model parameters. Note the general shape of the genealogies under the constant  $N_e$  model.

Now change **Growth Factor** to 100 and click **Draw Tree**. What does this value of **Growth Factor** mean biologically? Why does the genealogy have external long branches with few coalescence events near the present? How do the mismatch and haplotype frequency distributions change?

Note that the contrasting features of mismatch distributions and the haplotype frequency spectrum in populations that are constant or changing in size through time are generally more apparent when  $\theta$  is larger (more mutations occur) and there is a larger sample of lineages in the genealogy. The graph of theta by tau at the upper right shows the probability of mutation back through time as the genealogy approaches the MRCA. Larger values of tau represent deeper times in the past while larger values of theta represent a high probability of mutation. Simulate a genealogy with **Growth Factor** equal to 10 and then interpret the distribution of theta by tau.

strong directional selection, there is expected to be an excess of high-frequency haplotypes and very few rare haplotypes because most of the branch length lies in the internal branches of the genealogy.

Mismatch and haplotype frequency distributions can help identify instances of expansion or contraction in the effective population size if sequences are neutral. Alternatively, if sequences are known to come from a population with a constant effective size, then these distributions can be used to identify the action of natural selection. Several tests are available that use the haplotype frequency or mismatch distributions to evaluate the null hypothesis of constant effective population size through time using DNA sequences (Fu & Li 1993; Fu 1996, 1997; Schneider & Excoffier 1999; Mousset et al. 2004; Innan et al. 2005). It is important to note several limitations of these tests. First, recombination has the potential to impact the mismatch distribution along with population demography. Recombination events assemble novel sequence haplotypes from

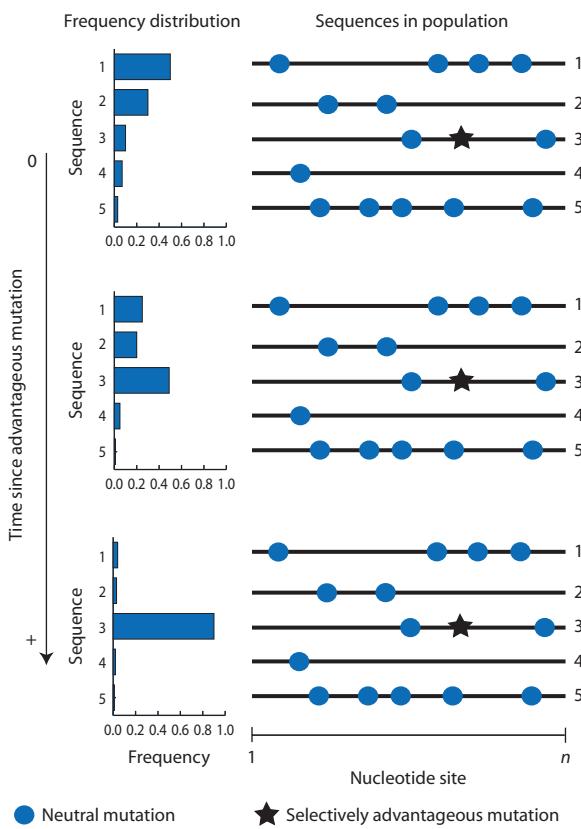
existing haplotypes and in doing so break up mutations that are associated due to identity by descent. Therefore, recombination obscures the history of mutations and in the extreme would lead to a uniform mismatch distribution. Second, coalescence is a stochastic process and there is an inherently large variance in times to coalescence (see Chapter 3). This leads to a large variance in the shape of mismatch distributions even when  $N_e$  is constant. Therefore, tests that utilize the mismatch distribution can only be expected to detect very large and sustained shrinkage or expansion of  $N_e$ .

## 8.6 Molecular evolution of loci that are not independent

- Gametic disequilibrium between neutral and selected sites influences polymorphism.
- Genetic hitch-hiking, selective sweeps, and background selection.
- Gametic disequilibrium and rates of divergence.

Up to this point in the chapter, loci or nucleotide sites have been considered completely independent entities that are not influenced by evolutionary processes at neighboring loci or nucleotide sites. This is equivalent to assuming that all alleles at all loci are in complete gametic equilibrium. Processes such as physical linkage covered in Chapter 2 may cause gametic disequilibrium among the alleles in any population. The impacts of gametic disequilibrium on new mutations is of particular interest in molecular evolution because the fate of new mutations dictates levels of polymorphism and rates of divergence. Because new mutations initially enter a population as a single copy, they must initially experience very high levels of gametic disequilibrium. A new mutation that is present in only one copy will be uniquely associated with the other alleles that just by chance occur on the same chromosome where the mutation occurred. The gametic disequilibrium experienced by new mutations has substantial consequences for neighboring sites in the genome if the new mutation is acted on by natural selection. First, let's explore changes in polymorphism caused by gametic disequilibrium between neutral nucleotide sites and nucleotide sites where mutations are acted on by natural selection. At the end of this section we will consider the implications for rates of divergence.

To see the consequences of gametic disequilibrium for new mutations, consider what happens when a favorable mutation arises in a population. Assume for now that the population is composed of haploid individuals that reproduce clonally so that there is no recombination. Figure 8.22 illustrates the changes to allele frequencies over time when a favorable mutation enters a population. Initially, the population contains five different haploid sequences. Each of these chromosomes bears a number of neutral mutations and each chromosome also has an intermediate frequency in the population that is the product of genetic drift. An advantageous mutation, indicated by a star on the figure, occurs by chance on one of the chromosomes. Over time, the chromosome bearing the favorable mutation will increase in frequency since it has a higher fitness and the other chromosomes will decrease in frequency. Eventually, depending on the relative fitness of the mutation, the chromosome bearing the advantageous mutation will approach fixation in the population. Because there is no recombination in this example, the advantageous mutation is only found on one type of chromosome. Thus, the advantageous mutation is



**Figure 8.22** The impact of natural selection on an advantageous mutation as well as on associated nucleotide sites, often called a selective sweep. Imagine a single population that contains five distinct DNA sequences without recombination because reproduction is clonal. Each DNA sequence is distinguished by a number of neutral mutations (blue circles) and has a frequency given by the histogram on the left. Initially, the population has polymorphism since the population is composed of intermediate frequencies of each DNA sequence. At time 0, the third DNA sequence experiences a mutation that is strongly advantageous, indicated by the star. Natural selection acts to increase the frequency of the advantageous mutation over time, until the population approaches fixation for the third DNA sequence. Once selection has swept the advantageous mutation to near fixation the population has very little polymorphism. This is because only those original neutral mutations that were linked to the advantageous allele on the same DNA sequence remain in the population. Thus, positive selection on one site also sweeps away polymorphism at linked nucleotide sites if gametic disequilibrium is maintained. The figure assumes that positive natural selection is strong and increases the frequency of the third DNA sequence rapidly such that no new mutations appear in the population.

in complete gametic disequilibrium with two neighboring neutral mutations that happened to be on that chromosome.

Maynard Smith and Haigh (1974) coined the term **genetic hitch-hiking** to describe the consequences

of gametic disequilibrium between a selected mutation and the alleles present at neighboring neutral nucleotide sites. As natural selection drives a favorable mutation to high frequency in a population, the neutral alleles in gametic disequilibrium with the selected mutation also reach high frequency because they happened to be on the chromosome where the advantageous mutation initially occurred. This also results in a loss of polymorphism in the population for the neutral alleles, since only the one set of neutral alleles in gametic disequilibrium with the advantageous mutation remains in the population once the advantageous mutation approaches fixation by natural selection. Levels of polymorphism decrease in Fig. 8.22 as the advantageous mutation becomes more and more frequent in the population. The reduction in polymorphism caused by genetic hitch-hiking is sometimes called a **selective sweep** because while an advantageous mutation and the neutral polymorphisms that are linked to it are swept to high frequency by natural selection, other neutral polymorphisms not in gametic disequilibrium with the selected site are swept out of the population at the same time. It is important to emphasize that the reduction in polymorphism seen in selective sweeps is an indirect consequence of natural selection since only the advantageous mutation itself has a selection coefficient that is not effectively zero.

**Genetic hitch-hiking** The process by which selectively neutral alleles increase or decrease in frequency due to their association with alleles that are under the influence of natural selection.

**Selective sweep** The reduction or elimination of polymorphism in a region of DNA sequence surrounding a site where a beneficial mutation has increased in frequency due to positive natural selection. The reduction of polymorphism is a result of gametic disequilibrium between a beneficial mutation and neighboring neutral sites that has not been broken down by recombination.

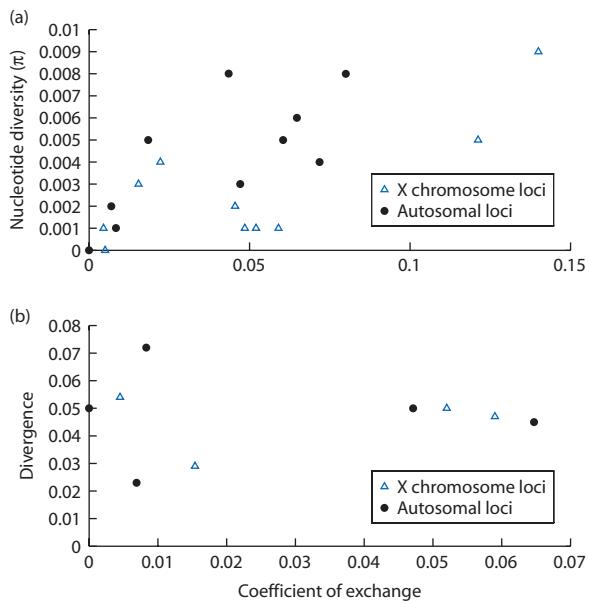
The genetic hitch-hiking and selective sweep process that leads to reduced polymorphism at nucleotide sites in gametic disequilibrium with selected sites can be thought of as analogous to genetic drift at neutral sites. Positive natural selection causes some

linked sets of alleles in the population to reach fixation faster than under genetic drift alone, in the same fashion that a reduction in the effective population size or genetic bottleneck would decrease the average time to fixation for those neutral alleles that do reach fixation. Gillespie (2000, 2001) has termed the selective sweep process **genetic draft** because neutral mutations in gametic disequilibrium are like an unpowered glider, floating their way to fixation by riding along on the up-currents of increased probability of substitution caused by selection. Gillespie has shown that genetic draft is a stochastic process because the neutral mutations that do reach fixation do so by random association with selected mutations. Thus, natural selection on infrequent beneficial mutations causes finite random sampling from the pool of available neutral mutations even if the effective population size is infinite.

Because mitochondrial genomes do not experience recombination, genetic hitch-hiking has the potential to cause strong selective sweeps. Evidence for selective sweeps in animal mitochondrial genomes comes from a comparison of polymorphism measured by nuclear allozyme loci, nuclear DNA sequences, and mitochondrial DNA sequences for a large sample of animal species (912, 417, and 1683 species, respectively). Using these data, Bazin et al. (2006) tested the neutral hypothesis that polymorphism for all three types of data should be correlated since each class of loci shares a similar effective population size. (Because mitochondrial genomes are haploid and uniparentally inherited, their effective population size is four times less than that of biparentally inherited nuclear loci.) Based on census population sizes, insects, echinoderms, and mollusks are expected to have larger effective population sizes than mammals, fish, reptiles, and birds. Neutral theory predicts that the taxa with larger effective population sizes should also have higher levels of polymorphism for the same loci. This neutral prediction is met for nuclear allozyme and DNA sequence data. Levels of polymorphism are higher for insects, echinoderms, and mollusks than for mammals, fish, reptiles, and birds. Levels of polymorphism estimated from the two types of nuclear loci are also highly correlated within each species. In contrast, levels of mitochondrial polymorphism were both low and nearly uniform across all the animal groups and did not show a correlation with levels of nuclear allozyme and DNA sequence polymorphism. This result can be explained by genetic hitch-hiking that has caused selective sweeps in the non-recombining mitochondrial genome.

In the extreme case of no recombination, an advantageous mutation can never become associated via recombination with other neutral alleles that exist in the population. But the absence of recombination is not necessary for selective sweeps to occur. When recombination does occur, selective sweeps will still happen as long as the increase in frequency of an advantageous mutation is rapid relative to the time that is required for recombination to break down the gametic disequilibrium between a new mutation and its neighboring sites. In genomes with recombination, we therefore expect that selective sweeps may impact only relatively small areas centered around the site where an advantageous mutation occurred. This is because recombination will happen more and more frequently as the distance from the site of an advantageous mutation grows larger. Only in the nucleotide sites relatively near the advantageous mutation will recombination be unlikely to occur, allowing gametic disequilibrium to persist for long enough for selection to make appreciable changes in the frequency of the advantageous mutation. We can also expect that stronger natural selection should lead to larger or more persistent areas of reduced polymorphism because an advantageous mutation will reach fixation faster, outpacing recombination that would distribute the advantageous mutation across chromosomes with different neutral mutations.

A common result from studies of genetic variation at numerous loci in natural *Drosophila* populations is that levels of polymorphism are positively correlated with rates of recombination (reviewed by Hudson 1994). The relationship between polymorphism and rates of recombination can be explained in several ways. A hypothesis consistent with selective neutrality is that the recombination rate at a locus is somehow related to its mutation rate. For example, the molecular processes that cause recombination might also cause point mutations. Another neutral hypothesis is that regions of low recombination might also have higher functional constraints and therefore lower neutral mutation rates. Recall that neutral theory predicts that levels of polymorphism and rates of divergence are correlated since both are ultimately products of the mutation rate. This leads to the prediction that both levels of polymorphism and divergence rates should correlate with the recombination rate if neutral processes explain the relationship between recombination and polymorphism at the *Drosophila* loci. Data to test this neutral hypothesis for the correlation between



**Figure 8.23** Plots of nucleotide diversity within *D. melanogaster* populations (a) and divergence between *D. melanogaster* and *D. simulans* (b) by the coefficient of exchange, a measure of the recombination rate, for numerous loci. The nucleotide diversity at a locus decreases along with the recombination rate (a). The correlation of polymorphism and the recombination rate at a locus could be explained by neutral theory if loci with lower recombination rates also happen to have lower mutation rates. Under this neutral hypothesis, divergence rates would also be correlated with recombination rates since both polymorphism and divergence increase as the mutation rate increases. An alternative explanation for the correlation of recombination and polymorphism is the action of natural selection on beneficial mutations that has caused hitch-hiking and a reduction of polymorphism due to selective sweeps. Divergence rates at a subset of the loci (b) suggest that divergence and recombination rates are independent, rejecting the neutral hypothesis for these data. Data from Begun and Aquadro (1992).

polymorphism and recombination rate in *Drosophila* are shown in Fig. 8.23. While polymorphism in *D. melanogaster* clearly increases with the recombination rate, levels of divergence between *D. melanogaster* and *D. simulans* for a subset of the same loci are independent of the recombination rate. These data therefore reject the neutral hypothesis. An alternative explanation for these data is the operation of natural selection on beneficial mutations that has resulted in selective sweeps. The strength of genetic hitch-hiking and the amount of polymorphism lost to selective sweeps decreases with increasing recombination because recombination reduces gametic disequilibrium between a selected mutation and neighboring sites. However, selective sweeps have no impact on

the rate of divergence so that divergence will show no relationship to the recombination rate. Therefore, the selective sweep hypothesis explains why polymorphism increases with the recombination rate but divergence is independent of recombination rate in the *Drosophila* DNA sequence data.

Levels of polymorphism and rates of divergence in human DNA sequences provide a counter-example to the uncoupling of polymorphism and divergence seen in *Drosophila*. Hellmann et al. (2003) used a large amount of DNA sequence data to estimate both polymorphism within humans and divergence between humans, chimps, and baboons. The recombination rates near the loci used to estimate polymorphism and divergence were also known for humans. In the human sequences polymorphism increases as the recombination rate increases. But in agreement with the neutral expectation that polymorphism and divergence are correlated, the degree of divergence increased with the recombination rate as well. Thus, in humans, correlated levels of polymorphism and divergence argue for the neutral evolution of mutations. It is possible that rates of recombination and mutation are not independent in humans because mutations increase the probability of a recombination event nearby or the recombination process produces mutations.

#### ***Genetic hitch-hiking due to background or balancing selection***

While genetic hitch-hiking may sometimes involve positive selection on beneficial mutations, it is also possible that new mutations may be deleterious. Negative or purifying natural selection removes deleterious mutations from the population by driving them to loss. Negative selection against deleterious mutations is also expected to reduce polymorphism through hitch-hiking in a process called **background selection** (Charlesworth et al. 1995). When a new deleterious mutation appears in a population, it is in gametic disequilibrium with other neutral mutations. Negative selection will cause the deleterious allele to go to loss, bringing its associated neutral mutations to loss as well. Therefore, background selection is expected to cause a reduction in polymorphism in populations. Background selection effectively lowers the mutation rate at neutral sites because it removes some portion of the new neutral mutations that appear very briefly in the population but just happen to be associated with a deleterious mutation and so go to loss quickly.

**Background selection** A reduction in polymorphism caused by the combination of negative or purifying selection against deleterious mutations that also leads to the loss of neutral alleles in gametic disequilibrium with the deleterious mutation.

A third possibility is that new mutations are acted on by balancing selection, which would eventually bring new beneficial mutations at the same site to intermediate frequencies and maintain them in the population for very long periods of time. Balancing selection is also expected to impact polymorphism at neutral sites that are in gametic disequilibrium with the selected site. When a new beneficial mutation appears at a site under balancing selection, the initial increase in its frequency has a effect like a selective sweep. However, long-term balancing selection leads to an increase in polymorphism because balancing selection maintains multiple alleles that persist in the population for long periods of time. These selected alleles can then accumulate mutations at neutral sites that are in gametic disequilibrium, gathering polymorphism over time. Compared to independent neutral sites, neutral sites in gametic disequilibrium with sites under balancing selection have greatly increased segregation times and so have a greater opportunity to experience mutation that leads to the accumulation of polymorphism (reviewed by Charlesworth 2006).

Hitch-hiking can be thought of as a process that alters the time to fixation or loss in the random walk of neutral genetic drift (refer to Figs 8.2 and 8.3). Hitch-hiking with beneficial mutations under positive selection greatly speeds up the time to fixation, reducing polymorphism. Hitch-hiking with deleterious mutations under negative selection also reduces polymorphism because the time to loss is accelerated. Hitch-hiking with mutations under long-term balancing selection increases polymorphism because fixation or loss are avoided for long periods during which neutral mutations can accumulate.

#### ***Gametic disequilibrium and rates of divergence***

As we have seen earlier in the chapter, when the probability of fixation of a new mutation is dictated by natural selection then divergence rates change for those sites under natural selection. Positive natural selection speeds up divergence because advantageous

mutations fix faster on average than they would under genetic drift alone. Alternatively, negative natural selection slows divergence rates since deleterious mutations go to loss rapidly and fewer neutral mutations remain that might fix by genetic drift. The question that remains, then, is whether or not the rate of divergence at neutral sites will be sped up or slowed down if they are in gametic disequilibrium with nucleotides that are acted on by natural selection. This question was first answered by Birky and Walsh (1988).

Earlier in the chapter we established that the rate of divergence was a function of the rate of substitution (see equation 8.32). The expected rate of substitution within a species is determined by the

scaled mutation rate  $\left(2N_e\mu = \frac{\theta}{2}\right)$  and the probability

of fixation for mutations ( $P_F$ ), which can be stated in an equation as

$$k = \frac{\theta}{2} P_F \quad (8.60)$$

As shown earlier in the chapter, for independent neutral mutations  $P_F = \frac{1}{2N}$ , as this is also the initial frequency of each new mutation. Here we use a new symbol to express the fixation probability because the probability of fixation for neutral mutations ( $P_F$ ) may well be different when a mutation is in gametic disequilibrium with a selected site than when a mutation is independent.

Assume that we have a neutral locus with two alleles A and a. The frequency of the A allele is  $x$  and the frequency of the a allele is therefore  $1 - x$ . Further assume that this neutral locus is completely linked to another locus where all alleles are under the influence of very strong positive natural selection. Let's imagine that a new mutation occurs at the selected locus that is infinitely advantageous and so goes to fixation instantly. What is the probability that the A allele at the neutral locus is also fixed due to hitch-hiking? The A allele at the neutral locus has a frequency of  $x$  and therefore there is also the probability  $x$  that the new advantageous mutation at the selected locus is linked to the A allele. Thus, there is the probability  $x$  that the A allele will sweep to fixation with the new mutation at the selected locus. However, the probability that the A allele is fixed by genetic drift is also  $x$  since that is the initial frequency of the A allele. Therefore, even complete linkage to the

selected allele does not alter the fixation probability for the A allele.

Birky and Walsh (1988) work through a more general analytical case and present the results of simulations, all showing that neither positive nor negative natural selection will change substitution rates at neutral sites. This occurs because the increased probability of substitution of the copies of a neutral allele linked to a selected site is counterbalanced by a decrease of exactly the same amount in the probability of fixation of all the neutral allele copies that are not linked to a selected site.

## Chapter 8 review

- The neutral theory is a widely used null hypothesis in molecular evolution, predicting patterns and rates of DNA sequence change under the assumption that all mutations have no fitness advantage or disadvantage. Even though genetic drift leads to fixation or loss, neutral alleles experience a random walk to these end points that results in transient genetic variation.
- Neutral theory predicts that polymorphism, or genetic variation within populations, is a function of the effective population size and the mutation rate. Larger effective population sizes or higher mutation rates result in higher levels of equilibrium polymorphism.
- Neutral theory predicts that the rate of divergence, the accumulation of fixed nucleotide differences between two species, is a function of only the mutation rate.
- Nearly neutral theory uses the assumption that many mutations are effectively neutral because their selection coefficients are less than the pressure of genetic drift. When  $4N_e s = 1$ , genetic drift and natural selection are equally likely to dictate the fate of a new mutation.
- Apparent divergence between two DNA sequences may be underestimated because of multiple hit mutations or homoplasy. Nucleotide substitution models serve to correct observed divergence for multiple hits, giving a better estimate of actual divergence.
- Nucleotide diversity ( $\pi$ ) and the number of segregating sites ( $S$ ) are two measures of DNA sequence polymorphism that can be used to estimate  $\theta = 4N_e\mu$ .
- The molecular clock hypothesis uses the neutral theory prediction that divergence occurs at a

- constant rate over time to estimate the time that has elapsed since the two sequences shared a common ancestor.
- Heterogeneity in the rate of divergence over time is common and leads to difficulty equating divergence with time since divergence.
  - Under a Poisson process model, the variance in substitution rates should equal the mean substitution rate to give an index of dispersion of one. The index of dispersion is often not equal to one, suggesting either that the substitution rate is not dictated by purely neutral processes or that the Poisson model is not an apt description of the neutral substitution process.
  - Rate heterogeneity can be consistent with neutral evolution such as when mutation rates are constant per generation but generations span different lengths of time. Alternatively, rate heterogeneity may be caused by natural selection that changes the probability of substitution for mutations depending on their fitness.
  - The HKA and MK tests examine the neutral prediction that polymorphism and divergence should be proportional since both are functions of the mutation rate. Tajima's  $D$  compares  $\theta$  estimated from nucleotide polymorphism and the number of segregating sites, which are expected to be equal under neutrality. Mismatch distributions are expected to be bimodal under the standard neutral model. These are not tests only for the action of natural selection, since population structure and changes in the effective population size through time also lead to rejection of the null hypothesis.
  - Mutations acted on by natural selection can alter levels of polymorphism at linked neutral nucleotide sites. Genetic hitch-hiking reduces polymorphism at neighboring sites because positive selection will bring a mutation and linked neutral mutations to fixation but drive other mutations to loss in the process. Negative selection against some new mutations also reduces polymorphism at linked sites in a process called background selection. Balancing selection can elevate polymorphism at linked neutral sites because long-lived allele copies accumulate numerous mutations over time.
  - Divergence rates of neutral nucleotide sites are not impacted by natural selection at linked nucleotide sites.

### Further reading

Motoo Kimura provided a readily accessible review of neutral theory in:

Kimura M. 1989. The neutral theory of molecular evolution and the world view of neutralists. *Genome* 31: 24–31.

For a review of the nearly neutral theory see:

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annual Reviews of Ecology and Systematics* 23: 263–86.

A non-technical overview of molecular clock concepts, methods, and applications can be found in:

Bromham L. and Penny D. 2003. The modern molecular clock. *Nature Reviews Genetics* 4: 216–24.

A review of possible explanations for the overdispersed molecular clock can be found in:

Culter DJ. 2000. Understanding the overdispersed molecular clock. *Genetics* 154: 1403–17.

Edward Hooper's popular 1999 book *The River: a Journey to the Source of HIV and AIDS* (Little, Brown and Co., Boston, MA) examined in detail evidence connecting the origin of HIV and polio vaccination campaigns. While all of these ideas have been discredited, the book makes interesting reading as science fiction.

John Gillespie's 1991 classic book *The Causes of Molecular Evolution* (Oxford University Press, New York) is a wide-ranging discussion and review of empirical data as well as models of molecular evolution and the molecular clock.

For a review of empirical studies where directional or balancing natural selection has been invoked to explain observed polymorphism, see:

Hedrick PW. 2006. Genetic polymorphism in heterogeneous environments: the age of genomics. *Annual Review of Ecology Evolution and Systematics* 37: 67–93.

## Problem box answers

### Problem box 8.1 answer

For the wheat–maize divergence 60 million years ago divergence rates are

$$\text{coxI} \quad \frac{0.0504 \text{ substitutions per site}}{2(60 \text{ million years})} = 0.00042 \text{ substitutions per site per million years}$$

$$\text{atp9} \quad \frac{0.1374 \text{ substitutions per site}}{2(60 \text{ million years})} = 0.001145 \text{ substitutions per site per million years}$$

$$\text{nad4} \quad \frac{0.0381 \text{ substitutions per site}}{2(60 \text{ million years})} = 0.000318 \text{ substitutions per site per million years}$$

Using these absolute divergence rates, the divergence time of the monocot–dicot split is estimated as

$$\text{coxI} \quad \frac{0.2060 \text{ substitutions per site}}{(2)0.00042 \text{ substitutions per site per million years}} = 245.2 \text{ million years}$$

$$\text{atp9} \quad \frac{0.4439 \text{ substitutions per site}}{(2)0.001145 \text{ substitutions per site per million years}} = 193.8 \text{ million years}$$

$$\text{nad4} \quad \frac{0.1101 \text{ substitutions per site}}{(2)0.000318 \text{ substitutions per site per million years}} = 173.1 \text{ million years}$$

In both cases there is a factor of 2 in the denominator because there are two lineages accumulating substitutions independently during divergence. The estimated divergence time clearly depends on the locus used since the molecular clock ticks at slightly different rates per million years for each locus. The average divergence time of 204 million years ago for these three loci matches the average of about 200 million years based on all available data (Laroche & Bousquet 1995).

### Problem box 8.2 answer

The Mayotte and Europe populations both contain a sample of 15 sequences for the *runt* locus so that  $n = 15$

$$q_1 = \sum_{k=1}^{n-1} \frac{1}{k} = \sum \left( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} \right) = 3.2516$$

and

$$e_1 = \frac{n+1}{3q_1(n-1)} - \frac{1}{q_1^2} = \frac{15+1}{3(3.2516)(15-1)} - \frac{1}{3.2516^2} = -0.02258$$

Next,

$$a_2 = \sum_{k=1}^{n-1} \frac{1}{k^2} = \sum \left( \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} + \frac{1}{8^2} + \frac{1}{9^2} + \frac{1}{10^2} + \frac{1}{11^2} + \frac{1}{12^2} + \frac{1}{13^2} + \frac{1}{14^2} \right) \\ = 1.576$$

and

$$c = \frac{2(n^2 + n + 3)}{9n(n - 1)} - \frac{n + 2}{a_1 n} + \frac{a_2}{a_1^2} = \frac{2(15^2 + 15 + 3)}{9(15)(15 - 1)} - \frac{15 + 2}{(3.2516)(15)} + \frac{0.576}{(3.2516)^2} = 0.04813$$

so that

$$e_2 = \frac{c}{a_1^2 + a_2} = \frac{0.04813}{3.2516^2 + 1.576} = 0.00396$$

In the European population there are 17 segregating sites out of a total of 556 sites so that

$$\hat{\theta}_S = \frac{p_S}{a_1} = \frac{17/556}{3.2516} = 0.0094$$

whereas in the Mayotte population there are 34 segregating sites out of a total of 538 sites so that

$$\hat{\theta}_S = \frac{S}{a_1} = \frac{34/538}{3.2516} = 0.0194$$

Tajima's  $D$  for the European population is

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{e_1 p_S + e_2 p_S(p_S - 1)}} = \frac{0.0124 - 0.0094}{\sqrt{(0.0226)0.0306 + (0.00396)(0.0306)(0.0306 - 1)}} \\ = 0.0030/0.02393 = 0.1254$$

whereas Tajima's  $D$  for the Mayotte population is

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{e_1 p_S + e_2 p_S(p_S - 1)}} = \frac{0.0135 - 0.0194}{\sqrt{(0.0226)(0.0632) + (0.00396)(0.0632)(0.0632 - 1)}} \\ = -0.0059/0.0345 = -0.0171.$$

Neither population has patterns of DNA sequence polymorphism that deviate from those expected under the standard neutral model.

## CHAPTER 9

# Quantitative trait variation and evolution

### 9.1 Quantitative traits

- Components of phenotypic variation.
- Components of genotypic variation ( $V_G$ ).
- Inheritance of additive ( $V_A$ ), dominance ( $V_D$ ), and epistasis ( $V_I$ ) components of genotypic variation.
- Genotype-by-environment interaction ( $V_{G\times E}$ ).
- Additional sources of phenotypic variation.

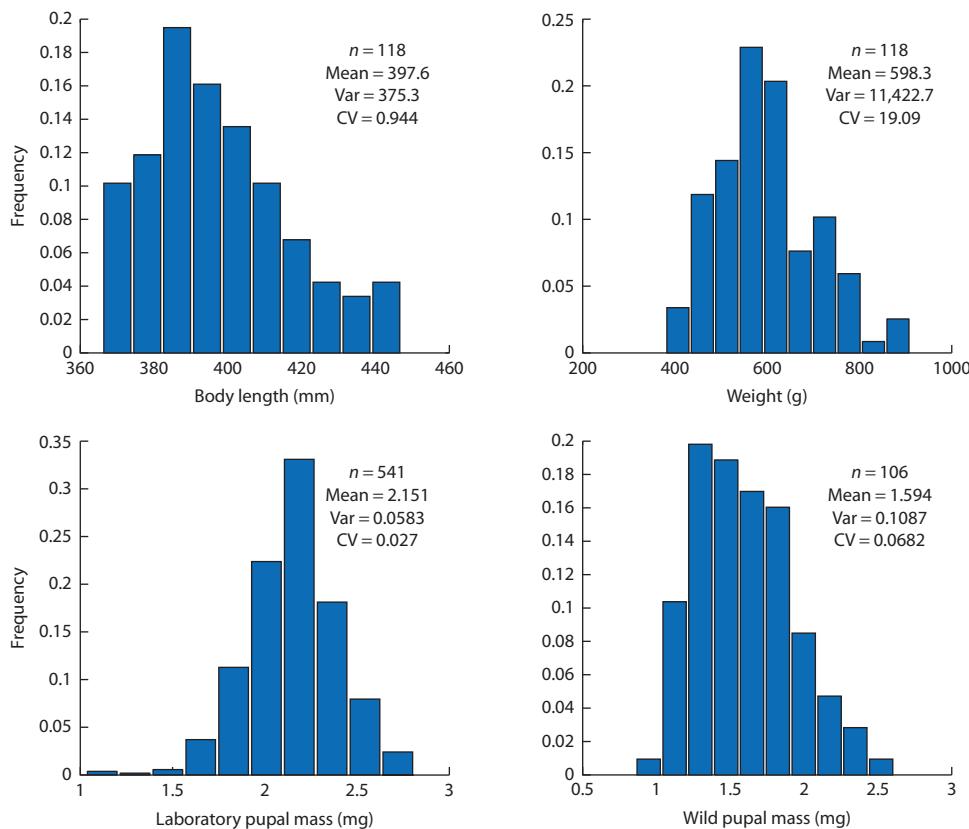
In the other chapters of this book, the concept of phenotype employed is somewhat simplistic. This is out of necessity because the emphasis in other chapters is on expectations for genotype and allele frequencies rather than on understanding the causes of variation in phenotype. Phenotypes were assumed to be completely determined by the genotype, and to have two or three discrete classes that correspond exactly to the three genotypes of a single locus with two alleles. (A minor exception is the two-locus model of natural selection where the phenotype is fitness.) While there certainly are examples of phenotypes in natural populations that fit this description, the majority of phenotypes are probably not well characterized by these assumptions. This chapter will expand the concept of phenotype and develop the concepts needed to understand the relationship between various types of genetic variation and phenotypic variation. The chapter will introduce the various components of quantitative trait variation, show how these components can be used to describe inheritance of phenotypes, and also explore the action of natural selection and genetic drift on complex phenotypes. The chapter will wrap up with a section devoted to genetic mapping methods used to identify and characterize individual loci that cause quantitative trait variation.

Think of variable phenotypes such as human height, the number of ears on a corn plant, daily milk production in domestic cows, wood density in a tree species, or the probability of onset of a disease such as diabetes or hypertension. Think next of complex behavioral phenotypes such as sexual preference or

propensity to substance addiction in humans, success in male–male contests for mates, or the quality of mates chosen by females. Also think of phenotypes related to Darwinian fitness such as individual size, number of gametes, number of progeny, or number of days an individual survives. While each of these classes of phenotypes seems unrelated, they all share features in common as **quantitative traits** (also called metric traits). Quantitative traits are sometimes called multifactorial traits because the variation in phenotype among individuals has multiple causes. Quantitative trait variation among individuals is a product of differences in genotype produced by multiple genes as well as differences in environmental conditions experienced by each individual.

The hallmark of quantitative traits is a broad range of variation characterized by a continuous distribution of individual phenotypes in a population (Fig. 9.1). **Continuous traits** have a scale of measurement that is naturally continuous, such as quantifying height in centimeters or weight in kilograms. **Meristic traits** exhibit a large number of discrete classes, such as the number of bristles on a fruit fly or the number of leaves on a tree, that forms a distribution of phenotypic values. **Threshold** or **liability traits** are continuously distributed phenotypes with some trait value that defines an upper or lower limit. Trait values above or below the threshold define qualitatively distinct categories such as “normal” and “symptomatic.” The production of insulin is one example, where human populations show a continuous distribution of insulin production and individuals are clinically recognized as diabetic when insulin production drops below a threshold level.

In quantitative genetics the words **phenotype**, **trait**, and **character** are all considered synonymous. The term **value** is used to refer to the phenotype in the same units that it is measured in. **Phenotypic value** refers to the observed phenotype of an individual, for example observing that an individual fish has a value of 400 mm for the phenotype of body length.



**Figure 9.1** Examples of continuous quantitative trait distributions. The top panels show the distributions of body length and weight in a sample of 3-year-old striped bass (*Morone saxatilis*). The bottom panels are pupal mass distributions for mosquitoes raised in laboratory or field conditions. Each panel gives the sample size (*n*), mean, variance, and coefficient of variation (CV) that quantify the phenotypic distribution. Striped bass data are from L. Pieper (unpublished results). Mosquito pupal mass data from Armbruster and Conn (2006) and P. Armbruster (unpublished results).

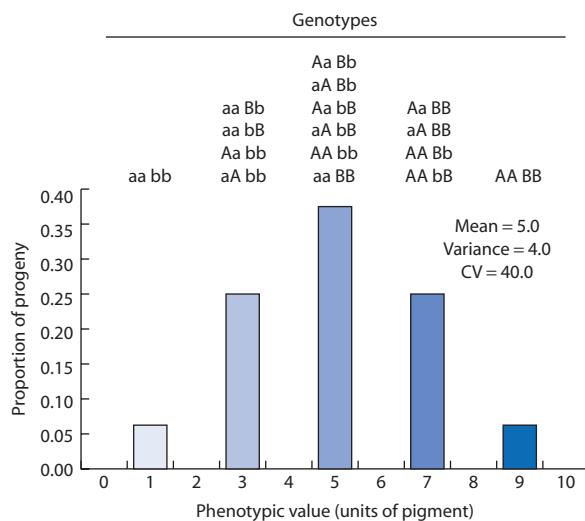
**Quantitative trait or character** A phenotype where values for numerous individuals in a population are continuously distributed and that variation has both genetic and environmental causes.

**Value** The phenotypic measurement of an individual in the units of trait measurement; the mean phenotypic measurement of a population.

Biologists have been aware of continuously distributed phenotypes since Mendel's time. After the recognition of Mendel's work in the early twentieth century there was a major controversy involving Mendelian genetics and quantitative genetics. The biometric school was a branch of genetics devoted to understanding the inheritance of continuously distributed phenotypes. Members of the biometric school pioneered the application of statistical methods to quantify and compare continuous phenotypic variation. Francis Galton (half-cousin of Charles Darwin) founded the biometric school through his study of human phenotypes and was an innovator in math

and statistics as well as the founder of the eugenics movement (see Gillham 2001). Adherents of the biometric school argued that continuous traits were due to a distinct set of biological causes and could not be explained by Mendel's theory of particulate inheritance. Galton tried unsuccessfully to develop a model that explained the inheritance of quantitative traits without reference to Mendelian genetics (see Provine 1971; Bulmer 1998). In 1918, Ronald A. Fisher (the same R.A. Fisher who contributed the fundamental theorem of natural selection) published a seminal paper showing definitively how single Mendelian loci that individually produced discrete classes of phenotypes could combine to result in continuously distributed phenotypes (Fisher 1918).

The continuous distribution of phenotypes under Mendelian inheritance is due to **polygenic variation**. The continuous variation in phenotype results from the simultaneous segregation of several to many independent Mendelian loci. Figure 9.2 shows the phenotypic distribution for a trait determined by two independent Mendelian loci. In this two-locus illustration, the expected frequency distribution of phenotypes in the population is stepped and not smooth. However, the distribution clearly resembles



**Figure 9.2** The phenotypic distribution for a trait determined by two Mendelian loci. This hypothetical phenotype might be something like flower color, ranging between nearly white and deep blue. The genotypes are those expected in a large number of progeny from a cross between two doubly heterozygous ( $AaBb$ ) parents. Alternatively, the genotype frequencies are those expected in a large number of progeny from a parental population with Hardy–Weinberg genotype frequencies where mating is random and all allele frequencies are  $\frac{1}{2}$ . Each  $a$  or  $b$  allele in a genotype causes  $\frac{1}{4}$  unit of pigment in the phenotype whereas each  $A$  or  $B$  allele in a genotype causes  $2\frac{1}{4}$  units of pigment in the phenotype. As the number of loci contributing to the trait increases, the expected frequency of any individual genotype decreases and the phenotypic distribution will become smoother. For this phenotypic distribution the mean = 5, the variance = 4, and the coefficient of variation (CV) = 40.

a normal distribution with symmetry about a single central mode. This hypothetical two locus pigment trait has five classes between 10 and 90% pigment that are evenly spaced as expected for a quantitative trait. One assumption in Fig. 9.2 is that the population of individuals is large. This ensures there is not a lot of chance sampling variation that would cause an observed distribution of phenotypes to differ greatly from its expected frequencies (e.g. due to chance no individuals with 10% pigment phenotypes are observed). This assumption is implicit in all of the expectations in quantitative genetics.

As the number of loci determining a trait increases, the phenotypic differences between adjacent genotype classes also decreases, resulting in a phenotypic distribution that is smoother. For a phenotype where there is codominance (also known as semi-dominance) and allele frequencies are all  $\frac{1}{2}$ , the expected frequencies of each class of phenotypic values can be found by taking the expected frequencies of the three geno-

types at a single locus and raising it to the power of the number of loci:

$$\left( \frac{1}{4} + \frac{1}{2} + \frac{1}{4} \right)^n \quad (9.1)$$

where the  $n$  is the number of loci that contribute equally to the quantitative trait. From this equation it is also apparent that a smoother distribution of phenotypic values would result from loci with more than two alleles because each locus would produce more than three genotypes under random mating. For the two locus phenotype shown in Fig. 9.2, the expected frequencies of the phenotypes are found by multiplying the frequencies of the Hardy–Weinberg genotype frequencies for each locus:

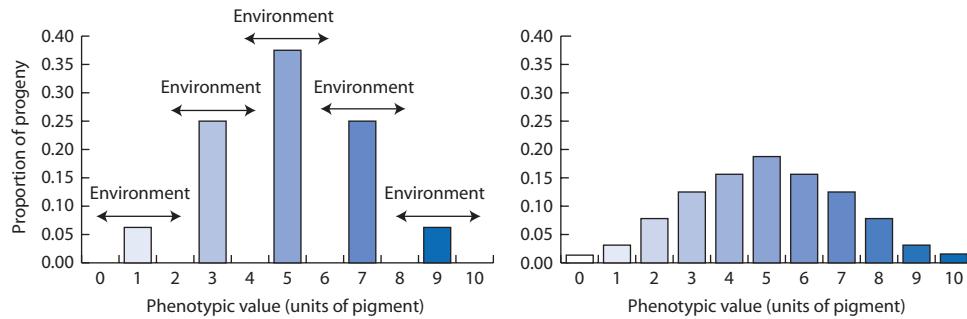
$$\left( \frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa \right) \left( \frac{1}{4}BB + \frac{1}{2}Bb + \frac{1}{4}bb \right) \quad (9.2)$$

and then summing the frequencies of those genotypes that have identical phenotypes.

### Problem box 9.1 Phenotypic distribution produced by Mendelian inheritance of three diallelic loci

Calculate the expected genotype frequencies for three locus genotypes in a population where mating is random, all loci have two alleles, and allele frequencies at all loci are  $\frac{1}{2}$ . Then construct a histogram of phenotypic values using the minimum and maximum phenotypic values used in Fig. 9.2 by assuming that alleles have phenotypic values of  $\frac{1}{6}$  (lower-case letter alleles) and  $1\frac{1}{2}$  (capital-letter alleles).

Another primary cause of variation among individuals in quantitative traits is the environment. Even if there is only a single genotype, the phenotype expressed by each individual in a population will vary somewhat depending on the environmental conditions each individual experiences. For example, the biomass and fruit production of plants is impacted by the amount of sunlight and nitrogen each individual receives. Another example of environmental variation in quantitative traits is the role



**Figure 9.3** The effect of environmental variation on phenotypic variation. The phenotypic distribution on the left is produced by two Mendelian loci with all allele frequencies equal to  $\frac{1}{2}$  as in Figure 9.2. If the environment causes some variation in the phenotype expressed by each genotype, then the distribution of phenotypes produced by polygenic variation becomes both smoother and wider. In this illustration, environmental variation causes 50% of the individuals of each genotype to randomly increase or decrease one unit in phenotypic value. Although the average effect of environmental variation here is a zero change in phenotypic value, the phenotypic variance increases.

of diet and exercise in human disease, where better conditions tend to lessen the frequency or severity of disease phenotypes. Figure 9.3 shows how environmental differences among individuals contribute to the continuous distribution of phenotypic variation. The left-hand panel of Fig. 9.3 shows the five phenotypes produced by a trait due to two diallelic loci. If each phenotypic class expressed by a genotype is modified by environmental variation, the distribution of phenotypes becomes both wider and smoother as shown in the right-hand panel of Fig. 9.3. The environmental variation experienced by individuals is also likely to be a truly continuous variable, unlike the discrete categories of genotypes produced by multiple loci, that then causes continuous variation in phenotypes.

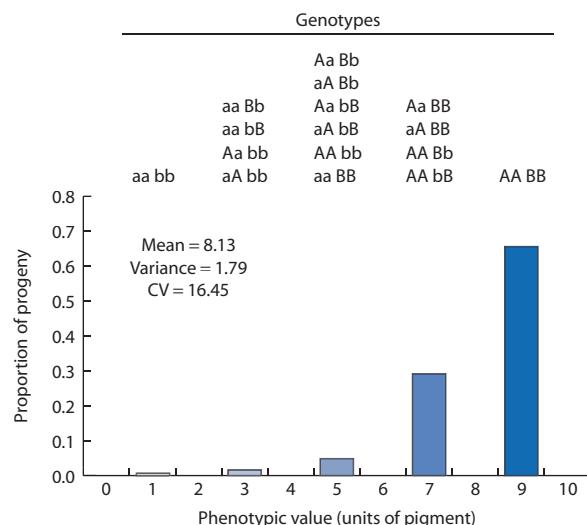
### Components of phenotypic variation

Now that we have seen how discrete Mendelian genetic variation for multilocus genotypes combined with continuous environmental variation produces continuous phenotypic distributions, let's represent the genetic and environmental causes of phenotypic variation in notation. In quantitative genetics it is customary to symbolize expected quantitative trait variation with a  $V$ . The  $V$  variable always bears a subscript to indicate a specific cause of phenotypic variation. The total variation in phenotype is represented by  $V_P$  and examples of total phenotypic variance are shown in Figs 9.1 and 9.3. This total phenotypic variation has both genetic and environmental causes. The phenotypic variation caused by variation in genotypes in the population is represented by  $V_G$ . Independently, the variation in phenotype

caused by the environment is represented by  $V_E$ . The equation

$$V_P = V_G + V_E \quad (9.3)$$

is used to represent that the total variation in phenotype in a population is the sum of phenotypic variation caused by genotype differences among



**Figure 9.4** Phenotypic distributions for a trait determined by two loci in populations with low additive genetic variation ( $V_A$ ). Genetic variation in phenotypes is additive since each  $a$  or  $b$  allele in a genotype contributes  $\frac{1}{4}$  unit of pigment while each  $A$  or  $B$  allele in a genotype contributes  $2\frac{1}{4}$  units of pigment irrespective of genotype. However, total genetic variation is relatively low since allele frequencies are near fixation and loss (the frequency of  $a$  and  $b$  alleles is 0.1 while the frequency of  $A$  and  $B$  alleles 0.9). Compare with Fig. 9.2 where allele frequencies are all equal to  $\frac{1}{2}$  and allelic effects are additive.

individuals and phenotypic variation caused by the different environments individuals experience.

Quantifying and comparing the variance of quantitative traits utilizes a set of summary statistics. Since quantitative trait distributions usually approximate hump-shaped normal distributions, statistics that describe normal distributions are useful (see the Appendix for a primer on basic statistics used in quantitative genetics). The middle or central tendency of a quantitative trait distribution is described by its average or mean. The spread of the observations around this central tendency is described by the variance of the distribution. The coefficient of variation or CV ( $CV = \frac{\sqrt{var}}{\bar{x}}(100)$  where  $\bar{x}$  is the trait mean and var is the trait variance) is used to compare the variances of distributions after correcting for differences due only to the value of the mean. The CV expresses the magnitude of the standard deviation as a percentage of the mean. As an illustration, look at the quantitative trait distributions in Fig. 9.1. Body length and weight for fish have much greater mean values than pupal weights for mosquitoes. But, using the CV, we can compare the variation in these traits to see that the standard deviation is about 5% of the mean for body length in striped bass and almost 21% of the mean for pupal mass in wild-reared mosquitoes. The CV also allows us to properly compare the distributions of pupal mass for laboratory-reared mosquitoes and mosquitoes grown in the wild to see that the spread of pupal mass values around the mean is about two times wider in the wild mosquitoes ( $CV = 20.68$ ) than in laboratory-reared mosquitoes ( $CV = 11.23$ ).

Biologically, equation 9.3 helps us to recognize and quantify the determinants of phenotypic variation. Equation 9.3 divides the causes of quantitative trait variation into those due to heredity and those due to the environment, or into the phenotypic variation caused by nature and that caused by nurture. Look again at the bottom panels of Fig. 9.1 that show pupal mass distributions for mosquitoes. Imagine that the laboratory and wild populations of mosquitoes have roughly the same genotype frequencies or  $V_G$ . The greater pupal mass variance or  $V_p$  in the wild population could then be explained by greater environmental variance or  $V_E$ , a common observation since laboratory conditions tend to be more uniform and benign. As an alternative, imagine that the laboratory population has less  $V_G$  since it was founded from a small sample of individuals and it also has less  $V_E$  since the conditions individuals experience are more uniform. Then the greater pupal mass variance ( $V_p$ ) in the wild could be caused by both more genetic variation ( $V_G$ ) and more environmental variation ( $V_E$ ) than experienced by individuals in the laboratory. In this fashion it is possible to quantify and compare the relative causes of phenotypic variation.

The  $V$  notation compactly expresses the multiple causes of total phenotypic variation,  $V_p$ , in quantitative traits. Genotypic variation ( $V_G$ ) and environmental variation ( $V_E$ ) are not the only causes of total phenotypic variation. Table 9.1 summarizes additional causes of total phenotypic variation. Notice that  $V_G$  is actually broken down into three distinct components due to the effects of alleles, the effects of dominance, and the effects of gene interaction or epistasis (epistasis is symbolized  $V_I$  since  $V_E$  is

**Table 9.1** Symbols commonly used to refer to categories or causes of variation in quantitative traits. Variation is indicated by  $V$  while the specific cause of that variation is indicated by a subscript capital letter (with one exception). Total genetic variation ( $V_G$ ) in phenotype can be divided into three subcategories.

Symbol	Definition
$V_p$	Total variance in a quantitative trait or phenotype
$V_G$	Variance in phenotype due to all genetic causes
$V_A$	Variance in phenotype caused by additive genetic variance or the effects of alleles
$V_D$	Variance in phenotype caused by dominance genetic variance or deviations from additive values due to dominance
$V_I$	Variance in phenotype caused by interaction genetic variance (epistasis between and among loci)
$V_E$	Variance in phenotype caused by environmental variation
$V_{G\times E}$	Variance in phenotype caused by genotype-by-environment interaction
$V_{Ec}$	Variance in phenotype caused by environmental variation shared in common by parents and offspring or by relatives

already taken to represent environmental variance). In Chapter 10, the Mendelian basis of quantitative traits is used to derive quantitative expectations for  $V_G$  and its components. For now, let's continue to develop an intuitive understanding of the causes of total phenotypic variation in addition to  $V_G$  and  $V_E$ .

### **Components of genotypic variation ( $V_G$ )**

Thus far, we have distinguished between the genetic and environmental causes of phenotypic variation. In quantitative genetics, a primary goal is to explain the hereditary causes of phenotypic variation. To fully understand the genetic contribution to total phenotypic variation it is necessary to recognize that  $V_G$  is itself made up of three separate causal components. The components of the total genotypic variation are

$$V_G = V_A + V_D + V_I \quad (9.4)$$

where  $V_A$  is additive genetic variation,  $V_D$  is dominance genetic variation, and  $V_I$  is interaction or epistasis genetic variation.

The **additive genetic variance** is the genotypic variance caused by the cumulative phenotypic effects of alleles when they are assembled into genotypes. **Additive** simply means that the phenotypic effect of each allele can be added together to determine the phenotypic value of any genotype. When alleles have additive effects, then the specific pairing of alleles in a genotype, be it a homozygote or a heterozygote, has no impact on the way alleles combine to produce a phenotype. In other words, additivity describes the situation when each allele has the same effect on the phenotypic value regardless of the context where it is found. Figure 9.2 shows an illustration of additive genetic variation in a quantitative trait. Each a or b allele in a genotype contributes  $\frac{1}{4}$  unit of pigment in the phenotype whereas each A or B allele in a genotype contributes  $2^{\frac{1}{4}}$  units of pigment in the phenotype. The phenotype of any genotype can be determined simply by adding together the phenotypic effects of the two alleles.

When gene action is additive, the amount of additive genetic variance ( $V_A$ ) in a population depends on allele frequencies. There is more additive genetic variance in phenotype when alleles are at intermediate allele frequencies than when they are near fixation and loss. This is because intermediate allele frequencies result in all possible genotypes (assuming random mating) being represented in the population, which in turn produces a wide range of phenotypes.

Figure 9.2 shows the wide range of phenotypes found in a population where genetic variation is additive and all alleles are at a frequency of  $\frac{1}{2}$ .

To see how genetic variation causes phenotypic variation when allelic effects are strictly additive, compare Figs 9.2 and 9.4. In both figures, phenotypes for each genotype are determined by adding together the phenotypic effects of the alleles in a genotype (each a or b contributes  $\frac{1}{4}$  unit of pigment and each A or B contributes  $2^{\frac{1}{4}}$  units of pigment). However, the allele frequencies in Fig. 9.4 are closer to fixation and loss so there is a much less even genotype frequency distribution in the population. The change in allele frequencies leads to two very common genotypes and three genotypes that are very rare. This reduction in genetic variation causes a reduction in phenotypic variation. With allele frequencies nearer fixation and loss, the frequency distribution of phenotypes is now narrower and clumped around values at the upper end of the range.

**Additive genetic variance ( $V_A$ )** The proportion of the total genotypic variance ( $V_G$ ) caused by the sum of phenotypic effects of alleles when they are assembled into genotypes.

In addition to the additive effects of alleles, quantitative trait variation is also caused by the effect of genotypes. Dominance and epistasis are properties of genotypes that can be thought of in two conceptually distinct ways (see Wade 1992; Cheverud & Routman 1995; Phillips 1998). In a physiological or functional sense, dominance and epistasis describe the way phenotypes map to genotypes. Both are forms of genetic interaction. With dominance, the phenotype depends on the combination of alleles within a locus that compose a genotype (allele interactions). With epistasis, the phenotype depends on the combination of genotypes at two or more loci (genotype interactions). Two diallelic loci can produce eight distinct ways by which genotypic values are determined by combinations of two-locus genotypes. These eight genetic effects are illustrated in Table 9.2. In general, there are a total of  $3^n - 1$  distinct genetic effects for  $n$  diallelic loci (Cockerham 1954; see also Goodnight 2000).

At the same time, both dominance and epistasis have a population-level meaning that is statistical. Both dominance and epistasis can be thought of as the “leftover” part of genotypic variance in the population

**Table 9.2** The eight uncorrelated (or orthogonal) types of genetic effects that can occur between two diallelic loci. Four of the eight types of genetic effects are interactions that give rise to  $V_I$ . The genotypic values assume all allele frequencies are  $1/2$ . Table after Goodnight (2000).

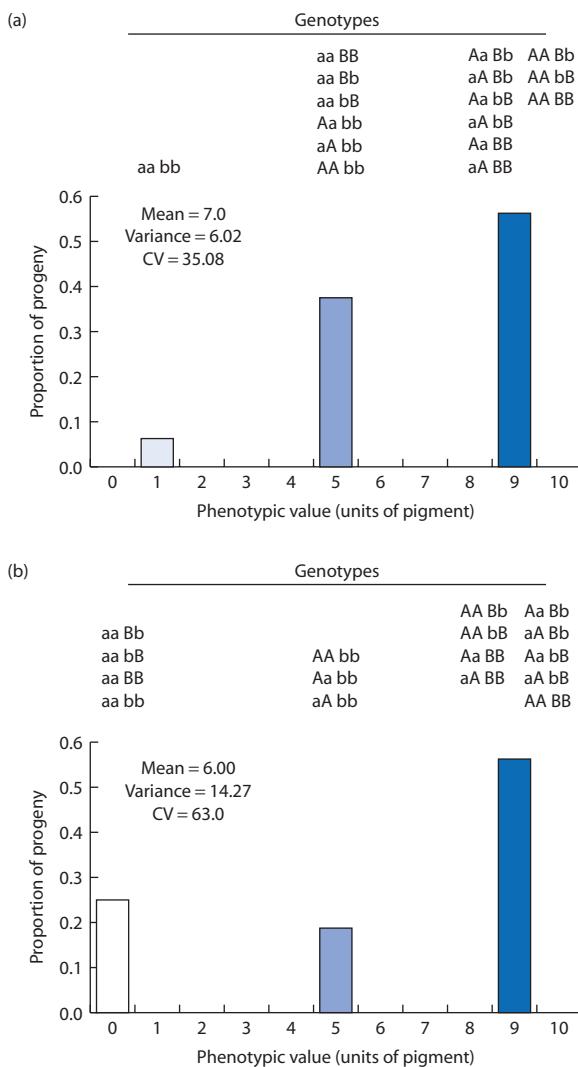
Genetic effect	Genotypes and phenotypes . . .	Genotypic value		
		$A_1A_1$	$A_1A_2$	$A_2A_2$
Additive A locus	$B_1B_1$	1	0	-1
	$B_1B_2$	1	0	-1
	$B_2B_2$	1	0	-1
Additive B locus	$B_1B_1$	1	1	1
	$B_1B_2$	0	0	0
	$B_2B_2$	-1	-1	-1
A locus dominance	$B_1B_1$	0	1	0
	$B_1B_2$	0	1	0
	$B_2B_2$	0	1	0
B locus dominance	$B_1B_1$	0	0	0
	$B_1B_2$	1	1	1
	$B_2B_2$	0	0	0
Additive-by-additive interaction	$B_1B_1$	1	0	-1
	$B_1B_2$	0	0	0
	$B_2B_2$	-1	0	1
Additive (A locus)-by-dominance (B locus) interaction	$B_1B_1$	1	-1	1
	$B_1B_2$	0	0	0
	$B_2B_2$	-1	1	-1
Dominance (A locus)-by-additive (B locus) interaction	$B_1B_1$	1	0	-1
	$B_1B_2$	-1	0	1
	$B_2B_2$	1	0	-1
Dominance-by-dominance interaction	$B_1B_1$	-1	1	-1
	$B_1B_2$	1	-1	1
	$B_2B_2$	-1	1	-1

that is not explained by the additive genetic variance due to differences in alleles among individuals. The magnitudes of  $V_D$  and  $V_I$  in the statistical sense are a function of both the genotype frequencies in the population as well as the relationship between genotype and phenotype. These non-additive parts of the phenotypic variance are sometimes called dominance and epistasis *deviations* since they are measured as differences from the variance that would be expected in a population if all genetic effects were additive.

The distinction between additive and dominance genetic variance in quantitative traits can be seen with a modification of the rules that specify the relationship between genotypes and phenotypes. Under additivity, each genotype's phenotypic value is determined by the sum of the phenotypic effects of the alleles that compose it. An alternative relationship between the genotype and phenotype is seen with complete dominance, where heterozygotes have the same phenotype as one of the homozygotes.

With dominance, the phenotypic value of the heterozygote is no longer determined by adding together the phenotypic effects of the two alleles that compose it. When gene action shows dominance, the phenotypic contribution of an allele depends on pairing of alleles in the genotype where it resides. Dominance causes the phenotype to be defined by the genotype context rather than being independent of the pairing of alleles in a genotype as it is under additivity.

The additivity rule that applies in Fig. 9.2 is changed to complete dominance in Fig. 9.5a. Dominance markedly changes the phenotypic frequency distribution, even though allele frequencies remain constant in the two panels. In Fig. 9.2 the phenotypic distribution is symmetric and exhibits every possible value of the phenotype. In contrast, Fig. 9.5a shows that about 56% of the individuals in the population have phenotypic values of 9. There are no longer any individuals with phenotypic values of 3 or 7 units of pigment. In this example, dominance



**Figure 9.5** Phenotypic distributions for a trait determined by two loci with either complete dominance or epistasis. (a) Due to dominance, the majority of phenotypes are 9 units of pigment and no individuals display phenotypes with 3 or 7 units of pigment. In this example of two diallelic loci, the frequency of heterozygotes is at a maximum because all allele frequencies are  $1/2$ . In general, total phenotypic variation in the population can be increased or decreased by dominance. (b) The bottom phenotypic distribution shows an example of epistasis identical to the two-locus system that determines coat color in Labrador retriever dogs. One dominant locus controls pigment color (BB and Bb genotypes have black coats and bb genotypes have brown coats) while a second completely dominant locus controls the presence or absence of pigment in hairs (AA and Aa genotypes have pigmented hair and aa genotypes have unpigmented hair). In this example of dominance by dominance epistasis, phenotypic expression of the genotype at the coat color locus depends on the genotype at the pigmentation locus. In both graphs, the genotype frequencies are those expected under Hardy–Weinberg assumptions where all allele frequencies are  $1/2$ . The mean and variance are based on a population of 1000 individuals.

increases  $V_p$  somewhat compared to additivity (see Fig. 9.2) because more genotypes (heterozygotes and dominant homozygotes) exhibit genotypic values at the upper extreme of the distribution. In general, dominance can either increase or decrease  $V_p$  depending on allele frequencies in the population. The Mendelian basis of dominance and the dominance variance are explained in detail in Chapter 10.

**Dominance genetic variance ( $V_D$ )** The proportion of the total genotypic variance ( $V_G$ ) caused by the deviation of genotypic values from their values under additive gene action caused by the combination of alleles assembled into a single-locus genotype.

**Epistasis or interaction genetic variance ( $V_I$ )** The proportion of the total genotypic variance ( $V_G$ ) due to the deviation of genotypic values from their values under additive gene action caused by interactions between and among loci.

Epistasis literally means “standing on” and denotes interactions between two or more loci that dictate the phenotypic value of a multilocus genotype. When epistasis is absent, the phenotypic value of a multilocus genotype is the sum of the phenotypic value of all single locus genotypes. With epistasis, the phenotypic contribution of the genotype at one locus depends on the genotypes of other loci that it is paired with. When there is epistasis then the *combination* of genotypes at two or more loci dictates the phenotypic value of a genotype. In the most extreme case, epistasis produces a phenotypic value for each multilocus genotype that is unique to that combination of genotypes. For example, with epistasis the phenotypic value of the AA genotype paired with the BB genotype (AABB) cannot be predicted from the phenotypic values of the AAB<sub>b</sub> or AA<sub>b</sub>B genotypes. The interaction genetic variance ( $V_I$ ) is caused by deviations of genotypic values from the values that would occur if each locus had additive effects.

Since the interactions between two loci can take many forms, a range of terminology has been used to describe the impact of locus interactions on phenotypic values (see Phillips et al. 2000). For example, the term **synergistic epistasis** describes the situation where a genotype has a larger effect on the phenotypic value in the presence of certain other genotypes than would be expected under additivity. In contrast,

**antagonistic epistasis** describes an interaction where a genotype has a smaller effect on the phenotypic value in the presence of certain other genotypes than would be expected under additivity.

A classic example of two-locus epistasis are the coat-color phenotypes of Labrador retriever dogs (see Fig. 9.5b). For the sake of illustration, assume one completely dominant locus controls pigment color, with BB and Bb genotypes having black coats and bb genotypes having brown coats (see Kerns et al. 2007 for a more complete description of the genetic basis of coat color in dogs). A second completely dominant locus controls the presence or absence of pigment in hairs. At the pigmentation locus, AA and Aa genotypes have pigmented hair and therefore exhibit the black or brown coat color determined by the B locus. (The pigmentation locus is often symbolized as the E locus but A is used here for consistency across examples.) However, if the pigmentation locus genotype is aa then the coat color locus has no effect and the coat is yellow because hair pigment is not produced. In a population of Labrador retriever dogs, interaction between the coat color and pigmentation loci will alter the mean and variance of coat color phenotypes relative to two loci that combine additively. The exact impact of the interaction on the phenotypic mean and variance will depend on the genotype frequencies at the two loci.

The distinction between phenotypic variation produced by additive gene action and dominance or epistasis helps clarify a subtle point of terminology in quantitative genetics. The problem is what exactly to call  $V_G$ . It is sometimes called genetic variation because multiple alleles in a population lead to variation that under additive gene action produces multiple phenotypes. But with dominance and epistasis, genetic variation is a product of multiple genotypes that then produce a range of phenotypes in a population. Calling  $V_G$  *genotypic variation* is probably best since it encompasses phenotypic variation due to both alleles and genotypes.

#### **Inheritance of additive ( $V_A$ ), dominance ( $V_D$ ), and epistasis ( $V_I$ ) genotypic variation**

Another way to appreciate the differences among the  $V_A$ ,  $V_D$ , and  $V_I$  components of the total genotypic variation is to consider an example of inheritance across a generation. Additive genetic variation ( $V_A$ ) has a distinct pattern of inheritance compared to dominance and epistasis genetic variation ( $V_D$  and  $V_I$ ). A critical distinction among the three components

of the total genotypic variance is that  $V_A$  is caused by the average phenotypic effects of *alleles* while  $V_D$  and  $V_I$  are caused by the average phenotypic effects of *genotypes*.

Additive genetic variation is the component of the genotypic variance that causes the phenotypic resemblance between relatives. For example, parents and their offspring or siblings (brothers and sisters) have a higher degree of phenotypic resemblance than two randomly sampled individuals in the same population. This average phenotypic resemblance comes about because relatives share alleles that are identical by descent. When alleles combine additively, then shared alleles translate into shared phenotypic values. Only when alleles have additive effects does genetic variation contribute to average resemblance between parents and offspring or among related individuals.

Examples of additive gene action across one generation are shown in Table 9.3. In the top half of the table alleles at one locus are assumed to act additively to determine phenotypic values. When crossing BB × bb parents or Bb × Bb parents, the parental mean phenotype and the progeny mean phenotype are always identical. The equality of mean phenotypes across one generation is remarkable giving that the genotypes of the parents and progeny are not identical. In fact, in the BB × bb cross, none of the progeny share their genotype with the parents since all progeny possess Bb genotypes. What is common between the parents and progeny in each case are allele frequencies. As long as alleles combine additively to determine phenotypic values, identical allele frequencies in two separate populations will produce identical mean phenotypes. This can be seen well in the second cross (B) for additive gene action, where a population of all Bb heterozygotes has the same allele frequencies and same mean phenotype as a population of  $\frac{1}{4}$  BB,  $\frac{1}{2}$  Bb, and  $\frac{1}{4}$  bb individuals.

In contrast to the additive effects of alleles are the genotype effects of dominance and epistasis. Dominance can be thought of as an interaction between the two alleles that make up a single-locus diploid genotype. With dominance, the genotypic value of the heterozygote is not just the sum of the two allelic effects but is some other value depending on how the two alleles interact when packaged into one genotype. While dominance is a continuous variable, complete dominance, where the heterozygote and the dominant homozygote have identical phenotypes, is a useful point of reference. In complete dominance, alleles behave differently in how they contribute to phenotype depending on whether a genotype is composed

<b>Additive gene action</b>			
Genotypes	BB	Bb	bb
Phenotypes	3	2	1
Cross		Mean phenotype	
(a) Parents	BB × bb		$\frac{3+1}{2} = 2$
Progeny	Bb		2
(b) Parents	Bb × Bb		2
Progeny	$\frac{1}{4}$ BB, $\frac{1}{2}$ Bb, $\frac{1}{4}$ bb		$\frac{1}{4}(3) + \frac{1}{2}(2) + \frac{1}{4}(1) = 2$

<b>Complete dominance</b>			
Genotypes	BB	Bb	bb
Phenotypes	3	3	1
Cross		Mean phenotype	
(a) Parents	BB × bb		$\frac{3+1}{2} = 2$
Progeny	Bb		3
(b) Parents	Bb × Bb		3
Progeny	$\frac{1}{4}$ BB, $\frac{1}{2}$ Bb, $\frac{1}{4}$ bb		$\frac{1}{4}(3) + \frac{1}{2}(3) + \frac{1}{4}(1) = 2.5$

**Table 9.3** Examples of parental and progeny mean phenotypes that illustrate the impacts of additive gene action (top) or complete dominance gene action (bottom). For both types of gene action, the phenotypic value of each genotype is given and the genotypes of two possible parental crosses are shown along with the genotypes in the progeny from each cross. Under additive gene action the mean phenotypic values are identical in the parents and progeny because phenotypic values are a function of allele frequencies and alleles are identical in parents and progeny. In contrast, under complete dominance parent and progeny mean phenotypic values differ because phenotypic values are a function of the genotype and genotype frequencies differ between parents and progeny.

of two identical alleles (homozygous genotypes) or two dissimilar alleles (heterozygous genotypes).

Two examples of the average phenotypic resemblance between parents and offspring under complete dominance are shown in the bottom half of Table 9.3. When crossing BB × bb parents (a) to yield a population of Bb progeny, the parental mean phenotype and the progeny mean phenotype are not identical. This lack of parent–offspring phenotypic resemblance occurs because the parental population and the progeny population do not share any genotypes in common. The parents are both homozygotes while all progeny are heterozygotes. The mean phenotype for parents and progeny is closer for the Bb × Bb parental cross (b). This occurs because 50% of the progeny have an identical genotype to the parents. Thus, with dominance, shared genotype frequencies will lead to phenotypic resemblance. Because particulate inheritance breaks up genotypes, alleles are inherited but diploid genotypes are not. Genotype frequencies are a consequence of how gametes combine to make progeny genotypes. Thus, the genotype effects of dominance ( $V_D$ ) do not contribute to average phenotypic resemblance between parents and offspring. Like dominance, epistasis ( $V_I$ ) is also a property of the genotype that does not contribute to average

phenotypic resemblance between parents and offspring because genotypes are not inherited. An exception is that additive by additive epistasis does contribute to resemblance between parents and offspring.

Whereas  $V_D$  does not contribute to the resemblance of parents and offspring, dominance variance can cause phenotypic resemblance between other types of relatives. In particular, dominance variance contributes to the phenotypic resemblance among full siblings (brothers and sisters). This occurs because full siblings can inherit identical genotypes since they share the same two parents in common. Thus, a shared heterozygote genotype would cause two full siblings to share a genotypic value caused by dominance. Resemblance among relatives is explored more fully in section 10.6.

#### **Genotype-by-environment interaction ( $V_{G\times E}$ )**

Phenotypic variation can be caused by the combination of genotypes and environments in a population. Up to this point we have assumed that genotypes are all equally sensitive to their environments, meaning that a change of environment would impact the phenotype of all genotypes to the same extent. In fact, genotypes very often have different degrees

of sensitivity to environmental conditions. This cause of phenotypic variance is called **genotype-by-environment interaction** and is symbolized by  $V_{G\times E}$ . This adds another term to the expression for the independent causes of total phenotypic variation in a population:

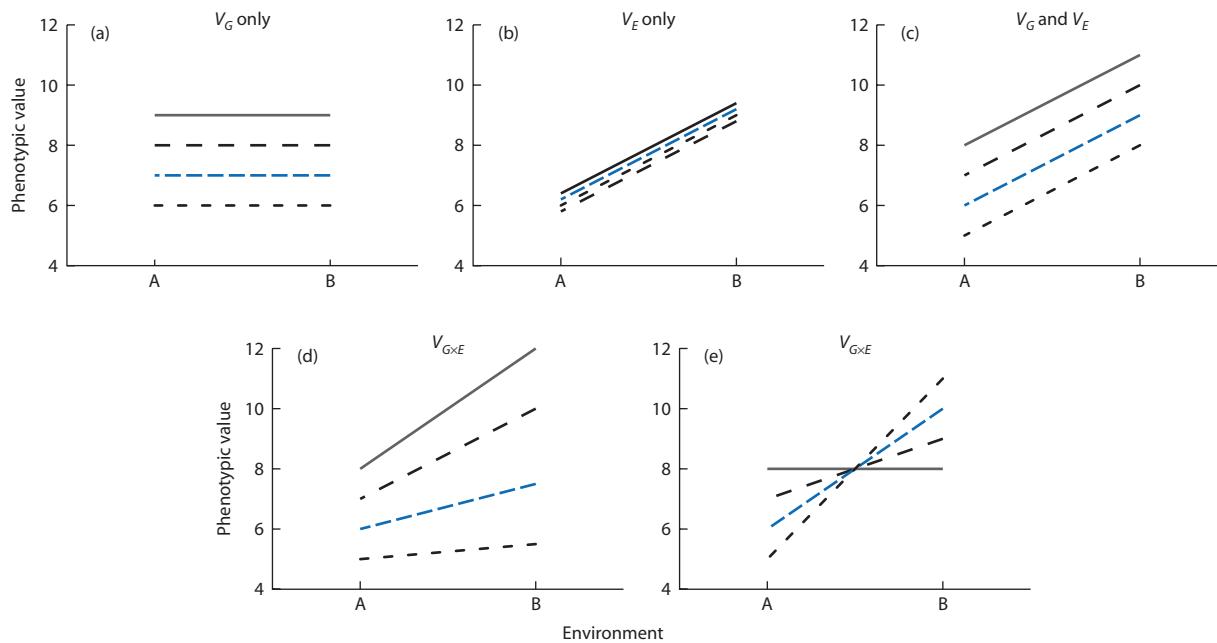
$$V_P = V_G + V_E + V_{G\times E} \quad (9.5)$$

In one form of genotype-by-environment interaction, genotypes are extremely sensitive to changes in the environment such that the total phenotypic variance changes markedly between two or more environments. In another form of genotype-by-environment interaction, genotypes change phenotypic rank in different environments. For example, genotype AA has a larger phenotypic value than genotype Aa in environment one but in environment two the order is reversed with genotype Aa having the larger phenotypic value.

### Genotype-by-environment interaction

( $V_{G\times E}$ ) The contribution to total phenotypic variation caused by genotypes that vary in their sensitivity to different environments. Also known as phenotypic plasticity.

Hypothetical genotypic ( $V_G$ ), environmental ( $V_E$ ), and genotype-by-environment interaction ( $V_{G\times E}$ ) contributions to total phenotypic variation are illustrated in Fig. 9.6. The figure illustrates the results of an imaginary experiment where individuals of four different genotypes are subjected to two environments. Lines connect the phenotype measured for the same genotype in the two environments. Genotypic variation only ( $V_G$ ; Fig. 9.6a) means that the four genotypes have different phenotypes but that no genotype changes its phenotype between the environments (notice that the spread of phenotypes does not change



**Figure 9.6** Examples of phenotypic variation due to genetic ( $V_G$ ), environmental ( $V_E$ ), and genotype-by-environment ( $V_{G\times E}$ ) causes shown in norm-of-reaction plots. In all graphs, the phenotypic values of four genotypes within each of two environments (here called A and B) are plotted. Lines connect the phenotypic values of one genotype measured in the two environments. (a) Genotypic variation, where the four genotypes have different phenotypic values but the phenotypic value of each genotype does not change between environments. (b) Environmental variation: all genotypes have identical phenotypes (lines are staggered so that each can be seen) but the phenotype changes between environments. (c) Both genotypic and environmental variation: genotypes differ in phenotype and genotypes have different phenotypes in the different environments. (d) and (e) Genotype-by-environment interaction, with genotypes differing in the phenotypic value expressed in two or more environments. One type of genotype-by-environment interaction is characterized by lines connecting the genotypes that are not parallel (d), leading to changes in the phenotypic variance. In a second type of genotype-by-environment interaction, the rank order of phenotypic values exhibited by genotypes changes across environments and leads to crossing lines in norm-of-reaction plots (e).

between the environments when there is only  $V_G$ ). Environmental variation only ( $V_E$ ; Fig. 9.6b) means that the four genotypes have identical phenotypes but the phenotype changes between the two environments (the four genotype lines are not drawn exactly on top of each other so each can be seen). A combination of both genotypic ( $V_G$ ) and environmental ( $V_E$ ) variation in phenotype means that the genotypes have different phenotypes and the phenotypes also change between environments (Fig. 9.6c). Notice that with  $V_G$  and  $V_E$  the lines are parallel, showing that each genotype has an identical change in phenotype caused by the change in environment.

Two types of genotype-by-environment interaction are shown in Fig. 9.6. Both examples illustrate the hallmark of genotype-by-environment interaction: different genotypes vary in their response to changes in their environment. One example shows that the range of genotypic values of the four genotypes depends on the environment (Fig. 9.6d). In environment A there is less variation in genotypic values and in environment B there is more variation in genotypic values. When the lines connecting genotypes are not parallel in a norm-of-reaction plot, then genotypic variance changes with environment. Another example shows different environmental sensitivity of genotypes that causes a change of phenotypic ranks between the two environments (Fig. 9.6e). Three of the genotypes respond to the change in environment and demonstrate increased genotypic values in environment B. The genotype with a value of 5 in environment A experiences the largest change in phenotype, showing a value of 11 in environment B. In contrast, the genotype indicated by the solid line is completely insensitive to the environmental change. Crossing lines indicate genotype-by-environment interactions as changes in rank order of genotypic values in norm-of-reaction plots.

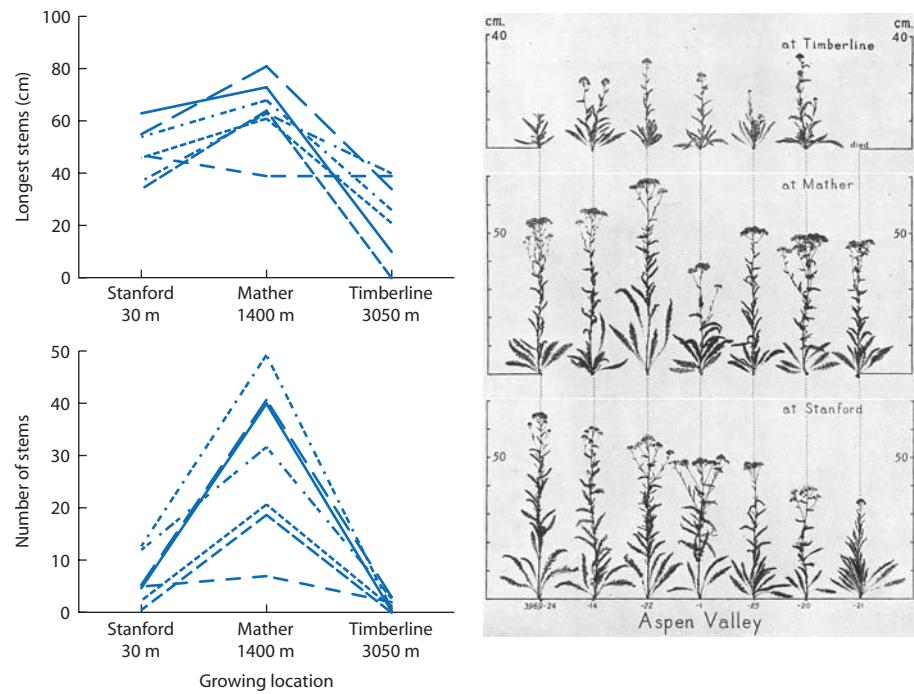
An early and still classic example of a genotype-by-environment interaction comes from Clausen, Keck, and Hiesey (1948; see Nunez-Farfán & Schlichting 2001). These researchers sampled a group of plants in the genus *Achillea* at a single location in Aspen Valley, California (elevation 1950 m). They then sprouted vegetative cuttings of each Aspen Valley plant to create multiple individuals of each genotype for the numerous sampled genotypes. The newly sprouted cuttings were planted at three sites with increasing elevations between sea level (Stanford) and high in the mountains above the treeline (Timberline). The sprouted cuttings of each of the original Aspen Valley plants were genetically identical, so plants with identical genotypes could be transplanted at each elevation.

The transplanted cuttings were monitored and their phenotypes were measured over several years. Figure 9.7 shows the values of two phenotypes (longest stem and number of stems) measured for seven different genotypes grown at each of the three elevations.

Genotypic, environmental, and genotype-by-environment interaction all contributed to phenotypic variation in *Achillea*.  $V_G$  is evident because the phenotypes for the different genotypes within an environment were clearly not identical.  $V_E$  was evident because the average phenotype varied across the environments, being highest at Mather, intermediate at Stanford, and lowest at Timberline.  $V_{G\times E}$  was also evident because the ranks of the phenotypic values for each genotype clearly changed across the three environments. In other words, genotypes that demonstrated phenotypic values above the mean in one environment had a below average phenotypic value in another environment and vice versa. Therefore, in *Achillea* the phenotypic variation seen across these three environments had three causes. It was due to a combination of differences in phenotypic values among genotypes, differences in phenotypic values among environments, and differences in the change in phenotypic value of genotypes across the three environments.

Genotype-by-environment interaction has been observed for diverse phenotypes and a wide range of organisms. An example related to human health involves the risk of colorectal adenoma, a disease characterized by pre-cancerous tumors of the colon, genotypes at the *UGT1A6* (UDP glycosyltransferase 1 family, polypeptide A6) locus, and use of aspirin. Regular aspirin intake reduces the risk of colorectal adenoma for individuals with all *UGT1A6* genotypes compared to no use of aspirin. However, individuals homozygous or heterozygous for the “slow” *UGT1A6* allele, a variant that leads to slower aspirin metabolism, show significantly reduced risk of colorectal adenoma when individuals are taking aspirin compared with other *UGT1A6* genotypes (Bigler et al. 2001; Chan et al. 2005; Hubner et al. 2006). Thus, individual risk of colorectal adenoma depends on combinations of *UGT1A6* genotype and aspirin “environments.” Additional examples of genotype-by-environment interactions in human disease are reviewed by Hunter (2005). See many more examples of genotype-by-environment interactions in Pigliucci (2001) and DeWitt and Scheiner (2004).

The genotype-by-environment interaction can also be thought of as a correlation between genotypic value and environmental impact on the phenotype. The decomposition of  $V_P = V_G + V_E$  without any contribution of  $V_{G\times E}$  makes the assumption that the



**Figure 9.7** The longest stem and number of stems phenotypes for seven *Achillea* genotypes originally sampled at Aspen Valley, California, cloned from vegetative clippings, and then transplanted at three elevations. *Achillea* phenotypes show that genetic ( $V_G$ ), environmental ( $V_E$ ), and genotype-by-environment ( $V_{G \times E}$ ) interaction contribute to total phenotypic variation ( $V_P$ ). Data are from Table 11 and the photograph from Figure 17 of Clausen et al. (1948). Used with permission of Carnegie Institution of Washington.

phenotypic value of a genotype and any environmental impact on the phenotype of that genotype are completely independent. The absence of a correlation between genotypes and effects of the environment is illustrated in Fig. 9.3, where phenotypic values are as likely to increase as to decrease due to their environment. In contrast, a correlation between genotypic value and environment is common in agricultural contexts. For example, domestic animals are often fed in proportion to their individual size or productivity, such as adjusting the amount of feed given to individual cows according to their production of milk. This feeding practice introduces a correlation between genotypic value and environmental variation that then has an impact on the total phenotypic variation. In the case of cows being fed based on milk production, total phenotypic variation should increase since individuals producing more milk (because of their genotypic value) are fed more (a non-random environment), which in turn makes their milk production even greater due to the impact of having more to eat.

#### Additional sources of phenotypic variance

In many organisms, progeny share an environment with their mother, their father or both parents for some period of time as embryos or during their development

and growth. This common environment, sometimes symbolized  $V_{Ec}$  for common environmental variance, can cause parents and offspring to resemble each other to a greater degree than they would if they each inhabited randomly sampled environments.

**Maternal effects**, the correlation between environment and genotype for mothers and their offspring, are common in mammals since mothers supply the pre-natal environment and often provide extensive postnatal care for young. As an example, consider mammals that nurse their progeny. A mother living in an environment rich with food is likely to have a greater body mass since she is well fed. Because the mother receives ample nutrition, she will also be able to produce ample milk for her offspring. If the offspring grow faster and larger due to their mother's ample milk supply, then there will be positive covariance between the mother's body size and their own body size. This positive covariance in body size is caused by the rich environment shared by the mother and her offspring. A shared common environment can also lead to an increase in phenotypic resemblance among full or half siblings since they too share the same environment. Thinking again of a mother providing an ample supply of milk due to a rich environment, the progeny of such a mother could exhibit positive covariance among their phenotypes

(a similar large body size or fast growth rate) due to their common milk supply environment compared to siblings raised in different random environments.

The total phenotypic variation is modified for non-independence of genotypic value and environmental variation by

$$V_p = V_G + V_E + 2(\text{cov}_{GE}) \quad (9.6)$$

where  $\text{cov}_{GE}$  is the covariance between genotypic value and environmental variation in phenotype. Since a covariance can be either positive or negative, the total phenotypic variance can be either increased or decreased by a correlation between the genotypic value and the impact of the environment. The  $\text{cov}_{GE}$  is almost never quantified in practice and is assumed to be zero. The contribution of  $\text{cov}_{GE}$  (if any) to  $V_p$  is therefore lumped in with non-additive causes of  $V_G$ . Despite this, it is important to remember that any covariance between genotypes and environments is a potential cause of increased or decreased variance in phenotype.

### Math box 9.1 Summing two variances

In quantitative genetics, it is common to sum variance components to obtain a total variance. For example, the total phenotypic variance is the sum of the genotypic variance and the environmental variance according to  $V_p = V_G + V_E$ . When summing variances of two variables, it is possible that they are not completely independent of each other. Therefore, it is necessary to account for the possibility of covariance between the variables and to adjust the total variance. When summing the variance of two variables  $X$  and  $Y$  to obtain the total variance,

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{covar}(X, Y) \quad (9.7)$$

A covariance can either be positive (e.g. when the value of  $X$  is large the value of  $Y$  tends to be large) or negative (when the value of  $X$  is large the value of  $Y$  tends to be small). Unless  $X$  and  $Y$  are independent (for any value of  $X$  the value of  $Y$  is random), the sum of two variances may be increased or decreased by covariance.

Another possible source of genotypic variance ( $V_G$ ) comes from gametic disequilibrium between loci as well as autozygosity within loci. As detailed in Chapter 2, physical linkage, finite population size, admixture of genetically diverged populations, mutation, and natural selection can all produce gametic disequilibrium. Additionally, consanguineous mating also causes an increase in gametic disequilibrium. When consanguineous mating occurs, one result is increased homozygosity for individual loci. An increase in homozygosity leads to gametic disequilibrium because genotypes at two or more loci within an individual will not be a random combination of all possible genotypes but will tend to be combinations of homozygous genotypes.

All forms of gametic disequilibrium contribute to the **disequilibrium covariance** of genotypic values that increases or decreases the total phenotypic variance (Cockerham 1956; Weir et al. 1980). The degree of gametic disequilibrium can be measured as a covariance *between individuals* of the correlation between genotypic values at two loci *within* an individual. To visualize this covariance, first imagine the possible correlations between the genotypic values at two loci within a single individual. For example, consider the two locus genotype AABB in a population. If the A and B alleles both contribute to larger phenotypic values, the AA genotypic value and the BB genotypic value are positively correlated in an individual with the AABB genotype. Under random mating and free recombination, the AA genotypic value and the BB genotypic value are found together in the same individual only occasionally since the two loci are independent. Free recombination and random mating mean that AA and BB single locus genotypes co-occur at random. In contrast, imagine the AABB genotype is common in the population due to consanguineous mating or strong gametic disequilibrium that causes high frequencies of AB gametes. In that case there will be a positive correlation of genotypic values between the A and B loci within individuals because the AA and BB single locus genotypes tend to occur together more frequently than expected under independent assortment.

Under random mating in a large population there should be little gametic disequilibrium for a trait not under selection and the fixation index ( $F$ ) should be approximately zero. This leads to a disequilibrium covariance of zero. Since the genotypic disequilibrium covariance is not estimated in practice, these conditions become implicit assumptions in the decomposition of  $V_p = V_G + V_E$ . As with correlations

between genotypic value and environment, the contribution (if any) of genotypic covariance to  $V_p$  is lumped in with non-additive causes of  $V_G$ .

## 9.2 Evolutionary change in quantitative traits

- Heritability.
- Changes in quantitative trait mean and variance due to natural selection.
- Estimating heritability by parent–offspring regression.
- Response to selection on correlated traits.
- Long-term response to selection.
- Neutral evolution of quantitative traits.

Evolutionary change in quantitative traits is caused by the processes that reduce, shape, or increase variation in the genotypes that underlie the genotypic portion of the total phenotypic variance. Just like the individual loci considered in earlier chapters, the multiple loci that compose a quantitative trait will experience genetic drift and mutation and will also be subject to natural selection. Change in quantitative trait means and variances will occur to the extent that these processes change the allele and genotypes frequencies at those loci that contribute to a quantitative trait. Because it is usually not possible to track the individual loci that underlie a quantitative trait, the genetic basis of quantitative traits is tracked by summary measures that describe a population. A critical distinction that we need to bear in mind is the difference between additive genetic variation ( $V_A$ ) that leads to resemblance between relatives, and dominance and interaction genetic variation ( $V_D$  and  $V_I$ ) that is not inherited. Because additive genetic variation does lead to parent–offspring resemblance, it is the basis of the action of natural selection on quantitative traits.

### **Heritability**

The **heritability** is used to express the proportion of the total phenotypic variance ( $V_p$ ) that is caused by either all types of genotypic variance ( $V_G$ ) or by only the additive genetic variance ( $V_A$ ). Utilizing the equation for the components of the total phenotypic variance  $V_p = V_G + V_E$ , both sides of the equation can be multiplied by  $1/V_p$  to obtain

$$\frac{V_p}{V_p} = \frac{V_G}{V_p} + \frac{V_E}{V_p} \quad (9.8)$$

This divides the total phenotypic variance, or 100% of  $V_p$ , into the proportion caused by genotypic variation and the proportion caused by environmental variation. The proportion of the total phenotypic variance caused by genotypic variance defines the **broad-sense heritability**:

$$h_{BS}^2 = \frac{V_G}{V_p} \quad (9.9)$$

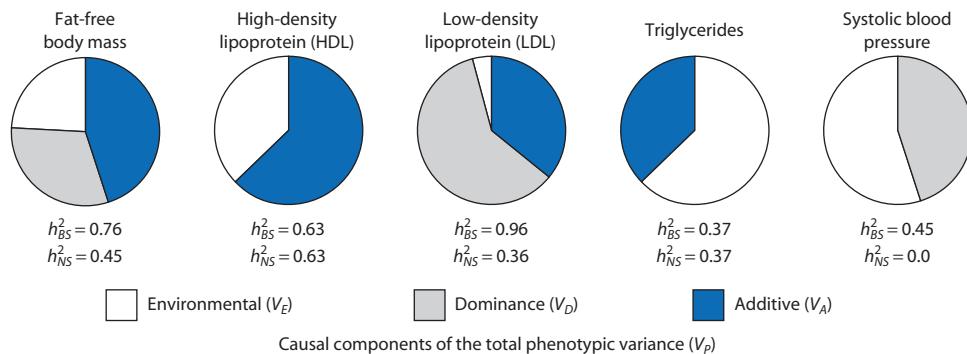
Since genotypic variance is composed of the separate components  $V_A + V_D + V_I$ , the proportion of the total phenotypic variance caused by only the additive genetic variance defines the **narrow-sense heritability**:

$$h_{NS}^2 = \frac{V_A}{V_p} \quad (9.10)$$

The narrow sense heritability expresses the proportion of the total phenotypic variance made up of genotypic variance that contributes to parent–offspring resemblance. The remaining proportion of  $V_p$  is caused by genotype components of the total genotypic variance ( $V_D + V_I$ ) and the environmental variance, none of which contribute to phenotypic resemblance between parents and offspring (with the exception of additive by additive epistasis).

Idealized heritabilities are proportions and therefore range between 0.0 and 1.0 (although *estimates* may fall outside this range due to estimation error). The symbol for the heritability is always  $h^2$  and there is no biological meaning in the square root of  $h^2$ . This is a convention that has held since Wright used the symbol in a 1921 paper (Wright 1921). Unless explicitly noted otherwise, the word heritability and  $h^2$  without a subscript will hereafter refer to the narrow-sense heritability, as is common practice.

As an illustration, broad-sense and narrow-sense heritabilities for five human phenotypes related to blood pressure are shown in Fig. 9.8. In these examples, only the additive and dominance components of genotypic variance were estimated but the epistasis component was not. Each pie chart divides the total phenotypic variation ( $V_p$ ) into the dominance ( $V_D$ ) and additive ( $V_A$ ) components of the total genotypic variance as well as environmental variance ( $V_E$ ). The broad-sense heritability for fat-free body mass shows that 76% of  $V_p$  was caused by genotypic variation. The remaining 24% of variation in fat-free body mass was caused by environmental



**Figure 9.8** Broad-sense and narrow-sense heritabilities for five blood-pressure-related quantitative traits in humans. Each pie chart divides the total phenotypic variation ( $V_p$ ) into its causal components of dominance ( $V_D$ ) and additive ( $V_A$ ) genotypic variance, as well as environmental variance ( $V_E$ ). These heritabilities were estimated in a small population of Hutterites, a self-reliant, communal group of Anabaptists that traces its origins to followers of Jakob Hutter who fled Austria in the sixteenth century to escape religious persecution. Today, Hutterites live in Canada and North America. The 806 individuals in this study are descendants of 64 ancestors so that many individuals have a non-zero probability of sharing a genotype that is identical by descent because their parents are distantly related. This improves the precision of estimates of dominance variance. Estimates from Abney et al. (2001).

differences among individuals. The genotypic variation can be further divided by comparing the broad-narrow and narrow-sense heritabilities. Additive genetic variation explained 45% of  $V_p$  while 31% (the difference between  $h_{BS}^2$  and  $h_{NS}^2$ ) of  $V_p$  was caused by dominance. Notice that the proportion of  $V_p$  explained by each of  $V_A$ ,  $V_D$ , and  $V_E$  varies considerably among the five traits in this population. While variation in low-density lipoprotein (LDL) levels among individuals is caused by a combination of  $V_A$ ,  $V_D$ , and  $V_E$  – as was the case for fat-free body mass – the relative contributions of each changed greatly, with dominance explaining 60% of  $V_p$  and environment only 5% of  $V_p$ . Variation in the high-density lipoprotein (HDL) and triglyceride level phenotypes were explained by  $V_A$  and  $V_E$  with no measurable contribution of  $V_D$ . Variation in systolic blood pressure was explained by about equal contributions of  $V_D$  and  $V_E$  with no measurable contribution of  $V_A$ . While variation in systolic blood pressure does have a genetic basis, there was very little phenotypic resemblance between relatives due to the additive effects of alleles.

It has long been observed that there is a correlation between the type of trait and the magnitude of its narrow-sense heritability. Life-history traits such as survival and number of progeny generally exhibit the lowest narrow-sense heritabilities whereas physiological (e.g. efficiency of food conversion, percentage of milk fat, serum cholesterol concentration), behavioral (e.g. parental care, mating, phototaxis), and morphological (e.g. body size, height, bristle number) phenotypes show progressively greater narrow-sense heritabilities (Mousseau & Roff 1987; Roff

& Mousseau 1987; Falconer & Mackay 1996). One hypothesis is that traits more closely related to fitness are likely to be under stronger and more consistent natural selection pressures than arbitrary morphological phenotypes, so that long-term response to selection has depleted additive genotypic variation for life-history traits but not for morphological traits. An alternative possible explanation for this pattern is that the relative magnitudes of  $V_A$ ,  $V_D$ , and  $V_E$  are different for morphological, behavioral, and life-history traits. In an analysis of 182 quantitative traits, Houle (1992) used a measure of additive genotypic variation standardized by the trait mean to show that life-history phenotypes actually have higher additive genotypic variance ( $V_A$ ) as well as residual variance (all remaining variance not due to additive genotypic variance) than traits presumably experiencing weaker natural selection. In the traits analyzed by Houle (1992), lower heritability of fitness-related traits was caused by higher levels of non-additive genotypic variance ( $V_D$  and  $V_I$ ) and environmental variance rather than by a lack of additive genotypic variation.

It is important to recognize that all heritability estimates are highly contextual. A heritability estimate made in one population may not be representative of heritability in another population of the same species because both the genetic and environmental causes of phenotypic and genotypic variation may differ among populations. Clearly, allele frequencies may differ among populations, causing  $V_A$  to differ. Genotype frequencies may also differ among populations, causing  $V_G$  to differ. Even if allele frequencies are identical, differences in mating system can impact  $V_D$  and

$V_p$  because homozygosity is increased or decreased by mating patterns. The range of environments or their quality may also differ among populations, so that even if genotype and allele frequencies are constant,  $V_p$  can increase or decrease and thereby cause the heritability to change as well.

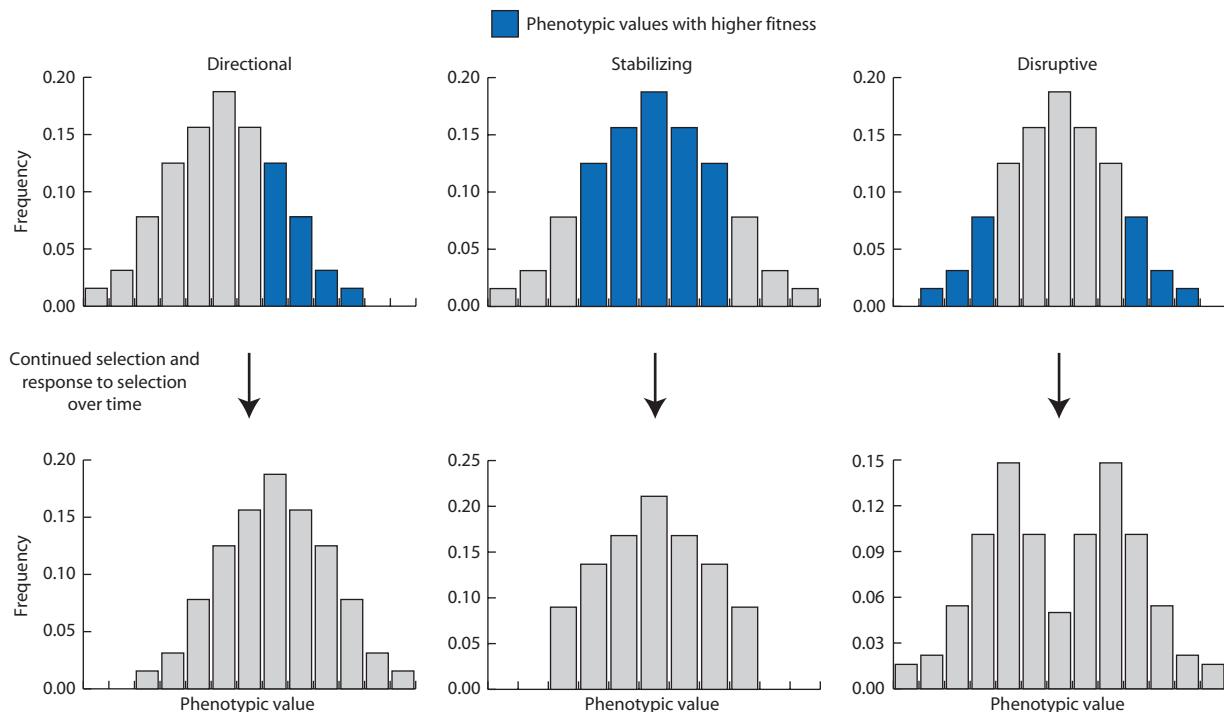
### Changes in quantitative trait mean and variance due to natural selection

To begin, it is important to understand that natural selection on quantitative traits operates on differences in phenotypic value. The phenotypic values of individuals dictate individual fitness, whereas the individual multilocus genotypes that underlie a quantitative trait are not directly perceived by natural selection and are unknown to an observer. In contrast, in the models of natural selection in Chapters 6 and 7 natural selection acted on different fitness values possessed by known genotypes. This distinction is biologically meaningful. In selection on phenotypic values, the manner in which phenotypic values are determined by alleles and genotypes (i.e. additivity, dominance, or epistasis) has a direct

impact on how effective selection is in changing the allele frequencies that underlie the mean phenotype in a population. As we will see in this section, selection is most effective when variation in phenotypic values is due to additive variation and least effective when phenotypic variation is due to dominance and epistasis.

There are three general types of natural selection experienced by quantitative traits, as illustrated in Fig. 9.9. Under **directional selection** individuals with phenotypic values at one edge of the phenotypic distribution have higher relative fitness, so that response to selection results in an increasing or decreasing mean phenotype. When **stabilizing selection** operates, phenotypic values around the population mean have the highest fitness. In contrast, under **disruptive selection** phenotypic values at the outer edges of the phenotypic distribution have higher relative fitness.

The response to **stabilizing** or **disruptive selection** on quantitative traits is distinct from the response to directional selection. When a trait responds to stabilizing selection, the mean phenotypic value in a population does not change. Instead, the variance



**Figure 9.9** General types of natural selection on quantitative traits. Directional selection occurs when phenotypic values at the upper or lower end of the distribution have the highest fitness. Stabilizing selection occurs when intermediate trait values have the highest fitness. Disruptive selection occurs when trait values at the edges of the phenotypic distribution have the highest fitness. Response to directional selection increases or decreases the population mean phenotype. Response to stabilizing or disruptive selection does not change the mean but decreases or increases the variance of the phenotypic distribution.

in phenotypic value decreases over time since individuals with phenotypic values at the upper and lower extremes have lower fitness values and do not contribute to future generations. Similarly, the mean phenotypic value in a population does not change with response to disruptive selection. Disruptive selection causes the variance in phenotypic values to increase since individuals at the upper and lower extremes of the distribution have higher fitness than individuals with phenotypic values near the mean. Disruptive selection results in the phenotypic distribution widening over time, and the distribution of phenotypic values can eventually become bimodal.

The heritability provides the basis for predicting the outcome of natural selection on quantitative traits according to

$$R = h^2 s \quad (9.11)$$

where  $R$  is the response to selection,  $h^2$  is the narrow-sense heritability, and  $s$  is the selection differential that measures the strength of natural selection.  $R$  and  $s$  are measured in the same units used to measure phenotypic value (e.g. kilograms, centimeters). Equation 9.11 is commonly called the **breeder's equation** because it predicts the change in mean phenotype in a population that will occur due to one generation of artificial selection as often employed by animal and plant breeders. The response to selection predicted by the breeder's equation is intuitive. Stronger natural selection or phenotypic variation that has a greater basis in additive genetic variance will result in a greater change in the mean phenotype in a population.

To better understand how natural selection changes the mean phenotypic value, let's work through an example of directional selection based on phenotypic value. The change in mean phenotype in a population caused by natural or artificial selection depends both on the force or amount of selection that is applied to the population and on genetic variation in the trait. The **selection differential** is one way to measure the strength of directional natural selection on quantitative traits. The selection differential is the difference between the phenotypic mean of the entire population and the phenotypic mean of that subset of individuals selected on the basis of their phenotypic value to be parents of the next generation. The **selection threshold** (sometimes called the truncation point) seen in Fig. 9.10 is the lower bound of phenotypic values in the group of parents selected to mate. The selection differential is com-

puted as the difference in the phenotypic mean of the selected parents ( $\mu_s$ ) and the phenotypic mean of the entire P1 population ( $\mu$ ):

$$s = \mu_s - \mu \quad (9.12)$$

As shown in Fig. 9.10, the selection differential for this case is

$$s = 12.5 - 10.0 = 2.5 \quad (9.13)$$

which expresses that the selected parents have a 2.5-unit greater average phenotypic value than the full population that they were sampled from. Larger selection differential values indicate stronger selection. Although not true in this example, selection differentials are often expressed in units of standard deviations by standardizing phenotypic values in the parental population to have a mean of zero and a variance of one.

Imagine now that those individuals with phenotypic values above the selection threshold in the P1 population mate at random and then produce a large population of progeny that are reared in the same environment as the parents. The phenotypic distribution of the progeny is shown in the lower panel of Fig. 9.10. The progeny have a mean phenotypic value of  $\mu' = 11.0$ . The difference between the F1 population phenotypic mean and the phenotypic mean of the entire P1 population expresses how much the phenotypic mean was changed by selection for larger trait values in the parents. The response to selection is

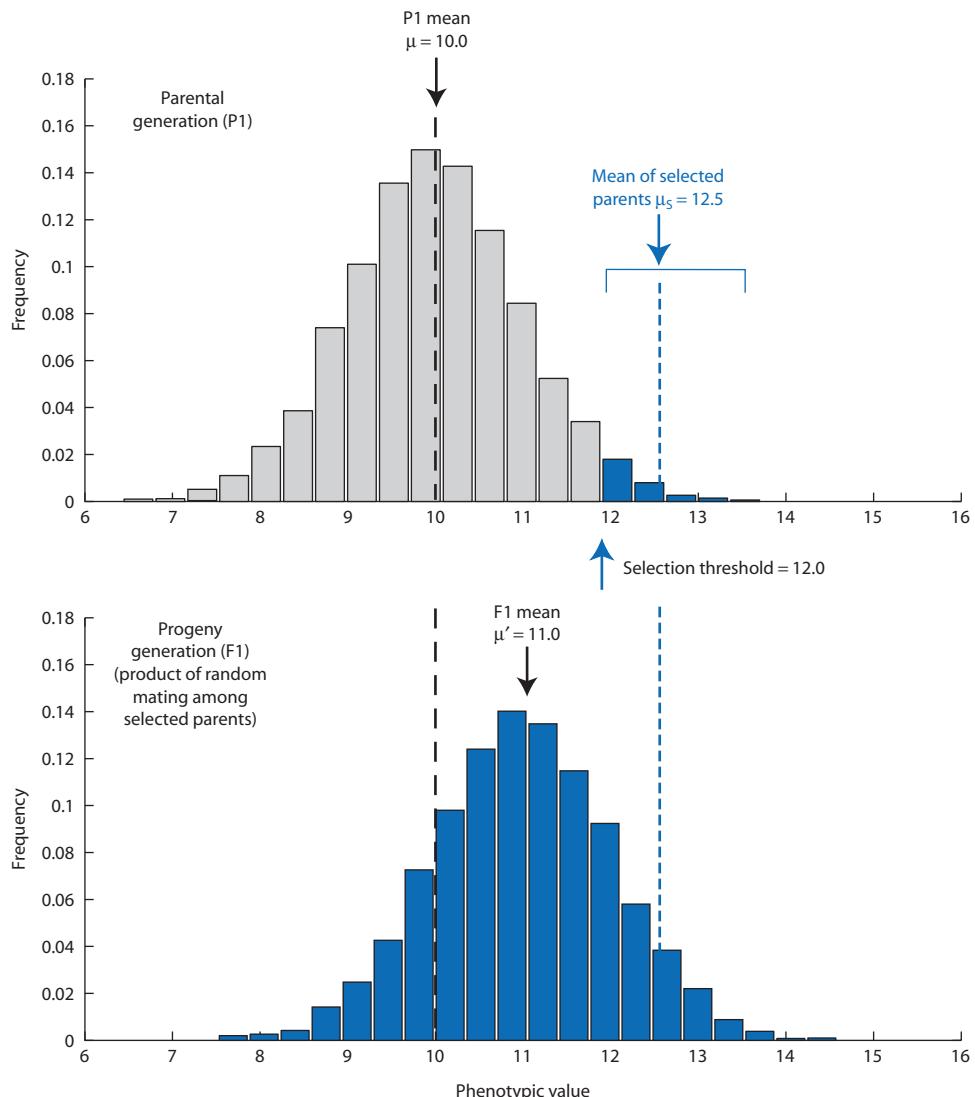
$$R = \mu' - \mu = 11.0 - 10.0 = 1.0 \quad (9.14)$$

The phenotypic mean in this example was increased one unit by directional selection in the P1 population.

The response to selection is a function of the amount of phenotypic variation that is due to additive genetic variation according to the breeder's equation. Since both the response to selection and the selection differential are known, it is then possible to estimate the heritability by rearranging the breeder's equation:

$$\hat{h}^2 = R/s \quad (9.15)$$

When estimated in this way,  $\hat{h}^2$  is called the **realized heritability** since it is estimated from the observed response to selection rather than predicted by resemblance of parents and progeny in the absence of selection. Using the selection differential



**Figure 9.10** A hypothetical example of directional selection and response to selection. In the parental generation those individuals with a phenotypic value of 12.0 or greater are allowed to mate. The selection differential is  $s = 12.5 - 10.0 = 2.5$ . Random mating among this subset of the parental population produces a distribution of progeny phenotypic values with a mean value of 11.0. The response to selection is  $R = 11.0 - 10.0 = 1.0$ . The realized heritability is therefore  $h^2 = 1.0/2.5 = 0.40$ . Response to selection is proportionate to the degree to which parental phenotypic values are caused by the phenotypic effects of alleles. Parental phenotypic values being caused by genotypes or the environment do not lead to response to selection since these causes of phenotypic value are not inherited by offspring.

and response to selection calculated for Fig. 9.10, the realized heritability is then

$$\hat{h}^2 = 1.0/2.5 = 0.40 \quad (9.16)$$

This tells us that 40% of the variance in trait values in the parental population was caused by additive genetic variation based on the definition of heritability.

Why was there a response to selection? Why did  $\mu'$  increase in value compared to  $\mu$ ? The phenotypic

value in the selected group of parents was greater than the rest of the parental population partly due to the alleles that they possessed in their multi-locus genotypes for this trait. When they bred, these alleles were passed down to their offspring. Selection changed the frequency of alleles that confer larger trait values because allele frequencies in the P1 individuals above the selection threshold were different than in the P1 population as a whole. Alleles that contributed to larger trait values became more frequent in the progeny population than they were

initially in the parental population. Therefore, the mean phenotypic value in the progeny population increased relative to the mean phenotypic value of the parental population.

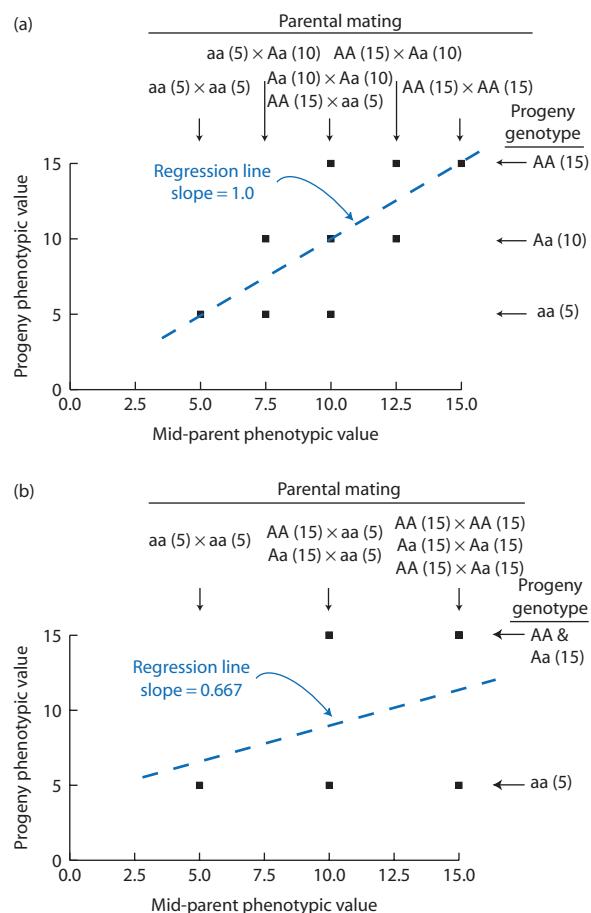
But why is the progeny mean phenotypic value ( $\mu'$ ) not equal to the mean value of the selected parents ( $\mu_s$ )? Parents in the selected group had phenotypic values above the truncation point partly due to causes other than the effects of the alleles in their multilocus genotypes for the trait. Part of the phenotypic variance in the parental generation ( $V_p$ ) was caused by factors that do not contribute to the resemblance of parents and offspring. The genotypic values of the selected parents were due to the combination of alleles in their genotypes. Such genotype effects cause dominance variance ( $V_D$ ) and interaction variance ( $V_I$ ) in quantitative trait values. However, these components of the genotypic variance are not inherited and do not contribute to resemblance between parents and offspring on average. In addition, some of the phenotypic variance in the parental population could have been caused by environmental variance ( $V_E$ ) that would also not contribute to resemblance between the selected parents and the offspring. In the example of Fig. 9.10, 60% (or  $1 - h^2$ ) of the variation in the P1 phenotypic values was caused by the combination of non-additive genetic variation ( $V_D + V_I$ ) and environmental variation ( $V_E$ ). This 60% is the percentage of the selection differential that did not produce a response to selection.

### **Estimating heritability by parent–offspring regression**

Another method used to estimate the heritability based on the resemblance between parents and their offspring is parent–offspring regression. Parent–offspring regressions predict  $V_A$  without actually carrying out a response-to-selection experiment. This method to estimate heritability is therefore applicable to populations where selection experiments cannot be carried out. Estimation of heritability by parent–offspring regression can even be carried out in natural populations if a reliable method to identify the parents of offspring, such as paternity analysis, is available. This method takes advantage of the fact that the phenotypes of offspring resemble the phenotypes of their parents to the extent that phenotypic values are caused by shared alleles. Neither dominance nor epistasis components of the genotypic variance are inherited by progeny so they do not cause progeny to resemble their parents. This

phenomenon was explained in the first section of the chapter and illustrated in Table 9.3. Now we will revisit the resemblance between parents and offspring in greater depth.

By comparing the phenotypes of parents and their offspring, it is possible to determine the degree of phenotypic resemblance. The necessary data come from the measures of phenotypic values for pairs of parents as well as the phenotypic values of all progeny produced by each pair of parents. Figure 9.11



**Figure 9.11** Parent–offspring regressions used to estimate heritability ( $h^2$ ) under the assumption of strict additivity (a) or complete dominance (b). The resemblance or covariance between the mid-parent phenotypic value and the mean progeny phenotypic value is greater on average with additivity (slope = 1.0) than with dominance (slope = 0.667). The slope of the regression line is equal to the heritability since it measures resemblance of parents and offspring. The mid-parent phenotypic value is the average phenotypic value of two parents that mate and produce progeny. Phenotypic values are given in parentheses next to each genotype. Each dot represents the mean phenotypic value of many progeny that result from a given parental mating. Both panels assume Hardy–Weinberg expected genotype frequencies in the parental populations, allele frequencies of  $p = q = 0.5$ , and that  $V_E = 0$ .

illustrates plots of parent–offspring regressions using one-locus genotypes assuming random mating among the parents. In both panels, the slope of the regression line between the phenotypic values of offspring and the average phenotypic values of their parents (called the **mid-parent value**) provides an estimate of phenotypic resemblance between parents and offspring. Because resemblance between parents and offspring must be due to inherited alleles rather than the effects of genotypes, the degree of resemblance is a function of the heritability. Therefore, the slope of the regression line estimates the heritability.

The parent–offspring regression for additive gene action is shown in Fig. 9.11a. With strictly additive effects of alleles, the phenotypic values of all progeny are determined by counting up the number of a and A alleles in their genotypes. Under additivity, the a allele contributes 2.5 and the A allele contributes 7.5 to the genotypic value. The mean phenotype of all progeny from one pair of parents corresponds exactly to the average number of A and a alleles transmitted to their progeny. For example, mating aa and Aa parents results in 75% a and 25% A alleles transmitted to progeny. The average progeny phenotype is then  $0.75(2.5) + 0.25(7.5) = 3.75$  for each allele in its genotype. Since there are two alleles, the average value among all progeny is  $2(3.75) = 7.5$ . This corresponds exactly to the aa  $\times$  Aa mid-parent value. In the additive case, this same logic can be applied to all parental matings. The result is always that the mid-parent phenotypic value equals the average progeny phenotypic value.

The parent–offspring regression for complete dominance is shown in Fig. 9.11b. We again see that when the combination of alleles in a genotype defines the phenotypic value of a heterozygote, the resemblance between parents and their offspring is not as large. In Fig. 9.11b, dominance causes the AA and Aa genotypes to have identical phenotypes. This means, for example, that while two Aa genotypes have an average phenotypic value of 15, their progeny have an average phenotypic value of  $0.25(5) + 0.5(15) + 0.25(15) = 12.5$ . Dominance masks the presence of a alleles in the heterozygotes. These a alleles result in the production of a portion of aa progeny in some parental matings with a phenotypic value of 5 that causes the progeny average to be less than 15. With dominance, the progeny mean phenotype no longer corresponds to a weighted average of the number of a and A alleles in their genotypes.

### Interact box 9.1 Estimating heritability with parent–offspring regression

Populus can be used to simulate parent–offspring regressions while varying the population allele frequencies, the degree of dominance, and the sample size of families. Go to the text web page to see a step-by-step guide as well as questions to answer based on simulation results.

It is possible to prove that the slope of the regression line for offspring values and mid-parent values estimates the heritability. The slope of a regression line,  $b$ , between variables  $x$  and  $y$  is defined by

$$b = \frac{\text{cov}(y, x)}{\text{var}(x)} \quad (9.17)$$

Based on the probability that half-sib progeny share alleles identical by descent (see section 10.6), the expected covariance between the mid-parent values and offspring values is

$$\text{cov}(\text{offspring}, \text{mid-parent}) = \frac{1}{2}V_A \quad (9.18)$$

It is also the case that the expected variance of the mid-parent values is equal to  $\frac{1}{2}V_P$ . This comes about since the variance of the phenotypic values of a large sample of pairs of parents (say  $n$  pairs where  $n$  is large) is expected to be half the variance of a large sample of individual parents (or  $2n$  individual parents) because of the smaller value of  $n$  used to compute the variance. In other words, if the phenotypic variance of a large number of individuals is  $x$ , then the expected phenotypic variance of one-half of these individuals is  $\frac{1}{2}x$ . Bringing these two points together, we can restate the regression coefficient as

$$\begin{aligned} b &= \frac{\text{cov}(\text{offspring}, \text{mid-parent})}{\text{var}(\text{mid-parent})} \\ &= \frac{\frac{1}{2}V_A}{\frac{1}{2}V_{P(\text{mid-parent})}} = \frac{V_A}{V_P} \end{aligned} \quad (9.19)$$

to prove that the slope of the regression line estimates the heritability.

### **Response to selection on correlated traits**

Our exploration of the heritability and the response to selection thus far has focused on single traits in isolation. While this approach helped to introduce these concepts, it is an unrealistic perspective on how quantitative traits evolve under natural selection in actual organisms. Phenotypes are not actually isolated, completely compartmentalized units that are unrelated. In contrast, many phenotypes are highly inter-related in the sense that they are not independent. Correlations among trait values can be manifest in two forms. The values for two traits may simply be correlated within individuals, a phenomenon called **phenotypic correlation**. The degree to which trait values are inherited in common is measured by the **genetic correlation**. As a simple example of these correlations, think of domestic cats and their traits of body weight and tail length. If in a population of individuals, heavier cats tend to have longer tails, weight and tail length show a phenotypic correlation. Imagine that artificial selection was applied to the cat population for greater body weight and that body weight responded to this selection because of additive genetic variation. An associated increase in tail length at the same time, in the absence of any selection on tail length, would demonstrate a genetic correlation between tail length and body weight.

**Genetic correlation** Non-independence of inherited values for two traits. A correlation between the breeding values of two quantitative traits.

**Phenotypic correlation** Non-independence of the values of two or more quantitative traits within individuals.

Phenotypic correlation between traits within the same individual may be caused by the environment. For example, environmental conditions may tend to have the same effect on several traits, resulting in a positive correlation, or have opposite effects resulting in a negative correlation. A genetic component of a phenotypic correlation, or a genetic correlation, can have two causes that are not mutually exclusive. One cause is pleiotropy, or the phenomenon where a single locus contributes to variation in two or more phenotypes. Another cause is genetic linkage such that two distinct loci, one locus affecting one trait only, are in strong gametic disequilibrium. Both situations lead to correlated changes in phenotypes

for two or more traits when genotype frequencies change in a population.

Both phenotypic and genetic correlations are represented mathematically in matrices that have as many rows and columns as there are phenotypes being considered. The values that represent the phenotypic variance and covariance are contained in a **P matrix**. The diagonal elements in **P** represent trait variances whereas the off-diagonal elements represent the covariance for each pair of traits. Similarly, the **G matrix** contains individual values that measure both the additive genetic variance and covariance. The diagonal elements in a **G** matrix are simply the narrow-sense heritabilities for each trait. The off-diagonal elements in **G** represent the additive genetic covariance for each pair of traits, or the degree to which trait values are co-inherited. Table 9.4a shows each element in **G** and **P** matrices for the case of two traits.

To see how the response to selection for one trait is a special case because there are no phenotypic or genetic correlations, the breeder's equation can be re-written as

$$R = (\mathbf{G}/\mathbf{P})\mathbf{s} \quad (9.20)$$

where **G** is the additive genetic variance, **P** is the total phenotypic variance, and **s** is the selection differential. In the breeder's equation for one trait, **P** is implicitly set to one so that the response to selection is then expressed per unit of phenotypic variance. For example, if **G** = 0.5, **P** = 1, and **s** = 0.5 phenotypic units for a single trait, then the population would have an increase of  $R = 0.25$  phenotypic units in the next generation. A larger value of **P** would yield a smaller response to selection (e.g. **P** = 2 yields  $R = 0.125$ ) whereas a smaller value of **P** would produce a greater response to selection (e.g. **P** = 0.5 yields  $R = 0.5$ ). The phenotypic variance plays a role in response to selection since the selection differential is relative to the trait variance. In order to compare traits or populations with different values of **P** and **G**, the additive genetic variance is scaled to the total phenotypic variance. This change of scale assures that the additive genetic variance is then equal to the heritability, a quantity that is independent of the magnitude of the total phenotypic variance. To see this scaling, imagine that **P**<sub>1</sub> = 1.0 for one trait and **P**<sub>2</sub> = 5 for another trait but that both traits have **G** = 0.3. The additive genetic variance relative to the total phenotypic variance is  $\mathbf{G}/\mathbf{P}_1 = 0.3/1 = 0.3$  for trait one and  $\mathbf{G}/\mathbf{P}_2 = 0.3/5 = 0.06$  for trait two.

**Table 9.4** Examples of response to selection for two phenotypes with the possibility of phenotypic or additive genetic covariance. The elements of the phenotypic variance/covariance matrix ( $P$ ), the additive genetic variance/covariance matrix ( $G$ ), the vector of selection differentials ( $s$ ), and the vector of predicted changes in mean phenotype ( $\Delta\bar{z}$ ) are shown in (a).

(a)	$G$	[Trait A]	[Trait B]
	Trait A	$h^2$	Genetic cov(A, B)
	Trait B	Genetic cov(A, B)	$h^2$
$P$		[Trait A]	[Trait B]
	Trait A	Variance(A)	Phenotypic cov(A, B)
	Trait B	Phenotypic cov(A, B)	Variance(B)
$s = [\text{selection differential trait A, selection differential trait B}]$			
$\Delta\bar{z} = [\text{change in mean of trait A, change in mean of trait B}]$			
(b)			
	$G = 0.5 \quad 0$	$P = 1.0 \quad 0$	
	$0 \quad 0.5$	$0 \quad 1.5$	
	$s = 0.5, 0.5$		
	$\Delta\bar{z} = 0.25, 0.1667$		
(c)			
	$G = 0.5 \quad 0$	$P = 1.0 \quad 0.6$	
	$0 \quad 0.5$	$0.6 \quad 1.5$	
	$s = 0.5, 0$		
	$\Delta\bar{z} = 0.3289, -0.1316$		
(d)			
	$G = 0.5 \quad 0.6$	$P = 1.0 \quad 0$	
	$0.6 \quad 0.5$	$0 \quad 1.5$	
	$s = 0.5, 0$		
	$\Delta\bar{z} = 0.25, 0.20$		

Dividing  $G$  by  $P$  expresses the additive genetic variance per unit of phenotypic variance. In this example, the heritabilities are  $h^2 = 0.3$  for trait one and  $h^2 = 0.06$  for trait two so clearly trait one would show a greater response to selection for a given value of the selection differential.

With the potential for different phenotypic variances for each trait, covariance between phenotypes, and genetic correlation, the breeder's equation needs to be extended. For two or more traits, the change in mean phenotype for each trait is predicted by

$$\Delta\bar{z} = GP^{-1}s \quad (9.21)$$

where  $P^{-1}$  indicates a matrix inverse (Lande 1979; Lande & Arnold 1983). For more than one trait, the change in the mean trait value is represented by the vector  $\bar{z}$  rather than the scalar  $R$  since there are as many means as there are traits. The selection differential is still symbolized as  $s$  even though it is now a vector.

**G matrix** The genetic additive variance/covariance matrix that expresses the heritability of each trait (the diagonal elements) as well as the genetic covariance between each trait (the off-diagonal elements).

**P matrix** The phenotypic variance/covariance matrix that expresses the variance of each trait (the diagonal elements) as well as the covariance between each trait (the off-diagonal elements).

**s** The selection differential, expressed as a vector of values when there are two or more traits.

**z** The phenotypic mean, expressed as a vector of values when there are two or more traits.

Some examples will help illustrate how equation 9.21 serves to combine the direct effect of selection on each trait with the indirect effects of genetic and

phenotypic correlations between traits to predict the total change in trait mean. In Table 9.4b, there is no genetic correlation and also no phenotypic correlation, making the traits completely independent. Both traits have heritabilities of 0.5 and selection differentials of 0.5. The response to selection is exactly as we would predict from the single-trait version of the breeder's equation for each trait separately. Note that trait B has a lower response to selection because it has more phenotypic variance, making the selection differential of 0.5 effectively weaker.

In the example of Table 9.4c, there is a fairly strong positive phenotypic correlation between the two traits and natural selection applied only to trait A but still no genetic correlation between traits. The mean of trait A is predicted to increase since it is experiencing natural selection and has a non-zero heritability. Trait B also shows response to selection, in this case a reduction in mean value. This change in mean is due to the correlation between the two traits alone and not due to any direct natural selection in trait B since its selection differential is zero.

In the final example in Table 9.4d, there is a strong positive genetic correlation between the two traits and natural selection is acting to increase the average of trait A. There is now no phenotypic correlation between the traits. The means of both traits are predicted to increase in this case. The mean of trait A will increase because of selection acting directly on it. At the same time the mean of trait B will also increase. This occurs not because selection is acting on trait B, after all it has a selection differential of zero, but rather because the two traits are genetically correlated. The change in genotype and allele

frequency caused by response to selection on trait A has also changed genotype and allele frequencies that influence the mean of trait B. Direct selection for an increase in the mean of A indirectly causes an increase in the mean of B due to a genetic correlation between the traits. To distinguish direct and indirect effects of natural selection, the change in a trait mean due to a direct effect is called **selection for** while **selection of** describes a change in a trait mean caused by an indirect response to selection on a genetically correlated trait.

When traits are genetically correlated, natural selection and any response is potentially not as simple as it would be with a single trait, which would be completely independent of all other traits. This is particularly true of quantitative traits related to Darwinian fitness of individuals. First, natural selection experienced by one trait could lead to a response to selection in another trait that does not directly experience natural selection. This means that natural selection on correlated traits is capable of indirectly changing traits that have little or no relationship with fitness. Second, when two traits respond to selection over time it may lead to the evolution of a negative genetic correlation that prevents further response to selection. If two traits are both related to fitness and each experience natural selection, then we expect alleles at any loci that independently cause variation in either trait to become fixed by long-term response to selection. However, any loci that have opposite effects on the two traits will not experience fixation or loss caused by natural selection. For example, if increased frequencies of AA genotypes increase fitness of trait one but simultaneously lead to decreases fitness for trait two, natural selection should result in neither the A nor a allele fixing. Alleles at such loci with contrasting effects cannot be fixed by selection because changing the allele frequency to increase fitness for one trait simultaneously causes a decrease in fitness for the other trait. Genetic variation at loci with such contrasting effects on traits causes a negative genetic correlation and prevents further change in trait means by natural selection.

Many examples of correlated responses to natural selection have been observed in agricultural organisms since artificial selection is routinely practiced to alter quantitative traits to increase yield and improve growth and harvest phenotypes. In one experiment, pigs were subjected to one generation of artificial selection for increased litter size in sows (Estany et al. 2002). During the course of this artificial selection experiment, the progeny of the selected females were

### Interact box 9.2 Response to natural selection on two correlated traits

Solving the equation  $\Delta\bar{z} = GP^{-1}s$  requires the use of matrix algebra. For those who have access to the program Matlab, the text web page has a short program that can be used to define  $G$ ,  $P$ , and  $s$  and then solve for  $\Delta\bar{z}$ .

As an exercise, change the sign of both the phenotypic and genetic covariances shown in the examples of Table 9.4. How does the predicted response to selection change?

measured for a number of morphological and behavioral phenotypes between the ages of 75 and 165 days old. A direct response to artificial selection was observed, increasing the litter size of the selected sows by an average of 0.46 piglets per litter compared to unselected controls. At the same time, the progeny of the selected sows showed a number of phenotypic differences from the control progeny of unselected sows. The pigs from selected sows deposited fat more rapidly, grew at different rates, and gained less weight per kilogram of feed. The pigs in the selected and control populations also showed behavioral differences, with the progeny of selected sows visiting feeding stations less frequently but spending more time eating and eating more per feeding bout. Since only the number of piglets per litter was under artificial selection, all of the other changes observed in quantitative traits were the result of correlated responses to selection caused by genetic correlations.

### **Long-term response to selection**

The breeder's equation predicts response to selection over single-generation intervals. Extrapolating the breeder's equation to longer time periods implicitly assumes that  $h^2$  and  $s$  remain constant through time. However, when natural or artificial selection continues for many generations, the assumptions of the breeder's equation may no longer hold. In particular, additive genetic variation may be consumed by response to selection over time as allele frequencies at the multiple loci that cause quantitative trait variation change over time. How rapidly additive genetic variation is exhausted by selection depends critically on the number of loci that underlie a quantitative trait as well as the percentage of trait variation that is caused by each locus (Fig. 9.12).

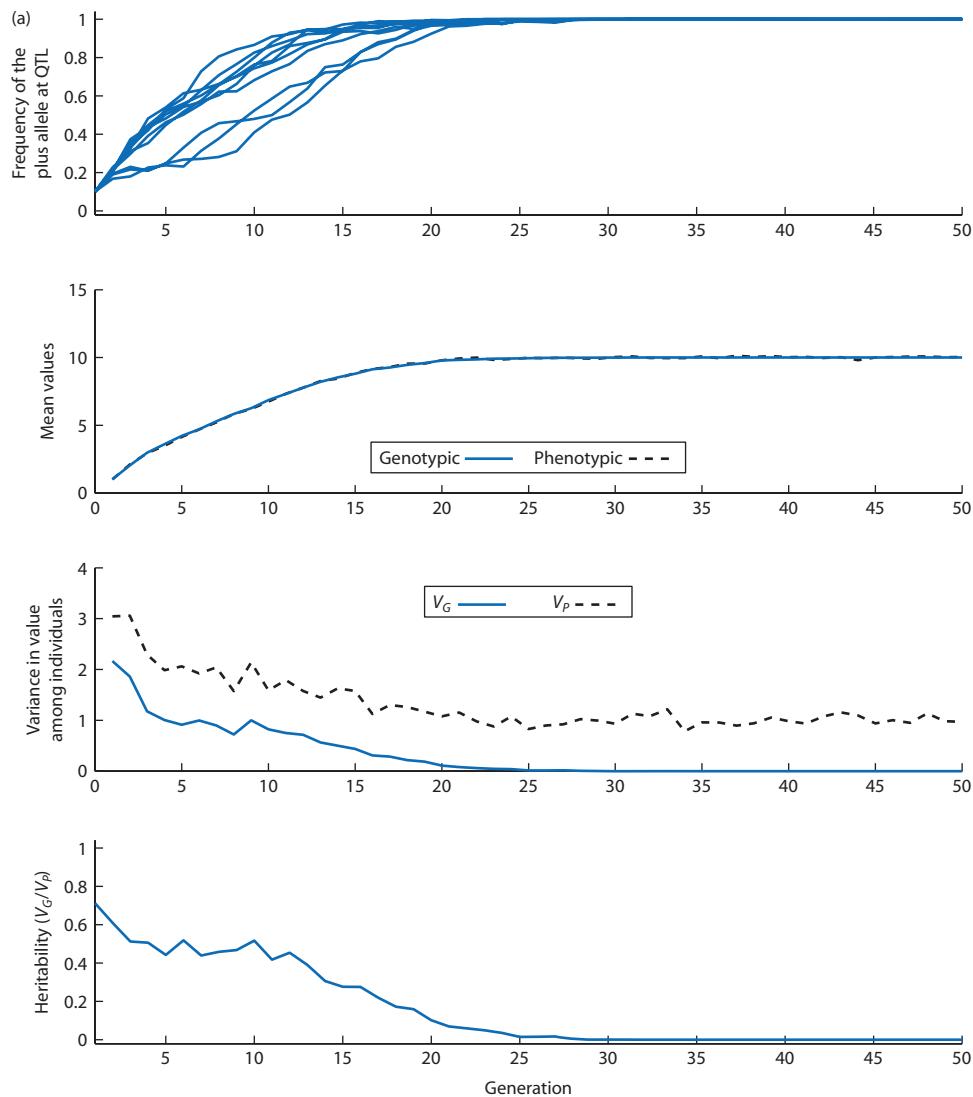
Genotypic variation in quantitative traits can sometimes be caused by the alleles segregating at a relatively small number of loci. When quantitative trait variation is caused by a small number of loci, additive genetic variance for a trait is depleted over time by response to selection. The decline in heritability occurs because response to selection causes allele frequencies at the loci that cause genotypic variation to move toward fixation and loss. When only a few loci explain genetic variation in a quantitative trait, changes in the trait mean from one generation to the next come about due to substantial changes in the allele frequencies of those few loci. For example, when there is artificial selection for an increased trait mean as in Fig. 9.10, alleles that

confer higher values of the trait at each locus are increased in frequency each generation. The greater a response to selection that has occurred, due to either continued selection over time or stronger selection, the more likely that allele frequencies at each locus will have been altered toward fixation and loss (Fig. 9.12a). In addition, genotypic variation in all traits will decrease over time in finite populations due to genetic drift, a process that is accelerated by selection since selection itself leads to reduced effective population sizes because not all individuals in the population contribute alleles to the next generation of progeny.

An alternative model is that the additive genetic variation in a quantitative trait is caused by a very large number of individual loci and all of these many loci have equal very small effects on a quantitative trait. Under these assumptions, response to selection may continue for a long time before changes occur in the amount of additive genetic variance (Fig. 9.12b). Under this **infinitesimal model**, as the number of loci grows very large then the amount of trait variation explained by any one locus approaches zero. Under the many loci assumption, when a quantitative trait responds to selection the trait mean changes but there is almost no change in the allele frequencies of the individual loci that cause genotypic variation because each locus explains such a small fraction of the additive genetic variation. Under the infinitesimal model then, response to selection can occur for many generations without causing substantial changes in the heritability required for response to selection. Note that, even under the infinitesimal model, response to selection acts to increase gametic disequilibrium for the loci that cause genotypic variation, potentially slowing response to selection over time.

**Infinitesimal model** A model of the genetic basis of quantitative traits that assumes that a very large number of independent loci contribute equally to trait genotypic variation, so that the impact of each locus on trait genotypic variation approaches zero.

The amount of additive genetic variation for a quantitative trait over many generations is not simply a function of the amount of standing genetic variation before natural selection starts. Rather, the amount of additive genetic variation over time is the net outcome of natural selection and genetic drift working



**Figure 9.12** Simulations of directional selection on a quantitative trait with strictly additive genetic variance under three scenarios (parts (a), (b), and (c)) for the genetic architecture of the trait. From top to bottom in each part, the graphs show the plus allele frequencies (in b and c 10 loci are shown of the total 100 loci that influence the trait), the genotypic and phenotypic mean values for the trait, genotypic and phenotypic trait variance ( $V_G$  and  $V_P$ ), and the narrow-sense heritability. In (a), there are 10 loci each with equal and large effects on the trait (10% of  $V_G$ ). In (b), 100 loci have equally small effects (each 1% of  $V_G$ ) on the trait. In (c), there are two loci with large effects (20% of  $V_G$  each) and 98 loci with small effects, as well as recurrent mutation between plus and minus alleles ( $\mu = 0.001$  mutations gamete $^{-1}$  generation $^{-1}$ ). In (c), the initial allele frequencies for the two loci of large effect (dashed allele frequency lines) and eight loci of small effect are initially 0.1 while the remaining 90 loci are fixed for the minus allele. A selection plateau is reached in (a) when all of the loci causing genotypic variation in the trait reach fixation ( $V_G$  and  $h^2$  drop to 0). Simulations in (b) and (c) never reach selection plateaus nor exhaust all trait variation because at least some loci that influence trait variation remain segregating. Genetic variation is maintained in (b) because selection in finite populations is not able to fix alleles at all loci with small effects. Even though the two loci with major effects fix rapidly in (c), recurrent mutation at the many loci with small effects maintains some additive genetic variation for continued response to selection. Selection was accomplished by forming the next generation with the 50 individuals with the largest phenotypic values. In all simulations the maximum genotypic value was 10.0,  $V_E = 0.1$ , and the truncation point for natural selection each generation was the 50th percentile of phenotypic value. QTL, quantitative trait loci.

to reduce genetic variation balanced by mutation introducing novel genetic variation. Depending on the number of loci influencing a trait and the mutation rate, depletion of additive genotypic variance

in quantitative traits caused by long-term selection can be counteracted by the addition of new alleles through mutation. Figure 9.12c shows an example of the impact of recurrent mutation on a quantitative

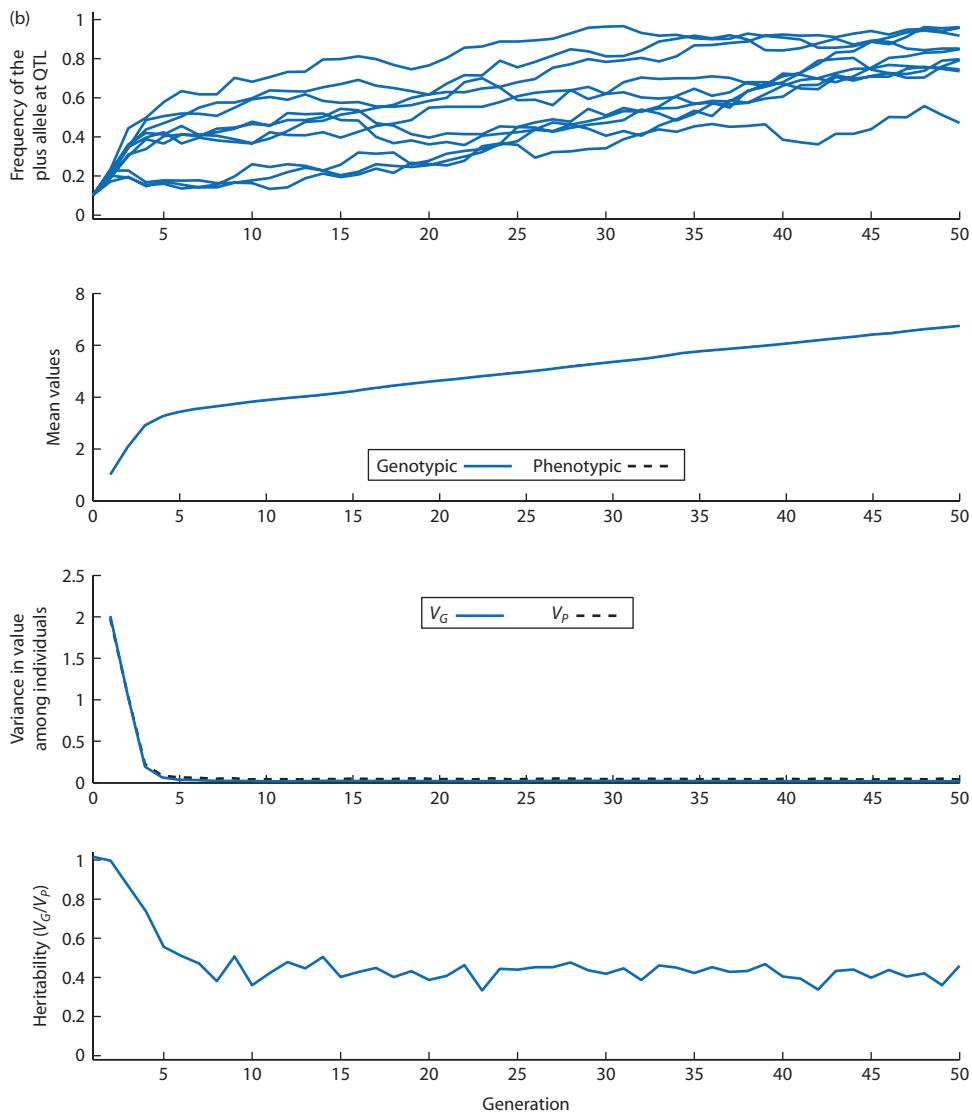
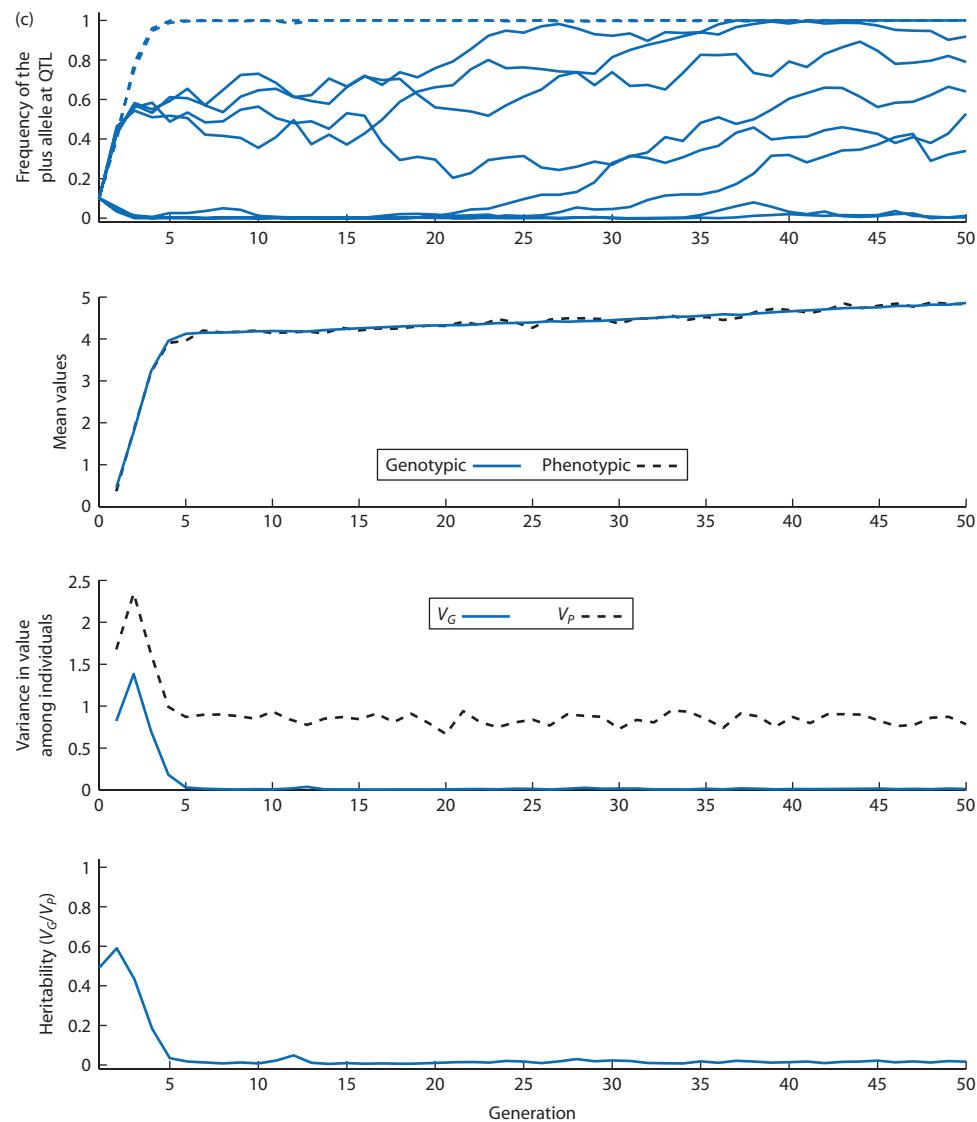


Figure 9.12 (continued)

### Interact box 9.3 Response to selection and the number of loci that cause quantitative trait variation

Use the **QTL** module in PopGene.S<sup>2</sup> to simulate response to selection for a quantitative trait that has genotypic variation caused by either few loci or by many loci. The key parameter to vary is the number of loci that underlie the quantitative trait, nQTL. Select the option for **All loci have equal effects** and try nQTL = 10 and nQTL = 100. Set the other simulation parameters at Natural selection phenotypic value truncation point = 0.5 (only individuals in the upper 50% of phenotypic value reproduce), N = 250, mutation rate = 0.0, generations = 100,  $V_E = 0.1$ , Maximum genotypic value = 10.0.

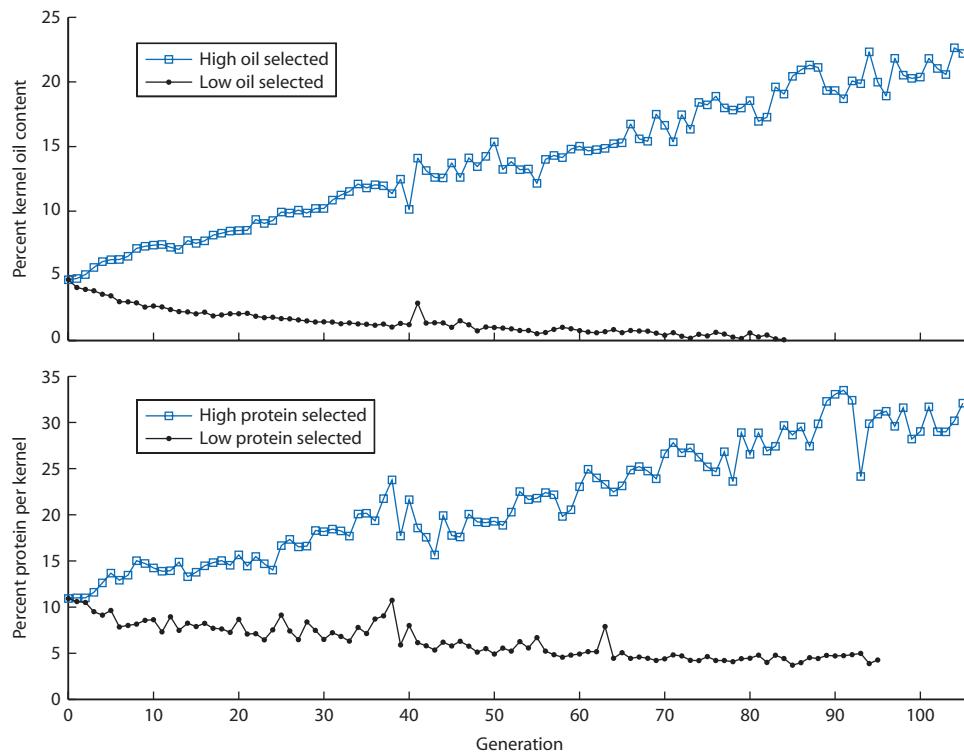
What is the heritability over time with nQTL = 10 or nQTL = 100? How does an increase in environmental variation ( $V_E = 0.1$  or, say,  $V_E = 1.0$ ) influence your conclusion?



**Figure 9.12** (continued)

trait that is the product of 100 loci. Even though 90 of the loci start out fixed for one allele, mutation at these loci over time produces enough additive genetic variation to maintain trait variation and a linear response to selection over many generations. Ultimately, the steady-state level of additive genetic variation for a quantitative trait is a product of genetic drift, mutation, and selection all acting on the loci that cause variation in the trait. Because of this, long-term response to selection will depend on numerous parameters in a population such as the effective population size, the mutation rate, and the selection differential along with the number and distribution of effects of the loci that underlie the quantitative trait.

The Illinois Long-Term Selection experiment provides one of the longest records of continuous response to selection for the phenotypes of oil content and percentage of protein in corn kernels (reviewed by Moose et al. 2004). This long-term selection experiment was initiated in 1896 with the goal of determining whether artificial selection could be used to develop strains of corn with kernel phenotypes that were improved from the perspective of animal feed and crop processing. Divergent selection for both higher and lower oil and protein content has been practiced for over 100 generations by using the highest and lowest scoring 20% of ears each generation to form the next generation (Fig. 9.13). From an initial value of 4.7% oil content, response to



**Figure 9.13**  
Phenotypic means for oil and protein content for high and low selected lines of the Illinois Long-Term Selection experiment initiated in 1896. Response to selection has been nearly linear over time, consistent with additive genetic variation that is caused by a relatively large number of loci each with a small effect on the phenotype. The low-oil-selected line was discontinued in generation 89 because oil content was too low to be measured reliably and plants had poor viability. Data kindly provided by J.W. Dudley and S.P. Moose.

selection steadily changed oil content to about 22% in the high line and to 0% in the low line. Response to selection was similarly linear for protein content, starting at 10.9% and reaching 32.1% in the high line and 4.2% in the low line.

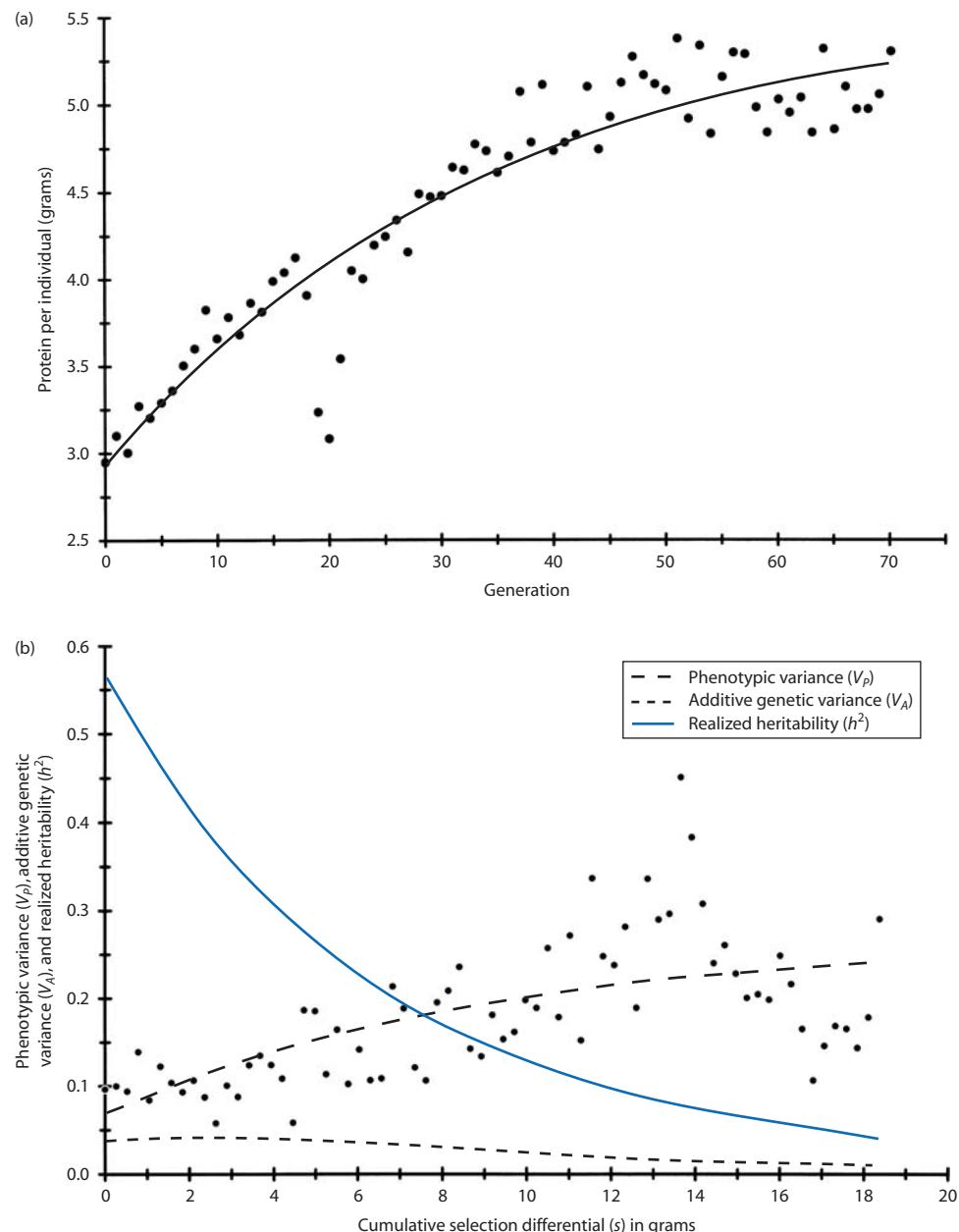
The Illinois Long-Term Selection experiment shows that response to selection is relatively steady and linear over a long period of time. (Slight fluctuations in the mean phenotype over time may be explained by some variation in the selection differential each generation as well as by environmental variation.) The observed change in phenotypic means in the Illinois Long-Term Selection experiment looks similar to the predicted response to selection on traits with many loci in the simulation results shown in Figs 9.12b and 9.12c. The results are therefore consistent with genetic variation in the oil- and protein-content phenotypes being caused by a relatively large number of loci and the possibility that mutation may have contributed some genetic variation over time.

Another example is long-term selection carried out for 70 generations to increase the percentage of body muscle, measured as protein content, in mice at 42 days old (Bünger et al. 1998). Figure 9.14a shows the mean amount of protein per individual over the course of the experiment. Selection increased protein content rapidly at first, but then response

to selection slowed as the experiment continued. This is a classic example of a **selection plateau** where continued natural selection shows a diminishing response over time such that the trait mean asymptotes toward a constant value even though selection is still being applied to increase the trait mean. The additive genetic variance ( $V_A$ ) and the heritability ( $h^2$ ) shown in Fig. 9.14b explain why response to selection decreased over time. The additive genetic variance, and therefore the heritability, decreased through time so that the response to selection was not constant over the 70 generations of the experiment. The observed reduction over time in the response to selection and the selection plateau are similar to the predicted response to selection on traits with few loci of large effects shown in Fig. 9.12a. The response to long-term selection for protein content in mice is therefore consistent with genetic variation caused by a relatively small number of loci having relatively large effects on the phenotype.

Two additional phenomena can cause limits to selection response. The first process is the accumulation of gametic disequilibrium that alters the amount of additive genetic variance. The additive genetic variance ( $V_A$ ) can be decomposed into two parts:

$$V_A = V_a + D \quad (9.22)$$



**Figure 9.14** Long-term selection for muscle mass in mice (measured as protein content per individual) over 70 generations. (a) The phenotypic mean over time, with a pronounced asymptote, or selection plateau, that indicates a diminishing response to selection over time. The total phenotypic variance, the additive genetic variance, and the realized heritability are shown in (b). Even though the selection differential is constant over time, the heritability declines steadily, as expected for a quantitative trait where genetic variation is caused by relatively few loci. The dip in protein content during generations 18–20 was an artifact due to an environmental effect. While the phenotypic variance increases over the experiment (b), this is caused largely by the increase in the phenotypic mean (the coefficient of variation of  $V_p$  stays nearly constant). Redrawn from Bünger, L., Renne, U., Dietl, G., and Kuhla, S. (1998) Long-term selection for protein amount over 70 generations in mice. *Genetical Research* 72: 93–109. Reprinted with the permission of Cambridge University Press.

where  $V_a$  is the additive genetic variance that is not impacted by gametic disequilibrium and  $D$  is the gametic disequilibrium coefficient. When there is gametic equilibrium ( $D = 0$ ) then the additive genetic variance is all caused by variance in alleles, or genic

variance. However, when there is some level of gametic disequilibrium then the additive genetic variance can be reduced (negative  $D$ ) or increased (positive  $D$ ). Directional and stabilizing natural selection tend to cause negative gametic disequilibrium

because individuals with similar phenotypic values tend to have a correlated set of alleles at each of the loci that contribute to the trait (also see the example in Chapter 2). In contrast, disruptive selection tends to cause positive gametic disequilibrium. If selection is strong relative to recombination, then gametic disequilibrium will alter the additive genetic variance and thereby the heritability. A reduction in the response to selection or a selection plateau can occur when negative gametic disequilibrium caused by selection has accumulated such that  $V_A$  declines even though the loci that underlie the trait have not all reached fixation and loss. For more details on this complex topic consult Bulmer (1985) and Lynch and Walsh (in preparation; draft chapters available at <http://nitro.biosci.arizona.edu/zbook/book.html>).

The other process that can limit response to selection is called **antagonistic pleiotropy**. It occurs when response to selection changes the mean of one trait (selection for) as well as the mean of a correlated trait (selection of). While selection for increased fitness drives change in the mean of one trait over time, the correlated change in the mean of the other trait may actually decrease fitness. After some response to selection has occurred, the fitness trade offs between the two traits may reach a point where further change in the mean of one trait that increases fitness is offset by the negative fitness consequences of change in the mean of a correlated trait. When such fitness trade offs for correlated traits exist, they will ultimately limit response to selection even when additive genotypic variation exists for both traits. Examples of such trade offs for correlated traits have been hypothesized or observed for a range of traits. One possible trade off maintained by antagonistic pleiotropy involves alleles that tend to increase reproduction at early ages but decrease lifespan (Williams 1957). In support of the hypothesis that survival and reproduction have a fitness trade off, Silbermann and Tatar (2000) have shown that a heat-induced protein expressed by the *hsp70* locus influences both egg hatching rate and survival rate in *Drosophila melanogaster*. Higher levels of *hsp70* expression lead to longer lifespans but at the same time lead to lower rates of egg hatching. If reproduction and survival in *D. melanogaster* are associated by antagonistic pleiotropy via *hsp70*, then response to any selection acting on these traits will not be able to exhaust all the additive genetic variation if the highest fitness is a balance of intermediate levels of both traits.

The genetic variance/covariance or **G** matrix has also become a focus of research to better understand

how the genetic basis of potentially non-independent quantitative traits changes over time with selection, genetic drift, and mutation (reviewed by Steppan et al. 2002; Jones et al. 2003). The **G** matrix represents the genetic inter-relationships among multiple traits, so estimating it in a range of species and for a range of traits is a prerequisite to understand how phenotypes might respond to long-term natural selection. Understanding how the **G** matrix changes over time is also fundamental to understanding long-term response to selection. The predicted trait means in equation 9.21 rely on the constancy of the **G** matrix over time. If **G** changes rapidly or unpredictably over time then predicted changes in trait means are not applicable over long periods of time. Rapid changes in **G** would also reduce the ability to infer past patterns of response to selection given a current estimate of **G**.

### **Neutral evolution of quantitative traits**

Considering a quantitative trait as selectively neutral is useful to predict the action of basic processes that will reduce as well as contribute to additive genetic variation. If a quantitative trait is neutral, each locus that contributes to variation in the trait should be neutral as well. Therefore, we can employ expressions already developed to predict the consequences of genetic drift. Additive genotypic variation is expected to decrease with genetic drift over time because the alleles at the loci that form the basis of  $V_A$  will progress toward fixation and loss. The expected rate of decrease in  $V_A$  by genetic drift is

$$V_A^t = V_A^0 \left( 1 - \frac{1}{2N_e} \right)^t \quad (9.23)$$

where  $V_A^0$  is the initial additive genotypic variation,  $N_e$  is the effective population size, and  $t$  is the number of generations that have elapsed. As explained in

Chapter 3, the  $\left( 1 - \frac{1}{2N_e} \right)$  term in this equation

predicts the decline in heterozygosity through time by genetic drift. Additive genetic variation is greater when heterozygosity is greater because alleles are at intermediate frequencies.

Next we can predict the balance between the additive genotypic variation lost by drift and new additive variation gained due to new mutations at the loci that influence a quantitative trait. Let's start by reworking the equation for decline in  $V_A$  over an

arbitrary time interval to instead predict the change in additive genotypic variation each generation due to drift:

$$\Delta V_A = -\frac{V_A}{2N_e} \quad (9.24)$$

This equation is obtained from equation 9.23 by setting  $t = 1$ , multiplying the right side out, and then subtracting  $V_A^o$  from both sides. Then assume that the amount of new additive genotypic variance caused by mutation each generation is  $V_A^M$ . The sum of additive variation lost by drift and additive variation gained by mutation is

$$\Delta V_A = -\frac{V_A}{2N_e} + V_A^M \quad (9.25)$$

This defines the net change in additive genotypic variation due to the action of both drift and mutation. If we assume that a population is at drift-mutation equilibrium then  $\Delta V_A = 0$ , allowing equation 9.21 to be rearranged to give

$$V_A = 2N_e V_A^M \quad (9.26)$$

(see Lynch & Hill 1986; Bürger & Lande 1994).

The expected additive genotypic variance for a neutral quantitative trait at drift-mutation equilibrium was the basis of a much debated recommendation that  $N_e = 500$  would be sufficient to maintain additive genotypic variation in populations of endangered species (Franklin 1980; Lande 1995). This recommendation came from rearranging equation 9.26 to solve for the effective population size:

$$N_e = \frac{V_A}{2V_A^M} \quad (9.27)$$

Assuming that total phenotypic variation is caused only by  $V_A$  and  $V_E$  and using the definition of the narrow-sense heritability,  $V_A$  can be expressed as  $V_A = \frac{h^2}{1-h^2}V_E$ . If  $h^2 = 0.5$ , then  $V_A$  in the numerator on the right side of equation 9.27 can be replaced with  $V_E$ . The final step was to utilize an estimate of the input of additive variation due to mutation available at the time that suggested  $V_A^M \approx 0.001V_E$ . Putting this all together results in

$$N_e \approx \frac{V_E}{2(0.001V_E)} \approx 500 \quad (9.28)$$

as an estimate of the effective population size expected to maintain a heritability of about one-half.

### Interact box 9.4 Effective population size and genotypic variation in a neutral quantitative trait

In PopGene.S<sup>2</sup>, open the QTL simulation module and then explore how the effective population size influences the level of genotypic variation for a selectively neutral quantitative trait. The key parameter to adjust is the number of individuals in the population. Try  $N = 50$  and then  $N = 500$  while the other parameters are set at nQTL = 10, mutation rate = 0.00001, generations = 100,  $V_E = 0.1$ , Maximum genotypic value = 10.0, and Selection threshold = 0.0 (this last parameter value sets the truncation point so that all individuals reproduce each generation, effectively removing natural selection).

How much genotypic variance ( $V_G$ ) do you find with these two effective population sizes? How variable are the levels of  $V_G$  over time with  $N = 50$  or  $N = 500$ ? How do the mutation rate and the environmental variance ( $V_E$ ) in the trait influence the long-term heritability for each effective population size?

This recommendation for an effective population size target of 500 sparked a debate that centered around the numerous assumptions about the nature and causes of quantitative trait variation in natural populations. Lande (1995) pointed out that  $V_A^M \approx 0.001V_E$  was probably too high since the mutation rate needed to be discounted for the number of mutations that were highly deleterious and a more realistic assumption was  $V_A^M \approx 0.0001V_E$  by counting only nearly neutral mutations. This change in the mutational input of quantitative genetic variance each generation substantially increases the required effective population size to approximately 5000. Franklin and Frankham (1998) countered that assuming a heritability of 0.5 was larger than necessary because heritabilities for fitness-related traits were often around 0.1 (see also Frankham

and Franklin 1998). They reasoned that even if  $V_A^M \approx 0.0001 V_E$ , an effective population size of about 550 was sufficient to maintain a heritability of 0.1. In contrast, inferred  $\frac{V_A}{V_A^M}$  ratios were observed in a range of organisms to be between 30 and 300 which implies that  $N_e$  is between 15 and 150 (Houle et al. 1996; Lynch & Lande 1998). Since this effective population size is definitely too low for some of the species studied by Houle et al. (1996), such as *Drosophila*, these findings then called into question the very assumptions of neutral drift–mutation equilibrium. An alternative explanation is that stabilizing natural selection is operating and causing reduced variation in quantitative traits and could explain the  $\frac{V_A}{V_A^M}$  ratios rather than such low effective population sizes.

The perspective of neutral evolution for quantitative traits has also been employed to make inferences about the nature of the total genotypic variation ( $V_G$ ). Genetic drift that occurs during genetic bottlenecks and founder events as well as consanguineous mating impacts the components of genotypic variation for neutral quantitative traits. In a seeming contradiction, genetic drift can produce a transient *increase* in additive genotypic variance for neutral quantitative traits that exhibit dominance ( $V_D$ ) or epistasis ( $V_I$ ) genotypic variance (Robertson 1952; Goodnight 1987, 1988; Willis & Orr 1993; Barton & Turelli 2004). For quantitative traits that exhibit only  $V_A$  and  $V_D$ , higher heterozygote frequencies produce more  $V_D$  and less  $V_A$  when the dominance coefficient ( $d$ ) increases. Genetic drift causes an increase in homozygosity as allele frequencies change toward fixation and loss on average under random sampling. An increase in homozygosity is also a decrease in heterozygosity and thereby a decrease in  $V_D$ . For a single generation bottleneck of  $N_e = 2$  and assuming no epistasis, Willis and Orr (1993) showed that the expected value of  $V_A$  over many replicate populations increases if  $d > 0.29$ . They also showed that as  $N_e$  during a bottleneck increases, the threshold dominance coefficient for  $V_A$  to increase approaches  $d > 0.20$ . The increase in  $V_A$  is greater for larger  $d$  and for lower initial frequencies of the recessive allele. These increases in  $V_A$  translate into increases in the heritability that also depend on the magnitude of  $V_E$  for the trait. Consanguineous mating also acts to increase homozygosity while not altering allele frequencies, and so can also cause a reduction in  $V_D$  that leads to a relative increase in  $V_A$ .

These predictions about changes in heritability after genetic drift or consanguineous mating lead to a test of whether quantitative traits have genotypic variation that is exclusively additive or is a combination of additive and non-additive genotypic variation. Van Buskirk and Willi (2006) reviewed the results of numerous studies that estimated heritability and  $V_A$  from both small or inbred populations as well as large randomly mated populations. They found that phenotypes closely associated with fitness (e.g. viability, fecundity, body size) often did show an increase in heritability after consanguineous mating or population bottlenecks. In contrast, phenotypes not associated with fitness (e.g. morphological traits, bristle number, oil content) showed only declines in heritability after consanguineous mating or population bottlenecks. This result is consistent with the explanation that  $V_G$  in fitness-related phenotypes was caused by both allele and genotype variation whereas  $V_G$  in non-fitness phenotypes was caused by alleles alone in the organisms studied.

### 9.3 Quantitative trait loci (QTL)

- QTL mapping with single marker loci.
- QTL mapping with multiple marker loci.
- Limitations of QTL mapping studies.
- Biological significance of QTL mapping.

Thus far in this chapter, quantitative traits have been described based on the mean and variance of values in a population. Quantities like the components of the total genotypic variance and the heritability illuminate population-level average qualities of phenotypes. However, the numerous loci that each contribute to such population variation have not been identified nor described. This final section of the chapter introduces the concepts and some methods needed to identify the individual loci that ultimately contribute to variation in quantitative traits.

Using the basic framework of phenotypic values and the population mean already established, and joining it with genetic marker data for individuals, it is possible to identify individual regions in a genome that contribute to quantitative trait variation. The genomic regions that contribute to variation in quantitative traits are called **quantitative trait loci** or **QTLs**. In the simplest idealized case, individual QTLs are single genes that contribute to the value of a quantitative trait and have alternate alleles with different effects on phenotypic value. The trait mean and variance would then be the sum of the mean

and variance contributed by each gene that affects the trait. In reality, QTLs are not necessarily individual genes but can be larger chromosomal regions held together by linkage that may contain several or many genes. It is these linkage blocks, each of which contains a genetic marker, that can be associated with an effect on the variance of a quantitative trait.

**Candidate loci** Loci that have known or inferred function and therefore are hypothesized to be causal contributors to genetic variation in a quantitative trait.

**Genetic architecture** The number of loci that underlie a quantitative trait and the magnitude of their contributions to quantitative trait genetic variation.

**Quantitative trait locus or QTL** A gene, or more often a genome region associated by linkage, possessing multiple alleles that affect the average value of a quantitative trait in a defined population associated through linkage with a genetic marker.

A major goal in identifying and describing the individual QTLs that cause quantitative genetic variation is to understand the **genetic architecture** of continuous phenotypes. By one definition, genetic architecture is all of the genetic and environmental factors that contribute to a quantitative trait, as well as their magnitude and their interactions (National Institute of General Medical Sciences 1998). More narrowly, genetic architecture often refers to the number of QTLs and the size of their effects on a quantitative traits. Identification of the number and phenotypic effects of QTLs has applications in many areas of biology. Identification of QTLs helps test the role, if any, of **candidate loci** (identified through independent molecular biology research or sequence analyses, for example) in explaining a portion of the genetic variance in quantitative traits. The reverse is also true, since loci identified by QTL mapping (which is described below) as causing some of the genetic variation in a quantitative trait are often further studied to better understand their function. In clinical settings, QTL mapping helps identify genes and alleles that cause disease conditions as well as genetic markers associated with disease QTLs that can be used to screen for disease risk. Identifying QTLs can improve the efficiency of animal and plant breeding, such as in the development of genetic markers

associated with QTLs that can be used to screen individuals early in life for traits that may only be manifest later in life. One example would be screening tree seedlings for mature wood characteristics and planting those individuals with genetic marker genotypes associated with desirable phenotypes that appear only after many years of growth.

The number of QTLs and the size of their effects on a quantitative trait also have profound implications for evolutionary change such as the response to natural selection and the amount of variation generated by mutation. QTL mapping has the potential to deconstruct phenomena in quantitative genetics that have traditionally only been analyzed via variance components. In principle at least, QTL mapping can distinguish the specific causes of genetic correlations (pleiotropy or linkage), identify the loci and alleles that demonstrate dominance and epistasis, and show how specific loci involved in genotype-by-environment interactions respond to their environments.

#### ***QTL mapping with single marker loci***

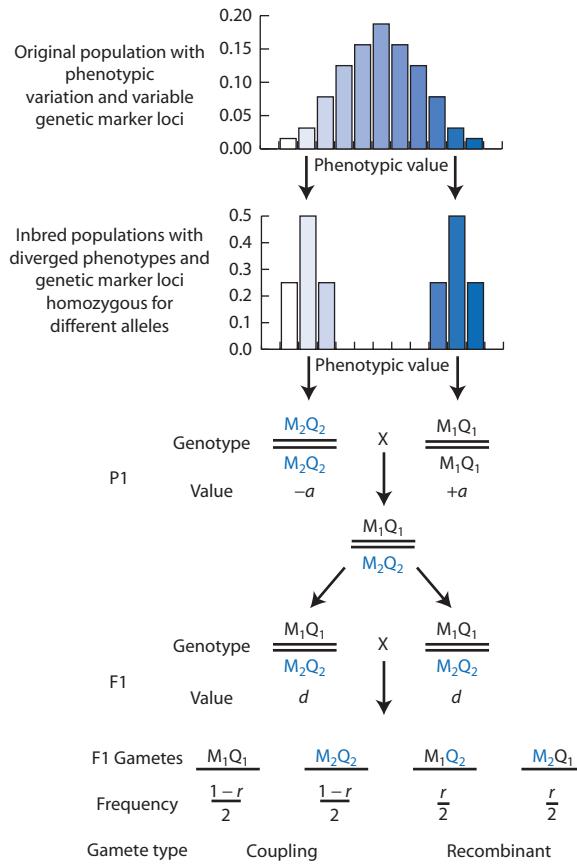
The process of identifying quantitative trait loci is called **QTL mapping** because it is based on the technique of linkage mapping that establishes the linear order of loci on chromosomes based on recombination frequencies. QTL mapping takes basic linkage mapping one step further by determining whether any of the mapped loci are associated with variation in a quantitative trait. QTL mapping requires variation for a quantitative trait within or between populations and numerous polymorphic genetic marker loci that are spread across the genome of the organism. The mapping is carried out for the marker loci, since after all the QTLs are unknown. QTLs are identified by differences in the mean phenotype of groups of individuals with different marker locus genotypes. Any association between phenotypic means and marker genotypes is caused by gametic disequilibrium between QTLs and marker loci when the two types of loci are close enough on a chromosome to be linked. Thus, QTL mapping uses known marker loci to detect the phenotypic signature of unknown QTLs that are in the same linkage block.

Genetic marker loci are a critical ingredient in QTL mapping. First and foremost, marker loci should be both independent of the phenotype(s) being mapped (i.e. not a QTL themselves) and also selectively neutral so that genotype frequencies are determined only by mating and recombination but not by viability

or fecundity selection. Marker loci must be polymorphic because they are only informative when individuals with different phenotypic values possess distinct marker locus genotypes. Codominant marker loci where all possible genotypes are detectable present the easiest case to consider, although mating designs exist that can utilize dominant marker loci. Before the present age of genomics, only loci with phenotypic effects or protein polymorphisms were used as markers and QTL mapping was generally limited by the small number and low polymorphism of such marker loci. In the present day, QTL mapping can employ a full battery of DNA markers such as restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), microsatellites, and single nucleotide polymorphisms (SNPs). The numerous genome-sequencing projects that have been completed or are planned lead to the identification of marker loci spread across the entire genome. Linkage maps based on many molecular markers are also now available for humans as well as model and agricultural species.

Experimental mating designs are a major component of QTL mapping. One of the most basic breeding schemes for QTL mapping in organisms that can be raised and mated in captivity is called the **F2 design or recombinant inbred line design** (Fig. 9.15). It starts with a population, perhaps in the wild, that has considerable phenotypic variation as well as genetic variation at numerous molecular marker loci. From this original population, individuals are sampled to start inbred lines or subpopulations. The inbred lines are maintained by some type of consanguineous mating (e.g. selfing, brother–sister mating) for numerous generations. The inbred lines eventually contain individuals with a high probability of homozygosity for different alleles at the loci that cause variation in the phenotype ( $Q_1$  and  $Q_2$ ) as well as for the different alleles present at the genetic marker loci ( $M_1$  and  $M_2$ ). An individual is sampled from each of two inbred lines that exhibit different values for the phenotype(s) of interest as well as different homozygous genotypes at the marker locus. These individuals form the P1 generation in Fig. 9.15.

When the P1 individuals are mated, they produce progeny that are all heterozygotes for both the QTL and the marker locus. Note that recombination events do not alter the gametes produced because each P1 individual is a double homozygote. The progeny of the P1 generation form the F1 generation. Figure 9.15 shows the four gametes that are produced by the F1 individuals and transmitted after



**Figure 9.15** The F2 or recombinant inbred line design for QTL mapping assuming one QTL (Q) and one genetic marker locus (M). The top phenotypic distribution represents the variance in value in the population that is subject to QTL mapping. Individuals at the edges of this phenotypic distribution are then sampled to start lines to be inbred for five (self-fertilization) to 10 (sib mating) generations to achieve high homozygosity. Individuals are then sampled from inbred lines with diverged phenotypes to form the P1 generation. The progeny of the P1 individuals are all double heterozygotes and form the F1 generation. Gamete frequencies produced by F1 individuals depend on recombination rates between loci. Random mating among F1 individuals produces the F2 individuals which are genotyped for the marker loci and measured for their phenotypic values.

random mating to their progeny in the F2 generation. Notice now that recombination between the QTL and the marker locus does influence gamete frequencies and thereby the genotype frequencies in the F2 generation. The QTL mapping analysis is based on two types of data for all F2 individuals: that is, (i) the marker locus genotype is known and (ii) the phenotypic value is known.

The 10 unique two-locus genotypes in the F2 population are shown in Table 9.5, grouped into three classes based on the marker-locus genotype.

**Table 9.5** Derivation of the expected phenotypic value for the three marker-locus genotypes when QTL mapping with a single marker locus associated with a single QTL. The three genetic marker genotypes may be associated with any of the three possible QTL genotypes because of recombination during the formation of F2 gametes. The difference between the  $M_1/M_1$  and  $M_2/M_2$  marker class means (expressions in the Marker-class mean value column) is equal to  $2\hat{a}$ . The phenotypic value of each marker locus genotype is a function of both the additive and dominance effects of the QTL ( $a$  and  $d$ ) as well as the recombination rate ( $r$ ). So unless there is no dominance and no recombination, estimates of QTL effects from single-marker-locus mapping are always minimum estimates. The gametes and expected gamete frequencies are given in Fig. 9.15.

Gametes	F2 genotype	Genotype frequency	Genotypic value	Frequency-weighted genotypic value	Marker genotype	Marker contribution to F2 population mean value	Marker genotype frequency in F2 population	Marker-class mean value
$c/c$	$\overline{\overline{M_1Q_1}}$	$\left(\frac{1-r}{2}\right)^2$	$+a$	$a\left(\frac{1-r}{2}\right)^2$	$\overline{M_1}$	$\bar{G}_{M_1M_1}^{pop} = \frac{a(1-2r)}{4} + \frac{2dr(1-r)}{4}$	$1/4$	$\bar{G}_{M_1M_1} = a(1-2r) + 2dr(1-r)$
$c/r$	$\overline{M_1Q_1}$	$(2)\frac{r}{2}\left(\frac{1-r}{2}\right)$	$d$	$2d\frac{r}{2}\left(\frac{1-r}{2}\right)$	$\overline{M_1}$	$\bar{G}_{M_1M_1}^{pop} = \frac{a(1-2r)}{4} + \frac{2dr(1-r)}{4}$	$1/4$	$\bar{G}_{M_1M_1} = a(1-2r) + 2dr(1-r)$
$r/r$	$\overline{M_1Q_2}$	$\left(\frac{r}{2}\right)^2$	$-a$	$-a\left(\frac{r}{2}\right)^2$	$\overline{M_1}$	$\bar{G}_{M_1M_1}^{pop} = \frac{a(1-2r)}{4} + \frac{2dr(1-r)}{4}$	$1/4$	$\bar{G}_{M_1M_1} = a(1-2r) + 2dr(1-r)$
$c/c$	$\overline{\overline{M_2Q_2}}$	$(2)\left(\frac{1-r}{2}\right)^2$	$d$	$2d\left(\frac{1-r}{2}\right)^2$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{d[(1-r)^2+r^2]}{2}$	$1/2$	$\bar{G}_{M_2M_2} = d[(1-r)^2+r^2]$
$r/r$	$\overline{M_1Q_2}$	$(2)\left(\frac{r}{2}\right)^2$	$d$	$2d\left(\frac{r}{2}\right)^2$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{d[(1-r)^2+r^2]}{2}$	$1/2$	$\bar{G}_{M_2M_2} = d[(1-r)^2+r^2]$
$c/r$	$\overline{M_2Q_1}$	$(2)\frac{r}{2}\left(\frac{1-r}{2}\right)$	$+a$	$2a\frac{r}{2}\left(\frac{1-r}{2}\right)$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{d[(1-r)^2+r^2]}{2}$	$1/2$	$\bar{G}_{M_2M_2} = d[(1-r)^2+r^2]$
$r/c$	$\overline{M_1Q_2}$	$(2)\frac{r}{2}\left(\frac{1-r}{2}\right)$	$-a$	$-2a\frac{r}{2}\left(\frac{1-r}{2}\right)$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{d[(1-r)^2+r^2]}{2}$	$1/2$	$\bar{G}_{M_2M_2} = d[(1-r)^2+r^2]$
$c/c$	$\overline{\overline{M_2Q_2}}$	$\left(\frac{1-r}{2}\right)^2$	$-a$	$-a\left(\frac{1-r}{2}\right)^2$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{d[(1-r)^2+r^2]}{2}$	$1/2$	$\bar{G}_{M_2M_2} = d[(1-r)^2+r^2]$
$c/r$	$\overline{M_2Q_2}$	$(2)\frac{r}{2}\left(\frac{1-r}{2}\right)$	$d$	$2d\frac{r}{2}\left(\frac{1-r}{2}\right)$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{-a(1-2r)}{4} + \frac{2dr(1-r)}{4}$	$1/4$	$\bar{G}_{M_2M_2} = -a(1-2r) + 2dr(1-r)$
$r/r$	$\overline{M_2Q_1}$	$\left(\frac{r}{2}\right)^2$	$+a$	$a\left(\frac{r}{2}\right)^2$	$\overline{M_2}$	$\bar{G}_{M_2M_2}^{pop} = \frac{-a(1-2r)}{4} + \frac{2dr(1-r)}{4}$	$1/4$	$\bar{G}_{M_2M_2} = -a(1-2r) + 2dr(1-r)$

Let's work through the steps needed to obtain the expected genotypic value of those F2 individuals with a M<sub>1</sub>/M<sub>1</sub> marker locus genotype. The mean genotypic value is obtained by determining the frequency and the genotypic value of each QTL genotype that is associated with an M<sub>1</sub>/M<sub>1</sub> marker genotype. There are three possible ways to make an M<sub>1</sub>/M<sub>1</sub> marker-locus genotype in an F2 individual: (i) combine two M<sub>1</sub>Q<sub>1</sub> coupling gametes, (ii) combine an M<sub>1</sub>Q<sub>1</sub> coupling gamete with an M<sub>1</sub>Q<sub>2</sub> recombinant gamete, and (iii) combine two M<sub>1</sub>Q<sub>2</sub> recombinant gametes. Each of these possible genotypes will have a frequency in the F2 population that is a product of the respective gamete frequencies. Expected F2 genotype frequencies are therefore: (i)  $\left(\frac{1-r}{2}\right)^2$  for the combination of two coupling gametes, (ii)  $r\left(\frac{1-r}{2}\right)$  for the combination of a coupling and a recombinant gamete, and (iii)  $\left(\frac{r}{2}\right)^2$  for the combination of two recombinant gametes.

Each F2 genotype frequency must also be weighted to account for a genotype's relative abundance. Constructing a 4 × 4 Punnett square using the four possible F1 gametes that are shown in Fig. 9.15 reveals that in the F2 population all double homozygote genotypes (e.g.  $\frac{M_1Q_1}{M_1Q_1}$ ) have a frequency of  $1/16$ , all Q locus heterozygotes (e.g.  $\frac{M_1Q_1}{M_1Q_2}$ ) have a frequency of  $2/16$ , and all double heterozygotes (e.g.  $\frac{M_1Q_2}{M_1Q_2}$ ) also have a frequency of  $2/16$ . Given an M<sub>1</sub>/M<sub>1</sub> marker-locus genotype, the Q locus homozygote genotypes  $\frac{M_1Q_1}{M_1Q_1}$  and  $\frac{M_1Q_2}{M_1Q_2}$  occur half as often as the Q locus heterozygote  $\frac{M_1Q_1}{M_1Q_2}$ . Therefore, the Q locus heterozygote frequency is weighted by a factor of 2, as shown in Table 9.5.

Each of the individuals carrying an M<sub>1</sub>/M<sub>1</sub> marker locus genotype will have a genotypic value that depends on the genotype at the QTL locus. It is customary to use an arbitrary scale of measurement for the genotypic values (see section 10.1). On this scale, the genotypic value of the Q<sub>1</sub>Q<sub>1</sub> genotype is assigned +a, the genotypic value of the Q<sub>2</sub>Q<sub>2</sub> genotype is assigned -a, and the genotypic value of the Q<sub>1</sub>Q<sub>2</sub> genotype is assigned d (refer to Fig. 10.1).

The point exactly mid-way between the genotypic values of the homozygotes, called the midpoint, is defined to be zero. When the genotypic value of the heterozygote is at the midpoint, d = 0. The degree of dominance can be expressed as the ratio of d/a. Using this scale to measure genotypic values, the  $\frac{M_1Q_1}{M_1Q_1}$  genotype has a value of +a, the  $\frac{M_1Q_1}{M_1Q_2}$  genotype a value of d, and the  $\frac{M_1Q_2}{M_1Q_2}$  genotype a value of -a. The expected frequency of each of the three genotypic values within the M<sub>1</sub>/M<sub>1</sub> marker-locus genotype is then the product of the genotype frequency and the corresponding genotypic value.

The portion of the mean genotypic value for the entire F2 population due to those individuals with an M<sub>1</sub>/M<sub>1</sub> marker genotype, call it  $\bar{G}_{M_1M_1}^{pop}$ , is the sum of the genotype-frequency-weighted genotypic values for each QTL genotype associated with an M<sub>1</sub>/M<sub>1</sub> marker genotype:

$$\bar{G}_{M_1M_1}^{pop} = a\left(\frac{1-r}{2}\right)^2 + 2d\left(\frac{r}{2}\right)\left(\frac{1-r}{2}\right) - a\left(\frac{r}{2}\right)^2 \quad (9.29)$$

Each term can be multiplied out and the first term separated to give

$$\bar{G}_{M_1M_1}^{pop} = \frac{a(1-2r)}{4} + \frac{ar^2}{4} + \frac{2dr(1-r)}{4} - \frac{ar^2}{4} \quad (9.30)$$

which after addition simplifies to

$$\bar{G}_{M_1M_1}^{pop} = \frac{a(1-2r)}{4} + \frac{2dr(1-r)}{4} \quad (9.31)$$

as the expected genotypic mean value of all individuals with an M<sub>1</sub>/M<sub>1</sub> marker locus genotype.

Since  $1/4$  of all individuals in the F2 population are expected to have an M<sub>1</sub>/M<sub>1</sub> marker genotype by Hardy–Weinberg, we can multiply by a factor of 4 to obtain the actual genotypic mean of that subset of individuals in the F2 population with M<sub>1</sub>/M<sub>1</sub> genotypes:

$$\bar{G}_{M_1M_1} = a(1-2r) + 2dr(1-r) \quad (9.32)$$

The same logic can be used to obtain expressions for  $\bar{G}^{pop}$  and  $\bar{G}$  for the M<sub>1</sub>/M<sub>2</sub> and M<sub>2</sub>/M<sub>2</sub> marker genotypes. Note that environmental variation in phenotype does not need to be accounted for because

it is not expected to change the *average* phenotypic value. However, variation in phenotype about the average will be caused by both  $V_G$  and  $V_E$ .

Examine the expression for the mean genotypic values for each marker genotype in Table 9.5. If there is free recombination between the QTL and the marker locus,  $r = 0.5$  and the marker locus segregates independently of the QTL that influences phenotype. With free recombination between the M and Q loci, the marker class genotypic means should not differ because the marker genotype is not associated with the QTL genotype. You can see that this is true by solving each marker class mean for  $r = 0.5$  and finding that the expected mean for all marker genotypes is  $0.5d$ . Therefore, when M is not associated by linkage with a QTL locus there should be no difference between the marker class means  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$ . If, on the other hand, the average phenotypic values of those individuals with different marker genotypes are different, then this is evidence that there is a QTL linked to the marker. In other words, a difference between  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  is evidence that the marker locus is located inside a linkage block that also contains a QTL.

Given a difference between  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$ , it is then possible to estimate the values of  $a$  and  $d$  for the linked QTL. We will need to make a distinction between the true values of  $a$  and  $d$  and the estimated values of  $\hat{a}$  and  $\hat{d}$  since recombination between the marker and QTL will cause  $a \neq \hat{a}$  and  $d \neq \hat{d}$ . The difference between the two most extreme phenotypic values in the F2 population defines the genotypic scale from  $+a$  to  $-a$  as in Fig. 9.15. Therefore, the difference between the  $M_1/M_1$  and  $M_2/M_2$  marker class means is equal to  $2\hat{a}$ . In an equation this can be stated as

$$2\hat{a} = \bar{G}_{M_1M_1} - \bar{G}_{M_2M_2} \quad (9.33)$$

Substituting the definitions of  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  from Table 9.5 gives

$$\begin{aligned} 2\hat{a} &= [a(1 - 2r) + 2dr(1 - r)] \\ &\quad - [-a(1 - 2r) + 2dr(1 - r)] \end{aligned} \quad (9.34)$$

which then simplifies to

$$\hat{a} = a(1 - 2r) \quad (9.35)$$

The expected degree of dominance of the QTL is also obtained using the definition of  $d$  from the genotypic scale of measurement as the difference between the heterozygote genotypic value and the midpoint

between  $+a$  and  $-a$ . For QTL mapping data, the value of the midpoint is half the difference between the  $M_1M_1$  and  $M_2M_2$  marker class mean values and the heterozygote genotypic value is the  $M_1M_2$  marker class mean value. The estimated degree of dominance is therefore

$$\hat{d} = \bar{G}_{M_1M_2} - \left( \frac{\bar{G}_{M_1M_1} + \bar{G}_{M_2M_2}}{2} \right) \quad (9.36)$$

Using the definitions of  $\bar{G}_{M_1M_2}$ ,  $\bar{G}_{M_1M_1}$ , and  $\bar{G}_{M_2M_2}$  from Table 9.5 yields

$$\begin{aligned} \hat{d} &= d[(1 - r)^2 + r^2] \\ &\quad - \frac{a(1 - 2r) + 2dr(1 - r) - a(1 - 2r) + 2dr(1 - r)}{2} \end{aligned} \quad (9.37)$$

which after simplifying the rightmost term gives

$$\hat{d} = d[(1 - r)^2 + r^2] - 2dr(1 - r) \quad (9.38)$$

This equation is then simplified by factoring out  $d$ , multiplying through, and adding terms to obtain the expression

$$\hat{d} = d(1 - 2r)^2 \quad (9.39)$$

A numerical example will help to illustrate the marker class phenotypic mean values and how they result in estimates of  $\hat{a}$  and  $\hat{d}$ . Different breeds of dog exhibit a very wide range of body sizes. An allele of the insulin-like growth factor 1 gene (*IGF1*) has been shown to be frequent in small dog breeds (<9 kg) but to have a frequency of near zero in large dog breeds (>30 kg; Sutter et al. 2007). Therefore, the *IGF1* locus is likely to be a **major gene** (a QTL that explains a large amount of quantitative trait variation) for body size in dogs. Let's assume there are two *IGF1* alleles segregating within a single randomly mating population of dogs and that body size ranges between a minimum of 9 kg and a maximum of 30 kg (see Fig. 10.1).

Imagine that an F2 QTL mapping design was carried out along the lines of Fig. 9.15 using large (30 kg) and small (9 kg) dogs as the P1 individuals. Suppose that the marker class means were  $\bar{G}_{M_1M_1} = 20$  kg,  $\bar{G}_{M_1M_2} = 18$  kg, and  $\bar{G}_{M_2M_2} = 12$  kg. Also, make the unrealistic assumption for the moment that there is no recombination between the QTL and the marker locus ( $r = 0$ ). The observed value of  $2\hat{a}$  is the difference between  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$ , which is

$2\hat{a} = 8 \text{ kg}$  or  $\hat{a} = 4 \text{ kg}$ . To determine whether the QTL shows dominance, we first need to compute the midpoint value, which is  $12 + (20 - 12)/2 = 16 \text{ kg}$ . The  $M_1M_2$  marker class mean is  $18 \text{ kg}$  and is  $2 \text{ kg}$  above the midpoint. The degree of dominance is  $\hat{d}/\hat{a}$ , which in this case is  $2/4 = 0.5$ . Therefore, in this example the QTL allele that increases body mass is 50% dominant to the allele that decreases body mass.

The difference in body mass between the  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  marker class means is  $20 \text{ kg} - 12 \text{ kg} = 8 \text{ kg}$ . The total difference in body weight in these dogs is  $30 \text{ kg} - 9 \text{ kg} = 21 \text{ kg}$ . The QTL linked to this genetic marker therefore accounts for  $8 \text{ kg}$  of the total  $21 \text{ kg}$  difference in body weight, or  $8 \text{ kg}/21 \text{ kg} = 38\%$  of the total body mass difference between large and small dog breeds. The percentage of the total difference in the phenotypic value of the two P1 individuals is called the **effect size** of the QTL. In this hypothetical example, the QTL would be considered a major gene because its effect size is large. The QTL mapping results also tell us that there is partial dominance for the two QTL alleles that have been examined in this study based on the mean phenotypic value of individuals heterozygous for the marker locus.

It is important to notice that the expressions for  $\hat{a}$  and  $\hat{d}$  are both functions of the recombination rate between the QTL and the marker locus. This is because the perceived true effect of the QTL is confounded with the recombination rate in a single-marker QTL mapping analysis. The true additive effect of a QTL (or  $a$ ) based on  $\hat{a}$  and the recombination rate can be found by rearranging equation 9.35 to give

$$a = \frac{\hat{a}}{(1 - 2r)} \quad (9.40)$$

Similarly, the true dominance effect of a QTL (or  $d$ ) based on  $\hat{d}$  and the recombination rate can be found by rearranging equation 9.39 to give

$$d = \frac{\hat{d}}{(1 - 2r)^2} \quad (9.41)$$

So unless the recombination rate between the marker locus and QTL is zero, which it almost never is,  $a$  and  $d$  are actually larger than the estimates of  $\hat{a}$  and  $\hat{d}$  from a single-marker QTL mapping experiment.

To see that  $\hat{a}$  and  $\hat{d}$  are minimum estimates, let's reconsider the example above but assume a different recombination rate. The estimated phenotypic

effect of the QTL ( $\hat{a}$ ) is a constant. The observed difference in body mass between the  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  marker class means is  $8 \text{ kg}$ , yielding an estimate of  $2\hat{a} = 8$  or  $\hat{a} = 4 \text{ kg}$  by equation 9.33. If the recombination rate is  $r = 0$ , then by equation 9.40  $a = 4 \text{ kg}/(1 - 2(0)) = 4 \text{ kg}$ . However, if the recombination rate between the marker locus and the QTL is  $r = 0.25$  instead, then by equation 9.35,  $a = 4 \text{ kg}/(1 - 2(0.25)) = 8 \text{ kg}$ . For dominance, if  $r = 0$  then by equation 9.41,  $d = 2 \text{ kg}/(1 - 2(0))^2 = 2 \text{ kg}$ . However, if  $r = 0.25$  then by equation 9.41,  $d = 2 \text{ kg}/(1 - 2(0.25))^2 = 8 \text{ kg}$ . Therefore, the inferred true phenotypic effect of the QTL ( $a$ ) and its degree of dominance ( $d$ ) are both a function of the recombination rate between the marker locus and the QTL. The degree of dominance ( $d/a$ ) inferred with  $r = 0$  is  $2/4 = 0.5$ . In contrast, the degree of dominance inferred with  $r = 0.25$  is  $8/8 = 1.0$ .

In this example, a recombination rate of 0.25 rather than zero doubles both the perceived true effect

### Problem box 9.2 Compute the effect and dominance coefficient of a QTL

Sax (1923) was the first to carry out a QTL mapping analysis. He showed evidence for a QTL explaining continuous variation in the trait of seed weight of the common bean (*Phaseolus vulgaris*) based on differences in the phenotypic means of plants that differed in seed color. The P1 individuals had seed weights of 48 and 21 centigrams (cg) and were homozygous for different alleles ( $P$  and  $p$ ) for a codominant gene that affects seed color. Using seed color patterns to determine genetic marker classes, the mean seed weights in the F2 individuals were

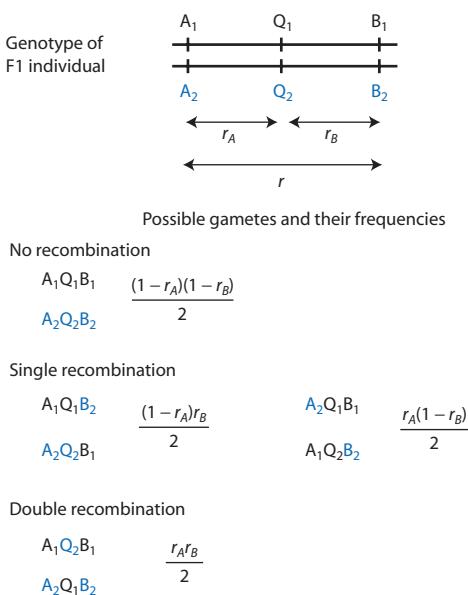
$$\begin{aligned}\bar{G}_{M_1M_1} &= 30.7 \text{ cg} & \bar{G}_{M_1M_2} &= 28.3 \text{ cg} \\ \bar{G}_{M_2M_2} &= 26.4 \text{ cg}\end{aligned}$$

Using the data collected by Sax, estimate  $\hat{a}$  and  $\hat{d}$  for the QTL for seed weight assuming that  $r = 0.2$  and  $r = 0$ . What is the effect of the QTL in terms of the percentage of phenotypic difference between the P1 individuals? Did Sax identify a major gene for seed weight?

of the QTL and the perceived degree of dominance. With a higher recombination rate, the QTL has a larger effect on the phenotype but not all of this effect is reflected in the marker class means since recombination breaks down the association between QTL genotypes and marker locus genotypes. As the randomizing effect of recombination between a QTL and a marker locus increases, the smaller the difference between  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  becomes for a QTL. With free recombination between the QTL and the marker locus,  $\bar{G}_{M_1M_1}$  and  $\bar{G}_{M_2M_2}$  are expected to be equal because there is no association between the QTL and the marker locus genotypes.

### QTL mapping with multiple marker loci

QTL mapping that utilizes numerous genetic marker loci is now routine. It is then possible to utilize all pairs of genetic markers for QTL mapping. The advantage of such **flanking-marker QTL analysis or interval mapping** is that the resulting estimates of  $\hat{a}$  and  $\hat{d}$  are not confounded with the recombination rate. Interval mapping is carried out with an F2 mating design like that shown in Fig. 9.15 to produce F1 individuals heterozygous at the two marker loci as well as the QTL. Figure 9.16 shows the arrange-



**Figure 9.16** Interval mapping utilizes two marker loci (A and B) that sit on either side of a QTL. An individual with the genotype shown can produce two types of gametes if there is no recombination, four type of gametes if there is a single recombination event, and two types of gametes from a double recombination event. The expected gamete frequencies are a function of the recombination rates.

ment of two marker loci that flank a QTL in an F1 individual along with the eight types of gametes (two coupling, four single recombinants, and two double recombinants) that can be produced by an F1 individual. These eight F1 gametes can be combined to make nine two-locus marker genotypes and 28 possible F2 genotypes. The recombination rate between the marker loci,  $r$ , can be estimated from the two-locus marker genotype frequencies in the F2 progeny. Assuming no interference (or that recombination rates  $r_A$  and  $r_B$  are independent), then  $r = r_A + r_B - 2r_Ar_B$ .

Using the expected gamete frequencies shown in Fig. 9.16, the mean phenotypic value of each marker genotype can be derived. Table 9.6 shows the derivation of the expected phenotypic value for the marker genotypes  $A_1A_1B_1B_1$  and  $A_1A_2B_1B_2$ . As for QTL mapping based on a single genetic marker, the portion of the mean genotypic value for the *entire* F2 population is referred to as  $\bar{G}_{A_1A_1B_1B_1}^{pop}$  and  $\bar{G}_{A_1A_2B_1B_2}^{pop}$ . These population mean genotypic values are the sum of the genotype-frequency-weighted genotypic values for each QTL genotype associated with a two-locus marker genotype. Since  $\frac{(1 - r)^2}{4}$  of all individuals in the F2 population are expected to have an  $A_1A_1B_1B_1$  marker genotype, we can multiply by a factor of  $\frac{4}{(1 - r)^2}$  to obtain the actual genotypic mean of that subset of individuals in the F2 population with  $A_1A_1B_1B_1$  marker genotypes or  $\bar{G}_{A_1A_1B_1B_1}$ . The expected frequency of the  $A_1A_2B_1B_2$  genotype is used to obtain the actual genotypic mean of F2 individuals with  $A_1A_2B_1B_2$  marker genotypes or  $\bar{G}_{A_1A_2B_1B_2}$ . The expected genotypic means of the remaining seven possible marker-class means as well as a regression method to estimate  $\hat{a}$  and  $\hat{d}$  are given in Haley and Knott (1992).

Like QTL mapping based on a single genetic marker, the phenotypic value of the  $A_1A_1B_1B_1$  marker genotype is a function of both the additive and dominance effects of the QTL. In contrast, the expected phenotypic value of the  $A_1A_2B_1B_2$  marker genotype is a function only of  $d$ . Therefore, the  $A_1A_2B_1B_2$  marker-class mean value provides an estimate of the dominance coefficient independent of  $a$ . Once the value of  $d$  is estimated, then other marker-class means can be used to estimate the value of  $a$ . As with single-marker QTL analysis, a statistically significant difference between marker-class mean values indicates that a QTL is present between a pair of genetic marker loci.

**Table 9.6** Derivation of the expected phenotypic value for two genetic marker genotypes when QTL mapping with pairs of marker loci that flank a QTL. There are a total of nine genetic marker genotypes possible with two genetic marker loci. Like QTL mapping based on a single genetic marker, the phenotypic value of the  $A_1A_1B_1B_1$  marker genotype is a function of both the additive and dominance effects of the QTL. In contrast, the expected phenotypic value of the  $A_1A_2B_1B_2$  marker genotype is a function only of  $d$ . Therefore, the  $A_1A_2B_1B_2$  marker class mean value provides an estimate of the dominance coefficient independent of  $a$ . The gametes and expected gamete frequencies are given in Fig. 9.16.

Marker genotype	Marker genotype frequency	F2 genotype frequency	F2 genotypic value	Frequency-weighted F2 genotypic value
$A_1A_1B_1B_1$	$\frac{(1-r)^2}{4}$	$\frac{A_1Q_1B_1}{\overline{A_1Q_1B_1}}$ $\frac{A_1Q_1B_1}{\overline{A_1Q_1B_1}}$ $\frac{A_1Q_1B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$	$\left(\frac{(1-r_A)(1-r_B)}{2}\right)^2$ $2\left(\frac{(1-r_A)(1-r_B)}{2}\right)\left(\frac{r_A'f_B}{2}\right)$ $\left(\frac{r_A'f_B}{2}\right)^2$	$+a$ $d$ $-a$
$\bar{G}_{A_1A_1B_1B_1}^{pop}$	$\frac{a(1-r_A)^2(1-r_B)^2}{4} + \frac{2dr_Ar_B(1-r_A)(1-r_B)}{4} + \frac{-ar_A^2r_B^2}{4}$	$\frac{A_1Q_1B_1}{\overline{A_1Q_1B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$	$\left(\frac{1-r_A)^2(1-r_B)}{2}\right)^2$ $\left(\frac{1-r_A)(1-r_B)}{2}\right)^2$ $\left(\frac{r_A(1-r_B)}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$	$\frac{a(1-r_A)^2(1-r_B)^2}{4}$ $\frac{d(1-r_A)^2r_B^2}{4}$ $\frac{dr_A^2(1-r_B)^2}{4}$ $\frac{dr_A^2r_B^2}{4}$
$A_1A_2B_1B_2$	$\frac{(1-r)^2+r^2}{4}$	$\frac{A_1Q_1B_1}{\overline{A_2Q_2B_2}}$ $\frac{A_1Q_1B_2}{\overline{A_2Q_2B_2}}$ $\frac{A_2Q_2B_1}{\overline{A_2Q_2B_1}}$ $\frac{A_2Q_1B_1}{\overline{A_1Q_2B_2}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$ $\frac{A_2Q_2B_2}{\overline{A_2Q_2B_2}}$ $\frac{A_1Q_1B_1}{\overline{A_1Q_1B_1}}$ $\frac{A_2Q_1B_2}{\overline{A_2Q_1B_2}}$ $\frac{A_2Q_2B_2}{\overline{A_1Q_2B_1}}$	$\left(\frac{(1-r)^2}{2}\right)^2$ $\left(\frac{(1-r_A)r_B}{2}\right)^2$ $\left(\frac{r_A(1-r_B)}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$ $\left(\frac{(1-r_A)(1-r_B)}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$ $\left(\frac{(1-r_A)(1-r_B)}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$	$d$ $d$ $d$ $d$ $+a$
$\bar{G}_{A_1A_2B_1B_2}^{pop}$	$\frac{d(1-r_A)^2(1-r_B)^2}{4} + \frac{d(1-r_A)^2r_B^2}{4} + \frac{dr_A^2(1-r_B)^2}{4} + \frac{2ar_Ar_B(1-r_A)(1-r_B)}{4} + \frac{-2ar_Ar_B(1-r_A)(1-r_B)}{4}$	$\frac{A_1Q_1B_1}{\overline{A_1Q_1B_1}}$ $\frac{A_2Q_1B_2}{\overline{A_2Q_1B_2}}$ $\frac{A_2Q_2B_2}{\overline{A_1Q_2B_2}}$ $\frac{A_1Q_2B_1}{\overline{A_1Q_2B_1}}$	$\left(\frac{1-r_A)(1-r_B)}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$ $\left(\frac{r_A'f_B}{2}\right)^2$	$\frac{-2ar_Ar_B(1-r_A)(1-r_B)}{4}$ $\frac{-2ar_Ar_B(1-r_A)(1-r_B)}{4}$
$\bar{G}_{A_1A_2B_1B_2}$	$\frac{4}{(1-r)^2+r^2} \left[ \frac{d(1-r_A)^2(1-r_B)^2}{4} + \frac{d(1-r_A)^2r_B^2}{4} + \frac{dr_A^2(1-r_B)^2}{4} + \frac{dr_A^2r_B^2}{4} + \frac{dr_A^2(1-r_A)^2r_B^2}{4} + \frac{dr_A^2(1-r_A)^2}{4} + \frac{dr_A^2(1-r_B)^2}{4} + \frac{dr_A^2r_B^2}{4} + \frac{dr_A^2(1-r_A)r_B^2}{4} + \frac{dr_A^2(1-r_A)^2}{4} \right] = d \frac{r_A'f_B + r_A(1-r_B)^2 + (1-r_A)r_B^2 + (1-r_A)(1-r_B)}{r^2 + (1-r)^2}$			

### Problem box 9.3

#### Derive the expected marker-class means for a backcross mating design

A commonly employed QTL mapping design is the backcross of an F1 individual to one of the P1 individuals shown in Fig. 9.15. One drawback of a backcross mating design for QTL mapping is that the resulting estimates of  $\hat{a}$  depend on the value of  $d$ . To see that this is the case, consider the backcross

$$\frac{A_1Q_1B_1}{A_2Q_2B_2} \times \frac{A_1Q_1B_1}{A_1Q_1B_1} \text{ and the marker-class means}$$

in the population of progeny. Derive the expected marker-class means for  $A_1A_1B_1B_1$  and  $A_1A_2B_1B_2$  marker genotypes (call these  $\bar{G}_{A_1A_1B_1B_1}^{BC}$  and  $\bar{G}_{A_1A_2B_1B_2}^{BC}$  respectively) and then compute the expected value of  $\hat{a} = \bar{G}_{A_1A_1B_1B_1}^{BC} - \bar{G}_{A_1A_2B_1B_2}^{BC}$ .

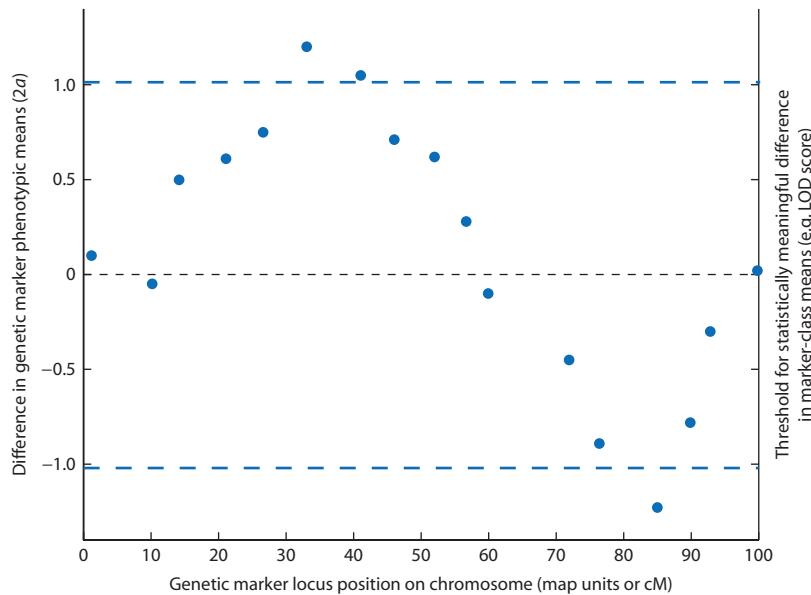
To work through this problem, first notice that the P1 individual  $\left(\frac{A_1Q_1B_1}{A_1Q_1B_1}\right)$  will produce only one type of gamete whereas an F1 individual  $\left(\frac{A_1Q_1B_1}{A_2Q_2B_2}\right)$  will produce eight types of gametes with the frequencies given in Fig. 9.16. This problem is made easier by the commonly invoked assumption that the marker loci are close enough on

the chromosome such that double recombination events are so rare they can be ignored. Therefore, the F1 parent gametes  $A_2Q_1B_2$  and  $A_1Q_2B_1$  can be left out of the marker-class means.

Start by constructing a table with two rows for the two F1 gametes produced by no recombination and four rows for the F1 gametes produced by a single recombination event. Then, following the model of Tables 9.5 and 9.6, fill in columns for the six categories of backcross progeny. The column headings are F1 parent gamete, expected F1 gamete frequency, backcross progeny genotype, backcross progeny genotypic value, frequency-weighted genotypic value, backcross progeny marker genotype expected frequency, and marker-class genotype mean. When adding the two terms that make up the  $A_1A_1B_1B_2$  or  $A_1A_2B_1B_1$  frequency-weighted genotypic values, any terms that contain  $r_A r_B$  can be crossed out because of the assumption that double recombination events are so rare that they can be ignored. The marker class means are obtained by dividing the frequency-weighted genotypic value by the marker genotype expected frequency.

A hypothetical example of data produced by interval QTL mapping in an F2 design is shown in Fig. 9.17. This example illustrates the difference in the value of the homozygous marker-class means (e.g. the difference between  $\bar{G}_{A_1A_1B_1B_1}$  and  $\bar{G}_{A_2A_2B_2B_2}$ ) for each of 17 genetic marker loci. The difference in marker-class means is given on the  $y$  axis while the position of each marker locus on a chromosome is given on the  $x$  axis. The marker-class mean differences are near zero for marker loci not in gametic disequilibrium with a QTL. Marker loci near QTLs show some difference in marker-class means, but the marker loci and the QTLs experienced frequent recombination and so are not strongly associated. As marker loci closer to the QTLs are considered, the amount of gametic disequilibrium between a marker locus and a QTL increases, producing a greater difference in marker-class means. In this hypothetical example, the genetic markers at 33 and 85 map units lie closest to QTLs,

as indicated by peak values for marker-class mean differences. The marker-class means at these marker loci differ by over one phenotypic standard deviation, indicating a statistically meaningful difference given sampling error and multiple statistical tests. (The threshold for statistical significance of marker-class mean differences is often judged by a log of odds or LOD score and differs depending on the details of each QTL study; see Van Ooijen (1999).) The QTL near 33 map units increases the mean value while the QTL near 85 map units decreases the mean value. These two hypothetical QTLs have opposite effects on the trait, a situation sometimes called dispersion. Dispersion can lead to downwardly biased estimates of QTL effects because each marker locus is associated with two QTLs with opposite phenotypic effects. The perceived marker-class mean difference is therefore the net phenotypic effect caused by association with two QTLs.



**Figure 9.17** The difference in phenotypic mean values for each of 17 genetic marker loci versus the position of each marker locus on a chromosome. In this hypothetical example, there are two genetic markers that lie closest to QTLs indicated by two peak values for marker-class mean differences. The phenotypes of the homozygote classes for the marker loci at 33, 41, and 85 map units (or centimorgans, cM) differ by over one phenotypic standard deviation. The QTL near 33 cM increases the mean value while the QTL near 85 cM decreases the mean value. Marker-class mean value differences greater than  $\pm 1$  phenotypic standard deviation (blue dashed lines) are considered statistically meaningful in this example. Marker-class mean differences smaller than one standard deviation could be different due to chance alone. The close proximity of these two QTLs with opposite effects on the trait would lead to reduced estimates of QTL effects at all marker loci. This occurs because each marker locus is partly linked to a QTL of both positive and negative effect, so the perceived marker class mean difference is the net effect of the two QTLs. Genetic marker locus positions are established by observed recombination rates. One map unit or centimorgan distance along a chromosome is equal to a 1% recombination rate. LOD score, log of odds score.

A large number of QTL mapping breeding designs and estimation methods exist. Some methods are related to interval mapping but utilize more than just pairs of marker loci (**composite interval mapping**) or utilize all of the linked markers on individual chromosomes (**multipoint mapping**). Another approach is to test for associations between marker genotypes and phenotypic means in only those individuals that show extreme phenotypic differences in an F2 population such as the individuals at the upper and lower tails of the phenotypic distribution. In such **bulked segregant analyses** the marker loci near QTLs are expected to be in gametic disequilibrium because of linkage. Therefore, the individuals in the lower tail of the phenotypic distribution would have one marker genotype and the individuals in the upper tail of the phenotypic distribution would have another marker genotype if a marker locus were very tightly linked to a QTL. In contrast, a marker locus independent of any QTL would show all genotypes at equal frequencies in the individuals that represented the upper and lower tails of the phenotypic distribution.

### Limitations of QTL mapping studies

QTL mapping results are highly context-sensitive, just like heritability estimates. The number and effects of QTLs identified in one population may not be representative of QTL effects in another population of the same species. Mapping only identifies QTLs that are segregating in the population at the time of mapping. Further, QTL mapping by the F2 design can only detect and estimate the phenotypic effects of two alleles at a QTL. The two QTL alleles detected are those fixed when inbreeding lines to form P1 in Fig. 9.15. In reality, there may be more than two alleles segregating in a population for any QTLs and detecting these requires screening replicate P1 crosses. Alleles at QTLs that are fixed or lost in the original population or become fixed or lost by genetic drift during the formation of inbred lines cannot be identified. Likewise, the estimates of QTL effect sizes are always relative to the other QTL loci segregating in a population. For example, QTL X could explain 10% of the phenotypic difference between marker-class means in one population whereas QTL Y has

even bigger effect but happens to be fixed for one allele. If QTL Y is segregating instead, then QTL X will have a smaller perceived effect on the phenotype in a mapping study. This occurs because when two alleles are segregating at QTL Y, there will be a greater difference in phenotype between P1 individuals than when QTL Y is fixed.

Several types of statistical power limitations impact QTL mapping results. Actual QTLs of small effect cannot be identified easily because it is difficult to show that a small difference between marker-class means is statistically meaningful given inherent environmental variation, experimental measurement error, and corrections required when carrying out a very large number of statistical tests. The effect sizes of QTLs that are identified as statistically meaningful can be substantially inflated since the QTLs with small effects are not identified. This so-called **Beavis effect** is pronounced if the number of progeny in a mapping study is about 100 and modest with about 500 progeny (Beavis 1994; Xu 2003).

The number of QTLs identified in mapping studies is likely to underestimate the true number of QTLs that cause variation in a trait (Otto & Jones 2000). Under conditions similar to many actual QTL mapping studies, no more than about a dozen QTLs will be identified as statistically meaningful (Hyne & Kearsey 1995). In addition, two or more QTLs that are adjacent in the genome may appear as a single QTL. If the effects of two linked QTLs are in the same direction, the perceived single QTL will have an inflated effect. On the other hand, if the two linked

QTLs have contrasting (or antagonistic) effects on the trait, then a single perceived QTL will be detected that has a downwardly biased effect size that is therefore less likely to be detected as statistically meaningful. The number and spacing of genetic markers also influences the perception of QTL numbers and effect sizes, since widely spaced markers (relative to the recombination rate) may miss QTLs or aggregate the effects of multiple QTLs. The term **quantitative trait region** or **QTR** is sometimes used to describe an association between a marker locus and a marker-class mean difference since multiple linked QTLs may exist within the genome interval mapped. Further fine-scale mapping with genetic markers spaced at smaller intervals along the chromosome in the specific chromosomal regions around QTR can be used to identify true single QTLs (e.g. Kroymann & Mitchell-Olds 2005).

#### **Biological significance of QTL mapping**

QTL mapping is now carried out routinely in model and domesticated species, facilitated by the large numbers of molecular markers and high-throughput genotyping techniques as well as genome sequencing and genetic linkage mapping projects. QTL mapping can also be carried out, albeit usually with more difficulty and less resolution, in some non-domesticated species using crosses within and between species (reviewed by Slate 2005). Table 9.7 shows some examples of the number of QTLs identified for various phenotypes in a range of species. In Table 9.7 the results range from one or a few QTLs with large

**Table 9.7** Examples of QTLs identified by mapping with genetic marker loci.

Organism	Phenotype	Number of marker loci	Number of QTLs	Reference
<i>Arabidopsis thaliana</i>	Days to first flower	65	7	Kearsey et al. 2003
	Number of buds at flowering		28	
	Rosette size at 21 days		4	
	Rosette size at flowering		10	
Dogs	Body size	116	1	Sutter et al. 2007
<i>Drosophila santomea</i> × <i>D. yakuba</i>	Prezygotic reproductive isolation	32	6	Moehring et al. 2006
Humans	Taste sensitivity to PTC	50	1	Kim et al. 2003
	Stature	> 253	3	Perola et al. 2007
Louisiana irises	Flowering time	> 414	17	Martin et al. 2007
Stickleback fish	Bony plates	160	4	Colosimo et al. 2004
<i>Zea mays</i>	Kernel oil concentration	488	> 50	Laurie et al. 2004

PTC, phenylthiocarbamide.

effects on phenotypic variation (dog body size, human stature) to a large number of QTLs with individual small effects on phenotypic variation (kernel oil content). This mirrors the overall trend in QTL mapping results, where the number of QTLs and their effect sizes are strongly dependent on the species, population, and phenotype studied. Whereas part of this diversity in QTL mapping results is likely caused by methodological and statistical power differences among studies, some of it likely reflects actual variability in the number and effects of QTLs that cause phenotypic variation.

QTL mapping studies not only identify effect sizes, but can also quantify the degree of dominance and epistasis for QTLs. QTLs have been observed to have dominance that ranges from zero (additive gene action), through all degrees of partial dominance to complete dominance, to cases of overdominance. One classic example is the wide range of the  $d/a$  ratio ( $a$  is the estimated additive effect of an allele and  $d$  is the estimated heterozygote value), spanning  $-2.0$  to  $+2.0$ , observed for 74 QTLs detected for 11 phenotypes in tomato (de Vincente & Tanksley 1993). Based on these results, dominance frequently contributes to the genotypic variation of quantitative traits. Evidence for interaction variance caused by epistasis has been equivocal. Interactions between or among loci did not often explain much of the observed genotypic variation in early QTL mapping studies. However, it is generally more difficult to detect epistasis for QTLs due to statistical and sample size limitations (see Carlberg & Haley 2004). Recently, more effort has been directed toward testing for epistasis in QTL studies and numerous empirical studies have identified statistically meaningful interactions between two or more QTLs as often as additive effects of QTLs (reviewed by Malmberg & Mauricio 2005).

The genetic architecture of quantitative traits – the effect size and gene action (additivity, dominance, or epistasis) of QTL underlying a trait – plays a crucial role in how quantitative traits will respond to natural selection. As an illustration, imagine that dog body mass is under natural selection for larger size such that  $s = 0.3$  with  $h^2 = 1.0$ . The breeder's equation  $R = h^2 s$  predicts response to selection without reference to the genetic architecture of the quantitative trait. In this example,  $R = (1.0)(0.3) = 0.3$ , or a predicted increase of 0.3 standard deviations per generation. Now imagine that the additive genetic variation in body mass is caused by a number of QTLs. The strength of selection on the trait is divided across the independent loci that cause the additive

genetic variation in the trait. The pressure of natural selection on one QTL is then only as large as its role in causing additive genetic variance in the phenotype. Said another way, the selection experienced by one QTL is proportional to its effect size. The breeder's equation can be modified for a single QTL by adjusting the selection differential for the proportion of additive genetic variation explained by a QTL according to

$$R = h^2(s(\text{QTL effect size})) \quad (9.42)$$

To see how this modified breeder's equation works, let's return to the example of the QTL that explains 38% of the total body mass difference between large and small dog breeds. That one QTL experiences 38% of the selection pressure that is exerted on the entire phenotype because it causes 38% of the additive genetic variation. Response to selection will be accomplished by relatively large changes in the allele frequencies at that one QTL with a 38% effect. In fact, the selection differential on that one QTL is  $s = (0.3)(0.38) = 0.114$ . This also leads to a predicted response to selection of 0.144 standard deviations for that one QTL alone.

The predictions of the nearly neutral theory (see Chapter 8) show the consequences of this division of the selection differential among QTL in proportion to their effect size. In finite populations, the balance between genetic drift and natural selection can be predicted with the quantity  $4N_e s$  where  $s$  is the selection differential. When  $4N_e s \gg 1$  then selection is the primary determinant of allele frequency, when  $4N_e s \ll 1$  then genetic drift is the main process influencing allele frequency, and when  $4N_e s$  is on the order of one then both selection and genetic drift determine allele frequency. For a QTL of 38% effect experiencing a selection differential of  $s = 0.114$ , the response to selection will depend on  $N_e$ . In the context of an effective population size greater than about 25, the frequencies of the alleles at a QTL with a 38% effect would be expected to be dictated exclusively by natural selection. Only if the effective population size were less than 10 would we expect genetic drift to exclusively dictate the fate of allele frequencies at the QTL.

Under more realistic circumstances, many QTLs will have effect sizes much smaller than 38% and traits will most commonly have heritabilities less than 1.0. To take another example, imagine a QTL with a 2% effect for a trait with  $h^2 = 0.3$  that is experiencing an identical selection differential of  $s = 0.3$ .

The effective response to selection for this QTL of small effect is  $R = (0.3)((0.3)(0.02)) = 0.0018$ , or a predicted increase of 0.0018 standard deviations per generation. This QTL of small effect experiences a very weak pressure from natural selection precisely because it is a weak cause of additive genetic variation in a trait that has a low heritability. According to the  $4N_e s$  rule, this QTL would have to experience selection in the context of an effective population size greater than about 2000 for natural selection alone to dictate allele frequencies over time.

### Interact box 9.5 Effect sizes and response to selection at QTLs

Use the **QTL** module in PopGene.S<sup>2</sup> to simulate response to selection for a quantitative trait that has QTL with variable effect sizes. The key parameter to vary is the percentage of genotypic variation explained by each QTL. Select the option for **Some loci have large effects and the rest have equal small effects**, set Number of large effect loci = 3, and  $V_G$  explained by each large effect locus = 0.20. Set the other simulation parameters at Natural selection phenotypic value truncation point = 0.5 (only individuals in the upper 50% of phenotypic value reproduce),  $N$  = 250, mutation rate = 0.0, generations = 100,  $V_E$  = 0.1, and Maximum genotypic value = 10.0.

Use the modified breeder's equation given in equation 9.42 to compute the net response to selection for the QTL with large and small effects in this simulation. When the selection truncation point is 0.50 in the simulation, the selection differential is  $s \approx 0.399$  (see the text web page for more details). The heritability is also  $h^2 \geq 0.5$ . Using these parameter values, what would you predict about the patterns of allele frequency change for QTL of large and small effect sizes?

Try out different values for the Natural selection phenotypic value truncation point or  $N$  and predict the consequences using equation 9.42. Then run the simulation to test your predictions.

These observations about how the effective population size influences response to selection on individual QTLs shed light on an implicit assumption of the infinitesimal model of the genetic basis of quantitative traits. It is only in the context of large effective population sizes that the allele frequencies of QTLs with small effect sizes will respond to natural selection. In a finite population, as the number of QTLs grows large and the effect size approaches zero, the net response to selection will shrink and each locus will be subject to genetic drift rather than natural selection. Therefore, the infinitesimal model must also assume that the effective population size is inversely related to the QTL effect size. Drift-selection balance for QTLs in finite populations also serves to explain the simulation results in Fig. 9.12 that relate to long-term response to selection. In Fig. 9.12b there are many QTLs of small effect that clearly remain segregating for a longer period of time than the QTLs of large effect in Fig. 9.12a. The reason why the QTLs of small effect remain segregating longer is that they experience a relatively weak net selection pressure. The allele frequencies at the loci in Fig. 9.12b are clearly spreading out toward fixation and loss as we would expect of many replicate neutrally evolving loci (see Chapter 3), although perhaps the probability of fixation is greater than it would be under pure genetic drift.

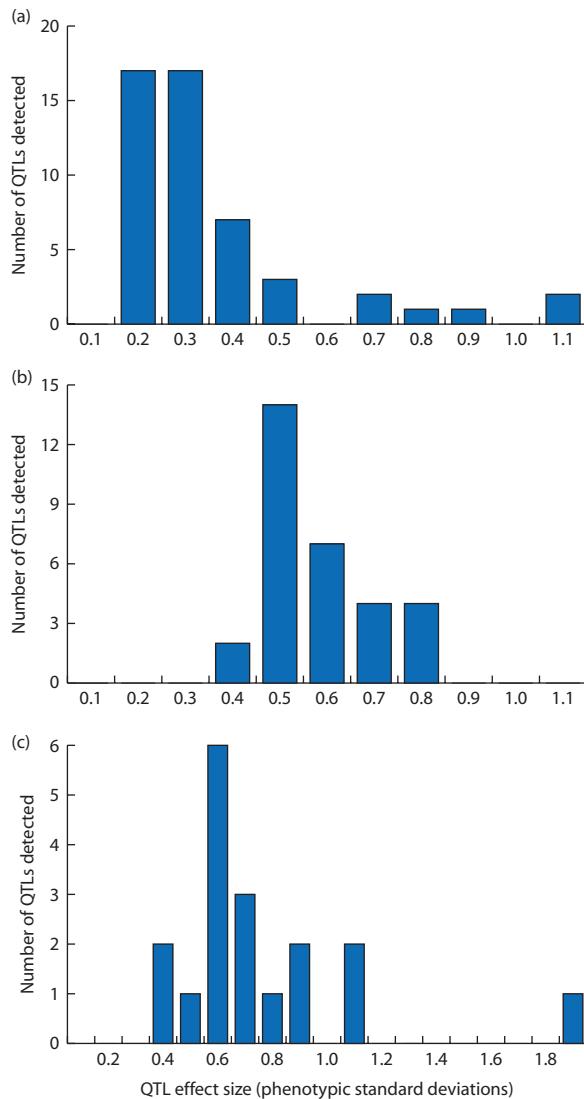
The expected behavior of QTLs under the processes of genetic drift and natural selection has lead to a possible test for the action of natural selection on quantitative traits. Under the influence of natural selection alone, QTL alleles that fix in a population should all cause the trait value to change in the same direction (Orr 1998b; Anderson & Slatkin 2003). For example, if natural selection is acting to increase the trait mean then only QTL alleles that increase the trait value should be fixed in the population. Alternatively, under pure genetic drift QTL alleles that both increase and decrease the trait value should fix with equal frequency. Based on this logic, Rieseberg et al. (2002) used QTL effects estimated from 42 traits with greater than six QTLs to test for the action of directional selection. They concluded that alleles at QTLs are very often under the influence of natural selection and that life-history traits experience stronger natural selection than morphological traits. Interpreting this test for natural selection on QTLs is problematic, however, since quantitative traits in general are not sampled at random for QTL mapping studies. Rather, quantitative traits in domesticated organisms that have been under

long-term selection are most likely to be subject to QTL mapping since the effort to map QTLs is justified by the potential for improvements in selective breeding. While Rieseberg et al. (2002) showed that QTL alleles do tend to cause the trait value to change in the same direction as expected under selection, it is not clear whether the sample of taxa and phenotypes used for the test is representative of taxa and phenotypes in general.

Other results suggest that QTLs experience a combination of both genetic drift and natural selection. The QTLs of small effect identified from the Illinois Long-Term Selection study clearly show the effects of both genetic drift and natural selection. Among the QTLs identified for kernel oil concentration, about 20% show effects that are opposite to the direction of natural selection, such as QTLs that reduce oil concentration segregating in the lines selected for high oil concentration (Laurie et al. 2004). Since selection is acting against these QTLs with opposite effects, they must have escaped loss due to strong genetic drift in the context of an effective population size of about 10. Results similar to kernel oil concentration have been observed for tomato QTLs.

The results of QTL mapping studies can also be used to estimate the distribution of effect sizes for numerous QTLs. Figure 9.18 shows distributions of QTL effect sizes estimated for numerous phenotypes in dairy cows, pigs, and *D. melanogaster*. The distributions show the difference in phenotypic values for alternative homozygous genotypes (or the equivalent of  $2a$ ) at QTLs in units of phenotypic standard deviations. QTLs with smaller effect sizes explain a smaller proportion of the additive variance in a quantitative trait. QTLs at the right-hand edge of the  $x$  axis have a large effect on the genotypic variance.

These distributions of QTL effect sizes can be related to the models of mutation considered in Chapter 5. In particular, Fisher's geometric model of mutation (refer to Fig. 5.5) predicts that mutations of smaller phenotypic effect are more likely to improve fitness. Extending this model to QTLs, we can now think of QTLs with larger and smaller effects on genotypic variation (refer to Fig. 5.6). Kimura's version of Fisher's geometric model of mutation predicts that the genetic drift–natural selection balance will lead to fixation of alleles with the same effect on trait value (all either increasing or decreasing the trait mean) for those QTL of medium to large effect. Orr's (1998a) version of the geometric model considers a series of QTL mutations over time (but only one mutation segregates at a time) that take the popula-



**Figure 9.18** Distributions of estimated QTL effect sizes from several animal species. The distributions show the difference in phenotypic values for alternative homozygous genotypes (or the equivalent of  $2a$ ) at a QTL in units of phenotypic standard deviations. QTLs with larger effect sizes explain a large proportion of the additive variance in a quantitative trait. The distribution in (a) represents QTLs identified from multiple phenotypes in dairy cattle while the distribution in (b) represents QTLs identified in multiple phenotypes in pigs. The distribution in (c) shows QTLs located on the third chromosome that influence sternopleural bristle number in *D. melanogaster*. Data in (a) and (b) from Hayes and Goddard (2001); data in (c) from Shrimpton and Robertson (1988).

tion closer to the optimum. Orr's model differs from Fisher's because it predicts that QTLs with larger effects can be fixed by selection. In addition, Orr's model suggests that the distribution of QTL effect sizes should have an increasing number of QTLs with small effects so that the number of QTLs increases rapidly as the QTL effect size decreases. Looking at the data provided

in Fig. 9.18, it is difficult to test the predictions of Fisher's or Orr's models. The QTLs from dairy cows in Fig. 9.18a have the largest number of QTLs with small effect. In contrast, the effect distribution in Figs 9.18b and 9.18c have more QTLs in the middle effect sizes. Because Fisher's and Orr's models make abstract rather than specific predictions they are difficult to test in practice. In addition, the potential biases and uncertainties in QTL data themselves leave us uncertain whether the distribution of effect sizes has been estimated accurately.

### Chapter 9 review

- Variation in quantitative trait values among individuals within a population ( $V_p$ ) has both genetic and environmental causes. The genetic causes are due to genotypic variance ( $V_G$ ) that can be partitioned into the distinct components of additive ( $V_A$ ), dominance ( $V_D$ ), and epistasis ( $V_I$ ) variance.
- Phenotypic variation can be caused by genotype-by-environment interaction ( $V_{G\times E}$ ) where genotypes express heterogeneous phenotypic values in response to different environmental conditions.
- The additive component ( $V_A$ ) of the genotypic variance ( $V_G$ ) is caused by the sum of the phenotypic effects of alleles when they are assembled into genotypes. Phenotypic effects of alleles cause the resemblance of parents and offspring as well as the resemblance among relatives.
- Dominance ( $V_D$ ) and epistasis ( $V_I$ ) components of  $V_G$  are caused by the effects of genotypes.  $V_D$  and  $V_I$  do not contribute to phenotypic resemblance between parents and offspring because particulate inheritance breaks up genotypes (additive by additive epistasis is an exception).
- The proportion of the total genotypic variance ( $V_G$ ) due to the additive effects of alleles is measured by the narrow-sense heritability  $h^2 = V_A/V_p$ . Parent–offspring regression is one method to estimate  $h^2$ .
- Response to selection over one generation depends on the force of natural selection and the heritability and is predicted by the breeder's equation  $R = h^2 s$ .
- Because traits show both genetic and phenotypic correlations, response to selection on one trait may change the mean of other correlated traits or be constrained by correlations with other traits.
- Long-term response to natural selection depends on the number of loci that underlie a quantitative trait. Linear response to selection over many generations is expected when many loci with small effects underlie a trait, consistent with the infinitesimal model. In contrast, selection plateaus are consistent with fewer loci of larger effect since selection causes fixation and loss at these loci. Depending on the rate, mutation may replace variation lost due to fixation and loss caused by response to selection.
- The neutral evolution of genotypic variance depends on the balance of genetic variation lost by genetic drift and mutation that replaces variation.
- The individual loci that cause variation in quantitative traits, or QTLs, can be identified by taking advantage of gametic disequilibrium between QTLs and genetic marker loci.
- In an F2 mating design, comparing the phenotypic means of F2 individuals bearing different marker genotypes identifies marker loci near QTLs.
- QTL mapping can only identify alleles that are segregating in the individuals used to found the mapped populations. QTL mapping tends to underestimate the true number of QTLs and overestimate the true effects of QTLs.
- QTL mapping quantifies the genetic architecture of quantitative traits by estimating the number of loci that cause quantitative trait variation, the distribution of QTL phenotypic effects, and the physical organization of QTLs on chromosomes.

### Further reading

For a review of the progress over the last one hundred years on fundamental questions in quantitative genetics and the evolution of quantitative traits see:

Roff DA. 2007. A centennial celebration for quantitative genetics. *Evolution* 61: 1017–32.

More detail on statistical estimators as well as experimental designs used to estimate heritability and map QTLs can be found in:

Lynch M and Walsh B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

The response to selection predicted by the breeder's equation is actually an approximation that neglects four other types of parent–offspring phenotype relationship, as explained in:

Heywood JS. 2005. An exact form of the breeder's equation for the evolution of a quantitative trait under natural selection. *Evolution* 59: 2287–98.

The concepts involved in predicting levels of quantitative genetic variation and the evolution of quantitative traits are reviewed by:

Barton NH and Keightley PD. 2002. Understanding quantitative genetic variation. *Nature Reviews Genetics* 3: 11–21.

Quantitative trait mapping methods and empirical observations from QTL mapping are reviewed by:

Mackay TFC. 2001. The genetic architecture of quantitative traits. *Annual Reviews of Genetics* 35: 303–39.

For a review of how epistasis can be detected in QTL mapping studies, see:

Carlberg Ö and Haley CS. 2004. Opinion: epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* 5: 618–25.

For a critical review of the evolutionary inferences about quantitative traits that can be drawn from QTL mapping data, see:

Erickson DL, Fenster CB, Stenoien HK, and Price D. 2004. Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology* 13: 2505–22.

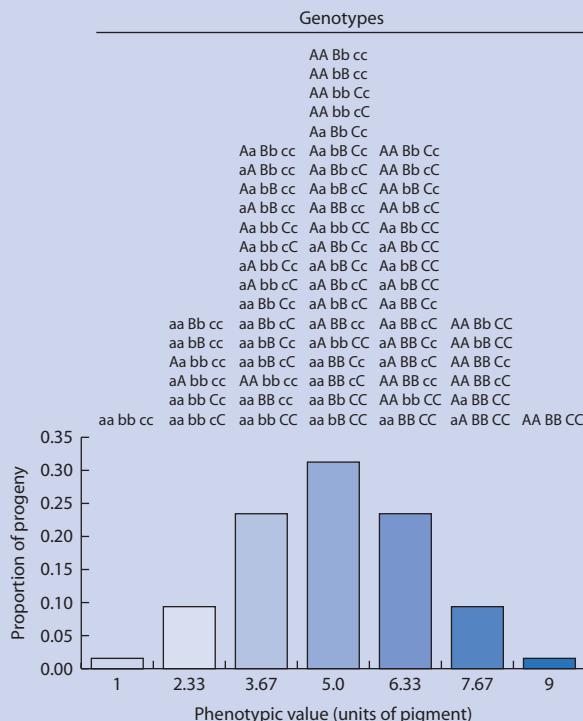
## Problem box answers

### Problem box 9.1 answer

For three diallelic loci each with codominance, the expected genotype frequencies in the population can be obtained by expanding

$$\left( \frac{1}{4}AA + \frac{1}{2}Aa + \frac{1}{4}aa \right) \left( \frac{1}{4}BB + \frac{1}{2}Bb + \frac{1}{4}bb \right) \left( \frac{1}{4}CC + \frac{1}{2}Cc + \frac{1}{4}cc \right)$$

or by constructing an  $8 \times 8$  Punnett square (the gametes are ABC, AbC, ABc, Abc, aBC, abC, aBc, and abc). The phenotypes are in an expected ratio of 1 : 6 : 15 : 20 : 15 : 6 : 1 for phenotypes of 1,  $\frac{2}{6}$ ,  $\frac{3}{6}$ , 5,  $\frac{2}{6}$ ,  $\frac{7}{6}$ ,  $\frac{4}{6}$ , and 9 as shown in Fig. 9.19.



**Figure 9.19** The genotypes and distribution of phenotypic values for a trait caused by three codominant diallelic loci in a randomly mating population where all alleles at each locus have a frequency of  $1/2$ .

**Table 9.8** Derivation of the expected phenotypic values for marker genotypes used to estimate  $\hat{\sigma}$  in a P1  $\times$  F1 backcross mating design when QTL mapping with pairs of marker loci that flank a QTL.

F1 parent	F1 gamete frequency	BC progeny genotype	BC progeny genotypic value	Frequency-weighted genotypic value	BC progeny marker genotype	Marker genotype frequency	Marker class mean ( $\bar{G}_{A_x B_x}^{BC}$ )
$A_1 Q_1 B_1$	$\frac{(1-r)}{2}$	$\overline{\overline{A_1 Q_1 B_1}}$	$+a$	$a \frac{(1-r)}{2}$	$A_1 A_1 B_1 B_1$	$\frac{(1-r)}{2}$	$a$
$A_1 Q_1 B_2$	$\frac{(1-r_A)r_B}{2}$	$\overline{\overline{A_1 Q_1 B_1}} \quad \overline{\overline{A_1 Q_1 B_2}}$	$+a$	$\frac{ar_B + ar_A r_B + dr_A - dr_A r_B}{2} \approx \frac{ar_B + dr_A}{2}$	$A_1 A_1 B_1 B_2$	$\frac{r}{2}$	$\frac{ar_B + dr_A}{r}$
$A_1 Q_2 B_2$	$\frac{r_A(1-r_B)}{2}$	$\overline{\overline{A_1 Q_1 B_1}} \quad \overline{\overline{A_1 Q_2 B_2}}$	$d$	$\frac{ar_A + ar_A r_B + dr_B - dr_A r_B}{2} \approx \frac{ar_A + dr_B}{2}$	$A_1 A_2 B_1 B_1$	$\frac{r}{2}$	$\frac{ar_A + dr_B}{r}$
$A_2 Q_2 B_1$	$\frac{(1-r_A)r_B}{2}$	$\overline{\overline{A_1 Q_1 B_1}} \quad \overline{\overline{A_2 Q_2 B_1}}$	$d$	$\frac{dr_A + ar_A r_B + dr_B - dr_A r_B}{2} \approx \frac{dr_B}{2}$	$A_1 A_2 B_1 B_2$	$\frac{(1-r)}{2}$	$d$
$A_2 Q_1 B_1$	$\frac{r_A(1-r_B)}{2}$	$\overline{\overline{A_1 Q_1 B_1}} \quad \overline{\overline{A_2 Q_1 B_1}}$	$+a$				
$A_2 Q_2 B_2$	$\frac{(1-r)}{2}$	$\overline{\overline{A_2 Q_1 B_1}} \quad \overline{\overline{A_2 Q_2 B_2}}$	$d$	$d \frac{(1-r)}{2}$			

**Problem box 9.2 answer**

The difference between the homozygote marker class means is  $30.7 - 26.4 = 4.3$  cg so that  $\hat{a} = 2.15$  cg. Assuming that  $r = 0$ ,  $a = 2.15$  cg. Assuming instead that  $r = 0.2$ , then  $a = 2.15 / (1 - 2(0.2)) = 3.58$  cg. The midpoint value is  $26.4 + (30.7 - 26.4)/2 = 28.55$  cg. The  $M_1M_2$  marker class mean is less than the midpoint so  $d = 28.3 - 28.55 = -0.25$ . With  $r = 0.2$ ,  $d = -0.25 / (1 - 2(0.2))^2 = -0.69$ . The coefficient of dominance is  $-0.25 / 2.15 = -0.12$  if  $r = 0$  and  $-0.69 / 3.58 = -0.19$  if  $r = 0.2$ . The difference in seed weight between the  $Q_1Q_1$  and  $Q_2Q_2$  genotypes accounts for  $4.3$  cg /  $27$  cg or  $16\%$  of the total seed weight difference between the two parental lines. The phenotypic effect is large enough to be considered a major gene.

**Problem box 9.3 answer**

The backcross design results in  $\hat{a} = a - d$  and therefore is a biased estimate of the additive effect of a QTL unless there is no dominance ( $d = 0$ ). See Table 9.8.

# CHAPTER 10

## The Mendelian basis of quantitative trait variation

### 10.1 The connection between particulate inheritance and quantitative trait variation

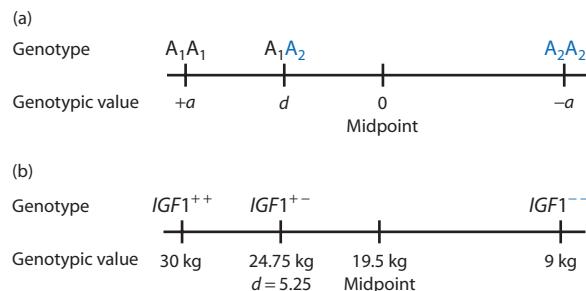
- Establishing a scale for genotypic values.
- Defining  $+a$ ,  $d$ , and  $-a$ .
- Phenotypic values as population averages.
- Why we can neglect environmental variation ( $V_E$ ).

This chapter will develop the concepts needed to understand the detailed connections between quantitative trait variation and particulate inheritance. Although the components of quantitative trait variation were described in Chapter 9 as population-level phenomena, the variance is ultimately caused by different alleles and genotypes possessed by individuals. The goal of this chapter is to show how additive and dominance components of variation in quantitative traits ( $V_A$  and  $V_D$ ) are caused by allele and genotype frequencies in a population as well as by the nature of gene action when alleles are combined into genotypes. To accomplish this goal, we will work with a hypothetical quantitative trait that is the product of a single locus with two alleles throughout the chapter. While use of a single diallelic locus as an example does not approximate the multilocus basis of quantitative traits and the multiallelic state of many loci, it greatly simplifies the resulting mathematical expressions while still illustrating key biological concepts. Bear in mind that the epistatic component of genetic variance ( $V_I$ ) arises due to interaction between two or more loci and therefore cannot be represented in a single-locus model. Therefore, the use of a single locus is an implicit assumption that  $V_I$  is zero. This chapter will start by constructing expressions that predict the phenotypic mean value of some type of population. The population of interest will initially be all individuals and later be only those individuals with genotypes that contain a certain allele. The population mean value will also be divided into components due to the additive action of alleles and the

dominance effect of genotypes. Ultimately, these mean values will be used to build expressions for the variance in phenotypic values expected in a population, specifically the additive and dominance components of genetic variation ( $V_A$  and  $V_D$ ). The chapter will conclude with a section on the expected phenotypic resemblance among populations of relatives based on the probabilities that related individuals share alleles or genotypes in common.

#### Scale of genotypic values

The hypothetical single locus used throughout this section will have two alleles,  $A_1$  and  $A_2$ . By convention, the  $A_1$  allele contributes to larger phenotypic values and the  $A_2$  allele to smaller phenotypic values. A conceptual scale of measurement is used to represent the genotypic values of each of the three genotypes (Fig. 10.1a). On this scale, the genotypic value of the  $A_1A_1$  genotype is  $+a$ , whereas the genotypic value of the  $A_2A_2$  genotype is  $-a$ . The genotypic



**Figure 10.1** The genotypic scale of measurement for quantitative traits. The variables  $+a$  and  $-a$  define genotypic values of the  $A_1A_1$  and  $A_2A_2$  homozygotes, respectively. The genotypic value of the heterozygote is defined as  $d$  and is measured relative to the midpoint ( $d = A_1A_2$  genotypic value minus the midpoint). The degree of dominance is expressed as the ratio  $d/a$ . These genotypic measurements are illustrated for the  $IGF1$  gene in dogs, which is a major gene contributing to body-size differences. Since the degree of dominance is unknown for  $IGF1$  genotypes, the genotypic value of the heterozygote is hypothetical.

value of the  $A_1A_2$  genotype is  $d$  and it is always measured relative to the midpoint as the  $A_1A_2$  genotypic value minus the midpoint value. The midpoint on this scale of genotypic measurement, or the point exactly mid-way between the genotypic values of the homozygotes, is defined to be zero. Notice that when the genotypic value of the heterozygote is at the midpoint,  $d = 0$ . This corresponds to the heterozygote having a genotypic value that is exactly the average of the two homozygotes as expected under additive gene action. In contrast, when the genotypic value of the heterozygote is not at the midpoint then there is some type of dominance. When  $d = +a$  or  $d = -a$  there is complete dominance. If the genotypic value of the heterozygote is outside the range of the homozygotes, then there is overdominance ( $d > +a$ ) or underdominance ( $d < -a$ ). The degree of dominance can be expressed as the ratio of  $d/a$ .

Throughout this section, body size in domestic dogs will be used as an example to illustrate concepts and computations as it was in Chapter 9. Comparing large ( $>30$  kg) and small ( $<9$  kg) dog breeds, Sutter et al. (2007) showed that an allele of the insulin-like growth factor 1 gene (*IGF1*) was common to all small breeds and had a frequency of near zero in large breeds. Thus, the *IGF1* locus is likely to be a **major gene** (a single locus that explains a large amount of quantitative trait variation) for body size in dogs. For the sake of illustration, assume there are two *IGF1* alleles segregating within a single randomly mating population of dogs and that body size ranges between a minimum of 9 kg and a maximum of 30 kg. Genotypic values for *IGF1* under these assumptions are shown on the genotypic scale of measurement in Fig. 10.1b.

In Chapter 9, the term **phenotypic value** was introduced to refer to the phenotype of an individual in the same units that it is measured in. Alternatively, phenotypic value refers to the *average phenotype* of a population of individuals. In the sense of the

### Problem box 10.1 Compute values on the genotypic scale of measurement for *IGF1* in dogs

Using the genotypic values of 30, 24.75, and 9 kg for the three *IGF1* genotypes, compute  $a$ , the midpoint,  $d$ , and the degree of dominance.

population average value,  $P$  refers to the average phenotypic value,  $G$  refers to the average phenotype for many individuals with a given genotype called the **genotypic value**, and  $E$  refers to the average deviation in phenotype among many individuals caused by all environmental factors combined. This leads to an equation for the causes of mean phenotypic value:

$$P = G + E \quad (10.1)$$

In words, this equation says that the mean phenotype in a population is the sum of the mean genotypic value plus the average change in phenotype due to the environment.

**Environmental deviation** The change in mean phenotype of a population caused by all non-genetic influences. Often assumed to be zero because environment is equally likely to cause an increase or decrease in individual phenotypic value, yielding a net change of zero.

**Genotypic value** The average or expected phenotype of a population of individuals all possessing an identical genotype.

**Phenotypic value** The average or expected phenotype of a population of individuals; the observed phenotype of a single individual.

It is common to assume that the mean environmental deviation is zero so that it can be ignored when deriving genotypic values. An example can be seen in Fig. 9.3, where the genotypic value of an *aabb* genotype is 1 unit of pigment and the genotypic value of a *AaBb* genotype is 5 units of pigment. The effect of the environment on individuals of a given genotype is an equal probability of deviating from the genotypic value by plus or minus 1 unit of pigment, which averages to an environmental deviation of 0 units of pigment.

With the definition of phenotypic value as a population average phenotype established, we can now develop expressions for the mean genotypic value based on genotypic values and population allele frequencies. Such mean values are prerequisites for determining how the phenotypic value of a population will change over time. As we will see eventually, mean phenotypic values are also needed to determine components of phenotypic variance since a variance is always a function of the mean.

**Table 10.1** The population mean phenotype ( $M$ ) obtained from genotype frequencies under random mating, genotypic values, and frequency-weighted genotypic values for a diallelic locus. These expectations assume that the environmental deviation is zero for each genotype.

Genotype	Frequency	Genotypic value	Frequency-weighted genotypic value
$A_1A_1$	$p^2$	$a$	$p^2a$
$A_1A_2$	$2pq$	$d$	$2pqd$
$A_2A_2$	$q^2$	$-a$	$-q^2a$
			$M = a(p - q) + 2pqd$

## 10.2 Mean genotypic value in a population

- Deriving the population mean phenotypic value.
- The population mean phenotypic value under random mating.
- The population mean phenotypic value with non-random mating.

The mean phenotypic value of a population (symbolized  $M$ ) is dictated by the genotypic values and the genotype frequencies. Table 10.1 shows the frequencies, genotypic values, and frequency-weighted genotypic values for the three genotypes under the assumption of random mating. The average phenotype of the entire population is just the sum of the phenotypic values of all individuals divided by the number of individuals. If the average environmental deviation for all genotypes is zero, then the mean phenotype of the population is the sum the three frequency-weighted genotypic values:

$$M = p^2a + 2pqd + (-q^2a) \quad (10.2)$$

This can be simplified by factoring an  $a$  out of the first and third terms:

$$M = a(p^2 - q^2) + 2pqd \quad (10.3)$$

Then notice that  $(p^2 - q^2) = (p + q)(p - q)$  and also that  $p + q = 1$  for a diallelic locus. Making this substitution leads to

$$M = a(p - q) + 2pqd \quad (10.4)$$

This population mean genotypic value is identical to the population mean phenotypic value if the mean environmental deviation is zero. It is important to note that  $M$  is measured as a deviation from the midpoint on the scale of genotypic values. Therefore,

$M$  must be added to the midpoint value to obtain the absolute population genotypic mean value.

This expression for the mean genotypic value in a population leads to two important biological conclusions. First, it shows that the mean of a quantitative trait depends on allele frequencies in the population because these dictate genotype frequencies. Second, the division of the expression for the mean phenotype into two terms is informative. The  $a(p - q)$  term shows that changes in allele frequencies shift the mean up or down by some fraction of  $a$  depending on which of the two homozygotes is more frequent. At the extremes of allele frequency, when  $p = 1$  then the mean phenotype is equal to  $+a$  and if  $q = 1$  then the mean phenotype is  $-a$ . The  $2pqd$  term shows that the frequency of the heterozygote alone determines the impact of dominance on the mean phenotype. When the phenotypic value of the heterozygote is exactly at the midpoint between the homozygotes ( $d = 0$ ), the heterozygote genotype has no impact on the population mean. Completely additive gene action exists when  $d = 0$ .

The role of allele frequency on the population mean can be seen in an example. Imagine a population of dogs where mating is random and the  $IGF1^+$  and  $IGF1^-$  alleles are both segregating. For  $IGF1$  in dogs,  $a = 10.5$  kg,  $d = 5.25$  kg, and the midpoint is 19.5 kg (see Problem box 10.1). If the two  $IGF1$  alleles are at equal frequencies then  $p = q = 0.5$ . The mean phenotypic value would be

$$\begin{aligned} M &= 10.5(0.5 - 0.5) + 2(0.5)(0.5)(5.25) \\ &= 2.625 \text{ kg} \end{aligned} \quad (10.5)$$

as a deviation from the midpoint. In absolute terms, the mean phenotype in the population would be  $2.625 + 19.5 = 22.125$  kg. Since both homozygotes are equally frequent their effect on the average cancels out. That leaves 50% of the population as

heterozygotes, shifting the mean above the midpoint by  $\frac{1}{2}d$ . If the two *IGF1* alleles are at unequal frequencies, say  $p = 0.9$  and  $q = 0.1$ , the mean phenotypic value would be

$$\begin{aligned} M &= 10.5(0.9 - 0.1) + 2(0.9)(0.1)(5.25) \\ &= 9.345 \text{ kg} \end{aligned} \quad (10.6)$$

as a deviation from the midpoint and 28.845 kg in absolute terms. At these allele frequencies the  $A_1A_1$  homozygote composes 81% of the population while the  $A_2A_2$  homozygote is only 1% of the population so the balance is strongly in favor of  $+a$ . The remaining 18% of the population is made up of heterozygotes, which also shift the average toward  $+a$  since  $d$  is positive.

The role of gene action on the population mean phenotype can also be seen in this example. Suppose that the *IGF1* alleles are  $p = 0.9$  and  $q = 0.1$  but that  $d = 0$ . The mean phenotypic value is then

$$\begin{aligned} M &= 10.5(0.9 - 0.1) + 2(0.9)(0.1)(0) \\ &= 8.4 \text{ kg} \end{aligned} \quad (10.7)$$

as a deviation from the midpoint and 27.9 kg in absolute terms. This population mean phenotype is lower than the result in equation 10.6 because the genotypic value of the heterozygotes is zero when  $d = 0$ . Therefore, the 18% of the population made up of heterozygotes does not contribute to any deviation from the midpoint in addition to that caused by the frequencies of the two homozygotes.

So far, random mating has been assumed when predicting the mean phenotype. However, consanguineous mating is also common in populations and will change the mean phenotype in predictable ways. Using the results of equation 2.20, the expected genotype frequencies can include the impact of non-random mating. The mean phenotype with the possibility of non-random mating is given by

$$M = (p^2 + fpq)a + (2pq - f2pq)d + (q^2 + fpq)(-a) \quad (10.8)$$

where  $f$  is the inbreeding coefficient. Changes in autozygosity ( $f \neq 0$ ) alter all genotype frequencies but will only impact the population mean phenotype when there is some degree of dominance ( $d \neq 0$ ). The impact of the inbreeding coefficient is identical on each of the homozygotes, increasing or decreasing them both by the same amount  $fpq$  with no impact on the mean. In contrast, changes in the autozygosity

will impact the mean phenotype if there is dominance because the frequency of heterozygotes will change by the amount  $-f2pq$ .

Extending the *IGF1* example further, imagine there is consanguineous mating so that  $f = 0.2$  when allele frequencies are  $p = 0.9$  and  $q = 0.1$ , and  $d = 5.25$ . This would lead to  $f2pq = 0.036$  or a 3.6% deficit of heterozygotes and a 1.8% excess of both homozygotes compared to random mating. The mean phenotype is then

$$\begin{aligned} M &= (0.81 + 0.018)(10.5) + (0.18 - 0.036)(5.25) \\ &\quad + (0.01 + 0.018)(-10.5) \end{aligned} \quad (10.9)$$

which equals 9.156 kg as a deviation from the midpoint, or 28.656 kg. This final example shows that changes in genotype frequencies caused by non-random mating, even when allele frequencies remain constant, can impact the population mean phenotype.

### 10.3 Average effect of an allele

- Deriving the average phenotypic effect of an allele in a population.
- The average effect as a deviation from the population mean.
- The average effect as substitution of one allele in a genotype for another.

To describe the inheritance of quantitative phenotypes across generations, it is necessary to think in terms of alleles. Although genotypes determine genotypic values, alleles and not genotypes are inherited by individuals. Thus, a new measure of value is needed that can be used to link the genotypic values of one generation with the genotypic values of their progeny in the next generation. The concept of **average effect** is used to assign a value to an *allele* and predict how it impacts the mean genotypic value of the population. As we will see, the average effect depends on the genotypic values  $a$  and  $d$  as well as the allele and genotype frequencies in the population.

**Average effect of an allele** The mean phenotypic deviation from the population mean of that group of individuals which received a particular allele from one parent and the other allele from a parent drawn at random from the population.

One way to visualize the basis of the average effect is to think of a slot machine used in gambling (also called a poker or fruit machine). Mechanical slot machines normally have three wheels that are each labeled with symbols having different frequencies on the wheels. When the wheels are spun, they will each stop and display a random symbol. An imaginary average-effect slot machine has just two wheels that represent the two alleles in a diploid genotype. One of the wheels is broken and does not spin, so that a single symbol is always displayed. For the average effect of the  $A_1$  allele for example,  $A_1$  would always appear on the broken wheel. The other wheel has symbols for all of the alleles in the population in proportion to their frequency in the population. Spinning this slot machine many times with one allele on the broken wheel and recording the value of the genotype for each spin would give expected frequencies of all of the genotypes in the population that contain an  $A_1$  allele under random mating. The average of all of the genotypic values from many spins would give the average value of all the genotypes that contain a given allele. This average value of genotypes containing one allele may very well be different than the average value of the population. This difference in averages is the average effect of the allele that is present on the broken wheel.

The derivation of the mean value of all genotypes that contain either an  $A_1$  ( $M_{A_1}$ ) or an  $A_2$  ( $M_{A_2}$ ) allele is shown in Table 10.2. This logic could be used for a locus with any number of alleles. If mating is random, we expect an  $A_1$  allele to be paired in a genotype with another  $A_1$  allele  $p$  percent of the time and with an  $A_2$  allele  $q$  percent of the time. The average value of all genotypes that contain an  $A_1$  allele is the frequency-weighted sum of the values of each genotype that has at least one  $A_1$  allele. In symbols, this mean value is

$$M_{A_1} = pa + qd \quad (10.10)$$

for the  $A_1$  allele. This quantity can be thought of as the mean value of a large number of individuals that all inherit the same allele from the same parent but that inherit the other allele in their genotype from a parent drawn at random from the population. Notice that Hardy–Weinberg expected genotype frequencies are a key assumption in Table 10.2, since expected genotype frequencies would be different if any of the Hardy–Weinberg assumptions did not hold.

The average effect is a *deviation* that measures the difference between the value of all genotypes that contain a given allele and the population mean. Therefore, the population mean must be subtracted from the mean values of genotypes produced by a given allele that are shown in Table 10.2. The average effect is then

$$\alpha_x = M_{A_x} - M \quad (10.11)$$

using the  $\alpha$  to represent the average effect and  $x$  to indicate any allele. The average effect is determined by a regression line (see Fig. 10.4, below) that minimizes the deviations (rather than the squared deviations) of individual values from the line. In many cases, when genotypes are in Hardy–Weinberg frequencies, the simple relationship of equation 10.11 suffices to describe the average effect. In more complicated situations such as when Hardy–Weinberg is not met ( $f \neq 0$ ), the line that minimizes the deviations is determined by a least-squares fit.

Substituting the expression for the population mean from equation 10.4 and the mean value of all genotypes containing the  $A_1$  allele from Table 10.2, the average effect of the  $A_1$  allele is

$$\alpha_1 = pa + qd - (a(p - q) + 2pqd) \quad (10.12)$$

which simplifies to

$$\alpha_1 = q(a + d(q - p)) \quad (10.13)$$

**Table 10.2** The mean value of all genotypes that contain either an  $A_1$  ( $M_{A_1}$ ) or an  $A_2$  ( $M_{A_2}$ ) allele. The average effect of an allele ( $\alpha_x$ ) is the difference between the mean value of the genotypes that contain a given allele and the population mean ( $\alpha_x = M_{A_x} - M$ ).

Allele	Genotype value			Mean value of all genotypes containing a given allele
$A_1$	$A_1A_1$ + $a$	$A_1A_2$ $d$	$A_2A_2$ − $a$	$M_{A_1} = pa + qd$ $M_{A_2} = pd - qa$
	$p$	$q$	0	
$A_2$	0	$p$	$q$	

as shown in Math box 10.1. The average effect of the  $A_2$  allele is

$$\alpha_2 = -p(a + d(q - p)) \quad (10.14)$$

based on the difference between the mean values of genotypes produced by  $A_2$  and the population mean.

### Math box 10.1 The average effect of the $A_1$ allele

Start with the difference between the mean value of all genotypes produced by the  $A_1$  allele in Table 10.2 and the population mean in equation 10.4:

$$\alpha_1 = pa + qd - (a(p - q) + 2pqd) \quad (10.15)$$

Then expand  $a(p - q)$  to give

$$\alpha_1 = pa + qd - pa + qa - 2pqd \quad (10.16)$$

Cancel terms to obtain

$$\alpha_1 = qd + qa - 2pqd \quad (10.17)$$

Then factor out  $q$ :

$$\alpha_1 = q(d + a - 2pd) \quad (10.18)$$

Inside the parentheses, two terms contain  $d$  in common that can be factored to give

$$\alpha_1 = q(a + d(1 - 2p)) \quad (10.19)$$

The final step is to notice that  $p + q = 1$  so that  $(p + q)$  can be substituted:

$$\alpha_1 = q(a + d(p + q - 2p)) \quad (10.20)$$

After adding  $p$  to  $-2p$ , the simplified equation for the average effect of the  $A_1$  allele is

$$\alpha_1 = q(a + d(q - p)) \quad (10.21)$$

Working through several examples based on the *IGF1* locus in dogs will help illustrate the average effect and what it measures. Table 10.3 gives four

examples of the average effect for the  $A_1$  allele for the four combinations of two allele frequencies and two levels of dominance. In all of the examples, bear in mind that the  $A_1$ -containing genotypes are  $A_1A_1$  and  $A_1A_2$  and that all values are relative to a midpoint value of 19.5 kg.

In example (a) in Table 10.3, the  $A_1$  and  $A_2$  allele frequencies are equal. Since there is no dominance, the heterozygotes have a mean value equal to the midpoint. The two homozygotes are equally frequent and their average values are equal but have opposite signs. This results in a population mean value of zero. Given an  $A_1$  allele, when sampling a second allele from this population to make genotypes it is equally likely to obtain either allele. Thus, 50% of the  $A_1$ -containing genotypes are  $A_1A_1$  with a value of  $+a = 10.5$  and 50% are  $A_1A_2$  with a value of  $d = 0$ . In total, the mean value of the  $A_1$ -containing genotypes is  $0.5(a)$  or 5.25 kg. This is the same as the average effect since the mean value of all three genotypes is zero. At these allele frequencies, the population mean is exactly at the midpoint and the  $A_1$ -containing genotypes serve to increase the average value of the population by 5.25 kg.

The situation is different in example (b) in Table 10.3 since the frequency of  $A_1$  is high and the frequency of  $A_2$  is low. Given an  $A_1$  allele, when a second allele is drawn at random from this population to make genotypes, it is much more likely to be another  $A_1$  allele rather than an  $A_2$  allele. Thus, the mean of the  $A_1$ -containing genotypes is nearly the same as  $a$  (9.45 kg), the genotypic value of  $A_1A_1$ . However, the average effect is small (1.05 kg) since the mean value of all three genotypes is also large (8.4 kg). At these allele frequencies, the population mean is near its upper limit due to the low frequency of the  $A_2$  allele and the resulting low frequency of  $A_2$ -containing genotypes that would reduce the average value of the population.

Comparing cases (a) and (c) in Table 10.3 is informative since allele frequencies are identical but the degree of dominance is different. With dominance in example (c), 50% of the  $A_1$ -containing genotypes are  $A_1A_1$  with a value of  $+a = 10.5$ , and 50% are  $A_1A_2$  with a value of  $d = 5.25$ . The mean value of the  $A_1$ -containing genotypes is now larger at  $0.5(a) + 0.5(d)$  or 7.875 kg. However, because of dominance, the mean value of the total population is also greater at 2.625 kg. Thus, the difference between the mean value of the  $A_1$ -containing genotypes and the mean value of the entire population remains at 5.25 kg, exactly as it was in example A. Comparing examples

(a) $d = 0.0, p = 0.5, q = 0.5$
$M = 10.5(0.5 - 0.5) + 2(0.5)(0.5)(0.0) = 0.0$
$A_1 \quad M_{A_1} = pa + qd = (0.5)(10.5) + (0.5)(0.0) = 5.25$
$\alpha_1 = M_{A_1} - M = 5.25 - 0.0 = 5.25$
$\alpha_1 = q(a + d(q - p)) = 0.5(10.5 + 0.0(0.5 - 0.5)) = 5.25$
$\alpha = a + d(q - p) = 10.5 + 0.0(0.5 - 0.5) = 10.5$
$\alpha_1 = q\alpha = (0.5)(10.5) = 5.25$
(b) $d = 0.0, p = 0.9, q = 0.1$
$M = 10.5(0.9 - 0.1) + 2(0.9)(0.1)(0.0) = 8.4$
$A_1 \quad M_{A_1} = pa + qd = (0.9)(10.5) + (0.1)(0.0) = 9.45$
$\alpha_1 = M_{A_1} - M = 9.45 - 8.4 = 1.05$
$\alpha_1 = q(a + d(q - p)) = 0.1(10.5 + 0.0(0.1 - 0.9)) = 1.05$
$\alpha = a + d(q - p) = 10.5 + 0.0(0.1 - 0.9) = 10.5$
$\alpha_1 = q\alpha = (0.1)(10.5) = 1.05$
(c) $d = 5.25, p = 0.5, q = 0.5$
$M = 10.5(0.5 - 0.5) + 2(0.5)(0.5)(5.25) = 2.625$
$A_1 \quad M_{A_1} = pa + qd = (0.5)(10.5) + (0.5)(5.25) = 7.875$
$\alpha_1 = M_{A_1} - M = 7.875 - 2.625 = 5.25$
$\alpha_1 = q(a + d(q - p)) = 0.5(10.5 + 5.25(0.5 - 0.5)) = 5.25$
$\alpha = a + d(q - p) = 10.5 + 5.25(0.5 - 0.5) = 10.5$
$\alpha_1 = q\alpha = (0.5)(10.5) = 5.25$
(d) $d = 5.25, p = 0.9, q = 0.1$
$M = 10.5(0.9 - 0.1) + 2(0.9)(0.1)(5.25) = 9.345$
$A_1 \quad M_{A_1} = pa + qd = (0.9)(10.5) + (0.1)(5.25) = 9.975$
$\alpha_1 = M_{A_1} - M = 9.975 - 9.345 = 0.630$
$\alpha_1 = q(a + d(q - p)) = 0.1(10.5 + 5.25(0.1 - 0.9)) = 0.630$
$\alpha = a + d(q - p) = 10.5 + 5.25(0.1 - 0.9) = 6.3$
$\alpha_1 = q\alpha = (0.1)(6.3) = 0.63$

**Table 10.3** Examples of the average effect for the *IGF1* locus in dogs. All cases assume that  $a = 10.5$  kg as shown in the genotypic scale in Fig. 10.1. For each set of allele frequencies and dominance, the table shows the population mean ( $M$ ), the mean value of all genotypes that contain an  $A_1$  allele ( $M_{A_1}$ ), the average effect of an allelic replacement ( $\alpha$ ), and the average effect of an  $A_1$  allele ( $\alpha_1$ ). Values are all in kilograms and relative to the midpoint value of 19.5 kg.

(b) and (d) in Table 10.3 illustrates that the average effect changes with a shift in the dominance value. Both the population mean and the mean of the  $A_1$ -containing genotypes change with a change in dominance at those allele frequencies.

A critical lesson to take from these examples is the contextual nature of the average effect. Average effects depend on allele and genotype frequencies and are therefore specific to the population and time point when they are measured. This is sometimes a difficult point to grasp when considering the genetic basis of phenotypic variation. Even though the genotypic values and the degree of dominance may remain constant among populations, the average effect of an allele shifts depending on the allele and genotype frequencies. This is a part of the major distinction between identifying genes and alleles that impact phenotype viewed from the perspective of Mendelian genetics (e.g. an allele causes a certain phenotype) and understanding how such alleles shape phenotypic distributions within and between populations (e.g. an allele has a large average effect).

Another way to understand the average effect when there are two alleles at a locus is to suppose that it is possible to randomly sample genotypes from a population and then be able to substitute *one* of the alleles in a genotype for another. When there are two alleles, it would be possible to replace one  $A_2$  allele initially present in the genotype with one  $A_1$  allele or vice versa. The average effect can be thought of in terms of the change in the mean value of the population when one allele replaces another in all the sampled genotypes. To measure the average effect in this manner requires determining the change in value that comes about by allelic replacement as well as the frequency with which the change in value occurs. Summing these frequency-weighted value changes will give the mean change in value that is the average effect.

Imagine we elect to replace an  $A_2$  allele with an  $A_1$  allele in genotypes sampled at random from the population *that contain at least one  $A_2$  allele* (see Fig. 10.2). Let  $p$  be the frequency of the  $A_1$  allele and  $q$  be the frequency of the  $A_2$  allele. Changing an  $A_1A_2$  genotype

	$A_2$	$\longrightarrow$	$A_1^*$	
Original allele				Replacement allele
$A_1A_2$	$\longrightarrow$		$A_1A_1^*$	
Genotypic value	$d$	$+a$	$-a$	$d$
Change in genotypic value		$a - d$		$d + a$
Frequency of allele not replaced		Frequency of $A_1 = p$		Frequency of $A_2 = q$
Frequency-weighted change in value of allele replacement		$p(a - d)$		$q(d + a)$
Average change in value of allele replacement		$p(a - d) + q(d + a)$ $= a + d(q - p)$		

**Figure 10.2** The derivation of the average change in value caused by replacing one  $A_2$  allele with an  $A_1$  allele in those genotypes sampled at random from the population that contain at least one  $A_2$  allele. The mean change in value caused by an allelic replacement forms the basis of the average effect once it is multiplied by the frequency of the allele that is replaced.

to an  $A_1A_1$  genotype will change the value from  $d$  to  $+a$ , so the difference in value is  $a - d$ . Since the frequency of the  $A_1$  allele that remains intact in the genotype is  $p$ , then the frequency of changing an  $A_1A_2$  genotype to an  $A_1A_1$  genotype is  $p$ . The frequency-weighted change in value when making this allelic substitution is therefore  $p(a - d)$ . Similarly, when changing an  $A_2A_2$  genotype to an  $A_1A_2$  via replacement of one  $A_2$  allele, the value changes from  $-a$  to  $d$ , so the difference in value is  $d - (-a)$  or  $d + a$ . The frequency of the  $A_2$  allele that remains intact in the second genotype is now  $q$ , so the frequency-weighted change in value when making this allelic substitution is therefore  $q(d + a)$ . The total average change in value when replacing an  $A_2$  allele with an  $A_1$  allele is the sum of these two separate changes in value or  $p(a - d) + q(d + a)$ . (Note that the same result, except multiplied by  $-1$ , can be obtained by electing to replace an  $A_1$  allele with an  $A_2$  allele. The negative sign comes about because  $A_2$  is the allele that decreases value.)

Using algebraic manipulation similar to that in Math box 10.1, the expression for the average change in value due to an allelic replacement simplifies to

$$\alpha = a + d(q - p) \quad (10.22)$$

with  $\alpha$  not bearing a subscript denoting the average change in value caused by an allele replacement.

Notice that  $a + d(q - p)$  also appears in equations 10.13 and 10.14. On a strictly mathematical basis, we could substitute  $\alpha$  in these two equations to restate the expressions for the average effects of the  $A_1$  and  $A_2$  alleles:

$$\alpha_1 = q\alpha \quad (10.23)$$

and

$$\alpha_2 = -p\alpha \quad (10.24)$$

However, this reformulation makes intuitive sense in terms of our imagined ability to replace one allele with another in genotypes used to derive equation 10.22. When replacing an  $A_2$  allele with an  $A_1$  allele, the frequency of the  $A_2$  allele in the population is  $q$  and this is the frequency of allelic replacements under random mating (see Fig. 10.2). Likewise,  $p$  is the frequency in the population of the  $A_1$  allele being replaced with  $A_2$  alleles. The expression for  $\alpha_2$  is negative because  $A_2$  is the allele that decreases value. Table 10.3 also gives computations of the average effect using equations 10.23 and 10.24, with identical results.

### Problem box 10.2

#### Compute the allele average effect of the *IGF1* $A_2$ allele in dogs

Using Table 10.3 as a guide, compute the average effect in three ways: (1) as the difference between  $M_{A_2}$  and the population mean  $M$ , (2) by the formula  $\alpha_2 = -p(a + d(q - p))$  from equation 10.14, and (3) by  $\alpha_2 = -p\alpha$  from equation 10.24. Be sure to compare your results with the average effects for the  $A_1$  allele given in Table 10.3 and explain why the average effects are the same or different with reference to dominance and the allele frequencies.

## 10.4 Breeding value and dominance deviation

- Deriving breeding values in a population.
- Breeding values under random and non-random mating.
- Deriving dominance deviations in a population.
- Dominance deviations under random and non-random mating.

The next step in understanding the Mendelian basis of quantitative trait variation is to move back to the level of the genotype. In Chapter 9, both additive ( $V_A$ ) and dominance ( $V_D$ ) genetic variation were identified as separate components of the total genetic variation in quantitative traits. In terms of population mean values, we can divide the genotypic mean value into its components due to additive effects of alleles and the dominance effects of genotypes

$$G = A + D \quad (10.25)$$

In words, this equation says that the mean genotypic value ( $G$ ) is the sum of the mean breeding value ( $A$ ) and the mean dominance deviation ( $D$ ). As pointed out above, this assumes that there is no interaction genetic variance due to epistasis because we are working with a one-locus example. If there were epistasis, then there would also be a mean value for an interaction deviation due to the mean value of interactions among loci that make up genotypes.

This subsection will address how the mean phenotype of a population of progeny depends on mating among the population of parents. It is important to first understand the motivation to predict what is called the **breeding value** of a genotype. When natural selection occurs, it is essentially the differential mating success of certain phenotypes. Humans achieve the same thing in domestic plants and animals through artificial selection by allowing only those individuals with preferred phenotypes to breed. To the extent that phenotype is a function of genotype, natural and artificial selection allow some genotypes to breed more often than others. A full understanding of how and why a given mean phenotype occurs in a population of progeny therefore requires an understanding of the consequences of mating in the parental generation. Here we will start with the genotypic value of an individual, and then track the frequencies and genotypic values of its progeny to predict the mean genotypic value of its progeny.

**Breeding value of an individual** Twice the mean value of the progeny that would be produced by a single genotype under random mating expressed as a difference from the population mean.

Parents pass on alleles and not genotypes to their progeny. (From the progeny point of view, individuals

inherit one allele from each of two parents rather than a diploid genotype.) The impact of a single parent on the mean value of the progeny population could be measured in a manner akin to the average effect. Now, instead of a special slot machine, we could just take a single individual and have it mate with many individuals drawn at random from the parental population. Each progeny from these matings would inherit one allele from the focal parent and another allele from an individual drawn at random. The resulting population of progeny would have a mean value that could be measured. The **breeding value** of a genotype is two times the difference between the progeny mean value and the parental population mean value ( $M$ ). Expressed as an equation:

$$\text{Breeding value} = 2(M_{\text{progeny } A_x A_x} - M) \quad (10.26)$$

where  $M_{\text{progeny } A_x A_x}$  is the mean value of progeny produced when an individual of genotype  $A_x A_x$  is mated to many individuals drawn at random from the parental population. The difference between the means is multiplied by 2 because a parent's genotype possesses two alleles but its progeny inherit only one allele at a time.

The components that lead to an expression for  $M_{\text{progeny}}$  for the  $A_1 A_1$  genotype are shown in Table 10.4. When the  $A_1 A_1$  genotype mates, it will encounter and mate with individuals of a given genotype in proportion to the frequency of that genotype in the population. An  $A_1 A_1$  individual is expected to mate with  $A_1 A_1$ ,  $A_1 A_2$ , and  $A_2 A_2$  individuals with frequencies of  $p^2$ ,  $2pq$ , and  $q^2$ , respectively, since these are the Hardy–Weinberg expected genotype frequencies in the population. Each of the matings between an  $A_1 A_1$  genotype and another genotype will produce progeny with one or two genotypes. The phenotypic values of each of the progeny genotypes is also known. To obtain an expression for the mean phenotypic value of the progeny that result from the  $A_1 A_1$  genotype mating at random in the population ( $M_{\text{progeny } A_1 A_1}$ ), add up all of the progeny phenotypic values after weighting each one by its relative frequency among all of the progeny and by the frequency of mating pairs to obtain

$$M_{\text{progeny } A_1 A_1} = p^2a + 2pq(1/2a + 1/2d) + q^2d \quad (10.27)$$

This expression can be simplified by first expanding the middle term

$$M_{\text{progeny } A_1 A_1} = p^2a + pqa + pqd + q^2d \quad (10.28)$$

**Table 10.4** The mean phenotypic value of progeny that result when an individual of the genotype  $A_1A_1$  mates randomly. All genotypes in the population have Hardy–Weinberg expected frequencies. Therefore, each of the mating pairs has an expected frequency of  $p^2$ ,  $2pq$ , or  $q^2$ . The mean value of all progeny produced by the  $A_1A_1$  genotype is the frequency-weighted sum of the progeny phenotypic values.  $M_{\text{progeny } A_1A_1}$  forms the basis of the breeding value since the breeding value for  $A_1A_1$  is  $M_{\text{progeny } A_1A_1} - M$ .

Focal genotype	$A_1A_1$	$A_1A_2$	$A_2A_2$	
Mate genotypes	$A_1A_1$ $p^2$	$A_1A_2$ $2pq$	$A_2A_2$ $q^2$	
Progeny genotype and relative frequency from each mating	$A_1A_1$ $+a$	$\frac{1}{2}A_1A_1$ $+a$	$\frac{1}{2}A_1A_2$ $d$	$A_2A_2$ $d$
Progeny mean value	$M_{\text{progeny } A_1A_1} = p^2a + 2pq(\frac{1}{2}a + \frac{1}{2}d) + q^2d = ap + dq$			

and then factoring to give

$$M_{\text{progeny } A_1A_1} = ap(p+q) + dq(p+q) \quad (10.29)$$

and then noticing that  $p+q=1$  so that

$$M_{\text{progeny } A_1A_1} = ap + dq \quad (10.30)$$

The same logic can be applied to obtain the progeny mean values for the  $A_1A_2$  and  $A_2A_2$  genotypes under random mating.

Prepared with the expression for  $M_{\text{progeny } A_1A_1}$ , we can now obtain an equation for the breeding value of the  $A_1A_1$  genotype. Using the definition of the breeding value given in equation 10.26,

$$\text{Breeding value } A_1A_1 = 2(M_{\text{progeny } A_1A_1} - M) \quad (10.31)$$

Substituting in the expression for mean progeny value in equation 10.30 and the expression for the population mean in equation 10.4 gives

$$\text{Breeding value } A_1A_1 = 2(ap + dq - (a(p-q) + 2pqd)) \quad (10.32)$$

Rearrangement of this equation yields

$$\text{Breeding value } A_1A_1 = 2q(a + d(q-p)) \quad (10.33)$$

(Readers carrying out the algebra can refer to Math box 10.1 for the steps that show  $d(1-2p) = d(q-p)$ ). Notice that the  $q(a + d(q-p))$  term is equal to the definition of the average effect of the  $A_1$  allele given in equation 10.13. Therefore, the breeding value of the  $A_1A_1$  genotype is simply equal to two times the average effect of the  $A_1$  allele.

Based on the definition of the breeding value as the difference between the progeny mean and the population mean, the breeding value of any genotype is simply the sum of the average effects of each of the alleles that make up the genotype. We can therefore use the definitions of the average effect in the last section to define breeding values for the three genotypes at a diallelic locus:

$$\begin{aligned} \text{Breeding value } A_1A_1 &= \alpha_1 + \alpha_1 = q\alpha + q\alpha \\ &= 2q\alpha \end{aligned} \quad (10.34)$$

$$\begin{aligned} \text{Breeding value } A_1A_2 &= \alpha_1 + \alpha_2 = q\alpha + -p\alpha \\ &= (q-p)\alpha \end{aligned} \quad (10.35)$$

$$\begin{aligned} \text{Breeding value } A_2A_2 &= \alpha_2 + \alpha_2 = -p\alpha + -p\alpha \\ &= -2p\alpha \end{aligned} \quad (10.36)$$

Because breeding values are made up of the average effects of the two alleles in a genotype, breeding values also depend on the population context in which they are measured. Thus, breeding values are not universal features of genotypes but rather depend on the population allele and genotypes frequencies.

The breeding value expressions tell us the mean value of the progeny that would be produced by each genotype under random mating. Look at the example breeding values for the three genotypes at the *IGF1* locus in dogs (Table 10.5). Under the allele frequencies and zero dominance in case A, a population of progeny from an individual with an  $A_1A_1$  genotype would have a mean of 10.5 kg relative to the midpoint. (In the absolute, the progeny would have a mean value of 30 kg, which is identical to the  $A_1A_1$  genotypic value.) In case (a) with  $p=q=0.5$ , an  $A_1$  allele would be paired with an  $A_1$  allele in half of the progeny and

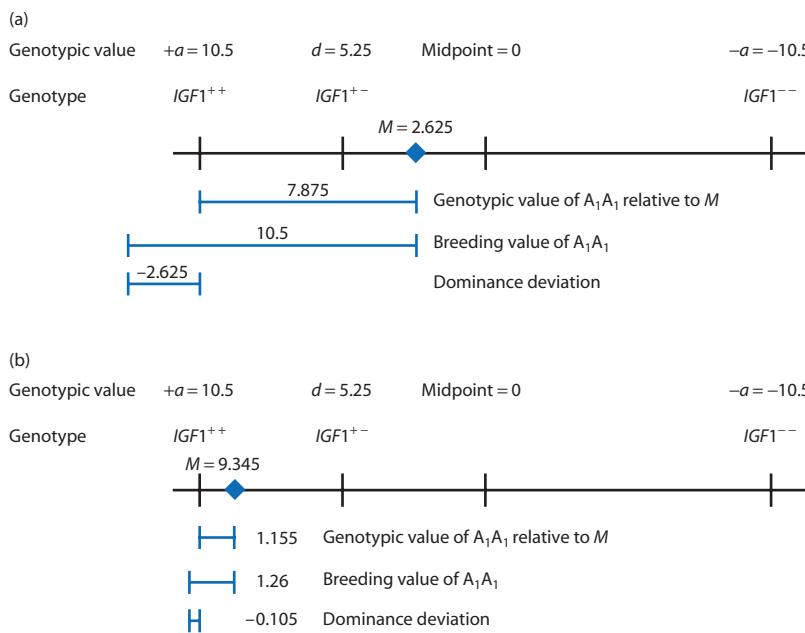
**Table 10.5** Examples of breeding values for the three *IGF1* locus genotypes in dogs. Values are all in kilograms and relative to the midpoint value of 19.5 kg.

	Breeding value		
	$A_1A_1$	$A_1A_2$	$A_2A_2$
(a) $d = 0.0, p = 0.5, q = 0.5, M = 0.0, \alpha = 10.5$	$2(0.5)(10.5) = 10.5$	$(0.5 - 0.5)(10.5) = 0.0$	$-2(0.5)(10.5) = -10.5$
(b) $d = 0.0, p = 0.9, q = 0.1, M = 8.4, \alpha = 10.5$	$2(0.1)(10.5) = 2.1$	$(0.1 - 0.9)(10.5) = -8.4$	$-2(0.9)(10.5) = -18.9$
(c) $d = 5.25, p = 0.5, q = 0.5, M = 2.625, \alpha = 10.5$	$2(0.5)(10.5) = 10.5$	$(0.5 - 0.5)(10.5) = 0.0$	$-2(0.5)(10.5) = -10.5$
(d) $d = 5.25, p = 0.9, q = 0.1, M = 9.345, \alpha = 6.3$	$2(0.1)(6.3) = 1.26$	$(0.1 - 0.9)(6.3) = -5.04$	$-2(0.9)(6.3) = -11.34$

with an  $A_2$  allele in half of the progeny to make equal frequencies of  $A_1A_1$  and  $A_1A_2$  progeny. The mean value of these progeny, and hence the average effect of an  $A_1$  allele, would be  $0.5(10.5) + 0.5(0.0) = 5.25$  kg. The breeding value of an  $A_1A_1$  genotype is twice the average effect of the  $A_1$  allele because it has two  $A_1$  alleles. Therefore, we double the average effect of an  $A_1$  allele to determine that the breeding value of the  $A_1A_1$  genotype is  $2(5.25) = 10.5$  kg. Figure 10.3 also shows two examples of the breeding value for the *IGF1* locus in dogs in relation to the genotypic scale of measurement and the population mean.

As another example, focus on cases (a) and (c) of Table 10.5 and consider why the heterozygotes both have breeding values of zero. When heterozygotes breed, they contribute equal proportions of  $A_1$  and  $A_2$  alleles to their progeny. With a dominance value of zero and equal allele frequencies, it makes intuitive sense in case (a) that the heterozygotes have a breed-

ing value of zero. Each of the  $A_1$  and  $A_2$  alleles that a heterozygote contributes to its progeny is paired with  $A_1$  and  $A_2$  alleles in equal frequency from the population. That would form an equal number of  $A_1A_1$  and  $A_2A_2$  genotypes with a mean value of zero whereas all heterozygous progeny would also have a mean value of zero since the heterozygote genotypic value is zero. In example (c), the progeny genotypes and frequencies are identical to example (a) since the allele frequencies are the same. However, there is now dominance such that the heterozygote has a genotypic value of 5.25 kg. Still the breeding value of a heterozygote is zero. An equal frequency of  $A_1A_1$  and  $A_2A_2$  genotypes in the progeny gives a mean value of zero. The 50% of progeny that are heterozygotes have a value of 5.25 because of dominance. Nonetheless,  $0.5(5.25) = 2.625$  exactly equals the population mean of 2.625 so the heterozygous progeny also do not alter the population mean.



**Figure 10.3** Illustration of dominance deviation for the *IGF1* gene in dogs. The dominance deviation is the difference between the genotypic value (measured relative to the population mean,  $M$ ) and the breeding value. The dominance deviation is a consequence of the heterozygote genotypic value not falling at the midpoint. Panels (a) and (b) correspond to cases (c) and (d), respectively, in Tables 10.3, 10.5, and 10.7.

### Interact box 10.1 Average effects, breeding values, and dominance deviations

Average effects, breeding values, and dominance deviations across all allele frequencies can be interactively graphed using an Excel spreadsheet. Parameter values of  $a$  and  $d$  can be set in the model. Use  $a = 10.5$  as in the *lGF1* example. Then set  $d$  to 0.0 and 5.25 view the graphs in each instance.

Another way to understand the breeding values is to determine the total breeding value in a population where all three genotypes are mating at random. Let the population of parents have Hardy–Weinberg expected genotype frequencies of  $p^2$ ,  $2pq$ , and  $q^2$ . To find the average breeding value of all three genotypes in the parental population, multiply the breeding value of each genotype by its corresponding genotype frequency. This gives

$$\begin{aligned} \text{Mean breeding value of all genotypes} \\ = p^2 2q\alpha + 2pq(q - p)\alpha - q^2 2p\alpha \quad (10.37) \end{aligned}$$

This equation can be simplified by factoring  $2pq$  from each term and expanding  $(q - p)\alpha$  in the middle term to give

$$\begin{aligned} \text{Mean breeding value of all genotypes} \\ = 2pq(p\alpha + q\alpha - p\alpha - q\alpha) = 0 \quad (10.38) \end{aligned}$$

The conclusion is that the mean breeding value of all three genotypes mating at random is zero. This result makes intuitive sense because when a large parental population composed of genotypes in Hardy–Weinberg expected frequencies mates at random, the mean value of the progeny population should be exactly the same since the progeny genotype frequencies are exactly the same as in the parental population. It is only when genotypes mate more or less often than expected by random mating that the progeny population mean value differs from the parental population mean value. Note that with natural or artificial selection, mating is by definition non-random and some genotypes mate more frequently than others. It is the over- or under-representation of parental genotypes in mating that causes the mean value of progeny to differ from the mean value of their parents.

### Dominance deviation

With the breeding value established, we can now focus on the dominance deviation that makes up the second portion of the total mean genotypic value in  $G = A + D$  (equation 10.25). While the breeding value measures the mean value of alleles passed to progeny by a given genotype, these same progeny also possess genotypes. Due to dominance, genotypes may not completely reflect the combinations of the alleles that make them up. For example, with complete dominance of the  $A_1$  allele, an  $A_1A_2$  genotype masks the fact that it has one  $A_2$  allele since its phenotype is indistinguishable from that of an  $A_1A_1$  genotype. When there is no dominance, the average value of progeny is a perfect representation of the parental genotypic value. However, dominance changes the average value of progeny and can make the breeding value different than the parental genotypic values. The difference between the genotypic value and the breeding value caused by dominance is called the **dominance deviation**.

**Dominance deviation** The difference between the genotypic value and the breeding value where the genotypic value is measured relative to the mean value of the population.

Expressions for the dominance deviations are obtained by taking the difference between the genotypic value and the breeding value for each genotype. We have already obtained all of the expressions needed for the dominance deviation. We do, however, need to obtain one new equation. Since breeding values are expressed relative to the population mean, we have to start by also expressing genotypic values relative to the population mean rather than relative to the midpoint. For the  $A_1A_1$  genotype, the difference between the genotypic value and the population mean in equation 10.4 is

$$a - (a(p - q) + 2pqd) \quad (10.39)$$

which simplifies to

$$2q(a - dp) \quad (10.40)$$

as the genotypic value of  $A_1A_1$  relative to the population mean.

**Table 10.6** Expressions for genotypic values relative to the population mean, breeding values and dominance deviations. Genotypic values can be expressed relative to the population mean by subtracting the population mean ( $M = a(p - q) + 2pqd$ ) from a genotypic value measured relative to the midpoint. The dominance deviation is the difference between the genotypic value expressed relative to the population mean ( $M$ ) and the breeding value.

Genotype . . .	Value		
	$A_1A_1$	$A_1A_2$	$A_2A_2$
Genotypic value relative to midpoint	$+a$	$d$	$-a$
Genotypic value relative to population mean	$2q(a - dp)$	$a(p + q) + d(1 - 2pq)$	$-2p(a - dp)$
Breeding value	$2q(\alpha - qd)$	$(q - p)\alpha + 2pqd$	$-2p(\alpha + pd)$
Dominance deviation	$2q(a + d(q - p))$	$(q - p)(a + d(q - p))$	$-2p(a + d(q - p))$
	$2q\alpha$	$(q - p)\alpha$	$-2p\alpha$
	$-2q^2d$	$2pqd$	$-2p^2d$

Using the  $A_1A_1$  genotype breeding value of  $2q\alpha$  where  $\alpha = a + d(q - p)$ , the difference between the genotypic value relative to the population mean and the breeding value is

$$\begin{aligned} A_1A_1 \text{ Dominance deviation} \\ = 2q(a - dp) - 2q(a + d(q - p)) \end{aligned} \quad (10.41)$$

This equation can be expanded to

$$\begin{aligned} A_1A_1 \text{ Dominance deviation} \\ = 2qa - 2qdp - 2qa - 2q^2d + 2qdp \end{aligned} \quad (10.42)$$

and then canceling terms gives

$$A_1A_1 \text{ Dominance deviation} = -2q^2d \quad (10.43)$$

Table 10.6 gives the expressions for all three genotypic values relative to the population mean as well as the dominance deviations derived using this same reasoning.

The dominance deviations for the four *IGF1* examples are shown in Table 10.7. Examination of Table 10.7 shows that the genotypic values of the heterozygotes (measured from the population mean) and breeding values in the table are always zero when there is no dominance. In both cases (a) and (b), the  $A_1A_2$  genotypic value is at the midpoint ( $d = 0$ ), giving dominance deviations of zero regardless of the allele frequencies. Another way to describe this is to say that when there is no dominance, a genotype has a breeding value that is identical to its genotypic value (measured relative to the population mean) since genotypic values are determined by the addition

of average effects alone. Without dominance, genotypic values in progeny as measured by the breeding value can be predicted perfectly from the combination of average effects.

In cases (c) and (d) the  $A_1A_2$  genotypic values are not at the midpoint ( $d \neq 0$ ), resulting in dominance deviations that are not zero. Let's examine the dominance deviation for the  $A_1A_1$  genotype in case (c) of Table 10.7 as illustrated in Fig. 10.3a. The genotypic value of the  $A_1A_1$  genotype when measured relative to the population mean is  $10.5 - 2.625 = 7.875$ . The breeding value of the  $A_1A_1$  genotype is  $10.5$  based on the average effects of its two alleles. With this breeding value, the mean of the progeny from an  $A_1A_1$  genotype would be outside the largest phenotypic value, which is physically impossible. The breeding value has in essence assumed that the population mean is zero, as it would be at  $p = q = 0.5$  without dominance. But there is dominance, so the population mean is greater than it would be without dominance. The  $A_1A_1$  breeding value needs to be adjusted downward so that it does not exceed the largest phenotypic value. This adjustment is the dominance deviation. In this particular case, the population mean is 2.625 rather than zero because of dominance shown by the  $A_1A_2$  genotype (the  $A_1A_2$  genotype has a value of 5.25 and makes up 50% of the population at  $p = q = 0.5$ ). Therefore, the dominance deviation is  $-2.625$ . Note that in general the magnitude of dominance deviations changes with population allele frequencies just as the magnitude of genotypic values measured from the population mean and breeding values do.

The dominance deviation for the  $A_1A_1$  genotype in case (d) of Table 10.7 is also diagrammed in Fig. 10.3.

**Table 10.7** Genotypic values, breeding values, and dominance deviations for the three *IGF1* locus genotypes in dogs. Genotypic values, breeding value and dominance deviation values are all given relative to the population mean,  $M$ . All values are in kilograms.

	Genotype		
	$A_1A_1$	$A_1A_2$	$A_2A_2$
(a) $d = 0.0, p = 0.5, q = 0.5, M = 0.0, \alpha = 10.5$			
Genotype frequency	0.25	0.5	0.25
Genotypic value	10.5	0.0	-10.5
Breeding value	10.5	0.0	-10.5
Dominance deviation	0.0	0.0	0.0
(b) $d = 0.0, p = 0.9, q = 0.1, M = 8.4, \alpha = 10.5$			
Genotype frequency	0.81	0.18	0.01
Genotypic value	2.1	-8.4	-10.5
Breeding value	2.1	-8.4	-10.5
Dominance deviation	0.0	0.0	0.0
(c) $d = 5.25, p = 0.5, q = 0.5, M = 2.625, \alpha = 10.5$			
Genotype frequency	0.25	0.5	0.25
Genotypic value	7.875	2.625	-13.125
Breeding value	10.5	0.0	-10.5
Dominance deviation	-2.625	-2.625	-2.625
(d) $d = 5.25, p = 0.9, q = 0.1, M = 9.345, \alpha = 6.3$			
Genotype frequency	0.81	0.18	0.01
Genotypic value	1.155	-4.095	-19.845
Breeding value	1.26	-5.04	-11.34
Dominance deviation	-0.105	0.945	-8.505

At the allele frequencies in case (d), the dominance deviation is smaller because the frequency of  $A_1A_2$  individuals in the population is smaller, causing less dominance contribution to the population mean. That ultimately results in a smaller difference between the genotypic value of  $A_1A_1$  measured relative to  $M$  and the breeding value of  $A_1A_1$ .

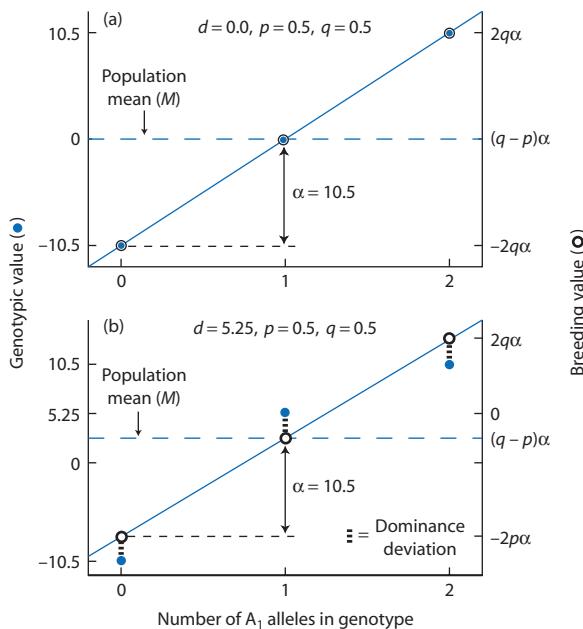
We can also determine the mean dominance deviation in a population where all three genotypes are mating at random. As for the average breeding value, the average dominance deviation of all three genotypes is obtained by multiplying the dominance deviation of each genotype by its corresponding genotype frequency. This gives

$$\begin{aligned} &\text{Mean dominance deviation of all genotypes} \\ &= p^2(-2q^2d) + 2pq(2pqd) - q^2(-2p^2d) \quad (10.44) \end{aligned}$$

which simplifies to zero since the first and third terms cancel with the middle term. The mean dominance deviation of all three genotypes mating at random is zero. Just as for the mean breeding value, a mean dominance deviation makes intuitive

sense because when a large parental population in Hardy–Weinberg expected genotype frequencies mates at random, the mean value of the progeny population should be exactly the same since the progeny genotype frequencies are exactly the same as in the parental population. Thus, the impacts of dominance on the progeny means of each genotype counteract each other to give an overall mean of zero under random mating.

One last illustration will help show the connection between the genotypic values and the genotype frequencies on one hand and the breeding values and dominance deviations on the other. Figure 10.4 shows the least-squares regression line between the genotypic values and the number of  $A_1$  alleles in a genotype for a population of 100 individuals (since genotype frequencies equal Hardy–Weinberg expectations, the population contains 25  $A_1A_1$ , 50  $A_1A_2$ , and 25  $A_2A_2$  individuals). The slope of the regression line gives the average phenotypic effect of a change in the number of  $A_1$  alleles in the genotype. So, for example, changing all  $A_2A_2$  genotypes to  $A_1A_2$  genotypes (replacing one  $A_2$  allele with an



**Figure 10.4** The relationship between average effect of an allele replacement, genotypic values, breeding values, and dominance deviations. The solid line represents the least-squares regression of genotypic value on number of A<sub>1</sub> alleles in a genotype in a population of 100 individuals with Hardy–Weinberg genotype frequencies. The slope of the line is equal to the effect of an allelic replacement or  $\alpha$ . There is no dominance in (a) and partial dominance in (b). The breeding value is equal to the genotypic value adjusted by the effects of dominance on the progeny population mean as measured by the dominance deviation. All values are deviations from the population mean.

A<sub>1</sub> allele) would change the phenotypic value of that group of individuals from  $-10.5$  to zero. Likewise, changing all A<sub>1</sub>A<sub>2</sub> genotypes to A<sub>1</sub>A<sub>1</sub> genotypes would change the phenotypic value of that group of individuals from zero to  $10.5$ . The average phenotypic effect of increasing the number of A<sub>1</sub> alleles is therefore  $10.5$  (or the slope of the regression line), so  $\alpha = 10.5$  as shown in Fig. 10.4.

Given the average effect of an allele replacement, we can then predict the expected phenotypic value of the progeny of any one genotype. Let's use the A<sub>2</sub>A<sub>2</sub> genotype as an example. When A<sub>2</sub>A<sub>2</sub> individuals mate, the mean value of the progeny population will decrease because progeny of A<sub>2</sub>A<sub>2</sub> individuals receive the A<sub>2</sub> allele that confers the lower genotypic value. The amount of the decrease in value is  $-2q\alpha$ , which in Fig. 10.4a equals  $-2(0.5)(10.5) = -10.5$ . Figure 10.4b shows the impact of dominance. Even with dominance, when A<sub>2</sub>A<sub>2</sub> individuals mate the mean value of the progeny population still decreases because their progeny receive the A<sub>2</sub> allele. The decrease in

phenotypic value is  $-2q\alpha = -10.5$ . But dominance has caused the population mean to be  $2.625$  rather than zero. Another consequence of dominance is that when A<sub>2</sub>A<sub>2</sub> individuals mate, some of the A<sub>2</sub> alleles will pair with A<sub>1</sub> alleles to make heterozygotes with the phenotype of  $d$ . Therefore, the average effects of two A<sub>2</sub> alleles need to be adjusted by the dominance deviation. The dominance deviation for A<sub>2</sub>A<sub>2</sub> is  $-2p^2d$ , which in panel (a) equals zero and in panel (b) equals  $-2(0.5)^2(5.25) = -2.625$ .

## 10.5 Components of total genotypic variance

- Deriving the additive ( $V_A$ ) and dominance ( $V_D$ ) components of genotypic variation.
- $V_A$  and  $V_D$  are related to allele frequencies and genotypic values.

We are at last in a position to obtain expressions for the components of variance in phenotype due to additive genetic variation and dominance genetic variation. Recall from earlier in the chapter that additive genetic variation ( $V_A$ ) contributes to the resemblance between parents and offspring. Using the terminology of this section, we can say that  $V_A$  is due to the *variance* in breeding values. Similarly, earlier in the chapter dominance genetic variation ( $V_D$ ) was described as being due to the effect, if any, of combining different alleles into a heterozygote genotype. We can now recognize  $V_D$  as the *variance* in dominance deviations. These variances describe the spread or range of values in a population caused by breeding value or by dominance.

The previous subsections devoted considerable effort to developing expressions for average values. In particular, we obtained expressions for the average breeding value and average dominance deviation of each genotype. Obtaining these averages was important because an average is a critical part of a variance (see the Appendix for the definition of a variance). As was shown above, in a randomly mating population both the mean breeding value and the mean dominance deviation are zero. These are extremely useful results because they greatly simplify the expressions for the variance in breeding value and the variance in dominance deviation.

The additive variance, or  $V_A$ , is the variance of breeding values in a randomly mating population. Since the mean breeding value taken over all genotypes is zero, the variance in breeding value is simply the square of the mean breeding value for each genotype multiplied by the frequency of each genotype:

$$V_A = p^2(2q\alpha)^2 + 2pq((q-p)\alpha)^2 + q^2(-2p\alpha)^2 \quad (10.45)$$

By expanding each term, factoring out  $2pq\alpha^2$  and then canceling, the equation simplifies to

$$V_A = 2pq\alpha^2 \quad (10.46)$$

or, after substituting the definition of  $\alpha$  from equation 10.22, to

$$V_A = 2pq(a + d(q-p))^2 \quad (10.47)$$

Similarly, the dominance variance, or  $V_D$ , is the variance in dominance deviation values in a randomly mating population. As before, the mean dominance deviation taken over all genotypes is zero so the variance in breeding value is the square of the mean dominance deviation for each genotype multiplied by the frequency of each genotype:

$$V_D = p^2(-2q^2d)^2 + 2pq(2pqd)^2 + q^2(-2p^2d)^2 \quad (10.48)$$

Expanding each term gives  $4p^2q^2d^2(p^2 + 2pq + q^2)$ , which then gives the equation for the dominance deviation variance as

$$V_D = (2pqd)^2 \quad (10.49)$$

The separate expressions for  $V_A$  and  $V_D$  give us the means to estimate the total genotypic variance or  $V_G$  as

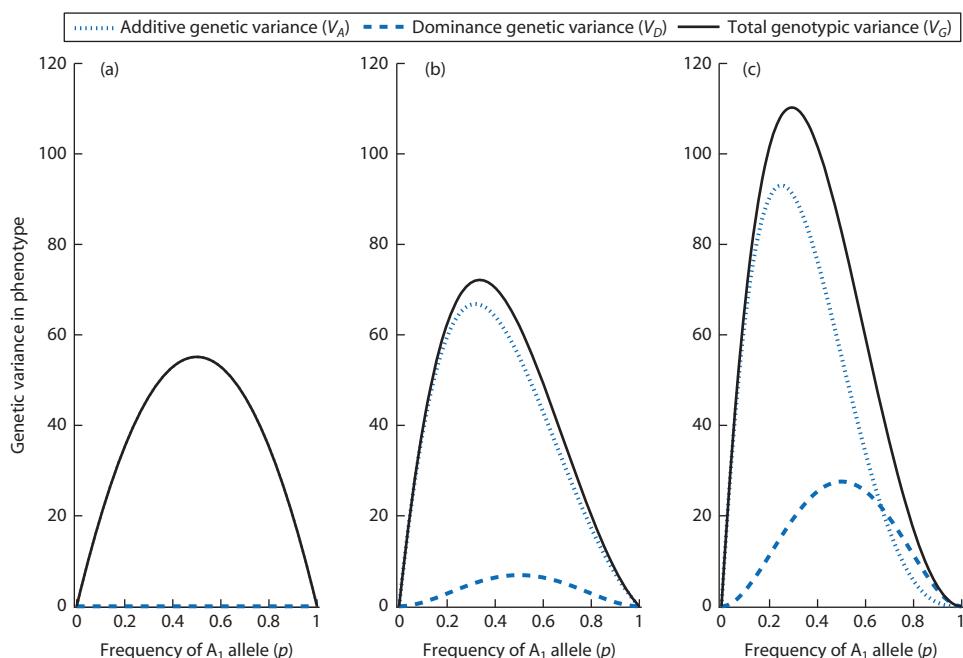
$$V_G = V_A + V_D \quad (10.50)$$

which after substitution of the expressions for  $V_A$  and  $V_D$  becomes

$$V_G = 2pq(a + d(q-p))^2 + (2pqd)^2 \quad (10.51)$$

Note that  $V_G$  is commonly referred to as the total genetic variance, even though it is the variance in genotypic values.

The components of the genetic variance and their relationship to the allele frequencies and the genotypic values can be seen in Fig. 10.5. In all cases,  $V_G$  is greatest at intermediate allele frequencies. When there is no dominance as in Fig. 10.5a,  $V_G$  is made up exclusively of additive genetic variation. Without any dominance  $V_A$  is greatest at intermediate allele frequencies where the total frequency of the two homozygotes equals the frequency of the heterozygotes under random mating. When there is dominance, as in panels (b) and (c),  $V_G$  is made up of both additive and dominance genetic variation. Dominance variance is greatest when the frequency of all heterozygotes is the greatest, or  $p = q = 0.5$  for a diallelic locus under random mating. Also notice that with dominance  $V_G$  is low when the frequency of the recessive allele ( $A_2$ ) is low because the population has very few individuals with the  $-a$  genotypic value. This means that  $V_G$  is made up mostly of the variance in  $+a$  and  $d$  genotypic values that are very similar or identical because of dominance. Since all of the illustrations



**Figure 10.5** The additive ( $V_A$ ) and dominance ( $V_D$ ) components of the total genotypic variance ( $V_G$ ). In (a) there is no dominance ( $d = 0.0$ ). Panel (b) has partial dominance ( $d = 5.25$ ) and panel (c) has complete dominance ( $d = 10.5$ ). In (b) and (c) the dominance component of the total genotypic variance is greatest when allele frequencies are equal because the frequency of heterozygotes is greatest. In all cases, the value of  $+a = 10.5$ .

in Fig. 10.5 use a genotypic value of  $a = 10.5$ , each panel represents one case of the components of genotypic variance for *IGF1* depending on the degree of dominance.

Across the three graphs in Fig. 10.5, the maximum  $V_G$  progressively increases and the allele frequency where maximum  $V_G$  occurs also shifts to lower values of  $p$ . This increase in maximum  $V_G$  corresponds to an

### Interact box 10.2 Components of total genotypic variance, $V_G$

The additive and dominance components of the total genotypic variance can be interactively graphed using an Excel spreadsheet model on the textbook website. Values of  $a$  and  $d$  can be set in the model. Set  $a = 10.5$  as in the *IGF1* example. Then set  $d$  to 0.0, 5.25, and 10.5 and view the graphs in each instance. The graphs should be identical to those in Fig. 10.5.

What would the graph of the components of the total genotypic variance look like with strong overdominance? Predict what the graph might look like and sketch it on paper. Then set  $d$  to a value that constitutes overdominance and view the graph. Was your prediction correct? What is the impact of strong overdominance on heritability? If a population with strong overdominance and allele frequencies near  $p = q = 0.5$  experienced genetic drift or consanguineous mating, how would  $V_A$  and the heritability change?

### Math box 10.2 Deriving the total genotypic variance, $V_G$

While the method of adding  $V_A$  and  $V_D$  to obtain the total genotypic variance seems reasonable, it does have an important assumption. The total genetic variance,  $V_G$ , is the sum of  $V_A + V_D$  plus twice the covariance between  $V_A$  and  $V_D$ :

$$V_G = V_A + V_D + 2\text{cov}_{AD} \quad (10.52)$$

This covariance can be estimated by summing the genotype frequency-weighted product of the breeding value and the dominance deviation for each of the genotypes:

$$\text{cov}_{AD} = p^2(2q\alpha)(-2q^2d) + 2pq((q-p)\alpha)(2pqd) + q^2(-2p\alpha)(-2p^2d) \quad (10.53)$$

After multiplying out each term to get

$$\text{cov}_{AD} = -4p^2q^3\alpha d + 4p^2q^2(q-p)\alpha d + 4p^2q^3\alpha d \quad (10.54)$$

and then factoring out  $4p^2q^2\alpha d$  to get

$$\text{cov}_{AD} = -4p^2q^2\alpha d(-q + q - p + p) \quad (10.55)$$

it is then apparent that the covariance between the breeding values and the dominance deviations is zero.

The same overall result of the total genetic variance being a function of  $V_A + V_D + 2\text{cov}_{AD}$  can be shown by starting with the expression for the total genotypic variance based on the frequency-weighted variances for each genotypic value

$$V_G = p^2(a - M)^2 + 2pq(d - M)^2 + q^2(-a - M)^2 \quad (10.56)$$

and then using definitional substitutions and algebra to obtain an equation with terms that represent the variance of the breeding values plus the variance of the dominance deviations and the covariance between the breeding values and dominance deviations.

increase in the maxima of both  $V_A$  and  $V_D$ . Complete dominance causes more total genotypic variation because the genotypic values in the population are only the extremes of  $+a$  and  $-a$  without a genotypic value that is intermediate. The variance in genotypic values is at a maximum when there are equal frequencies of the  $+a$  and  $-a$  genotypic values. At  $p \approx 0.29$  and  $q \approx 0.71$  the combined frequencies of the  $A_1A_1$  and  $A_1A_2$  genotypes are equal to the frequency of the  $A_2A_2$  genotypes, thus maximizing  $V_G$ .

## 10.6 Genotypic resemblance between relatives

- Additive ( $V_A$ ) and dominance ( $V_D$ ) components of genotypic variation and resemblance of relatives.
- Resemblance of relatives depends on the probability that alleles or genotypes are shared.
- The covariance between mid-parent and offspring genotypic values.

Obtaining expressions for the components of the total genotypic variation is ultimately useful because the variance components provide the basis to predict the resemblance of genotypic values between populations of related individuals. This relationship can also be reversed, and the phenotypic resemblance between relatives can be used to estimate the components of the total genotypic variance for specific phenotypes. This section shows how the variance within or covariance between relatives are related to components of the genotypic variance. In practice,  $V_A$  is usually estimated from the resemblance between relatives to then determine the heritability of a phenotype. See the Appendix for background on the covariance if necessary.

The expected covariance in genotypic values between relatives can be determined using the autozygosity of the relatives compared (Cotterman 1974, 1983; Crow and Kimura 1970). The expected covariance between related individuals is

$$\text{cov}(x, y) = rV_A + uV_D \quad (10.57)$$

where  $x$  and  $y$  represent the phenotypic values in a population of individuals and  $r$  and  $u$  are fractions determined by probabilities of identity by descent for the individuals in groups  $x$  and  $y$  and for their parents. If the parents of individuals in group  $x$  are A and B and the parents of individuals in group  $y$  are C and D then

$$r = 2f_{xy} \quad (10.58)$$

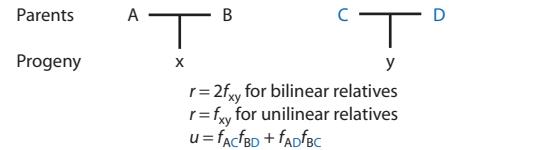
if the parental population is not inbred ( $f = 0$ ) and

$$u = f_{AC}f_{BD} + f_{AD}f_{BC} \quad (10.59)$$

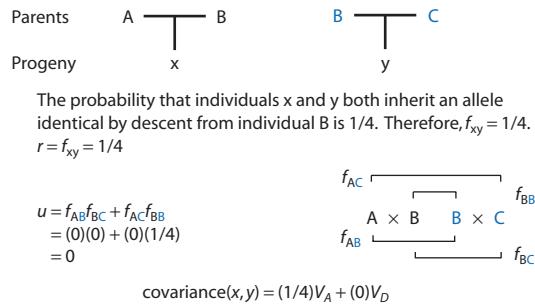
where  $f$  is the probability of autozygosity in a pedigree as explained in Chapter 2.

These coefficients have a clear biological interpretation based on the pedigrees of individuals  $x$  and  $y$  (Fig. 10.6). The  $r$  coefficient is twice the probability that individuals  $x$  and  $y$  inherit an allele that is identical by descent. The  $u$  coefficient is the probability that individuals  $x$  and  $y$  inherit the same genotype. The probability of inheriting the same *genotype* depends on the probability that *both* alleles in the genotypes of  $x$  and  $y$  are identical by descent. Two alleles could be identical by descent through the parents in the left

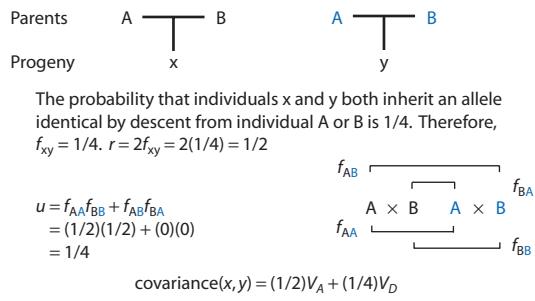
(a) General case



(b) Half siblings



(c) Full siblings



**Figure 10.6** The expected covariance in genotypic values for relatives based on probabilities that individuals share alleles ( $r$ ) and genotypes ( $u$ ) that are identical by descent. (a) The pedigree for the general case. (b) Half brothers and sisters share parent B in common and are thus unilinear relatives. (c) Full brothers and sisters share both parents in common and are thus bilinear relatives. In general, bilinear relatives include dominance components in their expected covariances.

**Table 10.8** Expected covariance in genotypic values between groups of relatives.

Relatives		Covariance in genotypic values
Offspring ( $x$ )	One parent ( $y$ )	$\frac{1}{2}V_A$
Offspring ( $x$ )	Mid-parent ( $y$ )	$\frac{1}{2}V_A$
Half siblings		$\frac{1}{4}V_A$
Full siblings		$\frac{1}{2}V_A + \frac{1}{4}V_D$
Nephew/niece ( $x$ )	Uncle/aunt ( $y$ )	$\frac{1}{4}V_A$
First cousins		$\frac{1}{8}V_A$
Monozygotic twins		$V_A + V_D$

positions of each pedigree ( $f_{AB}$ ) and at the same time through the parents in the right positions of each pedigree ( $f_{BD}$ ), giving  $f_{AC}f_{BD}$  as the probability that the genotype is identical by descent. Alternatively, two alleles could be identical by descent through the parents in the left and right positions of each pedigree ( $f_{AC}$ ) and through the two parents in the right and left positions of each pedigree ( $f_{BD}$ ), giving  $f_{AD}f_{BC}$  as the probability that the genotype is identical by descent. Since both outcomes can occur in the populations of  $x$  and  $y$  individuals, the total probability of genotypes being identical by descent is  $f_{AC}f_{BD} + f_{AD}f_{BC}$ . In all cases, if one parent of  $x$  and one parent of  $y$  are unrelated there is zero probability that the alleles they transmit to their progeny can be identical by descent.

Half and full siblings make instructive examples of how to determine the expected covariance between relatives. Figure 10.6 shows the pedigrees for these two cases. For both half and full siblings, the probability that individuals  $x$  and  $y$  inherit an allele identical by descent is  $\frac{1}{4}$ . Given that  $x$  and  $y$  have a parent in common, there is a probability of  $\frac{1}{2}$  that a given allele is transmitted from a common parent to individual  $x$  and an independent probability of  $\frac{1}{2}$  that a copy of the same allele is transmitted to individual  $y$ , giving  $f_{xy} = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ . For half siblings, parents A, B, and C are unrelated so there is zero probability that they transmitted alleles identical by descent to their offspring. There is, however, a probability of  $\frac{1}{2}$  that both half sibs inherited the same allele from parent B, but this is only one allele and not a genotype. For full siblings, while A and B are unrelated, A and B are the parents of both  $x$  and  $y$ . It is therefore possible that  $x$  and  $y$  inherited the same allele from parent A and the same allele from parent B to produce genotypes that are identical by descent.

Half siblings resemble each other since 50% of individuals share one allele that is identical by descent. Full siblings have an even greater degree of resemblance since 50% of individuals share an *allele* identical by descent and at the same time 25% of individuals share a *genotype* identical by descent. The variance in genotypic value caused by alleles and genotypes corresponds to  $V_A$  and  $V_D$ . Therefore, the genotypic values of half siblings have a covariance  $\frac{1}{2}V_A$  while the genotypic values of full siblings have a greater covariance of  $\frac{1}{2}V_A + \frac{1}{4}V_D$ . Table 10.8 gives additional examples of the expected covariance between various relatives based on the same logic used to obtain the covariances for half and full siblings.

Unilineal relatives such as half siblings can share only one allele in their genotypes that is identical by descent. This is the case since one of the parents of each individual are related and can provide an avenue for inheritance of an allele that is identical by descent. The other allele in the genotype cannot be identical by descent because they are inherited from two different parents who are unrelated. In contrast, bilineal relatives, such as full siblings, share both parents in common or have parents who are related. Bilinear relatives therefore have dominance components in their expected covariances because they have a chance of inheriting both alleles and genotypes that are identical by descent. Sharing alleles in common leads to phenotypic resemblance due to the additive phenotypic effects of alleles while sharing genotypes in common leads to phenotypic resemblance due to the phenotypic effects of genotypes (dominance and epistasis).

With knowledge of the genetic basis of covariance in genotypic values among relatives, we can look back to Fig. 9.8 to better understand why Abney et al. (2001) decided to estimate heritabilities in a

Hutterite population. The 806 individuals in that study descended from only 64 ancestors. That means that all individuals in the study had a non-zero probability of sharing two alleles that were identical by descent. The consequence is that the  $u$  coefficient in equation 10.57 was non-zero for all pairs of individuals in the study. This led to improved precision for estimates of dominance variance because comparisons between all pairs of individuals could be used to estimate the covariance between phenotype and the  $u$  coefficient. In contrast, the  $u$  coefficient is zero in randomly mating populations except for between pairs of full siblings, making estimates of dominance variance imprecise because the number of full siblings in one family is quite small.

Since the expected covariance between the mid-parent value and progeny values forms the basis of the parent–offspring regression, it is worth working through an additional method to obtain the covariance between mid-parent and offspring values. Deriving this covariance relies on a mathematical property of the covariance, or what some might call a math trick. The covariance can be expressed in a different form as

$$\begin{aligned}\text{cov}(x, y) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n (x_i y_i) - (\bar{x} \bar{y})\end{aligned}\quad (10.60)$$

In terms of the mid-parent and offspring values, the covariance is then

$$\text{cov}(O, \bar{P}) = (O\bar{P}) - M^2 \quad (10.61)$$

where  $O$  is the value of the offspring from each parental mating,  $\bar{P}$  is the mid-parent value of a mating between two parental genotypes, and  $M$  is the population mean. In the case of parents and offspring, the average taken by multiplication by  $1/n$  in equation 10.60 is instead accomplished by multiplying the product of  $\bar{P}$  and  $O$  by the expected frequency of each parental mating. Table 10.9 gives the expected frequencies for each union of parental genotypes under random mating, along with the expected mid-parent and progeny values. Multiplying the appropriate three quantities from each row in Table 10.9 and then summing across rows

$$\begin{aligned}\text{cov}(O, \bar{P}) &= p^4 a^2 + 4p^3 q \left( \frac{1}{4}(a^2 + 2ad + d^2) \right) \\ &\quad + 2p^2 q^2 (0) + 4p^2 q^2 \left( \frac{1}{2}d^2 \right) \\ &\quad + 4pq^3 \left( \frac{1}{4}(a^2 - 2ad + d^2) \right) + q^4 a^2 - M^2\end{aligned}\quad (10.62)$$

With some algebraic manipulation, canceling of terms and substitution of the definition of the population mean, the covariance between the mid-parent and offspring values is

$$\text{cov}(O, \bar{P}) = pq(a + d(q - p))^2 \quad (10.63)$$

Since  $\alpha = a + d(q - p)$  and  $V_A = 2pqa^2$ , after substitution the variance becomes

$$\text{cov}(O, \bar{P}) = \frac{1}{2} V_A \quad (10.64)$$

**Table 10.9** Frequencies and mean values for parents and progeny used to derive the covariance between the average value of parents (mid-parent value) and the average value of the progeny from each parental mating.

Parental mating	Parental mating frequency	Mid-parent value ( $\bar{P}_i$ )	Progeny genotype frequencies			Progeny value ( $O_i$ )
			$A_1A_1$	$A_1A_2$	$A_2A_2$	
$A_1A_1 \times A_1A_1$	$p^4$	$a$	1	–	–	$a$
$A_1A_1 \times A_1A_2$	$4p^3 q$	$\frac{1}{2}(a + d)$	$\frac{1}{2}$	$\frac{1}{2}$	–	$\frac{1}{2}(a + d)$
$A_1A_1 \times A_2A_2$	$2p^2 q^2$	$a + (-a) = 0$	–	1	–	$a + (-a) = 0$
$A_1A_2 \times A_1A_2$	$4p^2 q^2$	$d$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}d$
$A_1A_2 \times A_2A_2$	$4pq^3$	$\frac{1}{2}(-a + d)$	–	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}(-a + d)$
$A_2A_2 \times A_2A_2$	$q^4$	$-a$	–	–	1	$-a$

### Chapter 10 review

- Genotypic values can be expressed on an arbitrary scale with  $+a$  and  $-a$  representing the values of the homozygotes with respect to a midpoint between  $+a$  and  $-a$  of zero. The value of the heterozygote is represented by  $d$ .
- The mean phenotypic value of a population with Hardy–Weinberg expected genotype frequencies is  $M = a(p - q) + 2pqd$  where  $p$  is the frequency of the allele that increases value and  $q$  is the frequency of the allele that decreases value.
- The average effect of an allele is the difference between the mean value of that subset of genotypes that contain a given allele and the mean value of the entire population. Because alleles are inherited and genotypes are not, the average effect describes the mean value of those progeny that inherit a certain allele from one parent and the other allele sampled at random from the population.
- The genotypic mean is the sum of the breeding value and the dominance deviation.
- The breeding value of an individual is the mean value of the progeny produced by a given genotype assuming random mating. The breeding value is the sum of the average effects of the alleles in an individual's genotype.
- The dominance deviation is the difference between the genotypic value of a given genotype and the breeding value. When there is dominance ( $d \neq 0$ ), those progeny that are heterozygotes will have a value that is not the sum of the two alleles in their genotype.

- Since the mean breeding value of all genotypes is zero in a population under random mating, the mean value of a population should not change from one generation to the next under the many assumptions of Hardy–Weinberg.
- The variance in genotypic values around the population mean under random mating, commonly called the total genetic variance  $V_G$ , is the sum of the squared breeding values and the squared dominance deviations or  $V_G = V_A + V_D$ .
- The resemblance in genotypic value between relatives caused by  $V_A$  and  $V_D$  can be related to the probability of identity by descent. The expected covariance in genotypic values is a function of the probability that individuals share an allele that is identical by descent plus the probability that individuals share a genotype identical by descent. This covariance in genotypic values forms the basis of the parent–offspring method to estimate heritability.

### Further reading

For more details on the Mendelian basis of interaction variance ( $V_I$ ) as well as numerous perspectives on the role of epistasis in the evolutionary change of phenotypes, see chapters in:

Wolf JB, Brodie ED III, and Wade MJ (eds) 2000. *Epistasis and the Evolutionary Process*. Oxford University Press, Oxford.

## Problem box answers

### Problem box 10.1 answer

The entire range of genotypic values is  $2a = 30 - 9 = 21$ . Therefore,  $a = 10.5$  kg. The midpoint is then either  $30 - 10.5 = 19.5$  or  $9 + 10.5 = 19.5$ . The genotypic value of the heterozygote relative to the midpoint is  $d = 24.75 - 19.5 = 5.25$ . The degree of dominance is  $5.25/10.5 = 0.50$  or 50%.

### Problem box 10.2 answer

Case (a):

$$\begin{aligned} A_2 \quad M_{A_2} &= pd - qa = (0.5)(0.0) - (0.5)(10.5) = -5.25 \text{ kg} \\ \alpha_2 &= M_{A_2} - M = -5.25 - 0.0 = -5.25 \text{ kg} \\ \alpha_2 &= -p(a + d(q - p)) = -0.5(10.5 + 0.0(0.5 - 0.5)) = -5.25 \text{ kg} \\ \alpha &= a + d(q - p) = 10.5 + 0.0(0.5 - 0.5) = 10.5 \text{ kg} \\ \alpha_1 &= -p\alpha = -(0.5)(10.5) = -5.25 \text{ kg} \end{aligned}$$

Case (b):

$$\begin{aligned} A_2 \quad M_{A_2} &= pd - qa = (0.9)(0.0) - (0.1)(10.5) = -1.05 \text{ kg} \\ \alpha_2 &= M_{A_2} - M = -1.05 - 8.4 = -9.45 \text{ kg} \\ \alpha_2 &= -p(a + d(q - p)) = -0.9(10.5 + 0.0(0.1 - 0.9)) = -9.45 \text{ kg} \\ \alpha &= a + d(q - p) = 10.5 + 0.0(0.1 - 0.9) = 10.5 \text{ kg} \\ \alpha_1 &= -p\alpha = -(0.9)(10.5) = -9.45 \text{ kg} \end{aligned}$$

Case (c):

$$\begin{aligned} A_2 \quad M_{A_2} &= pd - qa = (0.5)(5.25) - (0.5)(10.5) = -2.625 \text{ kg} \\ \alpha_2 &= M_{A_2} - M = -2.625 - 2.625 = -5.25 \text{ kg} \\ \alpha_2 &= -p(a + d(q - p)) = -0.5(10.5 + 5.25(0.5 - 0.5)) = -5.25 \text{ kg} \\ \alpha &= a + d(q - p) = 10.5 + 5.25(0.5 - 0.5) = 10.5 \text{ kg} \\ \alpha_1 &= -p\alpha = -(0.5)(10.5) = -5.25 \text{ kg} \end{aligned}$$

Case (d):

$$\begin{aligned} A_2 \quad M_{A_2} &= pd - qa = (0.9)(5.25) - (0.1)(10.5) = 3.675 \text{ kg} \\ \alpha_2 &= M_{A_2} - M = 3.675 - 9.345 = -5.67 \text{ kg} \\ \alpha_2 &= -p(a + d(q - p)) = -0.9(10.5 + 5.25(0.1 - 0.9)) = -5.67 \text{ kg} \\ \alpha &= a + d(q - p) = 10.5 + 5.25(0.1 - 0.9) = 6.3 \text{ kg} \\ \alpha_1 &= -p\alpha = -(0.9)(6.3) = -5.67 \text{ kg} \end{aligned}$$

## CHAPTER 11

# Historical and synthetic topics

### 11.1 Historical controversies in population genetics

- The classical and balance hypotheses.
- How to explain levels of allozyme polymorphism.
- Genetic load.
- The selectionist/neutralist debates.

Although past debates in population genetics might not seem relevant today, it is usually the case that history has a strong influence on the present. This is certainly true in population genetics. Often, contemporary approaches to problems or accepted interpretations of a pattern are the products of rich and sometimes contentious debates. New data types such as DNA sequences and the sheer volume of genetic data available today have put to rest some open questions of the past. However, it is often the case that older controversies take new forms with some modifications, often because old problems are not completely resolved even though the field moves on to new topics. An appreciation of some of the ideas that have occupied population genetics in the past provides invaluable perspective on the present and future of population genetics. The brief and selective history presented here starts in the 1940s and 1950s and is meant to provide an overview of the spirit of past disagreements rather than a rigorous review. Note that the technical foundations involved in these topics are presented in earlier chapters as indicated but are not repeated in detail here. Readers interested in a history of early population genetics from Darwin through the 1930s should refer to Provine (1971).

#### *The classical and balance hypotheses*

The theoretical work of Fisher, Haldane, and Wright established the core principles of population genetics. These included that natural selection was able to alter allele frequencies rapidly, mutation and recombination supply genetic variation, that mating patterns

and gene flow shape the hierarchical organization of genetic variation, and that the effective population size regulates the process of genetic drift. Taken collectively, this body of theoretical expectations served to fuse Darwin's concept of natural selection with the principles of Mendelian particulate inheritance. These expectations form the foundation of population genetics and were labeled **neo-Darwinism** by Huxley (1942).

While the neo-Darwinian synthesis achieved by population genetics reached orthodoxy in the 1930s and 1940s, a long-running debate began to take shape. Under the logic of early neo-Darwinism, natural selection was the dominant evolutionary force in almost all aspects of evolutionary change. It was then a matter of debate as to what type of natural selection – directional or stabilizing – was most common in captive and natural populations. The answer to this question gradually turned into two broad points of view based on what one assumed and how one interpreted available data on genetic variation. Dobzhansky (1955) labeled these schools of thought the **classical hypothesis** and the **balance hypothesis**. Both hypotheses rely on natural selection as the principal process operating in populations, although they differ greatly in the predicted consequences of natural selection.

**Classical hypothesis** The point of view that directional natural selection is the dominant process in populations, predicting relatively little genetic variation except when selection pressures are heterogeneous in time or space or are frequency-dependent.

**Balance hypothesis** The point of view that balancing natural selection is the dominant process in populations, predicting extensive genetic variation caused by overdominance for fitness.

The classical hypothesis was that directional selection was the predominant process in populations and from this two major predictions arose as a consequence. The first prediction was that under random mating populations contained individuals homozygous at most loci. The second prediction was that populations harbored relatively little genetic variation since the equilibrium points for any sort of directional selection on a diallelic locus are fixation and loss or near fixation and loss (see Chapter 6). The classical school recognized the existence of “wild-type” alleles, or alleles at high frequency in a population because such alleles were of higher fitness and were brought to high frequency by directional selection. Alternative “mutant” alleles that appeared in populations were most often deleterious but on very rare occasions would have a higher fitness than the current wild-type allele and would then become the new wild-type allele. The classical school predictions were supported by a range of empirical observations, especially from laboratory populations of organisms such as *Drosophila*. In such populations, phenotypes are of the wild type (within some range of variation) and mutations with visible phenotypic effects appear rarely but are almost universally deleterious and do not reach high frequencies.

The classical hypothesis predicted that genetic variation in populations was produced in four ways (Dobzhansky 1955). First, deleterious mutations continually occur and segregate for a short time before they are eliminated by directional natural selection. Most of these deleterious mutations are likely to be recessive and thus exist mostly in heterozygote genotypes where they are sheltered from natural selection. (Dobzhansky pointed out that these are the sorts of mutations that cause hereditary diseases when homozygous.) Second, some proportion of mutations are adaptively neutral because they have marginal fitness values very near the mean. A third possibility is that rare beneficial mutations are found in a population before they have become established as the new wild-type allele. The fourth possibility is that some mutations are slightly beneficial in one environment but slightly deleterious in another environment. Alleles will then persist in a population exposed to environments that are heterogeneous in time or space. This last category motivated numerous population genetics models where directional selection occurs but fitness varies in time or space, or individual fitness is frequency-dependent (see Chapter 7).

The balance hypothesis took the alternate perspective that overdominance for fitness was the

general rule in most populations, so that balancing selection was the dominant process that regulated genetic variation. Under balancing selection, heterozygotes would have higher frequencies than in the absence of selection or under directional selection (see Chapter 6). The balance hypothesis, therefore, predicted that loci would maintain two or more alleles indefinitely. Owing to balancing selection, heterozygotes would be more frequent and homozygotes much less frequent than expected by Hardy–Weinberg or under directional selection. Of the new mutations that enter a population, only those that exhibited overdominance as heterozygotes were expected to be retained in the population.

As Dobzhansky (1955) explained, the balance hypothesis was also associated with predictions about the inter-relationship among loci. Natural selection on multiple loci can lead to the accumulation of gametic disequilibrium (see Chapter 2). High levels of gametic disequilibrium are expected under balancing selection in populations with all loci at intermediate allele frequencies since only that subset of gametes that produce multilocus heterozygote zygotes would have high fitness. (Note that under the classical hypothesis there is also strong natural selection, but relatively little gametic disequilibrium in absolute terms is expected because populations would be close to fixation for the wild-type allele.) Using this expectation for gametic disequilibrium and then assuming that most loci experience balancing selection leads to the prediction of **coadapted gene complexes** or **supergenes** within species (reviewed by Hedrick et al. 1978). A supergene is a haplotype or genotype at many loci that is held together and frequently inherited as an intact unit because gametic disequilibrium is very strong. A coadapted gene complex is a supergene where natural selection acts or has acted so that the alleles or genotypes at each locus have high fitness in the context of the alleles or genotypes at all other loci. Stated another way, selection will increase the frequency of new mutations that interact well with the alleles and heterozygous genotypes at other loci. In contrast, any mutations that have reduced relative fitness caused by interactions among loci will be reduced in frequency by natural selection. Thus, the notion of a coadapted gene complex assumes that epistasis for fitness is common.

Supergenes and coadapted gene complexes presented both a research agenda and a conceptual challenge to biologists of the classical/balance hypothesis era. A great deal of research was devoted to the study of multilocus genetic variation in laboratory

and natural populations. At the same time, numerous models of multilocus natural selection and recombination were developed and studied. Very high levels of gametic disequilibrium caused by balancing selection and epistasis for fitness served to negate the Mendelian process of independent assortment. What then was the source of genetic variation required for evolutionary change? The answer was often sought in population genetic mechanisms that had the potential to recombine or break up supergenes.

Although the notion that large supergenes or coadapted gene complexes are widespread is not popular now, contemporary population genetics has inherited an appreciation of the processes that cause gametic disequilibrium and evidence that multiple loci are not necessarily independent. We now have well-characterized examples of genome regions with high levels of gametic disequilibrium. One of the best examples is the major histocompatibility complex (*Mhc*) loci in mammals. These loci experience balancing selection because of their functional role in recognizing non-self peptide fragments and compose a large chromosomal region that has relatively high levels of gametic disequilibrium. The amorphous supergene prediction has now been refined into a series of much more specific hypotheses tailored to diverse areas of population genetics. Non-independence of loci is central to models of molecular evolution that seek to explain polymorphism within populations, as embodied by concepts such as hitchhiking, background selection, and genetic draft (see Chapter 8). Quantitative genetics recognizes non-independence of traits caused by phenotypic and genetic correlations. The idea that selection favors alleles that interact well across multiple loci is now called the **Dobzhansky-Muller model** and it serves as an explanation of how isolated populations might develop reproductive isolation that leads to speciation (reviewed by Coyne and Orr 2004).

The study of **ecological genetics** can be traced to efforts to test the classical and balance hypotheses with empirical data. Today, ecological genetics is defined as the study of genetic variation within species in the context of environmental and organismal interactions. Ecological genetics seeks to identify the causes of patterns of genetic variation, often with reference to the assumed or demonstrated pressures of natural selection imposed by ecological context. Early ecological genetics was focused on testing the classical and balance hypotheses for genetic variation. On the one hand was the classical school prediction that relative fitness of alleles varied in time and space.

On the other hand, the balance school predicted that overdominance for fitness was very common. Both of these possibilities were testable to some extent in natural populations, by measuring the relative fitness of phenotypes with a known genetic basis or observing the frequency of genetic polymorphisms. Dobzhansky was among the first to study “laboratory” organisms in the wild. He pioneered field research in *Drosophila* and established a tradition of empirical research that is now the norm in population genetics. Edmund B. Ford was also instrumental in the establishment of the field of ecological genetics. Ford studied wild butterflies and moths and wrote the enormously influential book *Ecological Genetics*, which was published in 1964.

Many of the widely known empirical studies in ecological genetics take on new meaning when viewed through the lens of the classical hypothesis/balance hypothesis debate. For example, industrial melanism in populations of spotted moth (also known as the peppered moth) in England was evidence for the classical-school position since it shows that directional selection pressures vary among populations based on proximity to industrial centers whose soot stained tree trunks black (reviewed by Majerus 1998). (It is no coincidence that Bernard Kettlewell, who performed much of the original spotted moth work as a research scientist at the University of Oxford, was supervised by E.B. Ford.) Another universally known example in ecological genetics, evidence that heterozygous blood-group protein genotypes have higher fitness in malarial areas of Africa, supports the balance hypothesis of overdominance for fitness as the force that maintains genetic variation.

### ***How to explain levels of allozyme polymorphism***

Another long-running controversy in population genetics grew out of the classical hypothesis/balance hypothesis debate. The new controversy revolved around how to explain genetic polymorphism within natural populations observed with a then radically new technique. The technique was gel electrophoresis of enzyme polymorphisms, or allozymes (see Box 2.2). Two papers published in 1966 ushered in the new controversy. Hubby and Lewontin (1966) presented allozyme estimates of heterozygosity for 21 loci estimated from multiple populations of 15–20 *Drosophila pseudoobscura* individuals. Nine of these 21 loci exhibited between two and six alleles segregating within populations. The Hubby and Lewontin paper showed a technique that could be used to

determine both the proportion of loci that possessed more than one allele and the level of heterozygosity for each polymorphic locus.

The controversy over the causes of allozyme polymorphism changed the focus of much of population genetics within the span of only a few years starting in the mid-1960s. Initially, the classical and balance hypotheses were considered as primary explanations. In fact, in a paper published along with the allozyme data themselves, Lewontin and Hubby (1966) argued that the level of heterozygosity observed (averaged over populations 30% of loci were polymorphic) were inconsistent with the balance hypothesis because of the segregation load that would have been required (see the Genetic load section, below). The remaining explanation for the allozyme polymorphism within the context of the time was directional natural selection consistent with the classical hypothesis.

The balance hypotheses experienced some setbacks from empirical data around the same time. Apparent overdominance in maize was shown to decline over multiple generations (Moll et al. 1964; reviewed by Crow 1993b). True overdominance should persist indefinitely as a function only of heterozygosity. These maize results, however, were consistent with the prediction that overdominance was actually caused by gametic disequilibrium between loci bearing beneficial dominant alleles and other loci bearing deleterious recessive alleles. When two individuals homozygous for different alleles at two such loci are crossed, the progeny will experience a great increase in fitness because the recessive deleterious phenotype will be masked by dominance. The maize results demonstrated that apparent overdominance phenomena were caused by the combination of simple dominance and linkage rather than by true overdominance.

The classical hypothesis/balance hypothesis debate soon receded. Selective neutrality was an explanation under the classical hypothesis that predicted a low level of genetic variation in populations. This idea of selectively neutral alleles, which was developed and mathematically formalized starting in the 1950s and 1960s, emerged as a primary hypothesis for genetic polymorphism. The neutral theory hypothesized that many loci have selectively neutral alleles and that polymorphism was a product of the non-equilibrium random walk that new neutral mutations experience because of genetic drift.

The waning of the balance hypothesis and the ascension of the two components of the classical hypothesis produced what was labeled the **neo-classical theory** of population genetics by Lewontin (1974).

This label came about because both elements of the classical hypothesis explanation for polymorphism – selectively neutral mutations and mutations under directional or purifying selection – are drawn from the early classical hypothesis. Under the neo-classical hypothesis, the debate became one about the relative contributions of neutral mutations or mutations acted on by natural directional selection to levels of genetic polymorphism. There was also continuing work on the selection element of the classical hypothesis, which was updated and bolstered with empirical support from more elaborate theoretical models and ecological genetic studies.

In the years after 1966, until DNA-based molecular techniques became available in the late 1980s, allozyme electrophoresis was perhaps the most widely used empirical technique in population genetics. The allozyme era of population genetics is sometime derisively referred to as the period of “find ‘em and grind ‘em”, in reference to collecting and then homogenizing samples in preparation for allozyme electrophoresis. It is certainly true that during this time a great deal of empirical research focused on gathering single-population estimates of heterozygosity via allozyme electrophoresis. However, these data were part of a very substantial shift in the conceptual focus of population genetics. The addition of allozyme data contributed to the shift in emphasis away from the entrenched positions of those who favored the classical or balance hypotheses to new models, data, and hypothesis tests.

### ***Genetic load***

For natural selection to change allele frequencies in a population, individuals of some genotypes must experience higher rates of death (either actual for viability selection or reproductive for fecundity selection) than other genotypes. Natural selection works by culling individuals of some genotypes in favor of individuals of other genotypes, increasing the mean fitness in the process. The amount of death or failed reproduction associated with natural selection was first called the **load** by Muller (1950). Genetic load comes in two forms. **Substitutional load** refers to the reduction in mean fitness caused during fixation of beneficial mutations or purging of deleterious mutations. Distinctly, the production of individuals with lower-fitness genotypes by Mendelian segregation during reproduction is called the **segregational load**. Sexual reproduction leads to segregational load because both recombination and independent

assortment produce novel progeny genotypes that have a range of fitnesses. Those progeny genotypes with a lower fitness perish (or do not reproduce) under viability selection. In principle, the genetic load places an upper limit on the ability of natural selection to change genotype frequencies in a population. Deaths due to viability selection cannot greatly exceed the total demographic excess of a population (the number of individuals produced each generation beyond those needed for demographic replacement) for long or the population will go extinct.

The genetic load has been a tool used to try to estimate the upper limits to the process of natural selection. The goal has been to determine how strong natural selection can be before an unrealistic genetic load occurs. The genetic load has been used for two main purposes. The substitutional load has been used to estimate mutation parameters in populations, such as the rate of fixation of beneficial mutations, the rate of deleterious mutations, or the decline in fitness associated with deleterious mutations. Substitutional load also played a prominent role in attempts to set acceptable thresholds for human radiation exposure during the 1950s. The segregational load was used as a counter-argument against the balance hypothesis during the 1960s in the early days of the neutral theory of molecular evolution. In these roles genetic load has been controversial and engendered passionate arguments over the span of many decades (see reviews by Wallace 1991; Crow 1993b). The genetic load has been employed widely in population genetics and evolutionary biology during the last three decades. For example, the concept has been invoked in the accumulation of deleterious alleles that might cause mutational “meltdowns” in small and endangered populations (e.g. Lynch and Gabriel 1990), and in arguments for the fitness advantage of sexual reproduction.

The genetic load concept (although not the term) was originated by Haldane (1937). Haldane’s result can be seen with the general dominance model of selection on a single diallelic locus (see Table 6.4). Assume a population at Hardy–Weinberg equilibrium, with complete dominance ( $h = 0$ ), and a maximum fitness that equals one. In such a population the fitness-weighted genotype frequencies are  $p^2$ ,  $2pq$ , and  $q^2 - sq^2$  where  $s$  is the selection coefficient. The mean fitness in the population is then the sum of the frequency-weighted fitness values or  $\bar{w} = p^2 + 2pq + q^2 - sq^2$ . The mean fitness is therefore  $\bar{w} = 1 - sq^2$  because  $p^2 + 2pq + q^2 = 1$ . The process of forward mutation (A to a) will make new recessive alleles in

the population each generation, transforming some number of Aa genotypes into aa genotypes. Assuming that natural selection and mutation are at equilibrium means that selection removes aa genotypes as fast as they are made by mutation. The rate at which new aa genotypes are made from Aa genotypes is the forward mutation rate or  $\mu$ . The mean fitness of the population is then  $\bar{w} = 1 - \mu$  (also assuming that reverse mutation can be ignored). With incomplete dominance ( $h \neq 0$ ),  $\bar{w} = 1 - 2pqhs - sq^2$  and the mean fitness can be shown to be approximately  $\bar{w} \approx 1 - 2\mu$ .

The genetic load is defined as

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}} \quad (11.1)$$

and expresses the difference between the maximum fitness ( $w_{\max}$ ), which corresponds to the most fit genotype, and the average fitness ( $\bar{w}$ ) in a population at a given point in time (Crow 1958). If  $w_{\max}$  is defined as one, then the genetic load is given more simply by

$$L = \frac{1 - \bar{w}}{1} = 1 - \bar{w} \quad (11.2)$$

Substituting the mean fitness from Haldane’s result above and recalling that  $w_{\max} = 1$ , the genetic load under complete dominance is  $L = 1 - (1 - \mu) = \mu$  and under incomplete dominance  $L = 1 - (1 - 2\mu) \approx 2\mu$ . The straightforward conclusion is that higher mutation rates lead to higher genetic loads because a larger number of deleterious mutations must be purged by natural selection to remain at selection/mutation equilibrium. This translates into some proportion of homozygous individuals that must die or fail to reproduce each generation. The load never disappears because while selection removes deleterious alleles from a population, mutation continually supplies new ones.

Later, Haldane (1957) used a mutational-load argument to estimate the rate of substitutions. His goal in this analysis was to understand the rate of *phenotypic* evolution, perceived at that time to be relatively slow, consistent with Darwinian emphasis on gradual change. Haldane concluded that natural selection could accomplish no more than about one beneficial substitution every 300 generations and that this would require the deaths of 30 times the population size present in one generation. This result made a major impact on some researchers in population genetics. Haldane’s conclusion was often

applied very generally as a fundamental limit on the rate of natural selection. As shown later by Ewens (2004), Haldane's result implicitly assumed that demographic excess in humans is limited to 10% of the population size and so was not as general as some had originally thought.

**Segregational load** The decrease in the mean fitness in a population caused by individuals with low-fitness genotypes that are introduced in a population by each generation of Mendelian segregation.

**Substitutional load** The decrease in the mean fitness in a population caused by the introduction of deleterious mutations or the culling of lower-fitness genotypes required for the eventual substitution of beneficial mutations.

To understand the segregational load, assume a standard diallelic locus model of overdominance for fitness where relative fitness values are  $1 - s$  for the AA genotype, 1 for the Aa genotype, and  $1 - t$  for the aa genotype, where  $s$  and  $t$  are selection coefficients. As shown in Chapter 6, the expected equilibrium allele frequencies for this model of natural selection are

$$p_{eq} = \frac{t}{s+t} \quad (11.3)$$

and

$$q_{eq} = \frac{s}{s+t} \quad (11.4)$$

Using these equilibrium allele frequencies, it is possible to express the equilibrium frequency of heterozygotes expected under balancing selection as

$$\begin{aligned} H_{eq} &= 2p_{eq}q_{eq} = 2\left(\frac{t}{s+t}\right)\left(\frac{s}{s+t}\right) \\ &= \frac{2st}{(s+t)^2} \end{aligned} \quad (11.5)$$

This shows that the frequency of heterozygotes in a population depends on the magnitude of the selection coefficients against the homozygous genotypes.

For example, if both homozygotes have 10% lower viability than the heterozygote,  $s = t = 0.1$  and at equilibrium half of the population is composed of heterozygotes ( $H_{eq} = \frac{2(0.1)(0.1)}{(0.1+0.1)^2} = 0.5$ ). Weaker selection results in less heterozygosity at equilibrium while stronger selection results in more equilibrium heterozygosity.

We can combine the expected frequency of heterozygotes at equilibrium and the mean fitness to get an expression for the genetic load at equilibrium under balancing selection in terms of the homozygote selection coefficients:

$$\bar{w} = 1 - \frac{st}{s+t} \quad (11.6)$$

as worked out in Math box 11.1. This population mean fitness can also be expressed in terms of the heterozygosity maintained by balancing selection by utilizing the equilibrium frequency of heterozygotes given in equation 11.6. Notice that

$\frac{st}{s+t} = \left(\frac{2st}{(s+t)^2}\right)\left(\frac{s+t}{2}\right)$  so that the mean fitness for one locus can be written as

$$\bar{w} = 1 - H_{eq} \left(\frac{s+t}{2}\right) = 1 - H_{eq} \bar{s} \quad (11.7)$$

This equation has the biological interpretation that the mean fitness at equilibrium in a population under balancing selection is one minus the product of the equilibrium heterozygosity and the mean selection coefficient. A population experiencing balancing selection therefore always has at least some genetic load since the mean fitness will never reach the maximum fitness of one (mating among heterozygotes will generate additional lower-fitness homozygotes every generation).

Let's use some data on heterozygosity and selection coefficients similar to those available in the 1960s to compute the segregational load caused by balancing selection. Allozyme surveys in *Drosophila* of the time estimated that average heterozygosity was around 0.3. A homozygote disadvantage of 10% was considered reasonable under the balance hypothesis, giving a selection coefficient of  $s = 0.10$ . Putting these two values into equation 11.7 gives the mean fitness as

### Math box 11.1

#### Mean fitness in a population at equilibrium for balancing selection

To solve the equation for the mean fitness in terms of the selection coefficients  $s$  and  $t$  on the homozygous genotypes, start with the standard expression for the mean fitness:

$$\bar{w} = p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa} \quad (11.8)$$

and then substitute the fitness values for each genotype:

$$\bar{w} = p^2(1-s) + 2pq(1) + q^2(1-t) \quad (11.9)$$

Then express the genotype frequencies as equilibrium allele frequencies given in terms of the selection coefficients (equations 11.3 and 11.4):

$$\bar{w} = \left( \frac{t}{s+t} \right)^2 (1-s) + 2 \left( \frac{t}{s+t} \right) \left( \frac{s}{s+t} \right) (1) + \left( \frac{s}{s+t} \right)^2 (1-t) \quad (11.10)$$

and then multiply through to give

$$\bar{w} = \frac{t^2(1-s)}{(s+t)^2} + \frac{2st}{(s+t)^2} + \frac{s^2(1-t)}{(s+t)^2} \quad (11.11)$$

Expanding the numerators of the first and last terms gives

$$\bar{w} = \frac{t^2 - t^2s + 2st + s^2 - s^2t}{(s+t)^2} \quad (11.12)$$

Notice that  $s^2t + t^2s = st(s+t)$  and  $t^2 + 2st + s^2 = (t+s)(t+s)$  in the numerator, and so making these substitutions:

$$\bar{w} = \frac{(s+t)^2}{(s+t)^2} - \frac{st(s+t)}{(s+t)^2} \quad (11.13)$$

which on simplifying gives

$$\bar{w} = 1 - \frac{st}{s+t} \quad (11.14)$$

$$\bar{w} = 1 - (0.3)(0.1) = 1 - 0.03 = 0.97 \quad (11.15)$$

so that the genetic load is 0.03. Allozyme surveys of the time also suggested that about one-third of all loci in *Drosophila* had more than one allele segregating. Extrapolated to the entire genome, thought to be composed of around 8000 to 10,000 loci, it was believed that perhaps 2000–3000 loci were variable. If each locus is completely independent, then the segregational load for the entire genome is

$$L = (1 - (0.3)(0.1))^{3000} = (0.97)^{3000} \\ = 2.07 \times 10^{-40} \quad (11.16)$$

This produces a conclusion that the genetic load would be enormous. Interpreted in biological terms, this genetic load means that an individual heterozygous at 3000 loci would have to produce  $10^{40}$  progeny (the inverse of the load) for each progeny produced by an individual homozygous at all loci. Even using lower average heterozygosity per locus

and selection coefficients leads to impossibly high genetic loads.

Segregational and substitutional loads played an important role in attempts to explain the proportion of segregating loci and levels of heterozygosity in the 1960s. Balancing selection was considered and rejected by Lewontin and Hubby (1966) as a hypothesis to explain the first allozyme polymorphism data in *Drosophila*. Expectations for the substitutional and segregational loads were explored and developed in a series of papers authored and coauthored by Kimura (Kimura 1960, 1967; Kimura et al. 1963; Kimura and Maruyama 1966). The genetic load ultimately played a key part of the argument given by Kimura in his proposition of the neutral theory of molecular evolution (Kimura 1968). Kimura argued that there was too much genetic polymorphism or too fast a rate of divergence for all genetic changes to be caused by natural selection. This is because natural selection imparts a genetic load. Kimura's alternative was that many polymorphisms are selectively neutral. The neutral explanation greatly reduces the genetic load since only a small portion of polymorphisms would be acted on by selection and accrue a genetic load.

There are a series of counter-arguments to the idea that natural selection is limited by the segregation load (reviewed in Wallace 1991; Crow 1993b; section 2.11 in Ewens 2004). One criticism revolves around the point of reference used for the maximum fitness in a population ( $w_{\max}$  above). Haldane and Kimura defined load relative to the most fit genotype in the population. In the case of the balance hypothesis, the most fit genotype would be one that is heterozygous at all loci. However, such a genotype that is heterozygous at *all* loci would be very, very infrequent in an actual population if we assume that fitness is based on numerous loci. For example, imagine that all allele frequencies are equal to 0.5 and fitness is determined by 100 loci of equal effect. The expected frequency of a 100 locus heterozygote is  $(0.5)^{100} = 7.89 \times 10^{-31}$  in a randomly mating population. Ewens (reviewed in 2004) showed that if genotypes with fitness values four standard deviations greater than the population mean of one are used as a reference point, then  $w_{\max}$  equals 1.98. This implies that the most fit individuals need to produce about two progeny for every one progeny produced by average fitness individuals. This cost of selection seems tolerable for many populations and species.

Another counter-argument focuses on the form that natural selection takes while it works to cull

individuals with less-fit genotypes from a population. Estimates of genetic load commonly assume that every locus is independent, so that natural selection must act against homozygous genotypes at each and every locus independently. This is equivalent to assuming multiplicative fitness across multiple loci, seen as an exponent in equation 11.16. This assumption has the consequence of maximizing the perceived genetic load. Another possibility is that natural selection can cull less-fit genotypes by acting on several loci simultaneously. For example, if selection results in the death of an individual because it bears a homozygous genotype at one locus, it also has the effect of culling any other homozygous loci in that individual from the population at the same time. Thus, counting each homozygous locus toward the genetic load, without regard to the fact that some homozygous loci occur in the same individual and can be selected against by only one selective death, overstates the total genetic load.

A final category of arguments against the cost of selection limits involves the way in which natural selection works. A viability selection model assumes that individuals selected against will die before reproduction. An obvious alternative is that selection instead takes the form of fecundity selection among reproducing adults. Some degree of fecundity selection (in lieu of some amount of viability selection) would reduce the number of selective deaths required for substitution or segregation. Instead, genetic load would take the form of differences in fecundity among individuals of different genotypes. This is often described as the distinction between hard (viability) and soft (fecundity) forms of selection. Further, truncation selection as an alternative form of natural selection (see Fig. 9.10) was modeled in detail as an alternative to assumption of independent loci in estimates of the genetic load. The genetic load is considerably less under truncation selection, making the cost of selection low enough to be approximately consistent with the balance hypothesis. However, it is not clear whether truncation selection ever actually occurs in nature.

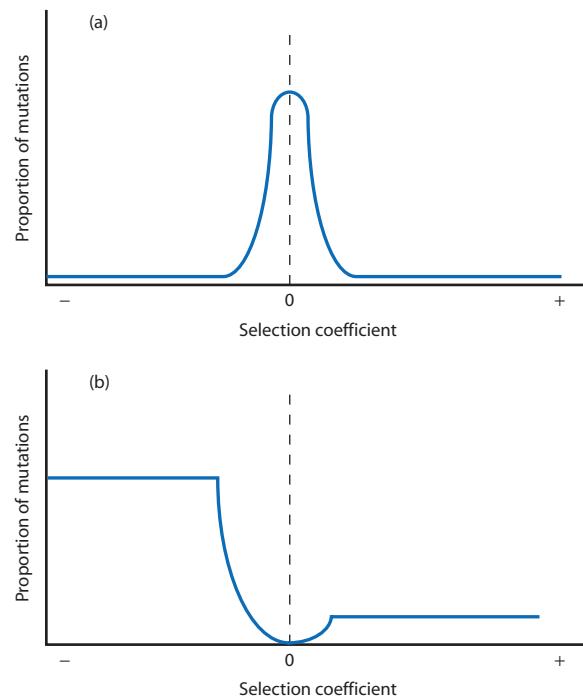
### **The selectionist/neutralist debates**

The neutral theory of molecular evolution was proposed by Kimura in 1968 (reviewed in Kimura 1983a), coupling his sophisticated knowledge of models of genetic drift with the then novel (and scarce) data on rates of amino acid divergence. Kimura used the amino acid divergence data in mammals

to estimate that the rate of nucleotide substitution genome-wide was about one site every other year. As discussed above, this implied a very large genetic load if natural selection was the principal process that governed the eventual substitution of new mutations. Kimura showed, instead, that if new mutations are neutral (meeting the condition that  $|2N_e s| \ll 1$ ) then the genetic load was low enough to seem reasonable. (In the process of making the load calculation he also showed that the rate of substitution of a neutral mutation was approximately the mutation rate.) In the same paper, Kimura considered the level of polymorphism in the data of Hubby and Lewontin (1966), estimating that an effective population size of between 2300 and 9000 would produce the levels of heterozygosity observed in *Drosophila* if alleles were selectively neutral.

The controversy caused by the neutral theory was both intense at times and long-standing. Defending and extending the neutral theory would occupy Kimura for the rest of his life. The proposal of the neutral theory ushered in a new era in population genetics that saw the development of numerous models constructed with the goal of explaining levels of polymorphism or rates of divergence (see Chapter 8). At the same time, there was an increasing volume of genetic data available to test population genetic models. New data often revealed patterns of polymorphism or divergence that were not strictly compatible with neutral theory, motivating continual extensions to the neutral theory. At the same time, numerous advances in the theory of natural selection at the molecular level (e.g. hitchhiking, codon bias, background selection) were developed as alternative hypotheses to the neutral theory. After Kimura's (1968) initial proposal of neutral theory, genetic load faded in importance while levels of polymorphism and rates of divergence became the primary issues.

**Pan-neutralism** and **pan-selectionism** are caricatures that illustrate some of the exaggerated stances taken in selectionist/neutralist debates. We can redraw the mutation fitness spectrum presented in Chapter 5 in a way that schematically represents the extreme positions of pan-neutralism and pan-selectionism (Fig. 11.1; compare with Fig. 5.1). Both of these points of view are extreme because they rely on a picture of the fitness of mutations that does not match most observations. The neutral theory was often misunderstood and taken by some as a proposal that *all* nucleotide or amino changes were selectively neutral (Fig. 11.1a). Pan-neutralism is



**Figure 11.1** The caricatures of the mutation fitness spectrum drawn to illustrate the extreme views of pan-selectionism and pan-neutralism. (a) Under a pan-neutralist view almost all mutations have little or no impact on fitness and so are selectively neutral. (b) Under a pan-selectionist view there would be almost no mutations that are selectively neutral, many mutations that are selected against, and a relatively high frequency of beneficial mutations. Neither of these extremes is supported by most observations, except perhaps in isolated cases. This illustration is inspired by figures in Turner (1992) and Crow (1972).

an inaccurate description of neutral theory and one that Kimura and others took pains to point out was not correct. Pan-neutralism cannot explain patterns of molecular evolution such as differences in divergence rates at synonymous and nonsynonymous sites within a gene (which share a common mutation rate and effective population size). Functional constraint due to natural selection (first suggested by King and Jukes 1969), an important part of neutral theory, can readily explain such observations. At the same time, the neutral theory helped illustrate implications of the classical and balance hypothesis perspectives, often to the discomfort of advocates of these points of view. Both of the classical and balance hypotheses emphasize natural selection and tend toward pan-selectionism. As Fig. 11.1b illustrates, arguing for a complete absence of neutral mutations is an unreasonable position because it requires a mutation fitness spectrum with a discontinuity around a selection coefficient of zero.

One part of the selectionist/neutralist debate involved the level of allozyme polymorphism (estimated by heterozygosity or gene diversity) in populations. Kimura and Crow (1964) had developed an expectation for the level of heterozygosity at drift–mutation equilibrium in the infinite alleles model (see Chapter 5). After the proposal of neutral theory, this expectation was applied to the growing body of protein polymorphism data. Observed polymorphism should match this prediction if neutral theory were correct. The problem at the time was that observed heterozygosities fell into too small a range (see Lewontin 1974). Based on mutation-rate estimates, the species sampled should have differed a great deal in their effective population size if mutations were neutral. Thus, there was perceived to be too little polymorphism to be consistent with neutral theory predictions. But at the same time, there was perceived to be too much polymorphism to be consistent with the classical hypothesis. Time has shown that estimates of key population parameters like mutation rates, assumed effective population sizes, and the estimates of polymorphism were sometimes imprecise enough to cause fairly large deviations from expectations. There was also the limitation that empirical data were not general since the sample of genes, population, and species was limited. It was during this period that Lewontin (1974) lamented that neutral theory expectations depended on the product of a large unknown number (the effective population size) and a small unknown number (the mutation rate), a comment he frequently repeated.

Numerous natural-selection-based explanations of polymorphism were developed as alternatives to or extensions of the neutral theory. One innovative hypothesis put forth to explain the deficit of polymorphism relative to neutral theory expectations was that of genetic hitchhiking around positively selected sites (Maynard Smith and Haig 1974). A reduction in polymorphism was predicted for neutral loci in gametic disequilibrium with positively selected loci, explaining a possible cause of reduced polymorphism. A decade later, Kreitman's (1983) DNA sequence data set showing polymorphism for the alcohol dehydrogenase or *Adh* locus in *Drosophila melanogaster* gave new life to the balance hypothesis. The *Adh* locus had originally been studied with allozyme techniques and was known to exhibit high levels of polymorphism. The *Adh* DNA sequences identified the basis of the allozyme polymorphism as a threonine/lysine difference and also exhibited synonymous variation around these nonsynonymous sequence changes. Overall, the

*Adh* sequences exhibited too much polymorphism to be consistent with strict neutrality but were consistent with a model of balancing selection. These data contributed to a renewal of the balance hypothesis and sparked continued debate about the processes regulating polymorphism at the DNA level.

Another part of the selectionist/neutralist debate revolved around rates of divergence. Neutral theory predicted that rates of divergence should be constant through time since the expected time between neutral substitutions is the inverse of the mutation rate (see Chapter 8). Amino acid sequence data collected in the 1970s and DNA sequence data collected in the 1980s showed that the variance in the number of substitutions (as measured by the index of dispersion; see Chapter 8) was too large to be compatible with neutral theory. One explanation for the variance was that the rates of fixation of new mutations was largely a consequence of natural selection. Since selection strength could vary for different mutations and also be variable through time and space, natural selection would cause variation in substitution rates. In particular, natural selection could cause periods of little or no substitution because deleterious mutations would be selected against while beneficial mutations would be fixed rapidly. The high variation in rates of substitution, called the overdispersed molecular clock, stimulated many extensions of the simple neutral model in an attempt to explain the degree of variation in substitution rates (see Gillespie 1991; Culter 2000a).

Originally, the neutral theory divided mutations into two discrete categories, those that were neutral and those that were either deleterious or advantageous and therefore acted on by natural selection. Kimura suggested that most mutations were in the neutral category. An alternative view was proposed by Ohta in an attempt to explain phenomena inconsistent with strict neutral theory (reviewed in Ohta 1992). Ohta suggested a third category of mutations that were weakly deleterious and therefore nearly neutral (see Chapter 8). A common misunderstanding of nearly neutral theory was the failure to distinguish between strict neutrality and near neutrality. A mutation is strictly neutral if there is no selection coefficient on it so that its fate is always dictated only by genetic drift. In contrast, the fate of a nearly neutral mutation depends on its effective population size context. With a selection coefficient that is small relative to the effective population size, the fate of a mutation is determined by genetic drift. Alternatively, the same mutation can be acted on by

natural selection if the selection coefficient is large compared to the effective population size.

The nearly neutral theory helps to explain why substitution rates are lower at nonsynonymous sites than at synonymous sites: a greater proportion of changes at nonsynonymous sites are deleterious and therefore fixation is prevented by natural selection. In contrast, synonymous sites are largely neutral and thus have rates of fixation determined by genetic drift. Codon bias (reviewed by Petrov 2008) is an alternative natural selection hypothesis for rates of synonymous substitution (unequal frequencies of tRNAs available during translation impose a fitness cost on infrequent codons because genes containing them are translated more slowly). The notion of constraint on DNA sequences is now accepted and used routinely to explain and predict rates of divergence and levels of polymorphism.

Nearly neutral theory also helped to explain the overdispersed molecular clock. Since rates of substitution for nearly neutral mutations would depend on the effective population size of a species, rates of substitution can vary among species because the effective population size also varies among species. A large part of the variance in substitution rates for synonymous substitutions is in fact explained by variance among species as predicted by nearly neutral theory. However, relatively little of the non-synonymous substitution rate variation is explained by variation among species, an observation that is consistent with natural selection acting on non-synonymous mutations.

The selectionist/neutralist debates produced a rich set of expectations for polymorphism and divergence that encompass the full range of population genetic processes and are able to explain many observed phenomena. Most contemporary population genetic models have the common feature that the strength of natural selection is expressed relative to the strength of genetic drift as measured by the effective population size. Natural selection and genetic drift are now seen as inseparably linked processes that lie on a continuum between the extremes of pure selection (large  $N_e s$ ) and pure drift (small  $N_e s$ ). While debate over the relative roles of natural selection and genetic drift in producing patterns of polymorphism or rates of divergence continues, discussions generally acknowledge the central role of genetic drift. The neutral theory is now almost universally utilized as a null hypothesis in all of population genetics and its varied empirical applications.

## 11.2 Shifting balance theory

- Sewall Wright's classic model of natural selection, genetic drift, gene flow, and mutation on an adaptive landscape.

If forced to choose a single model in all of population genetics that has had the longest-running impact, Sewall Wright's shifting balance model and its associated fitness surfaces would certainly be among the top picks. Wright first described the fitness surface in a 1932 presentation at the Sixth International Congress of Genetics. For that presentation, Wright (along with J.B.S. Haldane and R.A. Fisher) was asked to make a non-mathematical presentation of his theoretical work in population genetics. Wright's 1932 presentation and associated paper (Wright 1932) was an attempt to distill his very long and mathematically sophisticated 1931 paper (Wright 1931) that developed fundamental expectations for numerous population genetic processes. In a biography of Sewall Wright, Provine (1986) argues that Wright's invention and articulation of fitness surfaces as a heuristic aid to understand allele and genotype frequency dynamics "was one of his single most influential contributions to modern evolutionary biology." Indeed, the fitness surface is a commonly used metaphor in population genetics even today. The goal of this section is to introduce Wright's adaptive landscape metaphor and then explain Wright's interpretation of how genetic drift, gene flow, natural selection, and mutation might interact in the context of genetically subdivided populations known as the **shifting balance process**. After Wright's ideas are introduced, some of the criticisms of the adaptive landscape metaphor and controversies over the shifting balance process will be considered.

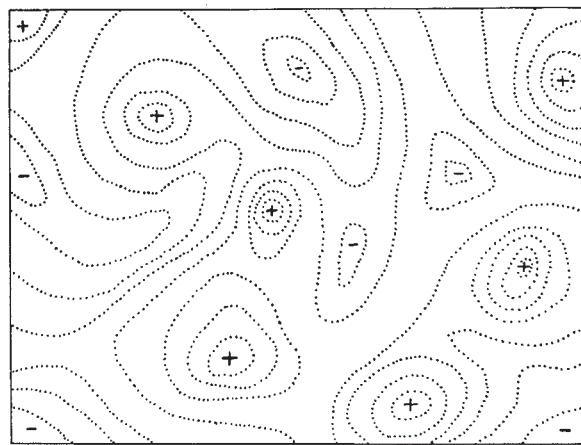
### **Allele combinations and the fitness surface**

At the beginning of his 1932 paper, Wright starts with the observation that there are a very large number of possible genotypes (what he called allelomorph combinations) in any given species. He gives the example that for 1000 loci with 10 alleles each, there are  $10^{1000}$  possible combinations of alleles that might make up a 1000-locus gamete haplotype. (Wright points out that this is a staggeringly large number, which he compares to the estimated number of electrons and protons in the visible universe.) He then assumes that wild-type alleles have a frequency of 0.99 at these 1000 loci, while the infrequent

alternate alleles at these loci confer phenotypes only slightly different than wild type. Wright reasoned that there will be only a very small proportion of individuals in a population exhibiting phenotypes caused by any combination of more than 20 non-wild-type alleles in a 1000-locus haplotype. For example, the expected frequency of a single haplotype containing 20 non-wild-type and 980 wild-type alleles is  $(0.01^{20})(0.99^{980}) = 5.3 \times 10^{-45}$  while the expected frequency of a haplotype containing all 1000 wild-type alleles is  $0.99^{1000} = 4.3 \times 10^{-5}$ . A haplotype composed of all wild-type alleles is 40 orders of magnitude more frequent than a haplotype with just 20 non-wild-type alleles.

Under these assumptions, the range of phenotypes exhibited in a single population will represent only a very small part of the possible range of phenotypes because haplotypes composed mostly of wild-type alleles are very common and haplotypes with numerous non-wild-type alleles are very infrequent. Wright draws the conclusion that “The population is thus confined to an infinitesimal portion of the field of gene combinations” even though there is only a very small chance that two individuals possess identical haplotypes. In other words, the individuals in any population represent only a small portion of the very large range of possible phenotypic variation because haplotypes composed mostly of wild-type alleles are very common whereas haplotypes composed of numerous non-wild-type alleles are very rare.

With this perspective on the large number of allele combinations in a haplotype or gamete, Wright goes on to describe what a surface or plot would be like “If the entire field of possible gene combinations be graded with respect to adaptive value under a particular set of conditions . . .” The **fitness surface** presented by Wright (1932) is shown in Fig. 11.2 as a two-dimensional representation of at least three dimensions. The  $x$  and  $y$  axes, although unlabeled by Wright in the original, represented the range of possible allelic combinations in a haplotype or genotype. The fitness of each genotype is represented by the height of each point (the dimension perpendicular to the page when the surface is drawn in two dimensions). In contemporary usage, the  $x$  and  $y$  axes represent allele frequencies that vary between zero and one while the height of each point on the surface is the mean fitness of a population at those allele frequencies. (The distinction between Wright’s surface and the contemporary interpretation of an adaptive landscape is taken up later in this section.) The dashed lines represent contour lines



**Figure 11.2** Sewall Wright’s original adaptive landscape diagram. The high-fitness points on the surface are indicated by + while the low fitness points are indicated by -. Original caption: “Diagrammatic representation of the field of gene combinations in two dimensions instead of many thousands. Dotted lines represent contours with respect to adaptiveness.” From Wright (1932); reproduced with permission.

of constant fitness. The high points on the surface are indicated by + while the low points are indicated by -. Since Wright’s fitness surface is exactly analogous to a topographic map with peaks and valleys, the term **adaptive landscape** is frequently used to describe three-dimensional graphs where allele or genotype frequencies are represented in two dimensions and a measure of fitness is represented in a third dimension.

Wright then imagined how evolution under natural selection would move a population on such an adaptive landscape. He considered the possibility that “a particular combination [of alleles] gives maximum adaptation and that the adaptiveness of other combinations falls off more or less regularly according to the number of removes.” In contemporary terminology, this is equivalent to additive gene action where the genotypic value is a linear function of the effects of the alleles in the genotype (see Chapters 9 and 10). Under additive gene action, a fitness surface is simply a plane where the highest point can be reached by moving through intermediate steps of increasing population average fitness. Alternatively, Wright also imagined fitness surfaces where “. . . it is possible that there may be two peaks . . .” Fitness surfaces with multiple peaks result from dominance and epistasis, types of gene action that can produce large changes in average fitness with only minor changes in allelic combinations because the genotypic value is a non-linear function of the effects of the

alleles in the genotype. Wright imagined that “In a rugged field of this character, selection will easily carry the species to the nearest peak, but there may be innumerable other peaks which are higher but which will be separated by ‘valleys’.” Refer to Chapter 7 for examples of fitness surfaces when allele combinations exhibit strict additivity as well as dominance.

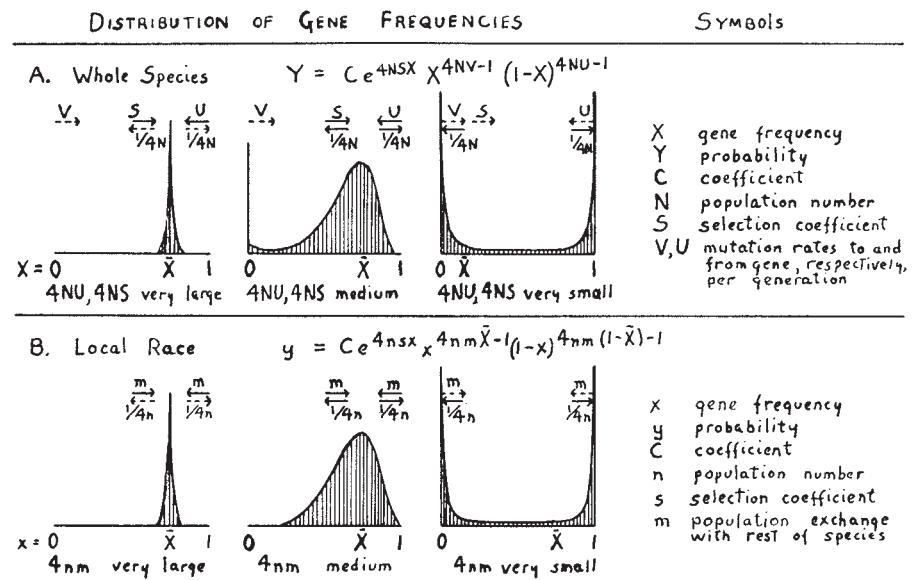
One of the basic challenges of proposing that evolutionary change occurs by natural selection exclusively can be visualized with the aid of Wright’s adaptive landscape diagram. Natural selection acts to increase the mean fitness of a population based on the slope of the adaptive landscape immediately around the current allele frequency position of a population. This leads to changes in allele frequency that, if unopposed by other processes, will eventually lead to the maximum mean fitness that can be achieved by continuously moving uphill on the adaptive landscape. The possible problem is that selection does not take a broad view of the adaptive landscape while causing generation-to-generation changes in allele frequencies. The process of natural selection “sees” only the slope of the adaptive landscape immediately around a population’s current position in allele frequency space. If the adaptive landscape has multiple peaks, the process of selection alone can lead to populations becoming “stranded” at points on the surface that are local mean fitness maxima but not among the highest levels of mean fitness that are possible. In terms of the landscape metaphor, natural selection is like a climber who must always go up in elevation and can never descend (even temporarily), nor cross a valley. Because of this increasing fitness requirement, natural selection may not be capable of reaching the highest peaks on a fitness surface.

Wright proposed his shifting balance model as a possible mechanism that would prevent populations from getting stuck forever on a limited number of fitness peaks. Wright (1932) states clearly that “The problem of evolution as I see it is that of a mechanism by which the species may continually find its way from lower to higher peaks” on a fitness surface. He goes on to say that “there must be some trial and error mechanism on a grand scale by which the species may explore the region surrounding the small portion of the [fitness surface] field it occupies. To evolve, the species must not be under strict control of natural selection. Is there such a trial and error mechanism?” These sentences capture what seems to have been the main question in Wright’s mind when he wrote his 1932 paper. It is a fundamental question in population genetics.

### ***Wright’s view of allele-frequency distributions***

After describing the fitness surface metaphor and asking what mechanisms might overcome the limitations of natural selection, Wright then attempted to summarize his work at that time on the shape of allele-frequency distributions that would be expected under the action of the basic processes of population genetics. He summarized these ideas very succinctly without reference to any equations in a total of three paragraphs that refer to one figure (see Fig. 11.3). The three panels of Fig. 11.3A refer to the expected equilibrium distributions of allele frequencies for one locus in many independent replicate finite populations in an entire species. The processes of natural selection, genetic drift, and mutation act simultaneously to shape these allele-frequency distributions. A narrow distribution of allele frequencies among independent populations is expected when  $4NU$  (equivalent to  $\theta$ ) and  $4NS$  (four times the product of the effective population size and the selection coefficient) are large (left-hand panel in Fig. 11.3A). This distribution results from a large selection coefficient in favor of the wild-type allele ( $S$  with solid arrow) with genetic drift that is relatively weak (dashed arrows labeled  $\frac{1}{4N}$  where  $N$  is the effective population size). Frequent forward mutation ( $U$  with solid arrow) prevents complete fixation of the wild-type allele, shifting the distribution to the left. A lower rate of reverse mutation ( $V$  with dashed arrow) increases the frequency of the wild-type allele a small amount, shifting the distribution to the right. A wider distribution of allele frequencies among independent populations is expected when  $4NU$  and  $4NS$  take intermediate values as shown in the middle panel of Fig. 11.3A. In that case, the forces of selection, mutation, and genetic drift (all solid arrows) are approximately equal to each other so that replicate populations exhibit a range of wild-type allele frequencies. In the right-hand panel of Fig. 11.3A the distribution of allele frequencies is horseshoe-shaped (most replicate populations near fixation or loss) when  $4NU$  and  $4NS$  are small because genetic drift is strong (solid arrows for  $\frac{1}{4N}$ ) relative to natural selection and mutation (dashed arrows for  $S$ ,  $U$ , and  $V$ ).

The three panels of Fig. 11.3B refer to the equilibrium distributions of allele frequencies at one locus in many finite subpopulations connected by some degree of gene flow. The forces that dictate the distribution of allele frequencies are now genetic drift (arrows labeled  $\frac{1}{4n}$  where  $n$  is the effective popula-



**Figure 11.3** Wright's schematic representation of the simultaneous action of multiple population genetic processes leading to equilibrium distributions of allele frequencies. Each distribution represents the allele frequencies of many replicate populations or an ensemble distribution (in A) numerous populations that make up a species are independent while in (B) subpopulations are interdependent in an island model. The magnitude and direction of the effects of a process on the distribution of allele frequencies is indicated by arrows bearing letters. Solid arrows indicate stronger processes and dashed arrows indicate weaker processes. For example, in the left-most distribution of (B) strong migration relative to drift maintains subpopulation allele frequencies with little divergence while weak genetic leads to a modest spread of subpopulation allele frequencies around the average allele frequency of the total population. In all panels, the frequency of the wild-type allele ( $x$ ) is given on the  $x$  axis and the frequency of populations with a given allele frequency is given on the  $y$  axis. The probability on the  $y$  axis is given by the equation at the top of (A) and (B) for the allele frequency on the  $x$  axis given values for the population parameters. Original caption: "Random variability of a gene frequency under various specified conditions." From Wright (1932); reproduced with permission.

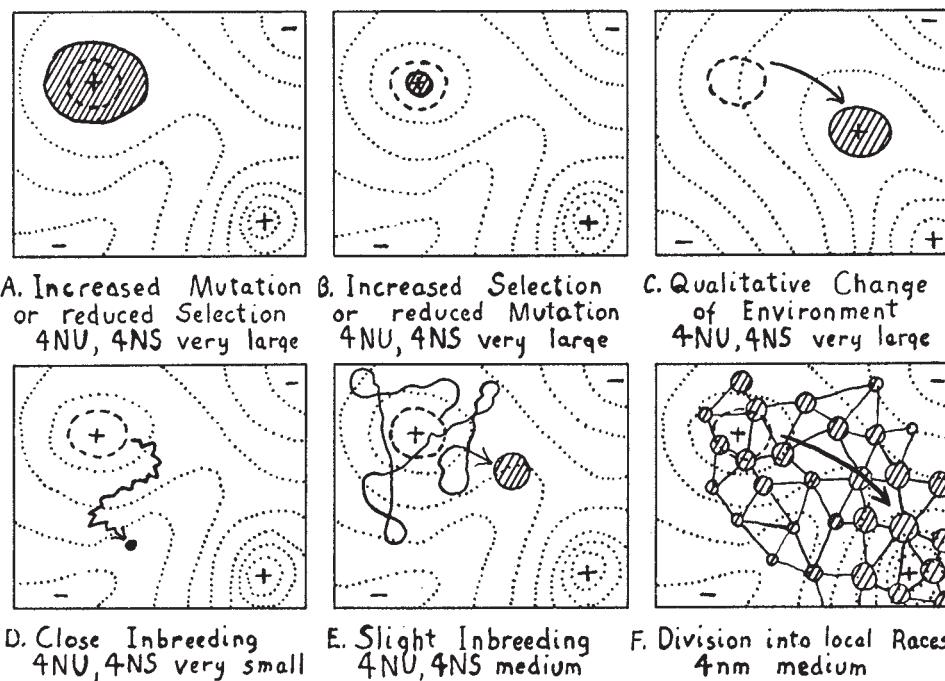
tion size of demes) and migration (arrows labeled  $m$  for the rate of migration in an island model). Here Wright assumed that migration was much stronger than mutation and selection so these later two processes could be ignored. At one extreme shown in the left-hand panel of Fig. 11.3B, a narrow distribution of allele frequencies among demes is expected when  $4nm$  is large because migration rates are high (solid arrows) and/or genetic drift is relatively weak due to large effective population size (dashed arrows). The middle panel of Fig. 11.3B shows the case when genetic drift and migration approximately balance, leading to a wide distribution of intermediate allele frequencies in demes. At the other extreme, shown in the right-hand panel of Fig. 11.3B, the distribution of allele frequencies is horseshoe-shaped (most populations near fixation or loss) when  $4nm$  is small because migration rates are low and and/or genetic drift is strong due to small effective population size.

#### Evolutionary scenarios imagined by Wright

With the allele-frequency distributions established, Wright then returned to the movement of populations on fitness surfaces given different parameters

for the processes of natural selection, mutation, genetic drift, and migration. Wright saw the position of populations on the landscape, the area occupied by populations on the landscape, and the very topography of the landscape itself as subject to change over time. His goal was to illustrate scenarios with sufficient trial and error (or genetic drift) that the tendency of natural selection to strand a population on a single fitness peak could be overcome. Figure 11.4 illustrates the six possibilities that Wright considered. In each of these six cases, the allele-frequency distributions in Fig. 11.3 define the range of variation in fitness expected among individuals in a population or deme. The connection arises if each locus in the allele-frequency distributions of Fig. 11.3 is interpreted as having a phenotypic effect and genotypes homozygous for the wild-type allele have the highest fitness. Broader allele-frequency distributions then lead to a larger number of possible allele combinations that would produce a wider range of fitness values.

Panels A and B in Fig. 11.4 show Wright's ideas about the area of the adaptive landscape that would be occupied by a species in the context of large  $4NS$  and  $4NU$  values. Imagine a population initially



**Figure 11.4** Wright's representation of the action of drift/mutation balance (dictated by the magnitude of  $4NU$ ), drift/selection balance (dictated by the magnitude of  $4NS$ ), and drift/migration balance in the island model (dictated by the magnitude of  $4nm$ ). Wright's parameters are  $N$  for effective population size,  $U$  for mutation rate,  $S$  for the selection coefficient in directional selection, and  $4nm$  for the effective migration rate in the infinite island model. The word "inbreeding" is used in the population sense where finite population size leads to genetic drift. Original caption: "Field of gene combinations occupied by a population within the general field of possible combinations. Type of history under specified conditions indicated by relation to initial field (heavy broken contour) and arrow." From Wright (1932); reproduced with permission.

occupies some area around a fitness peak, as shown by the dashed circle inside the shaded circle in Fig. 11.4A. If the selection coefficient against genotypes with non-wild-type alleles were to decrease or the forward mutation rate were to increase, the area a population occupies around the adaptive peak would spread out (larger shaded circle). This would correspond to the allele-frequency distribution changing from one like that in the left-hand panel of Fig. 11.3A to one like that in the middle panel of Fig. 11.3A. In contrast, when selection against genotypes with alleles other than the wild type becomes stronger or forward mutation rates decrease in the context of large  $4NS$  and  $4NU$ , the area on the fitness surface occupied by a population shrinks because populations take on a narrower range of allele frequencies (compare the circle made by a dashed line to the smaller shaded circle in Fig. 11.4B). This case corresponds to the allele-frequency distribution changing in the opposite direction, from one like the middle panel of Fig. 11.3A to one like that in the left-hand panel of Fig. 11.3A. In the case of Fig. 11.4A, the average fitness of the species decreases, making it possible that "... the spreading

of the [fitness surface] field occupied may go so far as to include another and higher peak ..." or that a fitness valley is crossed. It is also possible that the fitness peak itself could become taller if beneficial mutations occurred in a population and were fixed by selection. However, Wright pointed out that rates of mutation are very low so that evolutionary change of this sort would be very slow.

Wright also considered how the shape of the adaptive landscape itself might change. Figure 11.4C shows Wright's concept of how the adaptive landscape might change over time due to changes in the environmental context of a population (still in the context of large values of  $4NS$  and  $4NU$ ). Because genotypic fitness values are defined by the physical and biological environment a species experiences, the fitness values of allele combinations may very well change over time. This would lead to a reshaping of the fitness surface itself, with peaks and valleys changing elevation or the position of peaks on the adaptive landscape shifting over time. While such change in the fitness surface would cause populations to track high fitness peaks, Wright saw this as "change without advance in adaptation" because

populations were not necessarily occupying a number of peaks in the fitness landscape nor were populations necessarily evolving to higher levels of mean fitness.

This theme of constant environmental change driving a continual redefinition of the genotypes having highest fitness in a population was emphasized by Fisher (1999). Under this view, rugged adaptive landscapes are less problematic since a population can be thought of as occupying a region of allele-frequency space where the topography of the fitness elevations changes over time. If the fitness landscape is continually remodeling itself, a population will not be stranded on a fitness peak since eventually the peak itself will move or change position. This view is also the foundation of the so-called Red Queen or arms race model of Van Valen (1973), where a species must constantly experience adaptive change to keep pace with continually changing genotypic fitness values ultimately caused by a perpetually changing environmental context defined by other species that themselves are constantly changing.

Another set of possibilities that Wright considered, diagrammed in Figs 11.4D and 11.4E, focused on effective population size. Wright pointed out that if the effective population size were very small relative to the selection coefficient and the mutation rate (Fig. 11.4D), a population would likely experience fixation or loss at all loci due to genetic drift (see the allele-frequency distribution in the right-hand panel of Fig. 11.3A). As a consequence, a population would cease attraction to fitness peaks, would wander at random around the fitness landscape, and would also experience the fixation of deleterious alleles, leading to inbreeding depression. If the effective population size became small rapidly, then after fixation and loss at most loci, movement on the landscape would be very slow since most new mutations would be unlikely to segregate for long. In contrast, a finite population with a medium effective population size relative to the selection coefficient and the mutation rate (Fig. 11.4E) would occupy a fairly large area on the surface and would experience some random movement around a fitness peak but would not stray too far from the peak. This would occur because the population would experience an approximate balance between natural selection and genetic drift and would also have the input of new mutations over time (see the allele-frequency distribution in the center panel of Fig. 11.3A). Wright saw populations experiencing a balance of genetic drift, natural selection, and mutation (medium

values of 4NU and 4NS) as being able to shift fitness peaks and a means by which “the species may work its way to the highest peaks in the general field.” The limitation is that peak shifting by one such population was expected by Wright to be a very slow process that would only occur if the mutation rate was approximately equal to the reciprocal of the effective population size.

The final case Wright considered, shown in Fig. 11.4F, was the situation where a species was subdivided into a number of finite demes (or “small local races”) that were nearly genetically isolated but did experience some gene flow. Here Wright was thinking of the allele-frequency distribution in the center panel of Fig. 11.3B, where there is a balance between genetic drift leading to population differentiation and gene flow leading to homogenization of allele frequencies that produces a broad distribution of allele frequencies among demes. Wright’s idea was that many semi-independent finite demes would move positions on the fitness landscape more rapidly than a single panmictic population. Wright also conjectured that those demes that did reach higher fitness peaks would produce more migrants. The effect of more migrants would be for a deme on a higher fitness peak to shift the allele frequencies of the demes that received those migrants toward the positions of higher fitness peaks. This process of higher rates of gene flow from those demes on higher fitness peaks is often called **interdemic selection** since it is equivalent to natural selection acting at the level of demes with different levels of demographic productivity. Thus, Wright envisioned that a species made up of many subdivided demes experiencing approximately equal pressures of natural selection and genetic drift could explore more of the fitness surface and would be more likely to find more of the higher fitness peaks than natural selection alone. Wright concluded that “subdivision of a species into local races provides the most effective mechanism for trial and error in the field of gene combinations.”

The shifting balance process is often summarized by the simultaneous operation of three “phases” of population genetic change in a subdivided population. Phase I involves genetic drift within demes that causes the allele-frequency position of each deme to shift randomly with respect to the position of fitness peaks. Phase II is the operation of natural selection on demes such that the allele-frequency position of demes is shifted toward and higher up fitness peaks, with taller peaks exerting a stronger influence on allele frequencies. Phase III is interdemic selection

such that the rates of emigration from demes are proportional to the mean fitness of a population. Thus, demes at higher fitness peak elevations export more migrants and comprise a larger proportion of the immigrant pool of other demes, shifting allele frequencies of all demes toward the allele-frequency locations of taller fitness peaks.

### **Critique and controversy over shifting balance**

Although Wright's metaphor of the adaptive landscape and proposal of the shifting balance theory has stimulated the thinking of biologists for decades, his ideas have also generated sustained controversy. The fitness surface metaphor itself has been one focus of critique because the original fitness surface described by Wright is problematic in some respects. As Provine (1986) describes, Wright employed two distinct versions of the fitness surface. One version of the fitness surface illustrates the fitness of *each genotype* based on an ordering of the allele combinations in genotypes. In this version of the fitness surface, each combination of alleles has a relative fitness and defines one point on the landscape. This type of surface has been compared with the pixels that make up a photographic print or digital image (Ruse 1996). In the genotype version of the fitness surface, what is represented biologically by the dimensions other than that representing fitness is not clear since the genotype axes do not relate to the frequency of genotypes or alleles in a population. Another version of the fitness surface plots the *mean fitness of a population* for all possible allele frequencies (examples of population mean fitness surfaces can be found in Chapters 6 and 7). Contemporary usage of the fitness surface metaphor is often in the population mean fitness sense, with axes representing allele frequencies and one dimension representing the mean fitness of a population at those allele frequencies, although there are exceptions (e.g. Weinreich et al. 2005). Wright often switched back and forth between these two types of fitness surface in his writing, leading to ambiguity and confusion (Provine 1986). Fitness surfaces have been constructed and interpreted in a wide variety of ways since Wright's work (Gavrilets 2004; Skipper 2004).

Coyne et al. (1997) presented a detailed and vigorous critique of Wright's shifting balance theory that examined evidence for and against operation of the three phases in actual populations. They reexamined the theoretical basis of the shifting balance theory with the benefit of more than 60 years

of work in theoretical population genetics, and considered empirical evidence for the shape of fitness surfaces and operation of the stages of the shifting balance process. They concluded that "although there is some evidence for the individual phases of the shifting balance process, there are few empirical observations explained better by Wright's three phase mechanism than by simple mass selection." Other authors responded in defense of shifting balance theory or to offer alternative points of view (e.g. Peck et al. 1998; Wade and Goodnight 1998), generating a cascade of replies and counter-replies (Coyne et al. 2000; Goodnight and Wade 2000; Peck et al. 2000). While it is not possible here to consider in detail all of the points raised in that debate, disagreements over elements of the shifting balance process serve to highlight difficulties that arise when attempting to predict the outcome of multiple population genetic processes operating simultaneously.

The third phase of the shifting balance process, production of migrants in proportion to the population mean fitness and shifting of demes via differential contributions to the immigrant pool, is particularly problematic (see Crow et al. 1990). The difficulty is that the migration rate must be low enough to permit population subdivision into semi-isolated demes but at the same time high enough to permit the exchange of individuals (or gametes) among subpopulations that leads to interdemic selection. A general objection is that interdemic selection is a form of group selection, a process where there is greater survival or reproduction of a population of individuals compared to other populations such that some populations go extinct while others persist and expand. Williams (1966, 1992) has presented the classical arguments that natural selection on additive genetic variation among individual genotypes is expected to act more rapidly than selection on groups because the frequencies of individuals can change more rapidly than the frequencies of populations. Nonetheless, possible evidence for group selection in the context of differential migration has been shown in experiments with the flour beetle *Tribolium castaneum* (Wade and Goodnight 1991). Large changes in number of individuals per population were observed over nine generations by selecting individuals to be founders of the next generation based on the total number of individuals in a population. In contrast, the size of populations did not change over time when founding individuals were selected at random with respect to population size. The interpretation of numerous *Tribolium* experiments of this

type (reviewed by Goodnight and Stevens 1997) has been controversial since there is disagreement over what exactly constitutes group and individual selection and what type of selection is imposed by the experimental procedures (see Coyne et al. 1997; Getty 1999; Wade et al. 1999).

Another aspect of the disagreement over shifting balance theory involves the dual nature of the concept of epistasis (see Cheverud and Routman 1995; Whitlock et al. 1995; Fenster et al. 1997; Brodie 2000; Cordell 2002). Epistasis exists when genotypes at two or more loci result in a genotypic value that is greater or less than the sum of the genotypic effects of the loci when taken individually (see Chapter 9). The existence of an interaction between two or more loci indicates the existence of **physiological epistasis** (also known as functional or mechanistic epistasis). The term physiological epistasis simply recognizes that certain genotypes at two or more loci interact in the production of a phenotype. The contribution, if any, of such physiological epistasis to population level quantities is a function of the frequencies of interacting genotypes in a population. The term **statistical epistasis** is used to refer to the amount of standing population variation in genotypic values caused by interactions among loci. In the symbols and concepts of Chapters 9 and 10, statistical epistasis is  $V_I$ . The amount of statistical epistasis present in a population is a function of the frequencies of interacting multilocus genotypes and therefore a function of population allele frequencies, as it is for additive and dominance variance ( $V_A$  and  $V_D$ ), as well as a function of mating system and the rate of recombination.

Wright implicitly assumed that statistical epistasis was abundant in natural populations. While there is evidence that statistical epistasis exists in natural and laboratory populations (MacKay 2001; Cordell 2002; Carlborg and Haley 2004; see chapters in Wolf et al. 2000), statistical epistasis is not widespread in populations, although it remains difficult to estimate. There is currently no consensus over the relative contribution of epistasis to overall quantitative trait variation, although there is recognition that empirical detection of epistasis is limited by experimental designs and statistical power (see Whitlock et al. 1995). Some conclude that there is a lack of evidence for strong or frequent statistical epistasis in natural populations. Others suggest that there is some evidence for epistasis in natural populations, and since epistasis is difficult to detect, it is premature to draw a conclusion about the prevalence of epistasis. These

disparate views translate into difficulty summarizing the nature of adaptive landscapes in populations that are genetically variable.

When a population is at fixation and loss for all loci there can be no statistical epistasis since there is no variation in genotypic value, even though physiological epistasis may exist and even be very strong. An alternative definition of epistasis is useful for populations that may exhibit little statistical epistasis and yet have abundant physiological epistasis.

**Sign epistasis** is a special case of physiological epistasis in populations with little or no genetic variation (Weinreich et al. 2005). A locus exhibits sign epistasis when a new mutation exhibits higher-than-average fitness on some genetic backgrounds defined by other loci but lower-than-average fitness on other genetic backgrounds. The sign of the fitness value is therefore a function of the other loci that make up the genetic background of the allele. (This is in contrast to a more general definition of epistasis where a new mutation might always be deleterious, but it is more or less deleterious depending on the genetic background.) An example of sign epistasis in the context of a haploid system would be if the mean fitness of a new mutation at the B locus was positive if it was paired with an A allele and negative if it was paired with an a allele. This leads to an alternative view of how to define and test for epistasis that is amenable to empirical study.

Weinreich et al. (2006) examined five single-nucleotide mutations in the  $\beta$ -lactamase gene of bacteria. Four of these mutations result in missense versions of the  $\beta$ -lactamase gene, so that they are deleterious individually and selected against in antibiotic environments. The fifth mutation is a non-coding change 5' to the gene. However, when all five mutations occur simultaneously they lead to a version of the  $\beta$ -lactamase gene that confers resistance to  $\beta$ -lactam antibiotic drugs such as penicillin. When each mutation occurs on a genetic background with the other four mutations present its fitness is positive since the five-mutation version of the gene has high fitness in antibiotic environments. This situation constitutes an example of sign epistasis.

Even with these various objections and complications, the shifting balance theory has had an enduring impact on the imaginations of population geneticists. The basic problem that motivated Wright to propose the shifting balance theory – that natural selection cannot lead to decreases in mean fitness – remains a difficulty that continues to attract attention and motivate researchers many decades later.

## Chapter 11 review

- The synthesis of Darwin's concept of natural selection with Mendelian particulate inheritance that forms the basis of population genetics is called neo-Darwinism.
- The classical hypothesis predicted that directional natural selection was the dominant process. This led to a prediction that genetic variation was limited at most loci. What genetic variation that existed was thought to be caused mostly by deleterious mutations, along with some neutral mutations and very few beneficial mutations.
- The balance hypothesis predicted that balancing natural selection caused by overdominance for fitness was the dominant process. This led to the prediction that levels of genetic variation should be high for many loci since balancing selection maintains alleles at intermediate frequencies indefinitely. The balance hypothesis also predicted selection would cause gametic disequilibrium over large regions of the genome, producing supergenes or coadapted gene complexes.
- Genetic load results from the selective deaths (either reproductive or actual) that must occur as frequencies of mutations or genotypes change under natural selection. In principle, the amount of natural selection is limited by the genetic load a population can tolerate. Genetic load arguments were used to estimate the upper limit on the rate of substitution in mammals and to argue for the neutral theory of molecular evolution.
- The novel technique of allozyme electrophoresis first used in the mid-1960s revealed that about 30% of the loci in *Drosophila* were heterozygous. This observation set off a long-running debate over how to best explain levels of genetic polymorphism in natural populations.
- The proposal of the neutral theory of molecular evolution sparked a controversy between neutralists and those who favored selection-based explanations of rates of divergence and levels of polymorphism. This debate led to many innovations in the theory of both genetic drift and natural selection. Explanations such as the nearly neutral theory that combine the relative contributions of drift and selection and place them on a continuum are now common in population genetics.
- Wright's metaphor of the adaptive landscape is a heuristic device designed to articulate how the process of natural selection alone is con-

strained since it can only increase population mean fitness.

- Sewall Wright's shifting balance theory was designed as a hypothesis about how the simultaneous action of natural selection, genetic drift, mutation, and population subdivision might lead to the exploration of a larger portion of the adaptive landscape than would be possible under selection alone.
- While shifting balance theory and its key assumption of statistical epistasis remains controversial, adaptive landscapes are an enduring metaphor in population genetics.

## Further reading

For a history of early population genetics beginning with Darwin and Mendel and ending with Fisher, Haldane, and Wright, see:

Provine WB. 1971. *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago, IL (this book was originally published in 1971 while the 2001 edition has an afterword by Provine).

For a wide-ranging consideration and critique of aspects of the classical/balance hypothesis debate written in the midst of the allozyme era in population genetics, see:

Lewontin RC. 1974. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

For Kimura's view on why the classical and balance hypotheses failed to explain patterns of genetic variation see the following (especially the first chapter):

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge (Kimura also reviews the manifold arguments in support of the neutral and nearly neutral theories).

An illuminating history of the development of the neutral and nearly neutral theories is provided in:

Ohta T and Gillespie JH. 1996. Development of neutral and nearly neutral theories. *Theoretical Population Biology* 49: 128–42.

For a review of elements of the selectionism/neutralism debates see:

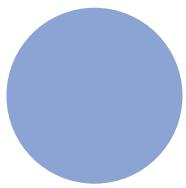
Nei M. 2005. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution* 22: 2318–42.

A highly readable review of the selectionist/neutralist debate written shortly after the proposal of the neutral theory can be found in:

Crow JF. 1972. Darwinian and non-Darwinian evolution. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. V, pp. 1–22. University of California Press, Berkeley, CA.

For background and explanation of the original fitness surface along with some response to criticisms, see:

Wright S. 1988. Surfaces of selective value revisited. *American Naturalist* 131: 115–23.



# Appendix

## Statistical uncertainty

- Quantities in population genetics are estimated with error.
- Parameters and parameter estimates.
- Introduction to variance, standard deviation, and standard error.

Statistical concepts arise in this book in several places. Chapter 1 points out the distinction between idealized parameters, which are exact values, and parameter estimates that have uncertainty obtained through sampling from populations. Both the variance and the covariance are important concepts that appear in a number of chapters, especially in Chapters 9 and 10 that cover quantitative genetics. This Appendix is meant to provide a basic introduction to statistical concepts relevant to these topics for readers without much prior background.

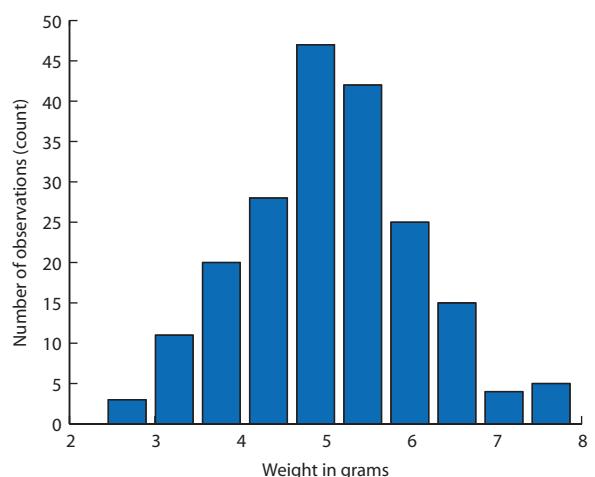
Imagine drawing a random sample of objects, say a handful of jelly beans from a candy dish, and weighing each one. The weights will not be identical, but there will be some value that occurs most often and some range of values. Plotting these values on a graph such as that in Fig. A.1 would show how often each one occurs between the lowest and highest values, or their frequency distribution (often truncated to just “distribution” in conversation). We often use the average (mean and average are used synonymously here) or the mode (the most frequently occurring value) to describe the central tendency or middle of a frequency distribution.

Let’s examine a hypothetical case that will show the distinction between a parameter and a parameter estimate as well as illustrate a common means to quantify uncertainty caused by sampling variance. Imagine we would like to estimate the frequency of the A allele (for a locus with two alleles) in a population of mice. These mice inhabit barns in an area where there are many isolated farms, each with a suitable barn. Therefore, the mice are found in many discrete populations that make up a larger total

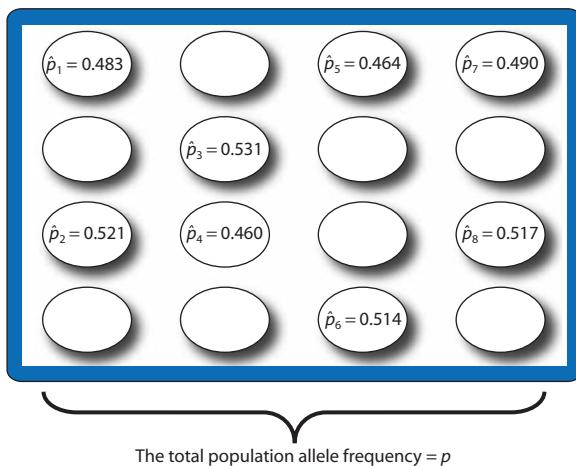
population. An example of this type of population is diagrammed in Fig. A.2.

We would like to estimate the frequency of the A allele in the entire population, which we will call  $\hat{p}$ . The entire population has an exact allele frequency, the parameter  $p$ , which we could only know if we determined the genotype of every mouse in the population. Since it would be very difficult to sample *every* mouse, we take samples from a number of distinct, independent populations to estimate the allele frequency within each population. The average of allele frequencies in sampled populations will be our estimate of the parameter  $p$ . Call the estimate of allele frequency for each barn  $\hat{p}_i$ , where the subscript  $i$  is just an index of which barn the value came from. For simplicity, we assume in this illustration that each value of  $\hat{p}_i$  within a barn is known without error.

A common quantitative tool to measure and express the range of values within a sample is called the **variance** (symbolized by  $\sigma^2$  or a lower-case “sigma” squared). In plain language, the variance is a standardized measure of the range of observed values relative to the average. It is simple to obtain the average allele frequency among all of the sampled



**Figure A.1** The frequency distribution of hypothetical weights for 200 jelly beans. The mean is 5.06 and the variance is 1.06.



**Figure A.2** An abstract representation of 16 mouse populations. The total population (the entire rectangle) is composed of a series of smaller, discrete populations (the individual ellipses), sometimes called subpopulations. Each subpopulation has its own frequency of the A allele, indicated by  $\hat{p}_i$ . The entire population also has an exact value for the frequency of the A allele, which can be estimated by sampling some subpopulations and then taking the average and variance of the  $\hat{p}_i$  values. Here eight subpopulations are sampled to estimate the allele frequency in the total population as  $\bar{p} = 0.4976$ , the standard deviation (SD) = 0.0272, and the standard error (SE) of the mean = 0.0096.

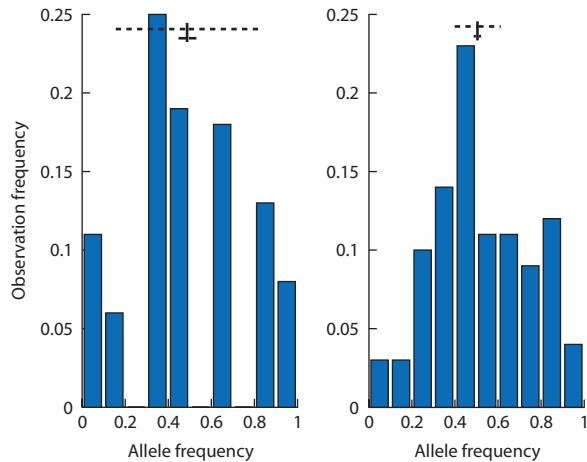
populations by adding up all of the  $\hat{p}_i$  observations and dividing by the total number of observations. In notation, the mean or average allele frequency among the populations sampled would be

$$\bar{p} = \frac{\sum_{i=1}^n \hat{p}_i}{n} \quad (\text{A.1})$$

where the bar over  $\bar{p}$  (pronounced “p bar”) indicates an average and the  $\sum_{i=1}^n$  means summing each of the  $\hat{p}_i$  observations starting with the first (there are a total of  $n$ , as indicated above the summation symbol, which is an upper-case “sigma”), and  $n$  is the sample size. The variance is an average of the square of how much each observation differs from the mean. The variance is taken by summing the squared differences between each estimate and the average and then dividing by one less than the sample size:

$$\text{var}(\hat{p}) = \sigma^2(\hat{p}) = \frac{\sum_{i=1}^n (\hat{p}_i - \bar{p})^2}{n-1} \quad (\text{A.2})$$

The square in the numerator comes about because  $(\hat{p}_i - \bar{p})$  for observations less than the mean will be



**Figure A.3** Two frequency distributions of 100 data points each with nearly identical means (0.498 on the left and 0.506 on the right) but different degrees of variance among the observations (the variance is 0.0293 on the left and 0.0025 on the right). The lines at the top indicate the position of the mean (vertical line) and two standard deviations on either side of the mean (horizontal dashed line) and two standard errors on either side of the mean (horizontal solid line). Like allele frequencies, each distribution is on the interval of 0 to 1.

negative for observations less than the mean but positive for observations greater than the mean. Squaring will make all differences positive so that positive and negative differences will not cancel each other out when summed. Figure A.3 shows two example distributions with variances that differ by more than 10-fold. You should also note that distributions with larger means will have larger variances even if the spread of observations is identical in the two distributions. This makes comparing variance values difficult without reference to the mean.

**Variance ( $\sigma^2$ )** The sum of the squared deviations from the mean divided by one less than the sample size.

**Standard deviation ( $\sigma$ )** The square root of the variance; the average deviation from the mean for a single observation. Quantifies the range of values around the mean seen in a sample.

**Standard error of an average (SE)** The product of the standard deviation and the square root of the sample size divided by the sample size; how far the true population average (a parameter) may be from the sample average (a parameter estimate) by chance.

The variance estimator shown in equation A.2 is sometimes called the sampling variance and it is an unbiased estimator of the variance in a very large population (unbiased means that the expected value of the variance is equal to the true value of the variance). An alternative form of the variance exists where the sums of squares are divided by  $n$  rather than  $n - 1$ . Dividing by  $n$  estimates what is sometimes called the parametric variance, or a variance for a finite population of size  $n$  where all  $n$  individuals have been used to compute the variance. The parametric variance is employed in idealized situations where every individual in a population can be measured or sampled. In practice, this distinction makes little difference for large  $n$ .

If we are willing to assume that our average is drawn from a frequency distribution called the normal distribution (which Interact box A.1 suggests is not unreasonable), then we can use the variance in equation A.2 as a measure of uncertainty in our estimate of the average allele frequency. The standard deviation is symbolized by  $\sigma$  (a lower-case “sigma”) or SD and is simply the square root of the variance (taking the square root returns the variance back to the original units of measurement). The SD measures the average deviation from the mean for a single observation. Figure A.3 illustrates the standard deviations that arise from the variances of two different frequency distributions. Normal distributions are very useful because the standard deviation corresponds to the probability that an observation is some distance from the average. For an ideal normal distribution, about 68% of the observations fall within plus or minus one SD of the mean, about 95% of the observations fall within two SDs of the mean, and about 99% of the observations fall within three SDs of the mean.

The standard deviation of a single observation can also be used to quantify uncertainty in the average of all observations. The standard error, or SE, for the sum of all observations (the denominator in equation A.1) is  $\sqrt{n}$  sample size multiplied by the SD. Since the SD measures the average spread of an observation from the mean, then the SE of the sum adds these individual deviations up. The square root of the sample size is the multiplier because the SE grows more slowly than the sample size itself as observations are added to the sum. The SE of the sum can be related to the average by taking the average of the SE of the sum, or

$$\text{SE of average} = \frac{(\sqrt{n})SD}{n} = \frac{SD}{\sqrt{n}} \quad (\text{A.3})$$

(the  $\sqrt{n}$  term in the numerator of the middle equation cancels because  $n = \sqrt{n^2}$  in the denominator). Notice that as the sample size increases, the SE of the mean decreases. Like the SD, the SE of the average defines a probability range around the mean due to chance events in the sample. The probability intervals defined by the SE serve to establish what are called confidence intervals. By convention, 95% confidence intervals are frequently used to quantify the chances that the parameter estimate ( $\bar{p}$ ) plus or minus 2 SEs covers the true parameter value in the population ( $p$ ). Again, Fig. A.3 illustrates the standard errors of the means for two frequency distributions.

To return to our mouse population example diagrammed in Fig. A.2,  $\bar{p} = 0.4976$ , the variance is 0.00074, and the standard deviation is 0.0272. The standard error of the sum is  $(\sqrt{8})(0.0272) = 0.0769$  and the standard error of the average is  $((\sqrt{8})(0.0272))/8 = 0.0096$ . Therefore, the 95% confidence interval for the mean is  $0.4976 - (2 \times 0.0096)$  to  $0.4976 + (2 \times 0.0096)$  or 0.4784 to 0.5168. Thus, we would expect 95 times out of 100 that the range of allele frequencies between 0.4784 and 0.5168 would include the actual allele frequency of the population,  $p$ . You will be reassured to note that the values in Fig. A.2 are random numbers generated by computer using a distribution with a mean of 0.50, equivalent to the true parameter value. The value of  $\bar{p}$  is very close to this parameter value and the parameter value also falls inside the 95% confidence interval for  $\bar{p}$  defined by plus or minus 2 SEs of the average.

### Problem box A.1 Estimating the variance

Refer to Fig. A.2 and replace the values in the figure for the allele frequency within each population with 0.6333, 0.5074, 0.4880, 0.3960, 0.5368, 0.3330, 0.5893, and 0.7029. What is the estimate of allele frequency in the entire population and what is the 95% confidence interval? How does the variance in this case compare with the variance for the original values given in Fig. A.2? How does the range of allele frequencies that include the true population parameter compare to the original values in Fig. A.2?

### Interact box A.1 The central limit theorem

The central limit theorem predicts that the distribution of means will approach a normal distribution regardless of the shape of the distribution of the original data sampled to compute the mean. The central limit theorem is the justification for using the standard deviation as an estimate of the confidence in a mean such as the average allele frequency (see main text). The central limit theorem also demonstrates that certainty in parameter estimates is directly proportional to the size of samples used to make such estimates. Instead of accepting the central limit theorem as a given, let's use simulation to explore it as a prediction that can be tested.

- Step 1** On the textbook web page under the heading Appendix, click on the link for **Central limit theorem simulation**.
- Step 2** A web page titled **Sampling Distributions** will open. After a moment to load, a **Begin** button will appear at the top left of the window. Press the button and a new simulation window will open. Once the Java code for the window loads, you will see four sets of axes with some buttons and pop-out menus down the right side of the window. The top set of axes will contain a normal distribution with the mean, median and standard deviation ("sd") given in three colors to the left of the distribution and their positions also indicated in those same colors below the x axis. Please read the **Instructions** text and refer to the simulation window so that you understand the controls.
- Step 3** This simulation samples individual data points from the distribution at the top. Pressing the **Animated** button on the right will show the sampling process with five data points at a time on the axes labeled **Sample Data**. Press the button and watch as five bars drop to indicate the five data points. Once the sample is complete a single data point, the mean of five individual data points, will appear on the axes labeled **Distribution of Means**. Press the **Animated** button a few times to get a sense of how the two middle windows display the data. Be sure you can distinguish between the sample size (the **N** = menu right of the axes) and the number of samples (tabulated as **Reps** = left of the axes).
- Step 4** Set **N** = 25 (the lower three graphs will clear) and press the **Animated** button five times (notice how **Reps** = increases by one each time). Now press the **5** button once. This is like sampling five more samples of 25 without the animation. Notice how reps is now 10 since the repetitions add up. Click on the **Fit normal** check box to compare the histogram in the distribution of means to an ideal normal distribution. Ideal normal distributions have skew and kurtosis (measures of asymmetry about the mean) values of zero, like the normal distribution at the top.
- Step 5** How do the sample size and the number of samples influence the distribution of means? To find out, simulate a range of values of one variable while holding the other variable constant (try **N** = 2 and 20 with 20, 50, 100, and 1000 reps). Look at the histograms and write down the skew and kurtosis values for each combination of the sample size and number of samples. Do larger samples improve the approach to normality of the distribution of means?
- Step 6** Is the distribution of means still a normal distribution even if the parent population (the top-most distribution) is not a normal distribution? Compare the skew and kurtosis values from samples taken from three different parent populations (changed with the pop-out menu to the right of the top-most axes) for a range of identical sample sizes and numbers of samples (try **N** = 5 with 20, 50, 100, and 1000 reps).
- Step 7** Feel free to explore the topic further by clicking on the **Exercises** link below the **Begin** button.

If at any point you would like to obtain a new simulation window with default starting values, just close the simulation window and then open a new one with the **Begin** button.

Note that in this example the allele frequencies for each population ( $\hat{p}_i$ ) can take on any value between zero and one, and are therefore a **continuous variable**. The estimate of allele frequency within each individual population is based on counting up alleles that can take only one of two forms, a and A, called a **binomial variable**.

### Covariance and correlation

- Quantifying the relationship between two variables.
- Introduction to covariance and correlation.
- The role of covariance in regression analysis.

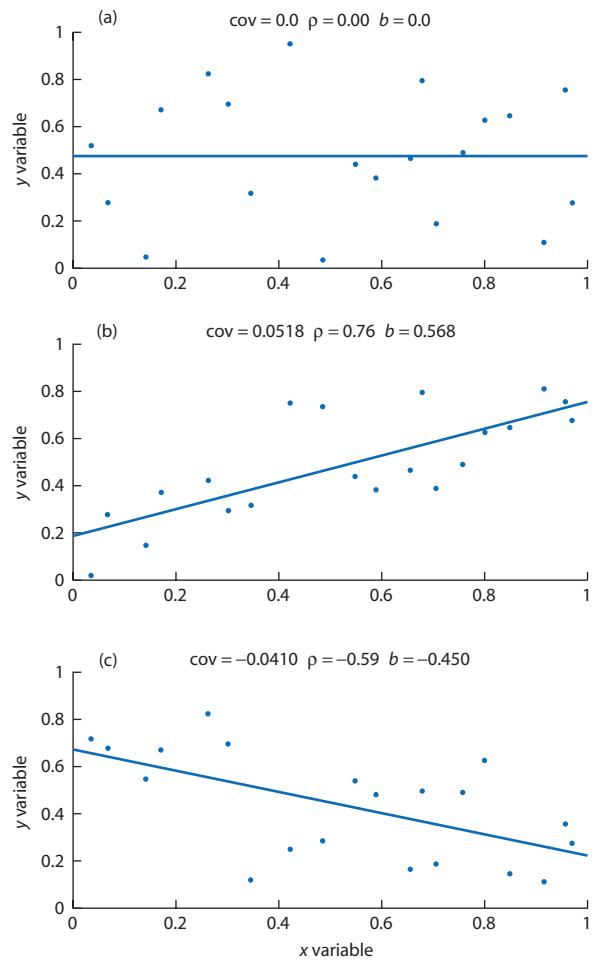
In population genetics and in all of biology, there are many situations where our goal is understand the relationship between two variables. To extend the jelly bean example used above, imagine that each jelly bean was weighed and its length was also measured. Each individual jelly bean then has two values for the two variables that describe it. One question we might ask is whether the length and mass of jelly beans are related or whether jelly beans of any mass can exhibit any length. These questions about jelly bean mass and length can be answered by determining the **covariance** and **correlation** between two variables.

The spread or scatter of two variables, call them  $x$  and  $y$ , viewed simultaneously is their **joint distribution**. Figure A.4 illustrates three different joint distributions of  $x$  and  $y$  values. The degree to which the values of two variables tend to vary in the same direction is measured by the covariance. The covariance is

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (\text{A.4})$$

where the deviations of each observation from the mean for each variable are multiplied, these products are summed over all observations, and then the sum is divided by the number of observations to achieve an average. Notice that the deviations from the mean are not squared as they are for the variance. This means that the covariance measures the direction of the deviations from the mean as well as their magnitude.

The joint distributions in Fig. A.4 illustrate a range of covariance values even though the variance of the  $x$  variable is constant. In Fig. A.4a, the



**Figure A.4** Examples of joint distributions between two variables,  $x$  and  $y$ . The covariance between  $x$  and  $y$  is zero in (a), positive in (b), and negative in (c). In all three panels the variance of  $x$  remains constant at 0.0912. Trends between the variables can be expressed as correlation coefficients ( $\rho$ ) or as the slopes of the least-squares regression of  $y$  on  $x$  (b).

$x$  and  $y$  values have zero covariance and the two variables are therefore independent. When a covariance is zero then the scatter of each variable can be described by its variance alone without reference to the other variable. In both Figs A.4b and A.4c the values of  $x$  and  $y$  are not independent but rather tend to vary together. A positive covariance between  $x$  and  $y$  is shown in Fig. A.4b, telling us that higher values of one variable tend to be associated with higher values of the other variable. Figure A.4c shows negative covariance between  $x$  and  $y$ , telling us that higher values of one variable tend to be associated with lower values of the other variable and vice versa. Using the jelly bean analogy, a zero covariance says that length and weight are independent of each other, a positive covariance says that heavier jelly

beans tend to be longer, and a negative covariance says that heavier jelly beans tend to be shorter.

The covariance forms the basis of the **correlation coefficient**, a summary measure of the strength and direction of the linear relationship between two variables. The Pearson product-moment correlation, symbolized by  $\rho$  (pronounced “roe”), is given by

$$\hat{\rho}_{x,y} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} \quad (\text{A.5})$$

where the  $\sqrt{\text{var}(x)}$  is really the same thing as the standard deviation of  $x$ . The correlation is a dimensionless quantity that takes on values between  $-1$  and  $+1$ . A perfect positive linear relationship between  $x$  and  $y$  gives a correlation of  $+1$ , a perfect negative linear relationship between  $x$  and  $y$  gives a correlation of  $-1$ , and a correlation of zero indicates independence of  $x$  and  $y$ . The correlation coefficients are also given for the joint distributions in Fig. A.4. A number of other non-parametric correlation measures exist that are appropriate for data that are not normally distributed, such as Spearman’s rank-order correlation.

It is important to recognize that the correlation coefficient measures only associations between variables but does not provide any information on how that association came into being. Unfortunately, correlations between variables are commonly misinterpreted as indicating a cause-and-effect relationship between variables. A hypothetical example might be a positive correlation between sales of lemonade and sales of baseballs. Consumption of lemonade does not cause people to also buy a baseball. Rather both events are tied to the weather in a cause-and-effect relationship; in warm weather people drink more lemonade and also play baseball. A classic demonstration of the weakness of the correlation as summary measure was made by Anscombe (1973), who concocted four data sets with identical means, standard deviations and correlation coefficients. These four data sets illustrate how non-normality, non-linearity, and outliers in data can produce high correlations but very poor summaries of the relationship between two variables.

The covariance also plays a fundamental role in regression analysis, which is used to estimate the resemblance between parents and offspring as described in Chapter 9. Assume that  $y$  is a response or dependent variable and  $x$  is an independent variable. The slope ( $a$ ) and intercept ( $b$ ) of a regression line for the variables  $x$  and  $y$  is represented by the equation

$$y = a + bx \quad (\text{A.6})$$

This equation can be rewritten using the average values of  $x$  and  $y$

$$\bar{y} = a + b\bar{x} \quad (\text{A.7})$$

The difference between any individual observation  $y$  and the average of  $y$  is then

$$y - \bar{y} = a + bx - a - b\bar{x} \quad (\text{A.8})$$

which simplifies to

$$y - \bar{y} = b(x - \bar{x}) \quad (\text{A.9})$$

Multiplying both sides of equation A.9 by  $(x - \bar{x})$  gives

$$(y - \bar{y})(x - \bar{x}) = b(x - \bar{x})^2 \quad (\text{A.10})$$

Notice that the left-hand side looks a lot like the covariance in equation A.4 and the right-hand side looks a lot like the variance in equation A.2. If we sum the quantities on both sides of equation A.10 then divide the sums by  $1/n$  to make them averages then A.10 becomes

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = b \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{A.11})$$

which is the same as

$$\text{cov}(x, y) = b \text{ var}(x) \quad (\text{A.12})$$

Solving for the slope of the regression line gives

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)} \quad (\text{A.13})$$

Therefore, we see that the slope of a regression line is the covariance between  $x$  and  $y$  divided by the variance in  $x$ .

The regression line slopes for the joint distributions in Fig. A.4 can be computed from the covariances of  $x$  and  $y$  along with the variance in  $x$ . In all three panels of Fig. A.4  $\text{var}(x) = 0.0912$ . In the top panel,  $\text{cov}(x, y) = 0$  so that the slope of the regression line is also zero. In the middle panel  $\text{cov}(x, y) = 0.0518$  so that  $b = 0.0518/0.0912 = 0.568$ . In the bottom panel  $\text{cov}(x, y) = -0.0410$  so that  $b = -0.0410/0.0912 = -0.450$ .

### Further reading

One approachable beginning statistics text is:

Freedman, D, Pisani R, and Purves R. 1997. *Statistics*, 3rd edn. Norton, New York.

Two classic biologically oriented statistics texts are:

Sokal RR and Rohlf FJ. 1995. *Biometry: the Principles and Practice of Statistics in Biological Research*, 3rd edn. W. H. Freeman & Company, New York.

Zar JH. 1999. *Biostatistical Analysis*, 4th edn. Prentice Hall, Upper Saddle River, NJ.

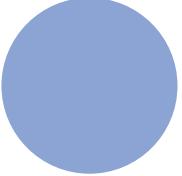
For a humorous take on how basic graphs and statistics can lead to miscommunication and a book that will improve your own presentation of statistical information see:

Huff D. 1954. *How to Lie with Statistics*. W. W. Norton, New York.

### Problem box answers

#### Problem box A.1 answer

Using the new values produces a mean of 0.5233, a variance of 0.0147, and a standard deviation of 0.1212. The SE of the sum is  $(\sqrt{8})(0.1212) = 0.3428$  and the standard error of the average is  $((\sqrt{8})(0.0272))/8 = 0.0429$ . Therefore, the 95% confidence interval for the mean is  $0.5233 - (2 \times 0.0429)$  to  $0.5233 + (2 \times 0.0429)$  or 0.4375 to 0.6091. Thus, we would expect 95 times out of 100 that the range of allele frequency between 0.4375 and 0.6091 would include the actual allele frequency of the population,  $p$ . The confidence interval about  $\hat{p}$  is much larger in this case since the variance is larger: all caused by the observations being more spread out around the mean. In this case, the allele frequency parameter is estimated with more uncertainty since the underlying observations used for the estimate have a much greater range of values.



## References

- Abbott RJ and Gomes MF. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* Heynh. *Heredity* 62: 411–18.
- Abney M, McPeek MS, and Ober C. 2001. Broad and narrow heritabilities of quantitative traits in a founder population. *American Journal of Human Genetics* 68: 1302–7.
- Allison AC. 1956. The sickle-cell and haemoglobin C genes in some African populations. *Annals of Human Genetics* 21: 67–89.
- Álvarez-Castro JM and Alvarez G. 2005. Models of general frequency-dependent selection and mating-interaction effects and the analysis of selection patterns in *Drosophila* inversion polymorphisms. *Genetics* 170: 1167–79.
- Anderson EC and Slatkin M. 2003. Orr's quantitative trait loci sign test under conditions of trait ascertainment. *Genetics* 165: 445–6.
- Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution* 18: 279–90.
- Anscombe, FJ. 1973. Graphs in statistical analysis. *American Statistician* 27: 17–21.
- Arbogast BS, Edwards SV, Wakeley J, Beerli P, and Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics* 33: 707–40.
- Armbruster P and Conn JE. 2006. Geographic variation of larval growth in North American *Aedes albopictus* (Diptera: Culicidae). *Annals of the Entomological Society of America* 99: 1234–43.
- Armbruster P, Hutchinson RA, and Linvell T. 2000. Equivalent inbreeding depression under laboratory and field conditions in a tree-hole-breeding mosquito. *Proceedings of the Royal Society of London Series B Biological Sciences* 267: 1939–45.
- Ashley-Koch A, Yang Q, and Olney RS. 2000. Sickle hemoglobin (*Hb S*) allele and sickle cell disease: a HuGE review. *American Journal of Epidemiology* 151: 839–45.
- Balloux F. 2004. Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution* 58: 1891–1900.
- Barton NH and Turelli M. 2004. Effects of genetic drift on variance components under a general model of epistasis. *Evolution* 58: 2111–32.
- Baudry E, Derome N, Huet M, and Veuille M. 2006. Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model *Drosophila simulans*. *Genetics* 173: 759–67.
- Bazin E, Glémén S, and Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science* 312: 570–2.
- Beavis WD. 1994. In *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*, pp. 250–66. American Seed Trade Association, Washington DC.
- Beck NR, Double MC, and Cockburn A. 2003. Microsatellite evolution at two hypervariable loci revealed by extensive avian pedigrees. *Molecular Biology and Evolution* 20: 54–61.
- Begin DJ and Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356: 519–20.
- Bergelson J, Stahl E, Dukek S, and Kreitman M. 1998. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* 148: 1311–23.
- Bernstein F. 1925. Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Z. Indukt. Abstammungs. Vererbungsl.* 37: 237–70.
- Bigler J, Whitton J, Lampe JW, Fosdick L, Bostick RM, and Potter JD. 2001. *CYP2C9* and *UGT1A6* genotypes modulate the protective effect of aspirin on colon adenoma risk. *Cancer Research* 61: 3566–9.
- Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, and Purvis A. 2007. The delayed rise of present-day mammals. *Nature* 446: 507–12.
- Birchler JA, Dawe RK, and Doebley JF. 2003. Marcus Rhoades, preferential segregation and meiotic drive. *Genetics* 164: 835–41.
- Birky CW and Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Sciences USA* 85: 6414–18.
- Bittles AH, Mason WM, Greene J, and Rao NA. 1991. Reproductive behavior and health in consanguineous marriages. *Science* 252: 789–94.

- Bittles AH, Grant JC, Sullivan SG, and Hussain R. 2002. Does inbreeding lead to decreased human fertility? *Annals of Human Biology* 29: 111–30.
- Bodmer WF. 1965. Differential fertility in population genetics. *Genetics* 51: 411–24.
- Bouteiller C and Perrin N. 2000. Individual reproductive success and effective population size in the greater white-toothed shrew *Crocidura russula*. *Proceedings of the Royal Society of London Series B Biological Sciences* 267: 701–5.
- Brodie ED III. 2000. Why evolutionary genetics does not always add up. In Wolf JB, Brodie ED III, and Wade MJ (eds), *Epistasis and the Evolutionary Process*, pp. 3–19. Oxford University Press, Oxford.
- Bromham L, Phillips MJ, and Penny D. 1999. Growing up with dinosaurs: molecular dates and mammalian radiation. *Trends in Ecology and Evolution* 14: 113–18.
- Bromham L, Penny D, Rambaut A, and Hendy MD. 2000. The power of relative rates tests depends on the data. *Journal of Molecular Evolution* 50: 296–301.
- Brown ADH. 1979. Enzyme polymorphism in plant populations. *Theoretical Population Biology* 15: 1–42.
- Brown KM, Baltazar G, and Hamilton MB. 2005. Reconciling nuclear microsatellite and mitochondrial marker estimates of population structure: breeding population structure of Chesapeake Bay striped bass (*Morone saxatilis*). *Heredity* 94: 606–15.
- Budowle B, Shea B, Niezgoda S, and Chakraborty R. 2001. CODIS STR loci data from 41 sample populations. *Journal of Forensic Science* 46: 453–89.
- Bulmer MG. 1985. *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, Oxford.
- Bulmer M. 1998. Galton's law of ancestral heredity. *Heredity* 81: 579–85.
- Bünger L, Renne U, Dietl G, and Kuhla S. 1998. Long-term selection for protein amount over 70 generations in mice. *Genetical Research* 72: 93–109.
- Buonagurio DA, Nakada S, Parvin JD, Krystal M, Palese P, and Fitch WM. 1986. Evolution of human influenza A viruses over 50 years: uniform rate of change in NS gene. *Science* 232: 980–2.
- Bürger R and Lande R. 1994. On the distribution of the mean and variance of a quantitative trait under mutation-selection-drift balance. *Genetics* 138: 901–12.
- Buri P. 1956. Gene frequency in small populations of mutant *Drosophila*. *Evolution* 10: 367–402.
- Byers DL and Waller DM. 1999. Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression. *Annual Review of Ecology and Systematics* 30: 479–513.
- Carlberg Ö and Haley CS. 2004. Opinion: Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics* 5: 618–25.
- Carr DE, and Dudash MR. 2003. Recent approaches to the genetic basis of inbreeding depression in plants. *Philosophical Transactions of the Royal Society of London B Biological Sciences* 358: 1071–84.
- Cavalli-Sforza LL and Bodmer WF. 1971. *The Genetics of Human Populations*. W. H. Freeman, San Francisco.
- Chakraborty R, Meagher TR, and Smouse PE. 1988. Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics* 118: 527–36.
- Chan AT, Tranah GJ, Giovannucci EL, Hunter DJ, and Fuchs CS. 2005. Genetic variants in the *UGT1A6* enzyme, aspirin use, and the risk of colorectal adenoma. *JNCI Journal of the National Cancer Institute* 97: 457–60.
- Chao L and Carr DE. 1993. The molecular clock and the relationship between population size and the generation time. *Evolution* 47: 688–90.
- Charlesworth B and Charlesworth D. 1997. Rapid fixation of deleterious allele can be caused by Muller's ratchet. *Genetical Research* 70: 63–73.
- Charlesworth B and Charlesworth D. 1998. Some evolutionary consequences of deleterious mutation. *Genetica* 102/103: 3–19.
- Charlesworth B and Charlesworth D. 1999. The genetic basis of inbreeding depression. *Genetical Research* 74: 329–40.
- Charlesworth B, Morgan MT, and Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289–1303.
- Charlesworth B, Charlesworth D, and Barton NH. 2003. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology and Systematics* 34: 99–125.
- Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* 2: 379–84.
- Charlesworth D, Charlesworth B, and Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141: 1619–32.
- Cheverud JM and Routman EJ. 1995. Epistasis and its contribution to genetic variance components. *Genetics* 139: 1455–61.
- Christensen K, Arnbjerg J, and Andresen E. 1985. Polymorphism of serum albumin in dog breeds and its relation to weight and length. *Hereditas* 102: 219–23.
- Clark AG and Feldman MW. 1986. A numerical simulation of the one-locus multiple allele fertility model. *Genetics* 113: 161–76.
- Clausen J, Keck DD, and Hiesey W. 1948. *Experimental Studies on the Nature of Species. III. Environmental Responses of Climatic Races of Achillea*. Publication 581. Carnegie Institution of Washington, Washington DC.
- Cockerham CC. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics* 39: 859–82.
- Cockerham CC. 1956. Effects of linkage on the covariances between relatives. *Genetics* 41: 138–41.
- Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, and Kingsley DM. 2004. The

- genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biology* 2: 635–41.
- Cordell HJ. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11: 2463–8.
- Cotterman CW. 1974. A calculus for statistico-genetics. In Ballonoff P (ed.), *Genetics and Social Structure: Mathematical Structuralism in Population Genetics and Social Theory*, pp. 157–272. Dowden, Hutchinson and Ross, Stroudsburg, PA.
- Cotterman CW. 1983. Relationship and probability in Mendelian populations. *American Journal of Medical Genetics* 16: 393–440.
- Cox JT and Durrett R. 2002. The stepping stone model: new formulas expose old myths. *Annals of Applied Probability* 12: 1348–77.
- Coyne JA and Orr HA. 2004. *Speciation*. Sinauer Associates, Sunderland, MA.
- Coyne JA, Barton NH, and Turelli M. 1997. Perspective: a critique of Sewall Wright's shifting balance theory of evolution. *Evolution* 51: 643–71.
- Coyne JA, Barton NH, and Turelli M. 2000. Is Wright's shifting balance process important in evolution? *Evolution* 54: 306–17.
- Crandall KA, Posada D, and Vasco D. 1999. Effective population sizes: missing measures and missing concepts. *Animal Conservation* 2: 317–19.
- Crawford TJ. 1984. The estimation of neighborhood parameters for plant populations. *Heredity* 52: 273–83.
- Crnokrak P and Barrett SC. 2002. Perspective: purging the genetic load. *Evolution* 56: 2347–58.
- Crow JF. 1958. Some possibilities for measuring selection intensities in man. *Human Biology* 30: 1–13.
- Crow JF. 1972. Darwinian and non-Darwinian evolution. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. V, pp. 1–22. University of California Press, Berkeley, CA.
- Crow JF. 1988. Eighty years ago: the beginnings of population genetics. *Genetics* 119: 473–6.
- Crow JF. 1993a. Felix Bernstein and the first human marker locus. *Genetics* 133: 4–7.
- Crow JF. 1993b. Mutation, mean fitness, and genetic load. *Oxford Surveys in Evolutionary Biology* 9: 3–42.
- Crow JF. 2002. Perspective: here's to Fisher, additive genetic variance, and the fundamental theorem of natural selection. *Evolution* 56: 1313–16.
- Crow JF and Kimura M. 1970. An Introduction to Population Genetics Theory. Harper and Row, New York.
- Crow JF and Aoki K. 1984. Group selection for a polygenic trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences USA* 81: 6073–7.
- Crow JF and Denniston C. 1988. Inbreeding and variance effective population numbers. *Evolution* 42: 482–95.
- Crow J, Engels WR, and Denniston C. 1990. Phase three of Wright's shifting-balance theory. *Evolution* 44: 233–47.
- Cummings CL, Alexander HM, Snow AA, Rieseberg LH, Kim MJ, and Culley TM. 2002. Fecundity selection in a sunflower crop-wild study: can ecological data predict crop allele changes? *Ecological Applications* 12: 1661–71.
- Culter DJ. 2000a. Understanding the overdispersed molecular clock. *Genetics* 154: 1403–17.
- Culter DJ. 2000b. Estimating divergence times in the presence of an overdispersed molecular clock. *Molecular Biology and Evolution* 17: 1647–60.
- Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
- Denny MW and Gaines SD. 2000. *Chance in Biology: Using Probability to Explore Nature*. Princeton University Press, Princeton, NJ.
- de Vicente MC and Tanksley SD. 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. *Genetics* 134: 585–96.
- Devlin B and Ellstrand NC. 1990. The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. *Evolution* 44: 248–59.
- DeWitt TJ and Scheiner SM (eds). 2004. *Phenotypic Plasticity: Functional and Conceptual Approaches*. Oxford University Press, Oxford.
- Dickerson RE. 1971. The structure of cytochrome c and the rates of molecular evolution. *Journal of Molecular Evolution* 1: 26–45.
- Dobzhansky T. 1955. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harbor Symposia on Quantitative Biology* 20: 1–15.
- Doukhan L and Delwart E. 2001. Population genetic analysis of the protease locus of human immunodeficiency virus type 1 quasispecies undergoing drug selection, using a denaturing gradient-heteroduplex tracking assay. *Journal of Virology* 75: 6729–36.
- Dow BD and Ashley MV. 1996. Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. *Molecular Ecology* 5: 615–27.
- Drake JW, Charlesworth B, Charlesworth D, and Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148: 1667–86.
- Dudash MR. 1990. Relative fitness of selfed and outcrossed progeny in a self-compatible, protandrous species, *Sabatia angularis* L. (Gentianaceae): a comparison in three environments. *Evolution* 44: 1129–39.
- Edwards AWF. 1994. The fundamental theorem of natural selection. *Biological Reviews* 69: 443–74.
- Edwards AWF. 2002. The fundamental theorem of natural selection. *Theoretical Population Biology* 61: 335–7.
- Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* 16: 551–8.
- Emara MG, Kim H, Zhu J, Lapierre RR, and Lakshmanan N. 2002. Genetic diversity at the major histocompatibility complex (B) and microsatellite loci

- in three commercial broiler pure lines. *Poultry Science* 81: 1609–17.
- Estany J, Villalba D, Tor M, Cubilo D, and Noguera JL. 2002. Correlated response to selection for litter size in pigs: I. Growth, fat deposition, and feeding behavior traits. *Journal of Animal Science* 80: 2556–65.
- Estes S, Phillips PC, Denver DR, Thomas WK, and Lynch M. 2004. Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. *Genetics* 166: 1269–79.
- Ewens WJ. 1982. On the concept of the effective population size. *Theoretical Population Biology* 21: 373–8.
- Ewens WJ. 2004. *Mathematical Population Genetics. I. Theoretical Introduction*, 2nd edn. Springer-Verlag, New York.
- Excoffier L, Smouse PE, and Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479–91.
- Fabiani A, Galimberti F, Sanvito S, and Hoelzel AR. 2004. Extreme polygyny among southern elephant seals on Sea Lion Island, Falkland Islands. *Behavioral Ecology* 15: 961–9.
- Fairhurst RM and Casella JF. 2004. Homozygous hemoglobin C disease. *New England Journal of Medicine* 350: 26.
- Falconer DS and Mackay TFC. 1996. *Introduction to Quantitative Genetics*. Longman, Harlow.
- Feldman MW, Christiansen FB, and Liberman U. 1983. On some models of fertility selection. *Genetics* 105: 1003–10.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17: 368–76.
- Fenster CB, Galloway LF, and Chao L. 1997. Epistasis and its consequences for the evolution of natural populations. *Trends in Ecology and Evolution* 12: 282–6.
- Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52: 399–433.
- Fisher RA. 1999. *The Genetical Theory of Natural Selection: a Complete Variorum Edition*. Oxford University Press, Oxford (originally published in 1930).
- Fitch WM. 1976. Molecular evolutionary clocks. In Ayala FJ (ed.), *Molecular Evolution*, pp. 160–78. Sinauer Associates, Sunderland, MA.
- Flint-Garcia SA, Thornsberry JM, and Buckler ES. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54: 357–74.
- Ford EB. 1964. *Ecological Genetics*. Chapman and Hall, London.
- Ford MJ. 2002. Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* 11: 1245–62.
- Franklin IR. 1980. Evolutionary change in small populations. In Soule ME and Wilcox BA (eds), *Conservation Biology: An Evolutionary-Ecological Perspective*, pp. 135–50. Sinauer Associates, Sunderland, MA.
- Franklin IR and Frankham R. 1998. How large must populations be to retain evolutionary potential? *Animal Conservation* 1: 69–70.
- Frankham R. 1995. Effective population size/adult population size ratios in wildlife: a review. *Genetical Research* 66: 95–107.
- Frankham R. 1998. Inbreeding and extinction: island populations. *Conservation Biology* 12: 665–75.
- Frankham R and Franklin IR. 1998. Response to Lynch and Lande. *Animal Conservation* 1: 73.
- Fry JD and Heinsohn SL. 2002. Environmental dependence of mutational parameters for viability in *Drosophila melanogaster*. *Genetics* 161: 1155–67.
- Fu Y-X. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* 145: 557–70.
- Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915–25.
- Fu Y-X and Li W-H. 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
- Fu Y-X and Huai H. 2003. Estimating mutation rate: how to count mutations? *Genetics* 164: 797–805.
- Gao LZ. 2005. Microsatellite variation within and among populations of *Oryza officinalis* (Poaceae), an endangered wild rice from China. *Molecular Ecology* 14: 4287–97.
- Garrigan D and Hedrick PW. 2003. Perspective: detecting adaptive molecular evolution, lesions from the MHC. *Evolution* 57: 1707–22.
- Gaut BS. 1998. Molecular clocks and nucleotide substitution rates in higher plants. *Evolutionary Biology* 30: 93–120.
- Gavrilets S. 2004. *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, NJ.
- Geertjes GJ, Postema J, Kamping A, van Delden W, Videler JJ, and van de Zande L. 2004. Allozymes and RAPDs detect little genetic population substructuring in the Caribbean stoplight parrotfish *Sparisoma viride*. *Marine Ecology Progress Series* 279: 225–35.
- Getty T. 1999. What do experimental studies tell us about group selection in nature? *American Naturalist* 154: 596–8.
- Gigord LDB, Macnair MR, and Smithson A. 2001. Negative frequency-dependent selection maintains a dramatic flower color polymorphism in the rewardless orchid *Dactylorhiza sambucina* (L.) Soò. *Proceedings of the National Academy of Sciences USA* 98: 6253–5.
- Gillespie JH. 1989. Lineage effects and the index of dispersion of molecular evolution. *Molecular Biology and Evolution* 6: 636–47.
- Gillespie JH. 1991. *The Causes of Molecular Evolution*. Oxford University Press, New York.
- Gillespie JH. 1994. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Evolution* 48: 1101–13.
- Gillespie JH. 1995. On Ohta's hypothesis: most amino acid substitutions are deleterious. *Journal of Molecular Evolution* 40: 64–9.

- Gillespie JH. 2000. The neutral theory in an infinite population. *Gene* 261: 11–18.
- Gillespie JH. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55: 2161–9.
- Gillespie JH and Langley CH. 1979. Are evolutionary rates really variable? *Journal of Molecular Evolution* 13: 27–34.
- Gillham NW. 2001. Sir Francis Galton and the birth of eugenics. *Annual Review of Genetics* 35: 83–101.
- Gillooly JF, Allen AP, West GB, and Brown JH. 2005. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences USA* 102: 140–5.
- Glazko GV and Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution* 20: 424–34.
- Goldstein DB and Pollock DD. 1997. Launching microsatellites: a review of the mutation processes and methods of phylogenetic inference. *Journal of Heredity* 88: 335–42.
- Goodman SJ. 1997. R<sub>ST</sub> Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* 6: 881–5.
- Goodnight CJ. 1987. On the effect of founder events on epistatic genetic variance. *Evolution* 41: 80–91.
- Goodnight CJ. 1988. Epistasis and the effect of founder events on the additive genetic variance. *Evolution* 42: 441–54.
- Goodnight CJ. 2000. Modeling gene interaction in structured populations. In Wolf JB, Brodie ED III, and Wade MJ (eds), *Epistasis and the Evolutionary Process*, pp. 129–45. Oxford University Press, Oxford.
- Goodnight CJ and Stevens L. 1997. Experimental studies of group selection: what do they tell us about group selection in nature? *American Naturalist* 150: 559–79.
- Goodnight CJ and Wade MJ. 2000. The ongoing synthesis: a reply to Coyne, Barton, and Turelli. *Evolution* 54: 317–24.
- Goudsmit J, de Ronde A, Ho DD, and Perelson AS. 1996. Human immunodeficiency virus fitness in vivo: calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. *Journal of Virology* 70: 5662–4.
- Graustein A, Gaspar JM, Walters JR, and Palopoli MF. 2002. Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*. *Genetics* 161: 99–107.
- Gu X and Li W-H. 1992. Higher rates of amino acid substitution in rodents than in humans. *Molecular Phylogenetics and Evolution* 1: 211–14.
- Haderlein KP and Liberman U. 1975. Selection models with fertility differences. *Journal of Mathematical Biology* 2: 19–32.
- Haldane JBS. 1924. A mathematical theory of natural and artificial selection. Part II. *Proceedings of the Cambridge Philosophical Society Biological Sciences* 1: 158–63.
- Haldane JBS. 1937. The effect of variation on fitness. *American Naturalist* 71: 337–49.
- Haldane JBS. 1940. The conflict between selection and mutation of harmful recessive genes. *Annals of Eugenics* 10: 417–21.
- Haldane JBS. 1957. The cost of natural selection. *Journal of Genetics* 55: 511–24.
- Haley CS and Knott SA. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69: 315–24.
- Hamilton MB and Miller JR. 2002. Comparing relative rates of pollen and seed gene flow in the island model using nuclear and organelle measures of population structure. *Genetics* 162: 1897–1909.
- Hamilton MB, Braverman JM, and Soria D. 2003. Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in New World species of the Lecythidaceae. *Molecular Biology and Evolution* 20: 1710–21.
- Hanski I and Simberloff D. 1997. The metapopulation approach, its history, conceptual domain, and application to conservation. In Hanski I and Gilpin ME (eds), *Metapopulation Biology: Ecology, Genetics, and Evolution*, pp. 5–26. Academic Press, San Diego, CA.
- Hardy GH. 1908. Mendelian proportions in a mixed population. *Science* 18: 49–50.
- Hasegawa M, Kishino H, and Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160–74.
- Hastings A. 1981. Disequilibrium, selection, and recombination: limits in two-locus, two-allele models. *Genetics* 98: 659–68.
- Hastings A. 1985. Stable equilibria at two loci in populations with large selfing rates. *Genetics* 109: 215–28.
- Hastings A. 1986. Limits to the relationship among recombination, disequilibrium and epistasis in two-locus models. *Genetics* 113: 177–85.
- Hastings A. 1987. Monotonic change of the mean phenotype in two-locus models. *Genetics* 117: 583–5.
- Hauser L, Adcock GJ, Smith PJ, Ramirez JHB, and Carvalho GR. 2002. Loss of microsatellite diversity and low effective population size in an overexploited population of New Zealand snapper (*Pagrus auratus*). *Proceedings of the National Academy of Sciences USA* 99: 11742–7.
- Hayes B and Goddard ME. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* 33: 209–29.
- Heckel G, Burri R, Fink S, Desmet JF, and Excoffier L. 2005. Genetic structure and colonization processes in European populations of the common vole, *Microtus arvalis*. *Evolution* 59: 2231–42.
- Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117: 331–41.
- Hedrick P. 2004. Estimation of relative fitnesses from relative risk data and the predicted future of haemoglobin alleles S and C. *Journal of Evolutionary Biology* 17: 221–4.

- Hedrick P, Jain S, and Holden L. 1978. Multilocus systems in evolution. In Hecht MK, Steere WC, and Wallace B (eds), *Evolutionary Biology Vol. 11*, pp. 104–84. Plenum Press, New York.
- Hein J, Schierup MH, and Wiuf C. 2005. Gene genealogies, variation and evolution. Oxford University Press, New York.
- Hellmann I, Ebersberger I, Ptak SE, Pääbo S, and Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics* 72: 1527–35.
- Heywood JS. 1986. The effect of plant size variation on genetic drift in populations of annuals. *American Naturalist* 127: 851–61.
- Hill WG and Robertson A. 1968. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38: 226–31.
- Hoffman EA, Schueler FW, and Blouin MS. 2004. Effective population sizes and temporal stability of genetic structure in *Rana pipiens*, the Northern leopard frog. *Evolution* 58: 2536–45.
- Holsinger KE and Feldman MW. 1985. Selection in complex genetic systems. VI. Equilibrium properties of two locus selection models with partial selfing. *Theoretical Population Biology* 28: 117–32.
- Hori M. 1993. Frequency-dependent natural selection in the handedness of scale-eating cichlid fish. *Science* 260: 216–19.
- Houle D. 1992. Comparing evolvability and variability of quantitative traits. *Genetics* 130: 195–204.
- Houle D, Morikawa B, and Lynch M. 1996. Comparing mutational variabilities. *Genetics* 143: 1467–83.
- Hu X-S and Ennos RA. 1999. Impacts of seed and pollen flow on population genetic structure for plant genomes with three contrasting modes of inheritance. *Genetics* 152: 441–50.
- Hubby JL and Lewontin RC. 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* 54: 577–94.
- Hubner RA, Muir KR, Liu J-F, Logan RFA, Grainge M, Armitage N, Shepherd V, Popat S, Houlston RS, and the United Kingdom Colorectal Adenoma Prevention Consortium. 2006. Genetic variants of *UGT1A6* influence risk of colorectal adenoma recurrence. *Clinical Cancer Research* 12: 6585–9.
- Hudson RR. 1990. Gene genealogies and the coalescent process. In Futuyma D and Antonovics J (eds), *Oxford Surveys in Evolutionary Biology*, pp. 1–44. Oxford University Press, Oxford.
- Hudson RR. 1994. How can low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination be explained? *Proceedings of the National Academy of Sciences USA* 91: 6815–18.
- Hudson RR and Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* 120: 831–40.
- Hudson RR, Kreitman M, and Agude M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–9.
- Hudson RR, Slatkin M, and Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132: 583–9.
- Hunter DJ. 2005. Gene-environment interactions in human diseases. *Nature Reviews Genetics* 6: 287–98.
- Husband BC and Schemske DW. 1996. Evolution of the magnitude and timing of inbreeding depression in plants. *Evolution* 50: 54–70.
- Huxley JS. 1942. *Evolution: The Modern Synthesis*. Allen and Unwin, London.
- Hyne V and Kearsey MJ. 1995. QTL analyses – further uses of marker regression. *Theoretical and Applied Genetics* 91: 471–6.
- Imhof M and Schlotterer C. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proceedings of the National Academy of Sciences USA* 98: 1113–17.
- Ingvarsson P. 2004. Population subdivision and the Hudson–Kreitman–Aguadé test: testing for deviations from the neutral model in organelle genomes. *Genetical Research* 83: 31–9.
- Innan H. 2006. Modified Hudson–Kreitman–Aguadé test and two-dimensional evaluation of neutrality tests. *Genetics* 173: 1725–33.
- Innan H, Kangyu Zhang K, Marjoram P, Tavaré S, and Rosenberg NA. 2005. Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* 169: 1763–77.
- Jacquard A. 1975. Inbreeding: one word, several meanings. *Theoretical Population Biology* 7: 338–63.
- Janecek LL, Honeycutt RL, Adkins RM, and Davis SK. 1996. Mitochondrial gene sequences and the molecular systematics of the Artiodactyl subfamily Bovinae. *Molecular Phylogenetics and Evolution* 6: 107–19.
- Jones AG and Ardren WR. 2003. Methods of parentage analysis in natural populations. *Molecular Ecology* 12: 2511–23.
- Jones AG, Arnold SJ, and Bürger R. 2003. Stability of the G-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. *Evolution* 57: 1747–60.
- Jones KN and Reithel JS. 2001. Pollinator-mediated selection on a flower color polymorphism in experimental populations of *Antirrhinum* (Scrophulariaceae). *American Journal of Botany* 88: 447–45.
- Jones MW and Hutchings JA. 2002. Individual variation in Atlantic salmon fertilization success: implications for effective population size. *Ecological Applications* 12: 184–93.
- Jorde LB. 1997. Inbreeding in human populations. In Dulbecco R (ed.), *Encyclopedia of Human Biology*, vol. 5, pp. 1–13. Academic Press, San Diego, CA.
- Joshi A, Bo MH, and Mueller LD. 1999. Poisson distribution of male mating success in laboratory populations

- of *Drosophila melanogaster*. *Genetical Research* 73: 239–49.
- Judson HF. 2001. Talking about the genome. *Nature* 409: 769.
- Jukes TH and Cantor CR. 1969. Evolution of protein molecules. In Munro HN (ed.), *Mammalian Protein Metabolism*, pp. 21–132. Academic Press, New York.
- Kaplan NL, Darden T, and Hudson RR. 1988. The coalescent process in models with selection. *Genetics* 120: 819–29.
- Kay KM, Whittall JB, and Hodges SA. 2006. A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BioMed Central Evolutionary Biology* 6: 36.
- Kearsey MJ, Poni HS, and Syed NH. 2003. Genetics of quantitative traits in *Arabidopsis thaliana*. *Heredity* 91: 456–64.
- Kerns JA, Cargill EJ, Clark LA, Candille SI, Berryere TG, Olivier M, Lust G, Todhunter RJ, Schmutz SM, Murphy KE, and Barsh GS. 2007. Linkage and segregation analysis of black and brindle coat color in domestic dogs. *Genetics* 176: 1679–89.
- Kerr WE and Wright S. 1954. Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila melanogaster*: I. Forked. *Evolution* 8: 172–7.
- Kim U-K, Jorgenson E, Coon H, Leppert M, Risch N, and Drayna D. 2003. Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. *Science* 299: 1221–4.
- Kimura M. 1953. "Stepping stone" model of population. *Annual Report of the National Institute of Genetics, Japan* 3: 62–3.
- Kimura M. 1955. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences USA* 41: 144–50.
- Kimura M. 1960. Optimum mutation rate and degree of dominance as determined by the principle of genetic load. *Journal of Genetics* 57: 21–34.
- Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47: 713–19.
- Kimura M. 1967. On evolutionary adjustment of spontaneous mutation rates. *Genetical Research* 9: 23–34.
- Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217: 624–6.
- Kimura M. 1970. The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genetical Research Cambridge* 15: 131–3.
- Kimura M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16: 111–20.
- Kimura M. 1983a. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kimura M. 1983b. The neutral theory of molecular evolution. In Nei M and Koehn RK (eds), *Evolution of Genes and Proteins*, pp. 208–33. Sinauer Associates, Sunderland, MA.
- Kimura M and Crow JF. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–38.
- Kimura M and Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49: 561–76.
- Kimura M and Maruyama T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54: 1337–51.
- Kimura M and Ohta T. 1969a. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61: 763–71.
- Kimura M and Ohta T. 1969b. The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63: 701–9.
- Kimura M and Ohta T. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences USA* 75: 2868–72.
- Kimura M, Maruyama T, and Crow JF. 1963. The mutation load in small populations. *Genetics* 48: 1303–12.
- King JL and Jukes TH. 1969. Non-Darwinian evolution. *Science* 164: 788–98.
- Korber B, Theiler J, and Wolinsky S. 1998. Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science* 280: 1868–71.
- Kraaijeveld-Smit FJL, Beebee TJC, Griffiths RA, Moore RD, and Schley L. 2005. Low gene flow but high genetic diversity in the threatened Mallorcan midwife toad *Alytes muletensis*. *Molecular Ecology* 14: 3307–15.
- Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304: 412–17.
- Kronforst MR and Flemming TH. 2001. Lack of genetic differentiation among widely spaced subpopulations of a butterfly with home range behavior. *Heredity* 86: 243–50.
- Kroymann J and Mitchell-Olds T. 2005. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* 435: 95–8.
- Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain-body size allometry. *Evolution* 33: 402–16.
- Lande R. 1995. Mutation and conservation. *Conservation Biology* 9: 782–91.
- Lande R and Arnold SJ. 1983. The measurement of selection on correlated characters. *Evolution* 37: 1210–26.
- Lande R and Schemske DW. 1985. The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* 39: 24–40.
- Langley CH and Fitch WM. 1974. An examination of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution* 3: 161–77.

- Laroche J, Li P, and Bousquet J. 1995. Mitochondrial DNA and monocot-dicot divergence time. *Molecular Biology and Evolution* 12: 1151–6.
- Latter BDH. 1973. The island model of population differentiation: a general solution. *Genetics* 73: 147–57.
- Laurie CC, Chasalow SD, LeDeaux JR, McCarroll R, Bush D, Hauge B, Lai, Clark CD, Rocheford TR, and Dudley JW. 2004. The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics* 168: 2141–55.
- Lehmann T, Graham DH, Dahl ER, Bahia-Oliveira LMG, Gennari SM, and Dubey JP. 2004. Variation in the structure of *Toxoplasma gondii* and the roles of selfing, drift, and epistatic selection in maintaining linkage disequilibria. *Infection, Genetics and Evolution* 4: 107–14.
- Leitner T and Albert J. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences USA* 96: 10752–7.
- Levin DA. 1977. The organization of genetic variability in *Phlox drummondii*. *Evolution* 31: 477–94.
- Levin DA. 1978. Genetic variation in annual *Phlox*: self-compatible versus self-incompatible species. *Evolution* 32: 245–63.
- Lewin B. 2003. *Genes VIII*. Prentice Hall, Upper Saddle River, NJ.
- Lewontin RC. 1974. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.
- Lewontin RC and Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54: 595–609.
- Li W-H, Tanimura M, and Sharp PM. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *Journal of Molecular Evolution* 25: 330–42.
- Li W-H, Ellsworth DL, Krushkal J, Chang BH-J, and Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Molecular Phylogenetics and Evolution* 5: 182–7.
- Lynch CB. 1977. Inbreeding effects upon animals derived from a wild population of *Mus musculus*. *Evolution* 31: 526–37.
- Lynch M and Hill WG. 1986. Phenotypic evolution by neutral mutation. *Evolution* 40: 915–35.
- Lynch M and Gabriel W. 1990. Mutation load and the survival of small populations. *Evolution* 44: 1725–37.
- Lynch M and Lande R. 1998. The critical effective population size for a genetically secure population. *Animal Conservation* 1: 70–2.
- Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, and Willis J. 1999. Perspective: spontaneous deleterious mutation. *Evolution* 53: 645–63.
- Mackay TFC. 2001. The genetic architecture of quantitative traits. *Annual Reviews of Genetics* 35: 303–39.
- Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–36.
- Majerus MEN. 1998. *Melanism: Evolution in Action*. Oxford University Press, New York.
- Malmberg RL and Mauricio R. 2005. QTL-based evidence for the role of epistasis in evolution. *Genetical Research Cambridge* 86: 89–95.
- Manchenko GP. 2003. *Handbook of Detection of Enzymes on Electrophoretic Gels*. CRC Press, Boca Raton, FL.
- Martin AP and Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences USA* 90: 4087–91.
- Martin NH, Bouck AC, and Arnold ML. 2007. The genetic architecture of reproductive isolation in Louisiana irises: flowering phenology. *Genetics* 175: 1803–12.
- Maruyama T and Kimura M. 1980. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *Proceedings of the National Academy of Sciences USA* 77: 6710–14.
- Maynard Smith J. 1978. *The Evolution of Sex*. Cambridge University Press, Cambridge.
- Maynard Smith J and Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* 23: 23–35.
- McCauley DE, Raveill J, and Antonovics J. 1995. Local founding events as determinants of genetic structure in a plant metapopulation. *Heredity* 75: 630–6.
- McDonald DB and Potts WK. 1997. DNA microsatellites as genetic markers at several scales. In Mindell DP (ed.), *Avian Molecular Evolution and Systematics*, pp. 29–48. Academic Press, San Diego, CA.
- McDonald JH and Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–4.
- Meagher S, Penn DJ, and Potts WK. 2000. Male-male competition magnifies inbreeding depression in wild house mice. *Proceedings of the National Academy of Sciences USA* 97: 3324–9.
- Meagher TR. 1986. Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents. *American Naturalist* 128: 199–215.
- Modiano D, Luoni G, Sirima BS, Simporé J, Verra F, Konaté A, Rastrelli E, Olivieri A, Calissano C, Paganotti GM et al. 2001. Haemoglobin C protects against clinical *Plasmodium falciparum* malaria. *Nature* 414: 305–8.
- Moehring AJ, Llopart A, Elwyn S, Coyne JA, and Mackay TFC. 2006. The genetic basis of prezygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to mating preference. *Genetics* 173: 215–23.
- Moll RH, Lindsey MF, and Robinson HF. 1964. Estimates of genetic variances and level of dominance in maize. *Genetics* 49: 411–23.
- Moose SP, Dudley JW, and Rocheford TR. 2004. Maize selection passes the century mark: a unique resource

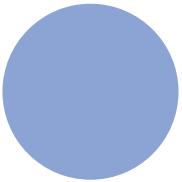
- for 21st century genomics. *Trends in Plant Science* 9: 1358–64.
- Morton NE. 1971. Population genetics and disease control. *Social Biology* 8: 243–51.
- Mousseau TA and Roff DA. 1987. Natural selection and the heritability of fitness components. *Heredity* 59: 181–97.
- Mousset S, Derome N, and Veuille M. 2004. A test of neutrality and constant population size based on the mismatch distribution. *Molecular Biology and Evolution* 21: 724–31.
- Mukai T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics* 50: 1–19.
- Mukai T, Chigusa SI, Mettler LE, and Crow JF. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics* 72: 335–55.
- Muller HJ. 1950. Our load of mutations. *American Journal of Human Genetics* 2: 111–76.
- Muller HJ. 1964. The relation of recombination to mutational advance. *Mutation Research* 1: 2–9.
- Muse SV and Weir BS. 1992. Testing for the equality of evolutionary rates. *Genetics* 132: 269–76.
- Nachman MW and Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156: 297–304.
- Narain P. 1970. A note on the diffusion approximation for the variance of the number of generations until fixation of a neutral mutant gene. *Genetical Research Cambridge* 15: 251–5.
- National Institute of General Medical Sciences. 1998. *The Genetic Architecture of Complex Traits Workshop*. [www.nigms.nih.gov/News/Reports/genetic\\_arch.htm](http://www.nigms.nih.gov/News/Reports/genetic_arch.htm).
- Nei M. 1972. Genetic distance between populations. *American Naturalist* 106: 283–92.
- Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA* 70: 3321–3.
- Nei M. 1978. The theory of genetic distance of the evolution of human races. *Japanese Journal of Human Genetics* 23: 341–69.
- Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Nei M and Roychoudhury AK. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76: 379–90.
- Nei M and Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences USA* 76: 5269–73.
- Nei M and Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
- Nei M, Chakravarti A, and Tateno Y. 1977. Mean and variance of  $F_{ST}$  in a finite number of incompletely isolated populations. *Theoretical Population Biology* 11: 291–306.
- Neuhauser C. 1999. The ancestral graph and gene genealogy under frequency-dependent selection. *Theoretical Population Biology* 56: 203–14.
- Neuhauser C and Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* 145: 519–34.
- Newby JC. 1980. *Mathematics for the Biological Sciences*. Oxford University Press, Oxford.
- Nordborg M. 1997. Structured coalescent processes on different time scales. *Genetics* 146: 1501–14.
- Nunez-Farfán J and Schlüchtung CD. 2001. Evolution in changing environments: the “synthetic” work of Clausen, Keck, and Hiesey. *Quarterly Review of Biology* 76: 433–57.
- Nunney L. 1993. The influence of mating system and overlapping generations on effective population size. *Evolution* 47: 1329–41.
- Ohta T. 1972. Evolutionary rate of cistrons and DNA divergence. *Journal of Molecular Evolution* 1: 150–7.
- Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annual Reviews of Ecology and Systematics* 23: 263–86.
- Ohta T. 1993. An examination of the generation-time effect on molecular evolution. *Proceedings of the National Academy of Sciences USA* 90: 10676–80.
- Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* 40: 56–63.
- Ohta T and Kimura M. 1969. Linkage disequilibrium due to random genetic drift. *Genetical Research* 13: 47–55.
- Ohta T and Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution* 1: 18–25.
- Ollivier L and James JW. 2004. Predicting annual effective size of livestock populations. *Genetical Research* 84: 41–6.
- Orel V. 1996. *Gregor Mendel: the First Geneticist*. Oxford University Press, Oxford.
- Orr HA. 1998a. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52: 935–49.
- Orr HA. 1998b. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait loci. *Genetics* 149: 2099–104.
- Orr HA. 2003. The distribution of fitness effects among beneficial mutations. *Genetics* 163: 1519–26.
- Otto SP and Jones CD. 2000. Detecting the undetected: estimating the total number of loci underlying a quantitative trait. *Genetics* 156: 2093–107.
- Ouma JO, Marquez JG, and Krafur ES. 2005. Macro-geographic structure of the tsetse fly, *Glossina pallidipes* (Diptera: Glossinidae). *Bulletin of the Entomological Research* 95: 437–47.
- Page RDM and Holmes EC. 1998. *Molecular Evolution: a Phylogenetic Approach*. Blackwell Science, Malden, MA.
- Peck SL, Ellner SP, and Gould F. 1998. A spatially explicit stochastic model demonstrates the feasibility

- of Wright's shifting balance theory. *Evolution* 52: 1834–9.
- Peck SL, Ellner SP, and Gould F. 2000. Varying migration and deme size and the feasibility of the shifting balance. *Evolution* 54: 324–7.
- Penn D and Potts W. 1998. MHC-disassortative mating preferences reversed by cross-fostering. *Proceedings of the Royal Society of London Series B Biological Sciences* 265: 1299–1306.
- Penrose LS. 1949. The meaning of fitness in human populations. *Annals of Eugenics* 14: 301–4.
- Perola M, Sammalisto S, Hiekkalinna T, Martin NG, Visscher PM, Montgomery GW, Benyamin B, Harris JR, Boomsma D, Willemsen G et al. and GenomEUtwin Project. 2007. Combined genome scans for body stature in 6,602 European twins: Evidence for common Caucasian loci. *PLoS Genetics* 3(6): e97.
- Petrov DA. 2008. Codon use bias. *Annual Review of Genetics* 42 (in press).
- Phillips PC. 1998. The language of gene interactions. *Genetics* 149: 1167–71.
- Phillips PC, Otto SP, and Whitlock MC. 2000. Beyond the average: the evolutionary importance of gene interactions and variability of epistatic effects. In Wolf JB, Brodie ED, and Wade MJ (eds), *Epistasis and the Evolutionary Process*, pp. 20–38. Oxford University Press, New York.
- Pigliucci M. 2001. *Phenotypic Plasticity*. Johns Hopkins University Press, Baltimore, MD.
- Pitra C, Hansen AJ, Lieckfeldt D, and Arctander P. 2002. An exceptional case of historical outbreeding in African sable antelope populations. *Molecular Ecology* 11: 1197–1208.
- Pollak E. 1978. With selection for fecundity the mean fitness does not necessarily increase. *Genetics* 90: 383–9.
- Posada D and Crandall KA. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817–818.
- Posada D and Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Systematic Biology* 50: 580–60.
- Pray LA, Goodnight CJ, Stevens L, Schwartz JM, and Yan G. 1996. The effect of population size on effective population size: an empirical study in the red flour beetle *Tribolium castaneum*. *Genetical Research* 68: 151–5.
- Provine WB. 1971. *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago, IL.
- Provine WB. 1986. *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago, IL.
- Przeworski M, Charlesworth B, and Wall JD. 1999. Genealogies and weak purifying selection. *Molecular Biology and Evolution* 16: 246–52.
- Ptak SE and Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics* 18: 559–63.
- Rand DM. 1994. Thermal habit, metabolic rate and the evolution of mitochondrial DNA. *Trends in Ecology and Evolution* 9: 125–31.
- Renwick A, Bonnen PE, Trikka D, Nelson DL, Chakraborty R, and Kimmel M. 2003. Sampling properties of estimators of nucleotide diversity at discovered SNP sites. *International Journal of Applied Mathematics and Computer Science* 13: 385–94.
- Rice SH. 2004. *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates, Sunderland, MA.
- Rieseberg LH, Widmer A, Arntz AM, and Burke JM. 2002. Directional selection is the primary cause of phenotypic diversification. *Proceedings of the National Academy of Sciences USA* 99: 12242–5.
- Robertson A. 1952. The effect of inbreeding on variation due to recessive genes. *Evolution* 37: 1195–1208.
- Rodriguez F, Oliver JF, Marin A, and Medina JR. 1990. The general stochastic model of nucleotide substitutions. *Journal of Theoretical Biology* 142: 485–501.
- Roff DA and Mousseau TA. 1987. Quantitative genetics and fitness: lessons from *Drosophila*. *Heredity* 58: 103–18.
- Rogers AR and Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution* 9: 552–69.
- Romero C, Pedryc A, Munoz V, Llacer G, and Badenes ML. 2003. Genetic diversity of different apricot geographical groups determined by SSR markers. *Genome* 46: 244–52.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. *Science* 298: 2381–5.
- Roughgarden J. 1996. *Theory of Population Genetics and Evolutionary Ecology: an Introduction*. Prentice Hall, Upper Saddle River, NJ.
- Roychoudhury AK and Nei M. 1988. *Human Polymorphic Genes: World Distribution*. Oxford University Press, New York.
- Rozen DE, de Visser JAGM, and Gerrish PJ. 2002. Fitness effects of fixed beneficial mutations in microbial populations. *Current Biology* 12: 1040–5.
- Ruse M. 1996. Are pictures really necessary? The case of Sewall Wright's "adaptive landscapes." In Baigrie B (ed.), *Picturing Knowledge: Historical and Philosophical Problems Concerning the Use of Art in Science*, pp. 303–37. University of Toronto Press, Toronto.
- Russell LB and Russell WL. 1996. Spontaneous mutations recovered as mosaics in the mouse specific-locus test. *Proceedings of the National Academy of Sciences USA* 93: 13072–7.
- Salemi M, Strimmer K, Hall WM, Duffy M, Delaporte E, Mboup S, Peeters M, and Vandamme A-M. 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB Journal* 15: 276–8.
- Sanjuan R, Moya A, and Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences USA* 101: 8396–401.

- Sarich VM and Wilson AC. 1967. Immunological time scale for hominid evolution. *Science* 158: 1200–3.
- Sax K. 1923. The association of size difference with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8: 552–60.
- Schaal BA. 1980. Measurement of gene flow in *Lupinus texensis*. *Nature* 284: 450–1.
- Schemske DW and Bierzychudek P. 2001. Perspective: Evolution of flower color in the desert annual *Linanthus parryae*: Wright revisited. *Evolution* 55: 1269–82.
- Schierup MH and Hein J. 2000. Recombination and the molecular clock. *Molecular Biology and Evolution* 17: 1578–9.
- Schierup MH, Vekemans X, and Christiansen FB. 1998. Allelic genealogies in sporophytic self-incompatibility systems in plants. *Genetics* 150: 1187–98.
- Schlager G and Dickie MM. 1971. Natural mutation rates in the house mouse: estimates for five specific loci and dominant mutations. *Mutation Research* 11: 89–96.
- Schneider S and Excoffier L. 1999. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 152: 1079–89.
- Selvin S. 1980. Probability of nonpaternity determined by multiple allele codominant systems. *American Journal of Human Genetics* 32: 276–8.
- Shaw RG, Byers DL, and Darmo E. 2000. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* 155: 369–78.
- Shrimpton AE and Robertson A. 1988. The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. II. Distribution of third chromosome bristle effects within chromosome sections. *Genetics* 118: 445–459.
- Silbermann R and Tatar M. 2000. Reproductive costs of transgenic *hsp70* overexpression in *Drosophila*. *Evolution* 54: 2038–45.
- Singh RS and Rhomberg LR. 1987. A comprehensive study of genetic variation in natural populations of *Drosophila melanogaster*. II. Estimates of heterozygosity and patterns of geographic differentiation. *Genetics* 117: 255–71.
- Skipper RA. 2004. The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. *Philosophy of Science* 71: 1176–88.
- Slate J. 2005. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology* 14: 363–79.
- Slatkin M. 1977. Gene flow and genetic drift in a species subject to frequent local extinctions. *Theoretical Population Biology* 12: 253–62.
- Slatkin M. 1987a. Gene flow and the geographic structure of natural populations. *Science* 236: 787–92.
- Slatkin M. 1987b. The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* 32: 42–9.
- Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genetical Research* 58: 167–75.
- Slatkin M. 1994. Linkage disequilibrium in stable and growing populations. *Genetics* 137: 331–6.
- Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–62.
- Slatkin M and Barton NH. 1989. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* 43: 1349–68.
- Slatkin M and Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555–62.
- Snow AA and Palma PM. 1997. Commercialization of transgenic plants: potential ecological risks. *BioScience* 47: 86–96.
- Steinberg EK, Lindner KR, Gallea J, Maxwell A, Meng J, and Allendorf FW. 2002. Rates and patterns of microsatellite mutations in pink salmon. *Molecular Biology and Evolution* 19: 1198–1202.
- Stephan SJ, Phillips PC, and Houle D. 2002. Comparative quantitative genetics: evolution of the G matrix. *Trends in Ecology and Evolution* 17: 320–7.
- Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117: 149–53.
- Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, Zhu L, Padhukasahasram B, Karlins E, Davis S, Jones PG et al. 2007. A single *IGF1* allele as a major determinant of small size in dogs. *Science* 316: 112–15.
- Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–95.
- Tajima F. 1989b. The effect of change in population size in DNA polymorphism. *Genetics* 123: 597–601.
- Tajima F. 1993a. Simple methods for testing molecular clock hypothesis. *Genetics* 135: 599–607.
- Tajima F. 1993b. Measurement of DNA polymorphism. In Takahata N and Clark AG (eds), *Mechanisms of Molecular Evolution*, pp. 37–59. Sinauer Associates, Sunderland, MA.
- Tajima F. 1996. The amount of DNA polymorphism maintained in a finite populations when the neutral mutation rate varies among sites. *Genetics* 143: 1457–65.
- Takahata N and Nei M. 1984.  $F_{ST}$  and  $G_{ST}$  statistics in the finite island model. *Genetics* 107: 501–4.
- Templeton AR and Read B. 1994. Inbreeding: one word, several meanings, much confusion. In Loeschke V, Tomiuk J, and Jain SK (eds), *Conservation Genetics*, pp. 92–105. Birkhauser Verlag, Basel.
- Tishkoff SA and Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics* 4: 293–340.
- Turelli M, Schemske DW, and Bierzychudek P. 2001. Stable two-allele polymorphisms maintained by fluctuating fitnesses and seed banks: Protecting the blues in *Linanthus parryae*. *Evolution* 55: 1283–98.
- Turner JRG. 1981. “Fundamental theorem” for two loci. *Genetics* 99: 365–9.

- Turner JRG. 1992. Stochastic processes in populations: the horse behind the cart? In Berry RJ, Crawford TJ, and Hewitt GM (eds), *Genes in Ecology*, pp. 29–53. The 33rd Symposium of the British Ecological Society, University of East Anglia. Blackwell Scientific, Oxford.
- Turner TF, Wares JP, and Gold JR. 2002. Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish (*Sciaenops ocellatus*). *Genetics* 162: 1329–39.
- Uyenoyama MK. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. *Genetics* 147: 1389–1400.
- Valdes AM, Slatkin M, and Freimer NB. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737–49.
- Van Buskirk J and Willi Y. 2006. The change in quantitative genetic variation with inbreeding. *Evolution* 60: 2428–34.
- Van Ooijen JW. 1999. LOD significance thresholds for QTL analysis in experimental populations of diploid species. *Heredity* 83: 613–24.
- Van Valen L. 1973. A new evolutionary law. *Evolutionary Theory* 1: 1–30.
- Varvio S-L, Chakraborty R, and Nei M. 1986. Genetic variation in subdivided populations and conservation genetics. *Heredity* 57: 189–98.
- Vekemans X and Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. *Genetics* 137: 1157–65.
- Viard F, Justy F, and Jarne P. 1997. Population dynamics inferred from temporal variation at microsatellite loci in the selfing snail *Bulinus truncatus*. *Genetics* 146: 973–82.
- Wade MJ. 1992. Epistasis. In Keller EF and Lloyd EA (eds), *Keywords in Evolutionary Biology*, pp. 87–91. Harvard University Press, Cambridge, MA.
- Wade MJ and McCauley DE. 1988. Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution* 42: 995–1005.
- Wade MJ and Goodnight CJ. 1991. Wright's shifting balance theory: an experimental study. *Science* 253: 1015–18.
- Wade MJ and Goodnight CJ. 1998. Perspective: The theories of Fisher and Wright in the context of metapopulations: when nature does many small experiments. *Evolution* 52: 1537–52.
- Wade MJ, Goodnight CJ, and Stevens L. 1999. Design and interpretation of experimental studies of interdemic selection: a reply to Getty. *American Naturalist* 154: 599–603.
- Wakeley J. 1998. Segregating sites in Wright's island model. *Theoretical Population Biology* 53: 166–74.
- Wakeley J. 2008. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Denver, CO.
- Wallace B. 1991. *Fifty Years of Genetic Load. An Odyssey*. Cornell University Press, Ithaca, NY.
- Wang M and Schreiber A. 2001. The impact of habitat fragmentation and social structure on the population genetics of roe deer (*Capreolus capreolus* L.) in central Europe. *Heredity* 86: 703–15.
- Waples RS. 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* 121: 379–91.
- Waples RS. 1991. Genetic methods for estimating the effective population size of cetacean populations. In Hoezel AR (ed.), *Genetic Ecology of Whales and Dolphins*, pp. 279–300. Report of the International Whaling Commission, Special issue no. 13. IWC, Cambridge.
- Watterson GA. 1975. On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* 7: 256–76.
- Wedekind C and Füri S. 1997. Body odour preferences in men and women: do they aim for specific MHC combinations or simply heterozygosity? *Proceedings of the Royal Society of London Series B Biological Sciences* 264: 1471–9.
- Weinberg W. 1908. Über vererbungsgesetze beim menschen. *Jaresh. Verein. f. vaterl. Naturk. Würtem.* 64: 368–82 (an English translation can be found in Boyer S (ed.). 1963. *Papers on Human Genetics*. Prentice-Hall, Englewood Cliffs, NJ).
- Weinreich DM, Watson RA, and Chao L. 2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59: 1165–74.
- Weinreich DM, Delaney NF, DePristo MA, and Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312: 111–14.
- Weir BS. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.
- Weir BS, Cockerham CC, and Reynolds J. 1980. The effects of linkage and linkage disequilibrium on the covariances of non-inbred relatives. *Heredity* 45: 351–9.
- Westneat DF and Stewart RK. 2003. Extra-pair paternity in birds: causes, correlates, and conflict. *Annual Review of Ecology, Evolution, and Systematics* 34: 365–96.
- Whitlock MC and McCauley DE. 1990. Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution* 44: 1717–24.
- Whitlock MC and McCauley DE. 1999. Indirect measures of gene flow and migration:  $F_{ST} \neq 1/(4Nm+1)$ . *Heredity* 82: 117–25.
- Whitlock MC, Phillips PC, Moore FB-G, and Tonsor SJ. 1995. Multiple fitness peaks and epistasis. *Annual Reviews of Ecology and Systematics* 26: 601–29.
- Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11: 398–411.
- Williams GC. 1966. *Adaptation and Natural Selection – a Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, NJ.
- Williams GC. 1992. *Natural Selection: Domains, Levels, and Challenges*. Oxford University Press, New York.

- Willis JH and Orr HA. 1993. Increased heritable variation following population bottlenecks: the role of dominance. *Evolution* 47: 949–57.
- Wolf JB, Brodie ED III, and Wade MJ (eds). 2000. *Epistasis and the Evolutionary Process*. Oxford University Press, Oxford.
- Wright S. 1921. Systems of mating. *Genetics* 6: 111–78.
- Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97–159.
- Wright S. 1932. The roles of mutation, inbreeding, cross-breeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* 1: 356–66.
- Wright S. 1943a. Isolation by distance. *Genetics* 28: 114–38.
- Wright S. 1943b. Analysis of local variability of flower color in *Linanthus parryae*. *Genetics* 28: 139–56.
- Wright S. 1946. Isolation by distance under diverse systems of mating. *Genetics* 31: 39–59.
- Wright S. 1951. The genetical structure of populations. *Annals of Eugenics* 15: 323–54.
- Wright S. 1978. *Evolution and the Genetics of Populations, vol. 4: Variability Within and Among Natural Populations*. University of Chicago Press, Chicago, IL.
- Wright S and Kerr WE. 1954. Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila melanogaster*: II. Bar. *Evolution* 8: 225–40.
- Wright SI, Lauga B, and Charlesworth D. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Molecular Ecology* 12: 1247–63.
- Xu S. 2003. Theoretical basis of the Beavis effect. *Genetics* 165: 2259–68.
- Zharkikh A. 1994. Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution* 39: 315–29.
- Zuckerkandl E. 1987. On the molecular clock. *Journal of Molecular Evolution* 26: 34–46.
- Zuckerkandl E and Pauling L. 1962. Molecular disease, evolution, and genetic heterogeneity. In Kasha M and Pullman B (eds), *Horizons in Biochemistry*, pp. 189–225. Academic Press, New York.
- Zuckerkandl E and Pauling L. 1965. Evolutionary divergence and convergence in proteins. In Bryson V and Vogel HJ (eds), *Evolving Genes and Proteins*, pp. 97–165. Academic Press, New York.



# Index

Page numbers in *italics* refer to figures, those in **bold** refer to tables and boxes, and those in **bold italics** refer to definitions. Note that figures, tables, and boxes are only indicated when they are separated from their text references.

$-a$  (genotypic value) 334–5  
 $+a$  (genotypic value) 334–5  
ABO blood groups 26–8  
*Achillea* 294, 295  
acquired characteristics, inheritance of 9  
adaptive landscapes (or topographies) 201, 208  
shifting balance model 368, 369–72  
Wright's concept 367–8  
*see also* fitness surfaces  
additive 288  
additive fecundities 217–18  
additive gene action 197–8  
effect on fitness surfaces 214, 215, 367  
genotypic scale of measurement 335  
parent–offspring regression 302, 303  
population mean phenotype and 336  
QTLs 327  
two-locus natural selection 215  
additive genetic variance ( $V_A$ ) **287**, 288, **288**, 334  
allele frequencies/genotypic values and 349–51  
change in mean fitness and 205–6  
covariance between relatives 352  
defining heritability 297–8  
derivation 348–9  
distinction from  $V_D$  289–90  
G matrix 304–5  
inheritance 291–2  
long-term response to selection and 307–10, 311–13  
neutral quantitative traits 313–15  
phenotypic distribution and 285, 286, 288  
response to selection and 300–1  
additive genetic variance ( $V_a$ ) 311–12  
admixture, population 47  
advantageous mutations 158  
effects of gametic disequilibrium 275–6, 277

time to loss/fixation 237, 238  
*see also* beneficial mutations  
alcohol dehydrogenase (*Adh*) gene 266–8, 365  
allele **12**  
average effect *see* average effect of an allele  
combinations 366–8  
counting method to estimate allele frequency 24, **30**  
fixation or loss *see* fixation/loss  
identity by descent *see* identity by descent  
mutant 158, 357  
wild-type 357, 366–7  
allele copy **89**  
allele frequencies  
average fitness and 201–3  
diffusion 68–70  
diffusion approximation 71–2  
distributions, Wright's view 368–9  
divergence among subpopulations 122–3  
DNA profiles **20**, 21  
*Drosophila* studies 66–7  
effective population size estimation 82, 83, 84–5  
equilibrium *see* equilibrium allele frequency  
estimates 24, 27–8, 30  
genetic drift effects 54, 56–7  
Hardy–Weinberg model 13  
inbreeding effects 33–4  
Markov chain model 63–5  
mean phenotypic value and 336  
mutations and 173–6  
under natural selection 193, **194**, 195–200  
population size effects 53–4  
random walk (for new mutations) 163, 236  
spatial structuring 107, 108–9, **110**  
 $V_A$  and  $V_D$  in relation to 349–51  
variability due to genetic drift 60–1  
variance in change *see*  $\Delta p$   
allelomorph combinations 366  
allozygous genotype 34, **37**  
finite population 78–9  
allozymes  
causes of polymorphism 358–9, 365  
electrophoresis **32**, 32

estimating fixation indices 31, **122**  
segregational load 361–3  
alternative hypotheses 4–5  
amplified fragment length  
polymorphisms (AFLPs) 317  
ancestor–descendant relationships *see* genealogies, gene  
ancestral polymorphism 257–60  
ancestral selection graph 228–30  
annual plants 77  
antagonistic epistasis 291  
antagonistic pleiotropy 313  
*Arabidopsis thaliana* 3  
gametic disequilibrium 48  
mutation-accumulation experiment 160  
QTLs **326**  
architecture, genetic 316, **316**  
arms race model 371  
artificial selection 300, 306–7, 342  
aspirin 294  
assortative mating 28, **30**  
genotype frequencies and 29–30  
negative 28, 29  
positive 28–9  
*see also* consanguineous mating;  
non-random mating  
assumptions 4–6  
autozygosity 34–7  
effective population size and 82  
finite population 78–80  
fixation index and 36–7  
infinite island model 136–7  
influence of mutations 177–8  
pedigree 34–7  
population mean phenotype 337  
*see also* inbreeding coefficient  
autozygous genotype 34, **37**  
average *see* mean  
average effect of an allele **337**, 337–41  
breeding value and 343, 344, **345**  
as deviation from population mean 338–9  
dominance deviation and 347–8  
substituting one allele for another 340–1  
average (mean) fitness ( $\bar{w}$ ) **188**  
under balancing selection 361, **362**  
clonal reproduction 188  
genetic load concept 360, 361–2  
increase by natural selection 200–6

- rate of change in allele frequency and 201–3  
 sexual reproduction 191–3  
 three-allele model of natural selection 210, 211  
 two-locus model of natural selection 214  
 average pairwise differences 248, 249–50  
*see also* nucleotide diversity
- backcross  
 impact on heterozygosity 35  
 mating design, QTL mapping 324  
 background selection 278, 278  
 bacteria, mutations 157, 160  
 balance hypothesis 356, 356–8, 359, 365  
 balanced polymorphism 200  
 balancing selection 198, 202  
 balance hypothesis 357–8  
 genealogies and 230–2  
 genetic hitch-hiking due to 278  
 genetic load concept 361, 362, 363  
 multi-allelic 232  
 natural selection–genetic drift balance 223  
 nonsynonymous mutations 268, 269  
 polymorphism under 237, 238, 278  
*see also* heterozygote advantage
- Bar* locus 223, 224
- bass  
 striped *see* *Morone saxatilis*  
 white 32
- Bateson, W 42
- Beavis effect 326
- beetle, flour 372–3
- behavioral traits 298, 307
- beneficial mutations 158  
 estimating frequency 159, 160  
 fixation by natural selection 164–6  
 loss due to Mendelian segregation 162  
 micromutationalism concept 164  
 time to loss 237, 238  
*see also* advantageous mutations
- Bernoulli (binomial) random variable 58–9, 60  
 variance 60, 62
- Bernstein, Felix 26
- $\beta$ -lactamase gene 373
- bi-directional (reversible) mutation model 174, 175–6, 177
- binomial probability distribution 58–61
- binomial random variable *see* Bernoulli random variable
- binomial variable 380
- biometric school 284
- biparental inbreeding 28, 30  
 effect on heterozygosity 34
- blending inheritance, model of 9
- blood groups 26–8, 358
- blood pressure 297–8
- bluebonnets, Texas 87
- body size of dogs 320–1, 326, 327  
 genotypic values 334, 335  
 mean phenotypic value 336–7  
*see also* *IGF1* gene
- bottlenecks, genetic 74, 75, 75  
 coalescent genealogies and 98–9  
 heritability changes after 315
- Boveri, Theodor 12, 41
- Brazil nut trees 111–15, 261
- breeder's equation 300, 304  
 extension for correlated traits 305  
 longer time periods 307  
 single QTL 327–8
- breeding effective population size ( $N_e^b$ ) 85–7, 87
- breeding sex ratio  
 coalescent model 97–8  
 effective population size and 75, 77
- breeding value 341–8  
 dominance deviation and 345–8  
 of an individual 342  
 mean, of all genotypes 345  
 under non-random mating 345  
 under random mating 342–5  
 variance 348–9
- bugs-in-a-box metaphor 88
- bugs-in-multiple-boxes metaphor 142
- bulked segregant analyses 325
- candidate loci 316, 316
- candidate parents 111, 116  
 example 112, 113–15, 116–17  
 exclusion 113, 116  
 exclusion probability 116, 116, 116, 117  
 inclusion 113–15  
 outcomes of parentage analysis 117–18
- carrying capacity (K) 220–1  
 density-dependent selection 221–2
- cause-and-effect relationships 2, 381
- census population size (N) 73, 73, 75–6  
 effective population size and 85
- centimorgan (cM) 325
- central limit theorem 379
- character 283
- Chi-squared ( $\chi^2$ ) test 24–5, 26  
 degrees of freedom (df) 25  
 gametic disequilibrium and 44  
 Tajima's 1D test 261
- chloroplast genomes 46, 87
- chromosome theory of heredity 12, 41
- chromosomes 41–2  
 crossing-over 42  
 genetic marker locus position 324, 325  
 rearrangements 155  
 sex determination 16, 46
- classical hypothesis 356, 356–8, 359
- clonal reproduction 45  
 natural selection with 185–9  
 population growth 185–6  
*see also* haploid populations
- coadapted gene complexes 357–8
- coalescent event (coalescence) 88, 89  
 combining migration events with 143–4, 145  
 haploid and diploid populations 89–90  
 probabilities 90–1, 93  
 times to *see* coalescent waiting times
- coalescent genealogies (trees) 87–96  
 average coalescent times 91–2  
 balancing selection and 230–2  
 building your own 94–5, 97, 181  
 directional selection and 227–30, 231  
 growing and shrinking populations 99–101  
 height 95, 96  
 with migration 144–9  
 mismatch distributions 272–4  
 with mutations 180–2  
 population bottlenecks and 98–9  
 structured populations with gene flow 142–9  
 Tajima's D test 269–70  
 total branch length 95–6  
 variation 91
- coalescent model  
 adding natural selection 227–32  
 effective population size in 96–101  
 gene genealogies 87–96  
 with mutation 178–82  
 number of segregating sites 248–9
- coalescent theory 88
- coalescent waiting times 89, 91–4  
 average 91–2  
 balancing selection 231–2  
 directional selection 229, 230  
 expressing population structure 149  
 growing and shrinking populations 99–101  
 solving two equations with two unknowns 148  
 structured populations with gene flow 146–9  
 variance in 91
- codominance 13, 197  
 marker loci for QTL mapping 317  
 phenotypic distribution 285
- codon bias 366
- coefficient of variation (CV) 287
- colonization 141–2
- colorectal adenoma 294
- common environmental variance ( $V_{Ec}$ ) 287, 295–6
- composite interval QTL mapping 325
- computer simulation 6–7
- consanguineous mating 28–9, 30, 33–41  
 fixation index and 31  
 heritability changes after 315  
 humans 38  
 impact on heterozygosity 33–4, 35  
 inbreeding coefficient and  
 autozygosity 34–7  
 natural selection–mutation balance 225–6  
 phenotypic consequences 37–40

- population mean phenotypic value 337  
*see also* inbreeding
- continent-island model 131, 131–4  
 irreversible mutation and 175  
 simulation 134
- continuing branch 227, 229
- continuous quantitative traits 283–6
- continuous variable 380
- controversies, historical 356–66
- corn (maize)  
 apparent overdominance 359  
 divergence time 254  
 heterosis 38  
 kernel phenotypes 28, 28, Plate 2.10  
 long-term selection experiment 310–11, 329  
 correlation 380–1  
 correlation coefficient ( $\rho$ ) 380, 381  
 squared ( $\rho^2$ ) 43
- Corythophora alta* 111–15, 116–17
- covariance 296, 380–1  
 disequilibrium 296–7  
 genotype and environment ( $\text{cov}_{GE}$ ) 296  
 genotypic values between relatives 351–3  
 mid-parent and offspring values 303, 353  
 $V_A$  and  $V_D$  ( $\text{cov}_{AD}$ ) 350
- cows, dairy 295, 329, 330
- Crow, James 169, 235, 351, 365
- cytochrome *c*, rate of evolution 251–2
- cytochrome *c* oxidase subunit II gene 243
- d* (genotypic value) 334, 335
- D* *see* diffusion coefficient; gametic disequilibrium parameter; genetic distance
- Darwin, Charles 9, 185
- Darwinian fitness *see* relative fitness
- De Finetti diagram 13, 14, 15  
 fitness surface 208  
 fundamental theorem of natural selection 205
- deductive reasoning 3–4, 4
- degree of relatedness 34
- degrees of freedom (df) 25
- deleterious mutations 158  
 background selection and 278  
 classical hypothesis 357  
 estimating frequency 159, 160  
 fixation 166–8  
 time to loss 237, 238
- deleterious recessive alleles 38, 39–40  
 natural selection–mutation balance 225–6
- deletion 154
- $\Delta p$  82, 83, 84–5
- $\Delta\bar{\alpha}$  305
- deme 87, 87
- Denny, MW 68, 76
- density-dependent natural selection 219–22
- desert snow (*Linanthus parryae*) 106–7, Plate 4.2
- deterministic processes 195
- Dickerson, Richard 251–2
- diffusion 68–70
- diffusion approximation of genetic drift 67–73, Plate 3.13
- diffusion coefficient ( $D$ ) 68, 69–70, 71
- diffusion equation *see* diffusion approximation of genetic drift
- dinosaurs 254
- dioecious individuals 31
- diploid populations  
 coalescence 89–90  
 effective population size 96–7  
 natural selection 189–93, 205–6
- directional selection  
 average fitness under 201–2  
 classical hypothesis 357, 358  
 coalescent genealogies 227–30, 231  
 polymorphism under 237, 238  
 quantitative traits 299, 300–2
- disassortative mating 28, 29
- disequilibrium covariance 296–7
- dispersal, gamete 86, 87
- dispersion  
 index of 257, 259–60, 262  
 QTL mapping 324
- disruptive selection 198, 299–300  
 mean fitness and 202  
*see also* heterozygote disadvantage
- distance, genetic ( $D$ ) 170–1
- distribution, frequency 376
- divergence 236, 236  
 dating with molecular clock 252–5  
 gametic disequilibrium and 278–9  
 measures 241–7  
 nearly neutral theory prediction 241  
 neutral theory prediction 237–40  
 recombination rate and 277–8  
 selectionist/neutralist debate 364, 365, 366  
 testing neutral theory predictions 265–7  
*see also* DNA sequence divergence
- DNA polymorphism 248–50  
 ancestral 257–60  
 nucleotide diversity ( $\pi$ ) 248, 249–50, 251  
 number of segregating sites ( $S$ ) 248–9, 251
- DNA profiling 19–22, 23  
 allele frequencies in loci used 20, 21  
 population structure adjustments 129–31
- DNA sequence divergence  
 ancestral polymorphism and 257–60  
 dating 252–5  
 measures 241–7  
 molecular clock and 250–5  
 nearly neutral theory prediction 241  
 neutral theory prediction 237–40  
 nucleotide-substitution models 244–7  
 saturation and 243–4  
 between species 242–3
- DNA sequences  
 alignment 242  
 insertions and deletions 154  
 molecular evolution 235  
 multiple alignment 248  
 mutation models 172–3  
 mutations *see* mutation(s)  
 testing neutral theory 265–74
- DNA sequencing 242, 242
- Dobzhansky, T 356, 357, 358
- Dobzhansky–Muller model 358
- dogs  
 body size *see* body size of dogs  
 coat color 290, 291  
 fixation indices 31
- dominance 13  
 average effect of an allele and 339–40  
 breeding value and 343–4  
 complete 13, 197, 291–2  
 effects on fitness surfaces 215, 367–8  
 general 193, 197–8, 360  
 genotypic scale of measurement 335  
 Mendel's studies 10–11  
 parent–offspring regression 302, 303  
 partial or incomplete 13  
 phenotypic effects of inbreeding and 37  
 population mean phenotype and 336, 337  
 QTL mapping 319, 320, 321, 322  
 QTLs 327  
*see also* codominance
- dominance coefficient ( $h$ ) 197
- dominance deviation 342, 345, 345–8  
*IGF1* locus in dogs 344, 346–7  
 mean, of all genotypes 347  
 under random mating 347  
 variance 348, 349
- dominance genetic variance ( $V_D$ ) 287, 288–90, 290, 334  
 allele frequencies/genotypic values and 349–51  
 covariance between relatives 352  
 derivation 348–9  
 heritability and 297–9  
 inheritance 291–2  
 neutral quantitative traits 315
- dominance hypothesis, inbreeding depression 38–9
- dominant phenotype, selection against 193, 196–7
- Drosophila*  
 genetic drift studies 66–7  
 genetic load 361–3
- Drosophila melanogaster* 3  
 effective population size 77–8  
 $hsp70$  313  
 mutation-accumulation experiments 159  
 natural selection–genetic drift balance 223, 224
- polymorphism and recombination rate 277–8

- QTL effect sizes 329  
testing neutral theory 266–8, 365
- Drosophila pseudoobscura* 358–9
- Drosophila sechellia* 266–7
- Drosophila simulans* 270, 277
- drug resistance, HIV 188–9
- dwell time, new mutations 236, 237, 238
- dynamics 6
- E* see environmental deviation
- ecological genetics 358
- effect size, QTL 321, 327–8  
distribution 329–30
- effective inbreeding coefficient ( $F_e$ ) 81
- effective migration rate ( $N_m$ ) 137–8, 139–41  
analogy to  $N_s$  223–5  
estimation 140–1
- effective population size ( $N_e$ ) 73, 73–8  
breeding ( $N_e^b$ ) 85–7, 87
- breeding sex ratio and 75, 77
- coalescent model 96–101
- different genomes 87
- estimation 77, 80–7
- family size variation and 76–7
- fate of new mutations and 163
- gametic disequilibrium and 48–9, 50
- haplotype frequency distributions and 273–4
- inbreeding ( $N_e^i$ ) see inbreeding  
effective population size
- mismatch distributions and 272–3, 274
- neutral evolution of quantitative traits and 314–15
- QTL response to selection and 327, 328
- rate of molecular evolution and 264
- shifting balance model 370, 371
- types 82–5
- variance ( $N_e^v$ ) 82, 82–5
- effective population size:census  
population size ratio ( $N_e/N$ ) 85
- effective population size-selection  
coefficient product ( $N_e s$ ) 222–5, 240
- electropherogram, DNA sequence 242, 242
- empirical evidence 2
- ensemble population 62
- environmental deviation ( $E$ ) 335, 335, 336
- environmental influences  
inbreeding depression 40  
phenotype 285–6, 304
- environmental variation ( $V_E$ ) 286, 287–8  
defining heritability 297–8  
 $V_{G \times E}$  and 293–5  
see also common environmental variance
- enzyme polymorphisms see allozymes
- epistasis ( $V_I$ ) 287, 290, 290–1  
antagonistic 291
- effects on fitness surfaces 215, 367–8
- heritability and 297–8
- inheritance 291–2
- neutral quantitative traits 315
- physiological (functional:  
mechanistic) 373
- population-level meaning 288–9
- QTLs 327
- sign 373
- single-locus model 334, 342
- statistical 373
- synergistic 290–1
- equations, mathematical 6
- equilibrium allele frequency 195  
additive gene action 197  
irreversible mutation model 174–5  
monomorphic 199
- natural selection–genetic drift  
combined 223
- natural selection–mutation balance  
225–6
- with overdominance 199, 200, 361
- polymorphic 199
- reversible mutation model 175, 176
- shifting balance theory 368–9
- two-island model 134–5
- unstable 198
- Escherichia coli* 160
- estimates, parameter 2–3, 3  
notation 2  
uncertainty in 376–80
- evolution  
modern synthesis 4, 185  
molecular 235–82  
quantitative traits 297–315
- evolutionary variance 271
- exclusion, candidate parents 113, 116
- exclusion probability 116, 116, 117
- expectations 1, 1–4  
comparison with observation 5–6
- exponential distribution 93
- extinction, metapopulation models  
141, 142
- f* see inbreeding coefficient
- F* see fixation index
- F1 generation 317, 322
- F2 design for QTL mapping 317
- F2 population 317–19, 322
- family size 76–7, 161
- fecundity 190  
additive 217–18  
average (mean) 217, 218  
selection 216–18, 363
- Felsenstein 81 (F81) nucleotide  
substitution model 247
- fibrinopeptides, rate of evolution 251–2
- finite island model 138–9
- finite rate of increase ( $\lambda$ ) 185, 186
- finite sites model 172, 173, 173
- first principles 2
- $F_{IS}$  121, 124, 125
- Fisher, Sir Ronald A 74, 204, 356  
continuously distributed phenotypes  
284
- fundamental theorem of natural  
selection 203–6
- geometric model of mutation 164–6,  
329–30
- model of fate of a new mutation  
160–2
- $F_{IT}$  121, 124, 125
- Fitch, WM 262
- fitness  
absolute 185–6, 188
- average see average (mean) fitness
- density-dependent 219–22
- frequency-dependent 218–19
- marginal see marginal fitness
- maximum ( $w_{max}$ ) 360, 363
- relative see relative fitness
- spectrum for mutations see mutation  
fitness spectrum
- fitness surfaces 201, 208–9  
critique and controversy 372
- genotype version 372
- multiple peaks 215, 367–8
- one locus with three alleles 209–11
- shifting balance model 368, 369–72
- two diallelic loci 214–15
- Wright's original concept 366–8
- fixation index ( $F$ ) 28–32, 30  
estimation 31, 122, 124, 140
- finite island model 138–9
- finite population 78–9
- inbreeding coefficient and 35, 36–7
- individual species 31
- infinite island model 136–8
- metapopulation model 142
- multiple subpopulations 118–24,  
125
- stepping-stone model 141  
see also  $F_{IS}$ ;  $F_{IT}$ ;  $F_{ST}$
- fixation/loss 54
- binomial probability distribution  
60–1
- coalescent model 88
- diffusion approximation 72–3, Plate  
3.13
- Drosophila* studies 66–7
- initial allele frequency and 57, 61
- mutation fitness and 158–9
- nearly neutral theory 240
- neutral theory 236–7, 238
- new mutations in a population  
160–8
- over multiple generations 56
- subpopulations 107
- fixation processes 258
- flanking-marker QTL analysis 322–5
- flux ( $J_x$ ) 68, 70–1, 71
- Ford, Edmund B 358
- forensic DNA profiling see DNA profiling
- founder effect 75
- founder event 74
- frequency dependence, negative  
218–19
- frequency-dependent natural selection  
218–19, 220, 220
- frequency distribution 376

- fruit fly *see Drosophila*  
 fruit machine analogy 338  
 $F_{ST}$  121, 123  
 divergent subpopulations 122, 123  
 estimation **124, 140**  
 finite island model 138–9  
 infinite alleles model 171, **172**  
 infinite island model 136, 137, **138**, **139**  
 metapopulation model 142  
 stepping-stone model 141  
 subdivided population 125  
 variation 124, 126  
 full siblings  
   covariance between 351, 352  
   degree of relatedness 34  
   impact of mating on heterozygosity 35  
   maternal effects 295–6  
 G matrix 304–5, **305**, 313  
 Gaines SD 68, 76  
 Galton, Francis 284  
 gametes 55  
   coupling 42–3  
   dispersal 86, 87  
   repulsion 42–3  
 gametic compatibility 190  
 gametic disequilibrium 41–9, **44**  
   chance effects 48–9, 50  
   classical/balance hypotheses 357–8  
   decay over time 44–5  
   divergence rates and 278–9  
   estimating **49**  
   estimators 43–4  
   genotypic variance due to 296–7  
   influence on polymorphism 275–8  
   long-term response to selection and 311–13  
   mating system effects 48  
   mutation effects 47  
   natural selection effects 46–7  
   physical linkage effects 45–6  
   population admixture effects 47  
   two-locus natural selection 212, 213, 216  
 gametic disequilibrium parameter ( $D$ ) 43–5  
 long-term response to selection and 311–13  
 terminology 45  
 gametic equilibrium 43  
 gel electrophoresis, allozyme **32**, 32  
 gene **12**  
   diversity 31–2  
   major 320, 321, 335  
 gene conversion 42, 154–5  
 gene copy **89**  
 gene duplication 154  
 gene flow 105, **107**  
   cryptic 116, **116**  
   direct estimates 111–18, 141  
   distinction from migration 111  
   impact on genealogies 142–9  
   indirect estimates 140–1  
 migrant-pool 141–2  
 models 131–42  
 off-plot 118  
 propagule-pool 141–2  
 reduced, subdividing populations 105, 106  
 shifting balance theory 368–9  
 subdivided population 123–4, 125  
 gene transfer, lateral or horizontal 155  
 genealogies, gene 87–96, **89**  
   accounting for mutation 178–82  
   natural selection on 226–32  
   prospective or time-forward model 88  
   retrospective or time-backward model 88  
   simulation **94–5, 97**  
   structured populations with gene flow 142–9  
   *see also* coalescent genealogies  
 Genepop on the Web **49**  
 general dominance model of natural selection **193**, 197–8, 360  
 general time reversible (GTR) nucleotide substitution model 247  
 generation time effects, nearly neutral theory 264  
 generation-time hypothesis 263, **263**  
 genetic additive variance/covariance matrix *see* G matrix  
 genetic architecture 316, **316**  
 genetic bottlenecks *see* bottlenecks, genetic  
 genetic correlation 304, **304**, 306  
 genetic distance ( $D$ ) 170–1  
 genetic draft 276  
 genetic drift **54–5**, 54–7  
   binomial probability distribution 58–61  
   decline in heterozygosity due to 79–80  
   diffusion approximation 67–73  
   *Drosophila* studies 66–7  
   fate of new mutations and 163–4, 166, 168  
   heritability changes after 315  
   influence of mutations on autozygosity and 177–8  
   initial allele frequency and 57, 61  
   Markov chain model 62–7  
   models 58–73  
   natural selection and 222–5, 366  
   nearly neutral theory 240–1  
   neutral quantitative traits 313–14, 315  
   neutral theory 236  
   over multiple generations 56–7  
   parallelism with inbreeding 78–80, **81**  
   QTls 329  
   shifting balance model 368–9, 370, 371  
   simulation **58, 65**  
   strength relative to natural selection **241**  
 subdivided population 123–4, 125  
 Wright–Fisher model *see* Wright–Fisher model  
 genetic hitch-hiking *see* hitch-hiking, genetic  
 genetic load *see* load, genetic  
 genetic markers  
   DNA profiling **22**  
   estimating fixation indices **124, 140**  
   mutation rates 157  
   parentage analysis 111–18  
   QTL mapping 316–17, 322  
 genetic neighborhood 86–7, **87**  
 genetic variation  
   additive *see* additive genetic variance  
   classical/balance hypotheses 357–8  
   defining population size 73  
   loss due to genetic drift 64, 66  
   multilocus 357–8  
   mutations as source 154, 157–8  
 genic variance 205  
   additive ( $V_A$ ) 311–12  
 genotype **12**  
   counting method to estimate allele frequency 30  
   resemblance between relatives 351–3  
 genotype-by-environment interaction ( $V_{G\times E}$ ) **287**, 292–5, **293**  
 genotype frequencies  
   breeding values/dominance deviation and 347–8  
   change over time 187–8  
   DNA profiles 21  
   genetic drift effects 54, 56–7  
   Hardy–Weinberg expected *see* Hardy–Weinberg expected genotype frequencies  
   inbreeding and 33–4  
   mean phenotypic value from 336–7  
   under natural selection 195–200  
   non-random mating effects 29–30  
   spatial structuring 107, 108–9  
 genotypic disequilibrium  
   covariance 296–7  
   estimating **49**  
 genotypic values **335**, **335**  
   components 342  
   mean phenotypic values from 336–7  
   population mean *see* mean genotypic value, population  
   scale of measurement 334–5  
    $V_A$  and  $V_D$  in relation to 349–51  
 genotypic variation (total) ( $V_G$ ) 286, 287–8  
   allele frequencies/genotypic values and 349–51  
   components **287**, 288–91, 296–7, 348–51  
   defining heritability 297, 298–9  
   derivation 348–9, **350**  
   inheritance 291–2  
   neutral quantitative traits 315  
   use of term 291  
    $V_{G\times E}$  and 293–5

- geometric model of mutation 164–6, 329–30  
 Gillespie, JH 258, 262, 276  
 group selection 372–3  
 $G_{ST}$  124, 138–9  
 guinea pigs 74
- $h^2$  see heritability  
 Haldane, JBS 74, 356, 360–1, 363  
 half siblings  
   covariance between 351, 352  
   maternal effects 295–6  
 haplo-diploid organisms 16  
 haploid populations  
   coalescence 89, 90  
   effective population size 97  
   natural selection 185–9, 204–5  
 haplotypes (haploid genotypes)  
   expected frequency 367  
   frequency distributions 272, 273–4  
   gamete 41, 42  
   paternal 113–15  
 hard selection 216, 363  
 Hardy, Godfrey H 13  
 Hardy–Weinberg equation 13  
 Hardy–Weinberg equilibrium 14  
 Hardy–Weinberg expected genotype frequencies 13–17  
   applications 16–17, 19–28  
   assumptions 14–15, 17  
   comparing two models of inheritance 26–8  
   fixation index and heterozygosity 28–32  
   forensic DNA profiling 19–22  
   graphical representation 13, 15  
   in haplo-diploid systems 16  
   more than two alleles 19  
   as null model 15  
   population size assumption 14, 53, 54  
   proof 17–19  
   simulation 14  
   testing for deviations 22–5  
 harmonic mean 74–5  
 Hasegawa–Kishino–Yano (HKY) nucleotide substitution model 247  
 hemoglobin  
    $\beta$  genotypes 209–12  
   protein, rate of evolution 251–2  
 heritability ( $h^2$ ) 297–9  
   breeder's equation 300  
   broad-sense 297–8  
   long-term response to selection and 307, 311, 312  
   narrow-sense 297–8  
   neutral evolution of quantitative traits and 314–15  
   parent–offspring regression for estimating 302–3  
   realized 300–1  
 hermaphrodites 31  
 heterosis 38, 38
- heterozygosity  
   allozyme estimates 358–9  
   decline due to genetic drift 79–80  
   deficits 31  
   effective population size estimation from 83, 84, 85  
   equilibrium ( $H_{equilibrium}$ ) 177–8, 237, 361  
   excess 31  
   expected ( $H_e$ ) 30, 31–2  
   fixation index and 28–32, 78–9  
   genetic load concept 361–3  
   inbreeding effects 33–4, 35, 37  
   individual ( $H_i$ ) 119, 124–6  
   island and mainland populations 80, 81  
   mutation effects 177–8  
   observed ( $H_o$ ) 32  
   subdivided populations 119–22, 125  
   subpopulation see  $H_S$   
   total see  $H_T$   
   Wahlund effect 124–8
- heterozygote advantage  
   (overdominance for fitness) 193, 199–200  
   apparent, decline over time 359  
   balance hypothesis 357  
   equilibrium allele frequency with 199, 200  
   inbreeding depression and 38, 39  
   mean fitness and 202–3  
   natural selection–genetic drift balance 223  
   see also advantageous mutations; balancing selection
- heterozygote disadvantage  
   (underdominance for fitness) 193, 198  
   mean fitness and 202–3  
   see also disruptive selection
- heterozygotes  
   breeding value 344  
   frequencies see genotype frequencies  
   genotypic value 334, 335
- $H_i$  119, 124–6  
 historical controversies 356–66  
 hitch-hiking, genetic 275–6, 276, 365  
   due to background or balancing selection 278  
   recombination rate and 277–8  
 HIV see human immunodeficiency virus  
 HKA test 265–7  
 $Hla$  genes 268, 269  
 homoplasy 169, 169  
 homozygote frequencies see genotype frequencies  
 horizontal gene transfer 155  
 $H_S$  119  
   divergent subpopulations 122, 123  
   infinite island model 136  
   Wahlund effect 125–7  
*hsp70* 313
- $H_T$  119, 120  
   divergent subpopulations 122, 123  
   infinite island model 136  
   Wahlund effect 125–6  
 Hubby, JL 358–9, 363  
 Hudson–Kreitman–Aguadé (HKA) test 265–7  
 human immunodeficiency virus (HIV)  
   date of origin 254–5  
   evolution of drug resistance 188–9  
 human leukocyte antigen ( $Hla$ ) B gene 268, 269  
 humans  
   chromosomes 41  
   date of origin 254  
   diseases caused by recessive alleles 128  
   inbreeding depression 38  
   mutation rates 157  
   polymorphism and divergence 278  
 Hunt, Thomas Morgan 42  
 Hutterite population 298, 353  
 Huxley, JS 356  
 hybrid vigor (heterosis) 38, 38  
 hybridization, sunflowers 218  
 hypotheses, alternative 4–5
- identity by descent (IBD) 37  
   finite population 78  
   infinite alleles model 169  
   pedigree 34–7  
   probability, in a population 82  
   resemblance between relatives and 351–2  
 identity in state 169  
 $IGF1$  gene 320–1  
   average effect of an allele 339–40, 341  
   breeding values 343–4  
   dominance deviation 344, 346–7  
   genotypic values 334, 335  
   mean phenotypic value 336–7  
 Illinois Long-Term Selection experiment 310–11, 329  
 inbreeding 33–41  
   biparental see biparental inbreeding  
   deleterious consequences 37–8  
   genotype and allele frequencies after 33–4  
   impact on heterozygosity 33–4, 35, 37  
   multiple meanings 40–1  
   parallelism with genetic drift 78–80, 81  
   phenotypic consequences 37–40  
   shifting balance model 370, 371  
   see also consanguineous mating
- inbreeding coefficient ( $f$ ) 34–7, 37  
   effective ( $F_e$ ) 81  
   fixation index and 35, 36–7  
   genotypic resemblance between relatives 351–2  
   pedigree 34–7  
   population mean phenotype and 337  
 Populus simulations 81

- inbreeding depression 31, 37–40, **38**  
 dominance and overdominance  
   hypotheses 38–9  
 environmental effects on expression  
   40  
 humans 38  
 model research organisms 39–40  
 inbreeding effective population size ( $N_e^i$ )  
   **82**, 82–5  
   coalescent model 96–8  
   estimation from change in  
     heterozygosity 83, **84**  
 inclusion, candidate parents 113–15  
 incoming branch 227, 229  
 indels 154  
 index of dispersion 257, 259–60, 262  
 inductive reasoning 3–4, **4**  
 industrial melanism 358  
 infer(ence) **5**, 5–6  
 infinite alleles model 169, **169**, 170–1  
   autozygosity and 177–8  
   coalescent genealogies 180–2  
   estimating population subdivision  
     171, **172**  
 infinite island model 132, **135**, 135–41  
   estimating number of migrants  
     140–1  
   levels of fixation 136–8  
 infinite sites model 172–3, **173**  
   coalescent genealogies 182  
 infinitesimal model 307, **307**, 328  
 influenza A virus, NS gene 251, 252  
 insertion 154  
 insulin-like growth factor 1 gene  
   see *IGF1* gene  
 interaction genetic variance ( $V_I$ )  
   see epistasis  
 interdemic selection 371  
 interval QTL mapping 322–5  
   composite 325  
 inversion, chromosomal 155  
 irreversible mutation (model) **174**,  
   174–5, **177**  
 island model of gene flow see infinite  
   island model  
 island populations  
   continent-island model 131–4  
   heterozygosity 80, **81**  
   infinite island model 135–41  
   two-island model 134–5  
 isolate breaking 128–9  
 isolation by distance **87**, **107**  
   among subpopulations 141  
   breeding effective population size 86  
   generating population structure  
     106–7, 108–9  
   Moran's *I* **110**, 110  
  
 joint distribution 380  
 Jukes–Cantor nucleotide substitution  
   model 244–7  
 $J_x$  see flux  
  
*K* see carrying capacity  
 $k$  alleles model 169, **169**  
 Kettlewell, Bernard 358  
 Kimura, Motoo 235  
   diffusion approximation of genetic  
     drift 68, 72  
   genetic load 363  
   geometric model of mutation 166,  
     329  
   infinite alleles model 169  
   molecular clock rate heterogeneity  
     261–2  
   natural selection–genetic drift balance  
     223  
   nearly neutral theory 240  
   neutral theory of molecular evolution  
     235–6, 363–5  
   stepping-stone model 151  
 Kimura 80 (K80) nucleotide substitution  
   model 247  
  
 Labrador retriever dogs, coat color 290,  
   291  
 Lamarck, Jean-Baptiste 9  
 $\lambda$  185, 186  
 Landsteiner, Karl 26  
 Langley, CH 258, 262  
 lateral gene transfer 155  
 lethal mutations 158  
 Lewontin, RC 358–9, 363, 365  
 liability traits 283  
 life cycle, reproductive 190, 216  
 life-history traits 298, 315, 328–9  
*Linanthus parryae* 106–7, Plate 4.2  
 lineage **89**  
 lineage branching 87–96  
 lineage effect, molecular clock 262,  
   **262**  
 linkage 41–2, **44**, 45–6  
   mapping 316  
   phenotypic correlation and 304  
 linkage disequilibrium parameter see  
   gametic disequilibrium parameter  
 load, genetic ( $L$ ) 359–63  
   purging 39, 40  
   segregational 359–60, **361**, 361–3  
   substitutional 359, 360, **361**, 363  
 local races, shifting balance model 370,  
   371  
 locus **12**  
 log of odds (LOD) score 324, 325  
 logistic population growth 220–1  
 loss, allele see fixation/loss  
*Lupinus texensis* 87  
  
*m* see migration rate; relative fitness  
*M* see mean phenotypic value,  
   population  
 maize see corn  
 major gene 320, 321, 335  
 major histocompatibility complex (MHC)  
   loci 29, 269, 358  
 malaria 209–10, 358  
 Malthusian fitness see relative fitness  
 mammals  
   evolution 254  
   maternal effects 295–6  
 marginal fitness **191**, 193  
   three-allele model of natural selection  
     210, **211**  
   two-locus model of natural selection  
     213–14  
 mark–recapture method 141  
 Markov chains 62–7, **66**  
 Markov property 65, **66**  
 maternal effects 295–6  
 maternity analysis 111  
 mating  
   among relatives see consanguineous  
   mating  
   designs, QTL mapping 317, **324**  
   mixed 34, 35  
   non-random see non-random mating  
   random see random mating  
   success 190  
 mating systems 31  
   effective population size and 75  
   fixation index and 31  
   gametic disequilibrium effects 48  
 Matlab **306**  
 McDonald–Kreitman (MK) test  
   267–9  
 mean 376  
   harmonic 74–5  
 mean fitness see average (mean) fitness  
 mean genotypic value, population  
   336–7  
   breeding value and 342–5  
   components 342  
   dominance deviation and 345–8  
 mean phenotypic value, population ( $M$ )  
   336–7  
   breeding value and 342–5  
   components 335  
   under non-random mating 337  
   under random mating 336–7  
 meiotic drive 190, 216  
 melanism, industrial 358  
 Mendel, Gregor 9  
 Mendelian genetics 9–12  
   continuous phenotypic distributions  
     284–6  
   gametic disequilibrium and 41, 42  
   loss of new mutations 160–2  
   quantitative trait variation  
     334–55  
 Mendel's first law 10–11, **11**  
 Mendel's second law 12, **12**  
 meristic traits 283  
 metabolic rate hypothesis 263–4  
 metapopulation models 141–2  
 metric traits see quantitative trait(s)  
 MHC loci see major histocompatibility  
   complex (MHC) loci  
 mice  
   inbreeding depression 40  
   long-term selection experiment 311,  
     312  
   mutation rates 155–6, **157**  
 micro-centrifuge tubes, sampling  
   experiment 53–4, 58–9,  
     Plate 3.1

micromutationalism 164, **164**  
 microsatellite loci  
   DNA profiling 22  
   estimating genotypic disequilibrium 49  
   mutation models 170  
   mutation rates 157  
   parentage analysis 112, 113, **114**  
 mid-parent value 302, 303  
 migrant-pool gene flow 141–2  
 migration 111  
   average length of a genealogy with 144–9  
   coalescence combined with 143–4, 145  
   different lineages 142, 143  
   distinction from gene flow 111  
   shifting balance model 369, 370  
 migration rate ( $m$ ) 111  
   average time to coalescence and 147, 149  
   continent-island model 131, 133, 134  
   different lineages 143  
   effective ( $N_e m$ ) *see* effective migration rate  
   infinite island model 135–7  
   two-island model 134–5  
*Mirounga leonina* 75  
 mismatch distributions 272, 272–4  
 missense mutations *see* nonsynonymous mutations  
 mitochondrial genomes 46, 87, 276  
 mixed mating 34, 35  
 MK test 267–9  
 MN blood group genotypes 22–5  
 mode 376  
 model organisms 3  
   inbreeding depression 39–40  
 models, population genetic 4  
 modern synthesis 4, 185  
 molecular clock 250–65  
   dating of events 252–5  
   examples 251, 252  
   hypothesis 250–2, **251**  
   overdispersed 257, **257**, 365, 366  
   Poisson process model *see* Poisson process molecular clock  
   testing and rate variation 255–65  
 molecular evolution 235–82  
   nearly neutral theory 240–1  
   neutral theory *see* neutral theory of molecular evolution  
   non-independent loci 274–9  
   null model 235  
   rate heterogeneity *see* rate heterogeneity, molecular evolution  
   rate homogeneity 255  
 monomorphic equilibrium 199  
*Moran's l* **110**, 110  
*Morone saxatilis* (striped bass) 32, **49**, 284, 287  
 morphological traits 298, 307, 315, 328–9  
 mosquito pupal mass 284, 287

most recent common ancestor (MRCA)  
   88, **89**  
   ancestral selection graph 229, 230  
   distribution of times to 95, 96  
   predicting time to *see* coalescent genealogies  
   moths, spotted (peppered) 358  
 MRCA *see* most recent common ancestor  
 Muller's Ratchet 166–8  
 multifactorial traits 283  
 multigene families 154  
 multiple hit mutations 173  
 multiple hit substitutions 243  
 multiple sequence alignment 248  
 multipoint QTL mapping 325  
 mutation(s) 154–83  
   advantageous *see* advantageous mutations  
   allele frequency and autozygosity effects 173–8  
   ancestral selection graph 228–30  
   beneficial *see* beneficial mutations  
   coalescent model with 178–82  
   detrimental *see* deleterious mutations  
   dwell time of new 236, 237, 238  
   fate of new 160–8  
   finite population 162–4  
   fixation by natural selection 164–6  
   forward 156  
   frameshift 154  
   gametic disequilibrium and 47  
   infinite population 162  
   irreversible 174, 174–5, **177**  
   lethal 158  
   loss due to Mendelian segregation 160–2  
   multiple hit 173  
   natural selection acting with 225–6  
   nearly neutral 158  
   neutral *see* neutral mutations  
   neutral quantitative traits 313–14, 315  
   nonsynonymous (missense) *see* nonsynonymous mutations  
   point 154  
   replication-dependent causes 263, **263**  
   replication-independent causes 262, **262**  
   reverse 156  
   reversible or bi-directional 174, 175–6, **177**  
   as source of genetic variation 154, 157–8  
   sublethal 158  
   synonymous (silent) *see* synonymous mutations  
   types 154–5  
 mutation-accumulation experiments 159–60  
 mutation fitness spectrum **158**, 158–9, 160  
 pan-neutralist/pan-selectionist view 364  
 mutation frequency 155  
 mutation models 168–73  
   coalescent genealogies 180–2  
   discrete alleles 169–71  
   DNA sequences 172–3  
   irreversible 174–5  
   QTL effect size distributions and 329–30  
 mutation pressure 174, **174**  
 mutation rates 155–7  
   coalescent model 179  
   equivalence to substitution rates 239, 250–1, 253  
   estimation 155–7  
   observed examples 156–7  
   scaled ( $4N_e \mu$ ) *see*  $\theta$   
   shifting balance model 368, 370, 371  
   synonymous vs nonsynonymous mutations 268  
   variability 262–3  
*N* *see* census population size  
 natural selection 185–207  
   adaptive landscape metaphor 368  
   alternative models 216–22  
   average fitness effects 200–6  
   balancing *see* balancing selection  
   classical/balance hypotheses 356–7  
   with clonal reproduction 185–9  
   combined with other processes 222–6  
   on correlated traits 304–7  
   deaths due to 359–60, 363  
   density-dependent 219–22  
   directional *see* directional selection  
   disruptive *see* disruptive selection  
   against dominant phenotype **193**, 196–7  
   Fisher's fundamental theorem 203–6  
   frequency-dependent 218–19, **220**, 220  
   gametic disequilibrium and 46–7  
   genealogical branching models 226–32  
   general dominance model **193**, 197–8, 360  
   genetic drift and 222–5, 366  
   geometric model of mutations fixed by 164–6  
   hard 216, 363  
   Hardy–Weinberg model and 15  
   inbreeding depression and 39  
   interdemic 371  
   loss of new mutations and 162  
   Muller's Ratchet and 168  
   mutation acting with 225–6  
   negative *see* negative selection  
   one diallelic locus 189–200, **203**  
   one locus with three alleles 209–12  
   positive *see* positive selection  
   purifying 268–9  
   on QTLs 327–9  
   on quantitative traits 299–302  
   against recessive phenotype **193**, 195–6

- restatement of Darwin's ideas 185  
with sexual reproduction 189–93  
shifting balance model 368–72  
soft 216, 363  
strength 199–200  
strength relative to genetic drift **241**  
two diallelic loci 212–16  
via different levels of fecundity  
216–18
- N<sub>e</sub>* *see* effective population size
- nearly neutral mutations 158, 240,  
365–6
- nearly neutral theory 240–1, 365–6  
rate heterogeneity and 264–5  
negative frequency dependence  
218–19
- negative selection  
nearly neutral theory 240  
polymorphism under 237, 238
- Nei, M 31–2, 170
- neighborhood, genetic 86–7, **87**
- neo-classical theory 359
- neo-Darwinian (modern) synthesis 4,  
185
- neo-Darwinism 356
- net neutral mutations 240
- neutral evolution, quantitative traits  
313–15
- neutral mutations 158, 235  
divergence due to 239–40  
fate in a finite population 162–4  
loss due to Mendelian segregation  
160–2  
polymorphism 236–7  
time to loss/fixation 237, 238
- neutral reference locus, HKA test 265,  
**266**, 267
- neutral theory of molecular evolution  
164, 235–40, 363  
compared to nearly neutral theory  
240  
divergence 237–40  
historical debate 363–6  
molecular clock hypothesis and  
250–1, 252  
polymorphism 236–7, 238  
rate heterogeneity and 262–4  
testing 265–74
- Newton, Isaac 3
- non-random mating 28  
breeding values under 345  
fixation indices accounting for  
120–2  
genotype frequencies and 29–30  
impact on genotype and allele  
frequencies 33–4  
population mean phenotypic value  
337  
*see also* assortative mating;  
consanguineous mating
- nonsynonymous mutations 154
- MK test 267–9
- nearly neutral theory testing  
264–5
- normal distribution 378, **379**
- nucleotide diversity ( $\pi$ ) 248, 249–50  
estimating **251**  
observed **250**  
Tajima's D test 269
- nucleotide polymorphisms 242
- nucleotide site 241–2
- nucleotide-substitution models  
244–7  
hierarchy of complexity 247  
relative rate tests incorporating  
261
- null model 15, **15**
- number of segregating sites *see*  
segregating sites, number of
- odds ratio 21, **21**
- off-plot gene flow 118
- Ohta, T 240, 261–2, 263, 264–5,  
365–6
- origination processes 258
- Orr, HA 166, 329–30
- outcrossing 34
- overdispersed molecular clock 257,  
**257**, 365, 366
- overdominance for fitness *see*  
heterozygote advantage
- overdominance hypothesis, inbreeding  
depression 38–9
- p* distance **242**, 242–3
- P matrix 304–5, **305**
- P1 generation 317
- pairwise differences, distributions of *see*  
mismatch distributions
- pan-neutralism 364
- pan-selectionism 364
- panmixia 105, **107**  
vs isolation by distance 107, 108
- parameters 2–3, **3**  
estimates *see* estimates, parameter  
statistical concepts 376–80
- parametric variance 378
- parentage analysis 111–18, 297–8  
applications 118  
cryptic gene flow 116  
example 111–15, 116–17  
four general outcomes 117–18
- parent–offspring regression 302–3,  
353
- parent–offspring resemblance 291,  
292, 302–3  
*see also* resemblance between relatives
- parents, candidate *see* candidate parents
- particulate genetics 4
- Hardy–Weinberg model 13–17
- Mendel's model 9–12
- quantitative trait variation and  
334–55
- Pasteur, Louis 5
- paternity analysis 111
- Pauling, L 255
- pea plants, Mendel's experiments  
9–12
- Pearson product-moment correlation ( $\rho$ )  
380, 381
- pedigrees
- ancestor–descendant relationships  
88
- inbreeding coefficient and  
autozygosity 34–7, 35
- Phaseolus vulgaris* **321**
- phenotype(s) **13**  
continuously distributed 283–6  
environmental influences 285–6,  
304
- impact of inbreeding 37–40
- quantitative genetics 283
- resemblance between relatives 291,  
292, 301–2, 352
- phenotypic correlation 304, **304**, 306
- phenotypic mean ( $\bar{z}$ ) 305, **305**
- phenotypic value 283, 335, **335**  
components 335  
natural selection on 299–302  
population mean *see* mean phenotypic  
value, population
- phenotypic variance/covariance matrix  
*see* P matrix
- phenotypic variation (total) ( $V_p$ ) 286–7  
components 286–97  
defining heritability 297–9  
P matrix 304–5
- $\varphi$  142
- $\varphi_{ST}$  **124**
- Phlox cuspidata* **122**
- phosphoglucomutase (PGM) 32
- phosphoglucomutase-2 gene (*Pgm-2*)  
**122**
- phosphoglucose isomerase-1 (*Pgi-1*) 87
- physiological traits 298, 307
- $\pi$  *see* nucleotide diversity
- pigs 306–7, 329
- pines, Ponderosa 31
- pioneer species 141
- plants
- gene flow 111  
monocot–dicot divergence **254**
- Plasmodium falciparum* malaria 209–10
- plastid genomes *see* chloroplast  
genomes; mitochondrial genomes
- pleiotropy 304
- antagonistic 313
- point mutations 154
- Poisson distribution 76  
fate of new mutations and 161  
variance effective population size and  
76, 77
- Poisson process molecular clock 256–7  
ancestral polymorphism and 257–60  
patterns and causes of rate  
heterogeneity 261–5  
substitutions over time 256–7  
*see also* molecular clock
- poker machine analogy 338
- polygenic variation 284–5
- polymorphic equilibrium 199
- polymorphism **236**  
ancestral 257–60  
balanced 200  
cause of allozyme 358–9

- under directional selection 237, 238  
*DNA* *see* DNA polymorphism  
 gametic disequilibrium and 275–8  
 genetic hitch-hiking 275–6  
 impact of background and balancing selection 278  
 marker loci for QTL mapping 317  
 measures 248–50  
 nearly neutral theory prediction 241  
 neutral theory prediction 236–7  
 nucleotide 242  
 recombination rate and 277–8  
 selectionist/neutralist debate 364, 365, 366  
 selective sweeps 275, 276–8  
 testing neutral theory predictions 265–74
- PopGene.S<sup>2</sup>** **14**
- population(s)**  
 actual 5–6  
 defining genetic 73  
 ensemble 62  
 finite nature 53  
 genetic 105–11  
 geographic barriers 105, 106  
 ideal 5–6  
 isolation by distance 106–7  
 mixing of diverged (admixture) 47  
**population growth**  
 genotypes with clonal reproduction 185–7, 188  
 logistic 220–1  
 simple model (unbounded) 185  
**population size** 73–8  
 allele frequencies and 53–4  
 census (*N*) *see* census population size  
 coalescent genealogies with changing 99–101  
 density-dependent selection and 221–2  
 diffusion of allele frequency and 70  
 effective *see* effective population size  
 exponential growth or shrinkage 100  
 fluctuations over time 73–5  
 genetic drift effects and 66  
 Hardy–Weinberg assumption 14, 53, 54  
 probability of coalescence and 91, 94  
**population structure (subdivision)** 105–11, **107**  
 balancing selection and 231  
 causes 105–7, 108–9  
 DNA profiling and 129–31  
 expressed in coalescent waiting times 149  
 fixation indices measuring 118–24  
 HKA testing and 267  
 impact on genealogical branching 142–9  
 implications 107–11  
 models 131–42  
 mutation models for estimating 171, **172**  
 shifting balance model 370, 371–2  
 Wahlund effect 124–31
- Populus** **58**
- positive selection**  
 nearly neutral theory 240  
 polymorphism under 237, 238  
 selective sweep 275, 276
- predator/prey dynamics** 73
- probability density** 71–2, Plate 3.13
- probability of identity** 21
- product rule** 21, **21**
- propagule-pool gene flow** 141–2
- prosecutor's fallacy** 21
- prospective model** 88
- protein locus genotyping** **32**
- proteins, rates of evolution** 251–2
- pseudogene** 154
- Punnett, RC** 42
- Punnett square** 9  
 Hardy–Weinberg principle 16, 18  
 Mendel's experiments 11
- purging of genetic load** 39, 40
- purifying selection** 268–9
- QTLs** *see* quantitative trait loci
- quantitative trait(s)** 283–97, **284**  
 continuously distributed 283–6  
 evolutionary change 297–315  
 long-term response to selection 307–13  
 meristic 283  
 natural selection on 299–302  
 neutral evolution 313–15  
 selection on correlated 304–7  
 threshold or liability 283
- quantitative trait loci (QTLs)** 315–30, **316**  
 candidate 316, **316**  
 effect size 321, 327–8, 329–30  
 flanking-marker analysis 322–5  
 major genes 320, 321, 335
- quantitative trait locus (QTL) mapping** 316–30  
 biological significance 326–30  
 composite interval 325  
 experimental mating designs 317, **324**  
 interval 322–5  
 limitations 325–6  
 multiple marker loci 322–5  
 multipoint 325  
 single marker loci 316–22
- quantitative trait region (QTR)** 326
- quantitative trait variation (V)** 286–8  
 environmental influences 285–6  
 Mendelian basis 334–55  
 notation 286–8  
 summary statistics 287
- r* *see* recombination fraction
- r* coefficient (in relatedness of relatives) 351
- random mating**  
 breeding values under 342–5  
 dominance deviation under 347  
 Hardy–Weinberg assumption 15, 17–18
- population mean phenotypic value 336–7
- population subdivision and 105
- rate heterogeneity, molecular evolution **255**, 255–7
- generation-time hypothesis 263
- lineage effects 262
- metabolic rate hypothesis 263–4
- patterns and causes 261–5
- relative rate tests 260–1
- residual effects 262
- rate homogeneity, molecular evolution 255
- realized heritability ( $h^2$ ) 300–1
- recessive allele/phenotype **13**  
 deleterious *see* deleterious recessive alleles  
 isolate breaking 128
- Mendel's studies 10–11  
 selection against **193**, 195–6
- recombinant inbred line design 317
- recombination 41–2, **43**  
 mismatch distributions and 274  
 natural selection acting with 46–7
- physical linkage and 45–6
- polymorphism and divergence and 277–8
- two-locus natural selection 212–13, 214–15, 216
- recombination fraction (*r*) 42–3, **44**  
 chance effects 48–9, 50  
 mating system effects 48  
 QTL mapping 320, 321–2
- two-locus natural selection 213, 214–15
- Red Queen model 371
- regression analysis 381
- relatedness **37**  
 degree of 34
- relative fitness (*w* or *m*) **188**  
 average 188  
 clonal reproduction 186–8  
 drug-resistant HIV **189**  
 selection coefficients from 193–5  
 sexual reproduction 191, **192**
- three-allele model of natural selection 209, 211
- two-locus model of natural selection 212, **213**
- relative rate tests, molecular clock 260–1
- relatives  
 mating among *see* consanguineous mating  
 resemblance between *see* resemblance between relatives
- replication-dependent causes of mutation 263, **263**
- replication-independent causes of mutation 262, **262**
- reproduction  
 genetic drift and 55  
 life cycle 190, 216  
 survival trade off 313  
*see also* clonal reproduction; sexual reproduction

- resemblance between relatives  
 genotypic 351–3  
 phenotypic 291, 292, 301–2, 352
- residual effects, molecular evolution 262, **262**
- restriction fragment length  
 polymorphisms (RFLPs) 317
- retrospective model 88
- retroviruses, mutation rates **157**
- reversible (bi-directional) mutation model **174**, 175–6, **177**  
 $p$  380, 381  
 $p^2$  43  
 $\rho_{ST}$  see  $R_{ST}$   
 RNA viruses, mutations **157**, 160
- rose pink 40
- $R_{ST}$  **124**, 171, **172**
- runt locus **271**
- $s$  see selection coefficients; selection differential  
**s** 304, 305, **305**
- $S$  see segregating sites, number of  
*Sabatia angularis* (rose pink) 40
- salmon 77
- sample configurations, lineages and demes 145, 146
- sample size  
 binomial probability distribution 59, 60  
 genetic drift and 56  
 simple experiment 54
- sampling 3, 53–8  
 simple experiment 53–4  
 Wright–Fisher model see Wright–Fisher model  
 sampling error **54**, 54–7  
 over multiple generations 56–7
- sampling variance 271, 377, 378
- Sanger method, DNA sequencing **242**
- saturation 243–4
- seals, elephant 75
- segregating sites, number of ( $S$ ) 248–9, **251**  
 Tajima's  $D$  test 269, **271**  
 segregational load 359–60, **361**, 361–3
- selection  
 artificial 300, 306–7, 342  
 on correlated traits 304–7  
 “for” and “of” distinction 306  
 long-term response 307–13  
 see also natural selection
- selection coefficients ( $s$  and  $t$ ) 193–5, 199–200
- frequency-dependent selection 219, **220**  
 genetic load 360, 361–3  
 molecular clock 255
- natural selection–genetic drift balance 222, 223
- natural selection–mutation balance 225
- nearly neutral theory 240
- shifting balance model 368, 370
- selection differential ( $s$ ) 300–1, 302  
 QTLs 327–8  
 as a vector ( $\mathbf{s}$ ) 304, 305, **305**
- selection plateau 311, 312
- selection threshold (truncation point) 300, 301, 302, 363
- selectionist/neutralist debates 363–6
- selective neutrality 235, 359
- selective sweeps 275, **276**, 276–8
- self-fertilization (selfing) 28–9, **30**, 31  
 complete 33–4, 35  
 gametic disequilibrium and 48  
 parentage analysis 116–17  
 partial (mixed mating) 34, 35  
 two-locus natural selection and 216
- semi-dominance see codominance
- sex determination, chromosomal 16, 46
- sex-linked traits 16
- sex ratio, breeding see breeding sex ratio
- sexual autogamy 28–9, **30**  
 see also self-fertilization
- sexual reproduction  
 natural selection with 189–93  
 segregational load 359–60  
 see also diploid populations
- shifting balance process 366
- shifting balance theory, Wright's 366–73
- allele-frequency distributions 368–9
- critique and controversy 372–3
- evolutionary scenarios 369–72
- siblings see full siblings; half siblings
- sickle cell anemia 209
- $\sigma$  see standard deviation
- $\sigma^2$  see variance
- sign epistasis 373
- Silene alba* 142
- silent mutations see synonymous mutations
- simple sequence repeats (SSR) see microsatellite loci
- simple tandem repeats (STR) see microsatellite loci
- simulation 6–7
- single nucleotide polymorphisms (SNPs) 242, 317
- slot machine analogy 338
- soft selection 216, 363
- species divergence 242–3, 258, 260
- spontaneous generation 5
- stabilizing selection 299–300
- standard deviation ( $\sigma$ ; SD) **377**, 378
- standard error of mean (SE) **377**, 378
- stepping-stone models 132, 141–2
- stepwise mutation model **169**, 170  
 estimating population subdivision **171**, **172**
- stochastic process **55**
- sublethal mutations 158
- subpopulations **105**, **107**  
 generation 105–7, 108–9  
 genetic characteristics 107–11  
 see also population structure
- substitution rate  
 absolute 253, **253**  
 equivalence to mutation rate 239, 250–1, 253  
 Haldane's estimate 360–1  
 heterogeneity see rate heterogeneity, molecular evolution
- nearly neutral mutations 366
- observed 263
- prediction 239–40, 241
- testing relative 260–1
- variation over time 255–7
- substitutional load 359, 360, **361**, 363
- substitutions **239**  
 multiple hit 243  
 nearly neutral theory 241  
 neutral theory 237–40  
 see also nucleotide-substitution models
- sunflowers, hybridization 218
- supergenes 357–8
- survival, reproduction trade off 313
- Sutton, Walter 12, 41
- symmetrical (SYM) nucleotide substitution model 247
- synergistic epistasis 290–1
- synonymous mutations 154  
 codon bias 366  
 MK test 267–9  
 nearly neutral theory testing 264–5
- Tajima's  $1D$  test 260–1
- Tajima's  $D$  test 269–72
- terminology, genetics 12–13
- theory 4–6, **5**
- $\theta(4N_e\mu)$  177–8, 249  
 ancestral polymorphism 258, 259  
 estimated from number of segregating sites 248, 249  
 nearly neutral theory 240
- Tajima's  $D$  test 269, **271**
- $\theta_{ST}$  **124**
- threshold traits 283
- time-backward model 88
- time-forward model 88
- tomato 327, 329
- Toxoplasma gondii* 45
- trait 283
- transition probability 62–4  
 matrix construction **66**
- transitions, nucleotide 154
- translocations 155
- transposable elements 155
- transversions, nucleotide 154
- TreeToy **274**
- Tribolium castaneum* 372–3
- truncation point (selection threshold) 300, 301, 302, 363
- two-island model 134–5  
 bi-directional mutation and 175
- two-way mutation model see reversible mutation model
- $u$  coefficient (in relatedness of relatives) 351, 353
- UGT1A6 genotypes 294

- uncertainty, statistical 376–80  
 underdominance for fitness  
     *see* heterozygote disadvantage  
 uniparental inheritance 87  
 unstable equilibrium 198
- V* *see* quantitative trait variation  
 $V_a$  (additive genic variance) 311–12  
 $V_A$  *see* additive genetic variance  
 value 283, **284**  
 Van Valen, L 371  
 variance ( $\sigma^2$ ) 376, **377**  
     change in allele frequencies *see*  $\Delta p$   
     estimation 376–8  
     parametric 378  
     sampling 271, 378  
     summing two **296**  
 variance effective population size ( $N_e^v$ )  
**82**, 82–5  
 $V_D$  *see* dominance genetic variance  
 $V_E$  *see* environmental variation
- $V_{Ec}$  *see* common environmental variance  
 $V_G$  *see* genotypic variation  
 $V_{G\times E}$  *see* genotype-by-environment interaction  
 $V_I$  *see* epistasis  
 viability 190  
 viability selection 190–1, **191**  
     alternative models 216  
     genetic load and 360, 363  
     relative fitness values 193–5  
     three alleles or two loci 208–16  
 vicariance events 105  
 $V_p$  *see* phenotypic variation
- w* *see* relative fitness  
 Wahlund, Sten GW 126  
 Wahlund effect 124–31, **126**  
     applied to DNA profiling 129–31  
     isolate breaking 128–9  
 waiting times, coalescent *see* coalescent waiting times
- Watterson, GA 248–9, 258  
 Weinberg, Wilhelm 13  
 wheat **254**  
 Wright, Sewall 73, 74, 356  
     fitness surface concept 366–8  
     heritability ( $h^2$ ) 297  
     infinite island model 135  
     isolation by distance 106–7  
     natural selection–genetic drift combined 222, 223, 224  
     shifting balance theory 366–73  
 Wright–Fisher model **55**, 55–7  
     adding natural selection 227  
     binomial probability distribution and 59, 60–1
- X-linked traits 16
- $\bar{z}$  305, **305**
- Zuckerkandl, E 255  
 zygotes 55, 190