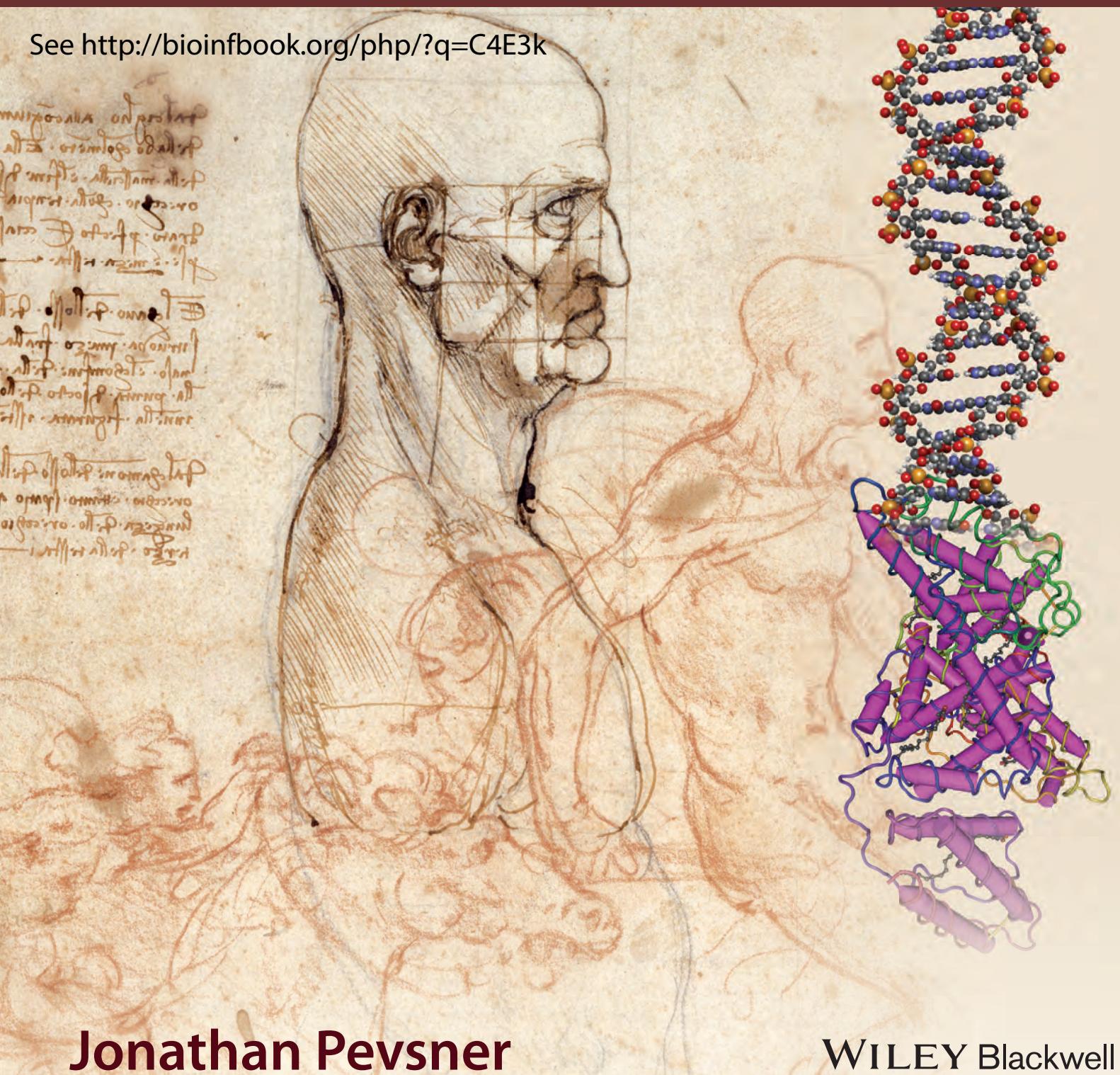


# BIOINFORMATICS AND FUNCTIONAL GENOMICS

third edition

See <http://bioinfbook.org/php/?q=C4E3k>



**Jonathan Pevsner**

**WILEY Blackwell**



BIOINFORMATICS AND  
FUNCTIONAL GENOMICS



# Bioinformatics and Functional Genomics

*Third Edition*

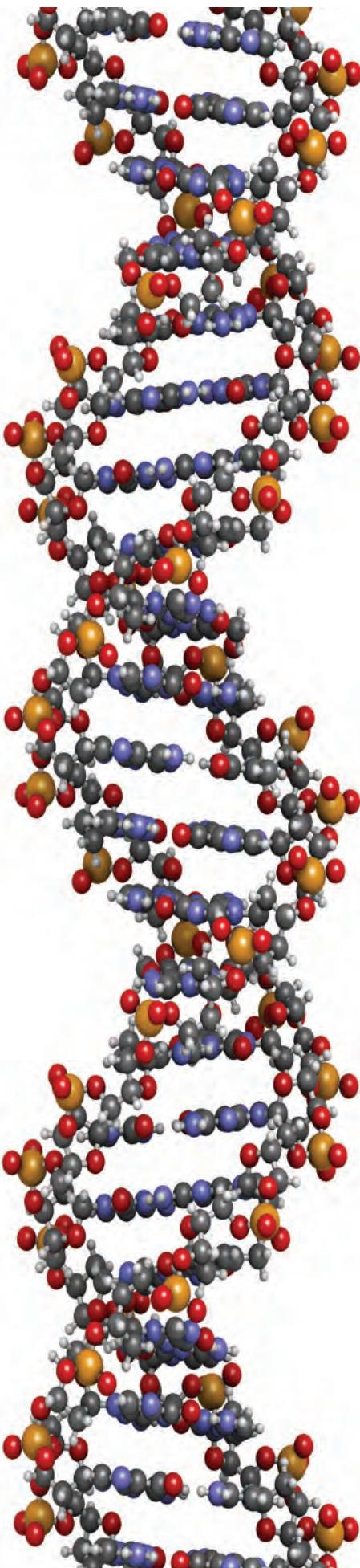
**Jonathan Pevsner**

Department of Neurology, Kennedy Krieger Institute,  
Baltimore, Maryland, USA

and

Department of Psychiatry and Behavioral Sciences,  
The Johns Hopkins School of Medicine, Baltimore,  
Maryland, USA

**WILEY Blackwell**



This edition first published 2015 © 2015 by John Wiley & Sons Inc

*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 UK 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.

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.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author(s) have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

*Library of Congress Cataloging-in-Publication Data*

Pevsner, Jonathan, 1961-, author.

Bioinformatics and functional genomics / Jonathan Pevsner.—Third edition.

p. ; cm.

Includes bibliographical references and indexes.

ISBN 978-1-118-58178-0 (cloth)

I. Title.

[DNLM: 1. Computational Biology—methods. 2. Genomics. 3. Genetic Techniques. 4. Proteomics. QU 26.5]

QH441.2

572.8'6—dc23

2015014465

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

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

The cover image is by Leonardo da Vinci, a study of a man in profile with studies of horse and riders (reproduced with kind permission of the Gallerie d'Accademia, Venice, Ms. 7r [236r], pen, black and red chalk). To the upper right a DNA molecule is shown (image courtesy of Wikimedia Commons) and a protein (human serum albumin, the most abundant protein in blood plasma, accession 1E7I, visualized with Cn3D software described in Chapter 13). Leonardo's text reads: "From the eyebrow to the junction of the lip with the chin, and the angle of the jaw and the upper angle where the ear joins the temple will be a perfect square. And each side by itself is half the head. The hollow of the cheek bone occurs half way between the tip of the nose and the top of the jaw bone, which is the lower angle of the setting on of the ear, in the frame here represented. From the angle of the eye-socket to the ear is as far as the length of the ear, or the third of the face." (Translation by Jean-Paul Richter, *The Notebooks of Leonardo da Vinci*, London, 1883.)

Set in Times LT Std 10.5/13 by Aptara, India

Printed in Singapore

1 2015

*For three generations of family: to my parents  
Aihud and Lucille; to my wife Barbara; to my daughters Kim,  
Ava, and Lillian; and to my niece Madeline*



# Contents in Brief

## PART I Analyzing DNA, RNA, and Protein Sequences

- 1 Introduction, 3
- 2 Access to Sequence Data and Related Information, 19
- 3 Pairwise Sequence Alignment, 69
- 4 Basic Local Alignment Search Tool (BLAST), 121
- 5 Advanced Database Searching, 167
- 6 Multiple Sequence Alignment, 205
- 7 Molecular Phylogeny and Evolution, 245

## PART II Genomewide Analysis of DNA, RNA, and Protein

- 8 DNA: The Eukaryotic Chromosome, 307
- 9 Analysis of Next-Generation Sequence Data, 377
- 10 Bioinformatic Approaches to Ribonucleic Acid (RNA), 433
- 11 Gene Expression: Microarray and RNA-seq Data Analysis, 479
- 12 Protein Analysis and Proteomics, 539
- 13 Protein Structure, 589
- 14 Functional Genomics, 635

## PART III Genome Analysis

- 15 Genomes Across the Tree of Life, 699
- 16 Completed Genomes: Viruses, 755
- 17 Completed Genomes: Bacteria and Archaea, 797
- 18 Eukaryotic Genomes: Fungi, 847
- 19 Eukaryotic Genomes: From Parasites to Primates, 887
- 20 Human Genome, 957
- 21 Human Disease, 1011

GLOSSARY, 1075

SELF-TEST QUIZ: SOLUTIONS, 1103

AUTHOR INDEX, 1105

SUBJECT INDEX, 1109



# Contents

Preface to the Third Edition, xxxi

About the Companion Website, xxxiii

## PART I ANALYZING DNA, RNA, AND PROTEIN SEQUENCES

### 1 Introduction, 3

Organization of the Book, 4

Bioinformatics: The Big Picture, 5

A Consistent Example: Globins, 6

Organization of the Chapters, 8

Suggestions For Students and Teachers: Web Exercises, Find-a-Gene, and

Characterize-a-Genome, 9

Bioinformatics Software: Two Cultures, 10

Web-Based Software, 11

Command-Line Software, 11

Bridging the Two Cultures, 12

New Paradigms for Learning Programming for Bioinformatics, 13

Reproducible Research in Bioinformatics, 14

Bioinformatics and Other Informatics Disciplines, 15

Advice for Students, 15

Suggested Reading, 15

References, 16

### 2 Access to Sequence Data and Related Information, 19

Introduction to Biological Databases, 19

Centralized Databases Store DNA Sequences, 20

Contents of DNA, RNA, and Protein Databases, 24

Organisms in GenBank/EMBL-Bank/DDBJ, 24

Types of Data in GenBank/EMBL-Bank/DDBJ, 26

Genomic DNA Databases, 27

DNA-Level Data: Sequence-Tagged Sites (STSs), 27

DNA-Level Data: Genome Survey Sequences (GSSs), 27

DNA-Level Data: High-Throughput Genomic Sequence (HTGS), 27

RNA data, 27

RNA-Level Data: cDNA Databases Corresponding to Expressed Genes, 27

RNA-Level Data: Expressed Sequence Tags (ESTs), 28

RNA-Level Data: UniGene, 28

Access to Information: Protein Databases, 29
UniProt, 31
Central Bioinformatics Resources: NCBI and EBI, 31
Introduction to NCBI, 31
The European Bioinformatics Institute (EBI), 32
Ensembl, 34
Access to Information: Accession Numbers to Label and Identify Sequences, 34
The Reference Sequence (RefSeq) Project, 36
RefSeqGene and the Locus Reference Genomic Project, 37
The Consensus Coding Sequence CCDS Project, 37
The Vertebrate Genome Annotation (VEGA) Project, 37
Access to Information via Gene Resource at NCBI, 38
Relationship Between NCBI Gene, Nucleotide, and Protein Resources, 41
Comparison of NCBI's Gene and UniGene, 41
NCBI's Gene and HomoloGene, 42
Command-Line Access to Data at NCBI, 42
Using Command-Line Software, 42
Accessing NCBI Databases with EDirect, 45
EDirect Example 1, 46
EDirect Example 2, 46
EDDirect Example 3, 46
EDDirect Example 4, 47
EDDirect Example 5, 48
EDDirect Example 6, 48
EDDirect Example 7, 48
Access to Information: Genome Browsers, 49
Genome Builds, 49
The University of California, Santa Cruz (UCSC) Genome Browser, 50
The Ensembl Genome Browser, 50
The Map Viewer at NCBI, 52
Examples of How to Access Sequence Data: Individual Genes/Proteins, 52
Histones, 52
HIV-1 pol, 53
How to Access Sets of Data: Large-Scale Queries of Regions and Features, 54
Thinking About One Gene (or Element) Versus Many Genes (Elements), 54
The BioMart Project, 54
Using the UCSC Table Browser, 54
Custom Tracks: Versatility of the BED File, 56
Galaxy: Reproducible, Web-Based, High-Throughput Research, 57
Access to Biomedical Literature, 58
Example of PubMed Search, 59
Perspective, 59
Pitfalls, 60
Advice for Students, 60

Web Resources, 60
Discussion Questions, 61
Problems/Computer Lab, 61
Self-Test Quiz, 63
Suggested Reading, 64
References, 64

### 3 Pairwise Sequence Alignment, 69

Introduction, 69
Protein Alignment: Often More Informative than DNA Alignment, 70
Definitions: Homology, Similarity, Identity, 70
Gaps, 78
Pairwise Alignment, Homology, and Evolution of Life, 78
Scoring Matrices, 79
Dayhoff Model Step 1 (of 7): Accepted Point Mutations, 79
Dayhoff Model Step 2 (of 7): Frequency of Amino Acids, 79
Dayhoff Model Step 3 (of 7): Relative Mutability of Amino Acids, 80
Dayhoff Model Step 4 (of 7): Mutation Probability Matrix for the Evolutionary Distance of 1 PAM, 82
Dayhoff Model Step 5 (of 7): PAM250 and Other PAM Matrices, 84
Dayhoff Model Step 6 (of 7): From a Mutation Probability Matrix to a Relatedness Odds Matrix, 88
Dayhoff Model Step 7 (of 7): Log-Odds Scoring Matrix, 89
Practical Usefulness of PAM Matrices in Pairwise Alignment, 91
Important Alternative to PAM: BLOSUM Scoring Matrices, 91
Pairwise Alignment and Limits of Detection: The “Twilight Zone”, 94
Alignment Algorithms: Global and Local, 96
Global Sequence Alignment: Algorithm of Needleman and Wunsch, 96
Step 1: Setting Up a Matrix, 96
Step 2: Scoring the Matrix, 97
Step 3: Identifying the Optimal Alignment, 99
Local Sequence Alignment: Smith and Waterman Algorithm, 101
Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST, 103
Basic Local Alignment Search Tool (BLAST), 104
Pairwise Alignment with Dotplots, 104
The Statistical Significance of Pairwise Alignments, 106
Statistical Significance of Global Alignments, 106
Statistical Significance of Local Alignments, 108
Percent Identity and Relative Entropy, 108
Perspective, 110
Pitfalls, 112
Advice for Students, 112
Web Resources, 112
Discussion Questions, 113
Problems/Computer Lab, 113

- Self-Test Quiz, 114
- Suggested Reading, 115
- References, 116

## 4 Basic Local Alignment Search Tool (BLAST) , 121

- Introduction, 121
- BLAST Search Steps, 124
  - Step 1: Specifying Sequence of Interest, 124
  - Step 2: Selecting BLAST Program, 124
  - Step 3: Selecting a Database, 126
  - Step 4a: Selecting Optional Search Parameters, 127
  - Step 4b: Selecting Formatting Parameters, 132
- Stand-Alone BLAST, 135
- BLAST Algorithm Uses Local Alignment Search Strategy, 138
  - BLAST Algorithm Parts: List, Scan, Extend, 138
  - BLAST Algorithm: Local Alignment Search Statistics and *E* Value, 141
  - Making Sense of Raw Scores with Bit Scores, 143
  - BLAST Algorithm: Relation Between *E* and *p* Values, 143
- BLAST Search Strategies, 145
  - General Concepts, 145
  - Principles of BLAST Searching, 146
    - How to Evaluate the Significance of Results, 146
    - How to Handle Too Many Results, 150
    - How to Handle Too Few Results, 150
  - BLAST Searching with Multidomain Protein: HIV-1 Pol, 151
- Using Blast For Gene Discovery: Find-a-Gene, 155
  - Perspective, 159
  - Pitfalls, 160
  - Advice for Students, 160
  - Web Resources, 160
    - Discussion Questions, 160
    - Problems/Computer Lab, 160
    - Self-Test Quiz, 161
    - Suggested Reading, 162
    - References, 163

## 5 Advanced Database Searching , 167

- Introduction, 167
- Specialized BLAST Sites, 168
  - Organism-Specific BLAST Sites, 168
  - Ensembl BLAST, 168
  - Wellcome Trust Sanger Institute, 170
- Specialized BLAST-Related Algorithms, 170
  - WU BLAST 2.0, 170
  - European Bioinformatics Institute (EBI), 170

Specialized NCBI BLAST Sites, 170
BLAST of Next-Generation Sequence Data, 170
Finding Distantly Related Proteins: Position-Specific Iterated BLAST (PSI-BLAST) and DELTA-BLAST, 171
PSI-BLAST Errors: Problem of Corruption, 177
Reverse Position-Specific BLAST, 177
Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST), 177
Assessing Performance of PSI-BLAST and DELTA-BLAST, 179
Pattern-Hit Initiated BLAST (PHI-BLAST), 179
Profile Searches: Hidden Markov Models, 181
HMMER Software: Command-Line and Web-Based, 184
BLAST-Like Alignment Tools to Search Genomic DNA Rapidly, 186
Benchmarking to Assess Genomic Alignment Performance, 187
PatternHunter: Nonconsecutive Seeds Boost Sensitivity, 188
BLASTZ, 188
Enredo and Pecan, 191
MegaBLAST and Discontinuous MegaBLAST, 191
BLAST-Like Tool (BLAT), 192
LAGAN, 192
SSAHA2, 194
Aligning Next-Generation Sequence (NGS) Reads to a Reference Genome, 194
Alignment Based on Hash Tables, 194
Alignment Based on the Burrows–Wheeler Transform, 196
Perspective, 197
Pitfalls, 197
Advice For Students, 198
Web Resources, 198
Discussion Questions, 198
Problems/Computer Lab, 198
Self-Test Quiz, 199
Suggested Reading, 200
References, 201

## 6 Multiple Sequence Alignment, 205

Introduction, 205
Definition of Multiple Sequence Alignment, 206
Typical Uses and Practical Strategies of Multiple Sequence Alignment, 207
Benchmarking: Assessment of Multiple Sequence Alignment Algorithms, 207
Five Main Approaches to Multiple Sequence Alignment, 208
Exact Approaches to Multiple Sequence Alignment, 208
Progressive Sequence Alignment, 208
Iterative Approaches, 214
Consistency-Based approaches, 218
Structure-Based Methods, 220
Benchmarking Studies: Approaches, Findings, Challenges, 221

Databases of Multiple Sequence Alignments, 222
Pfam: Protein Family Database of Profile HMMs, 223
SMART, 224
Conserved Domain Database, 226
Integrated Multiple Sequence Alignment Resources: InterPro and iProClass, 226
Multiple Sequence Alignment Database Curation: Manual Versus Automated, 227
Multiple Sequence Alignments of Genomic Regions, 227
Analyzing Genomic DNA Alignments via UCSC, 229
Analyzing Genomic DNA Alignments via Galaxy, 229
Analyzing Genomic DNA Alignments via Ensembl, 231
Alignathon Competition to Assess Whole-Genome Alignment Methods, 231
Perspective, 234
Pitfalls, 234
Advice for Students, 235
Discussion Questions, 235
Problems/Computer Lab, 235
Self-Test Quiz, 237
Suggested Reading, 238
References, 239

## 7 Molecular Phylogeny and Evolution, 245

Introduction to Molecular Evolution, 245
Principles of Molecular Phylogeny and Evolution, 246
Goals of Molecular Phylogeny, 246
Historical Background, 247
Molecular Clock Hypothesis, 250
Positive and Negative Selection, 254
Neutral Theory of Molecular Evolution, 258
Molecular Phylogeny: Properties of Trees, 259
Topologies and Branch Lengths of Trees, 259
Tree Roots, 262
Enumerating Trees and Selecting Search Strategies, 263
Type of Trees, 266
Species Trees versus Gene/Protein Trees, 266
DNA, RNA, or Protein-Based Trees, 268
Five Stages of Phylogenetic Analysis, 270
Stage 1: Sequence Acquisition, 270
Stage 2: Multiple Sequence Alignment, 271
Stage 3: Models of DNA and Amino Acid Substitution, 272
Stage 4: Tree-Building Methods, 281
Distance-Based, 282
Phylogenetic Inference: Maximum Parsimony, 287

Model-Based Phylogenetic Inference: Maximum Likelihood,	289
Tree Inference: Bayesian Methods,	290
Stage 5: Evaluating Trees,	293
Perspective,	295
Pitfalls,	295
Advice for Students,	296
Web Resources,	297
Discussion Questions,	297
Problems/Computer Lab,	297
Self-Test Quiz,	298
Suggested Reading,	298
References,	299

## PART II GENOMEWIDE ANALYSIS OF DNA, RNA, AND PROTEIN

### 8 DNA: The Eukaryotic Chromosome, 307

Introduction,	308
Major Differences between Eukaryotes and Bacteria and Archaea,	308
General Features of Eukaryotic Genomes and Chromosomes,	310
C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly,	312
Organization of Eukaryotic Genomes into Chromosomes,	310
Analysis of Chromosomes Using Genome Browsers,	314
Analysis of Chromosomes Using BioMart and biomaRt,	314
Example 1,	317
Example 2,	319
Example 3,	319
Example 4,	319
Example 5,	320
Analysis of Chromosomes by the ENCODE Project,	320
Critiques of ENCODE: the C Value Paradox Revisited and the Definition of Function,	322
Repetitive DNA Content of Eukaryotic Chromosomes,	323
Eukaryotic Genomes Include Noncoding and Repetitive DNA Sequences,	323
Interspersed Repeats (Transposon-Derived Repeats),	325
Processed Pseudogenes,	326
Simple Sequence Repeats,	331
Segmental Duplications,	331
Blocks of Tandemly Repeated Sequences,	333
Gene Content of Eukaryotic Chromosomes,	334
Definition of Gene,	334
Finding Genes in Eukaryotic Genomes,	336
Finding Genes in Eukaryotic Genomes: EGASP Competition,	339
Three Resources for Studying Protein-Coding Genes: RefSeq, UCSC Genes, Gencode,	340
Protein-Coding Genes in Eukaryotes: New Paradox,	342

Regulatory Regions of Eukaryotic Chromosomes, 342
Databases of Genomic Regulatory Factors, 342
Ultraconserved Elements, 345
Nonconserved Elements, 345
Comparison of Eukaryotic DNA, 346
Variation in Chromosomal DNA, 347
Dynamic Nature of Chromosomes: Whole-Genome Duplication, 347
Chromosomal Variation in Individual Genomes, 349
Structural Variation: Six Types, 351
Inversions, 351
Mechanisms of Creating Duplications, Deletions, and Inversions, 351
Models for Creating Gene Families, 353
Chromosomal Variation in Individual Genomes: SNPs, 354
Techniques to Measure Chromosomal Change, 355
Array Comparative Genomic Hybridization, 356
SNP Microarrays, 356
Next-Generation Sequencing, 359
Perspective, 359
Pitfalls, 359
Advice to Students, 360
Web Resources, 360
Discussion Questions, 361
Problems/Computer Lab, 361
Self-Test Quiz, 364
Suggested Reading, 365
References, 366

## 9 Analysis of Next-Generation Sequence Data, 377

Introduction, 378
DNA Sequencing Technologies, 377
Sanger Sequencing, 379
Next-Generation Sequencing, 379
Cyclic Reversible Termination: Illumina, 382
Pyrosequencing, 384
Sequencing by Ligation: Color Space with ABI SOLiD, 385
Ion Torrent: Genome Sequencing by Measuring pH, 387
Pacific Biosciences: Single-Molecule Sequencing with Long Read Lengths, 387
Complete Genomics: Self-Assembling DNA Nanoarrays, 387
Analysis of Next-Generation Sequencing of Genomic DNA, 387
Overview of Next-Generation Sequencing Data Analysis, 387
Topic 1: Experimental Design and Sample Preparation, 389
Topic 2: From Generating Sequence Data to FASTQ, 390
Finding and Viewing FASTQ files, 392
Quality Assessment of FASTQ data, 393
FASTG: A Richer Format than FASTQ, 394

- Topic 3: Genome Assembly, 394  
    Competitions and Critical Evaluations of the Performance of Genome Assemblers, 396  
    The End of Assembly: Standards for Completion, 398
- Topic 4: Sequence Alignment, 399  
    Alignment of Repetitive DNA, 400  
    Genome Analysis Toolkit (GATK) Workflow: Alignment with BWA, 401
- Topic 5: The SAM/BAM Format and SAMtools, 402  
    Calculating Read Depth, 405  
    Finding and Viewing BAM/SAM files, 405  
    Compressed Alignments: CRAM File Format, 406
- Topic 6: Variant Calling: Single-Nucleotide Variants and Indels, 408
- Topic 7: Variant Calling: Structural Variants, 409
- Topic 8: Summarizing Variation: The VCF Format and VCFtools, 410  
    Finding and Viewing VCF files, 413
- Topic 9: Visualizing and Tabulating Next-Generation Sequence Data, 413
- Topic 10: Interpreting the Biological Significance of Variants, 417
- Topic 11: Storing Data in Repositories, 421
- Specialized Applications of Next-Generation Sequencing, 421  
Perspective, 422  
Pitfalls, 423  
Advice for Students, 423  
Web Resources, 424  
    Discussion Questions, 424  
    Problems/Computer Lab, 424  
    Self-Test Quiz, 425  
    Suggested Reading, 425  
    References, 425

## 10 Bioinformatic Approaches to Ribonucleic Acid (RNA), 433

- Introduction to RNA, 433
- Noncoding RNA, 436  
    Noncoding RNAs in the Rfam Database, 436  
    Transfer RNA, 438  
    Ribosomal RNA, 441  
    Small Nuclear RNA, 445  
    Small Nucleolar RNA, 445  
    MicroRNA, 445  
    Short Interfering RNA, 447  
    Long Noncoding RNA (lncRNA), 447  
    Other Noncoding RNA, 448  
    Noncoding RNAs in the UCSC Genome and Table Browser, 448
- Introduction to Messenger RNA, 450  
    mRNA: Subject of Gene Expression Studies, 450  
    Low- and High-Throughput Technologies to Study mRNAs, 452

Analysis of Gene Expression in cDNA Libraries, 455
Full-Length cDNA Projects, 459
BodyMap2 and GTEx: Measuring Gene Expression Across the Body, 459
Microarrays and RNA-Seq: Genome-Wide Measurement of Gene Expression, 460
Stage 1: Experimental Design for Microarrays and RNA-seq, 461
Stage 2: RNA Preparation and Probe Preparation, 461
Stage 3: Data Acquisition, 464
Hybridization of Labeled Samples to DNA Microarrays, 464
Data acquisition for RNA-seq, 465
Stage 4: Data Analysis, 465
Stage 5: Biological Confirmation, 465
Microarray and RNA-seq Databases, 465
Further Analyses, 465
Interpretation of RNA Analyses, 466
The Relationship between DNA, mRNA, and Protein Levels, 466
The Pervasive Nature of Transcription, 467
eQTLs: Understanding the Genetic Basis of Variation in Gene Expression through Combined RNA-seq and DNA-seq, 468
Perspective, 469
Pitfalls, 470
Advice to Students, 470
Web Resources, 470
Discussion Questions, 471
Problems/Computer Lab, 471
Self-Test Quiz, 471
Suggested Reading, 472
References, 473

## 11 Gene Expression: Microarray and RNA-seq Data Analysis, 479

Introduction, 479
Microarray Analysis Method 1: GEO2R at NCBI, 482
GEO2R Executes a Series of R Scripts, 482
GEO2R Identifies the Chromosomal Origin of Regulated Transcripts, 485
GEO2R Normalizes Data, 486
GEO2R uses RMA Normalization for Accuracy and Precision, 488
Fold Change (Expression Ratios), 490
GEO2R Performs >22,000 Statistical Tests, 490
GEO2R Offers Corrections for Multiple Comparisons, 494
Microarray Analysis Method 2: Partek, 495
Importing Data, 496
Quality Control, 496
Adding Sample Information, 497
Sample Histogram, 498
Scatter Plots and MA Plots, 498

Working with Log <sub>2</sub> Transformed Microarray Data, 498
Exploratory Data Analysis with Principal Components Analysis (PCA), 498
Performing ANOVA in Partek, 501
From <i>t</i> -test to ANOVA, 503
Microarray Analysis Method 3: Analyzing a GEO Dataset with R, 504
Setting up the Analyses, 504
Reading CEL Files and Normalizing with RMA, 506
Identifying Differentially Expressed Genes (Limma), 508
Microarray Analysis and Reproducibility, 510
Microarray Data Analysis: Descriptive Statistics, 511
Hierarchical Cluster Analysis of Microarray Data, 511
Partitioning Methods for Clustering: k-Means Clustering, 516
Multidimensional Scaling Compared to Principal Components Analysis, 517
Clustering Strategies: Self-Organizing Maps, 517
Classification of Genes or Samples, 517
RNA-Seq, 519
Setting up a TopHat and CuffLinks Sample Protocol, 523
TopHat to Map Reads to a Reference Genome, 524
Cufflinks to Assemble Transcripts, 525
Cuffdiff to Determine Differential Expression, 525
CummeRbund to Visualize RNA-seq Results, 526
RNA-seq Genome Annotation Assessment Project (RGASP), 527
Functional Annotation of Microarray Data, 528
Perspective, 529
Pitfalls, 530
Advice for Students, 531
Suggested Reading, 531
Problems/Computer Lab, 532
Self-Test Quiz, 532
Suggested Reading, 533
References, 534

## 12 Protein Analysis and Proteomics, 539

Introduction, 539
Protein Databases, 540
Community Standards for Proteomics Research, 542
Evaluating the State-of-the-Art: ABRF analytic challenges, 542
Techniques for Identifying Proteins, 543
Direct Protein Sequencing, 543
Gel Electrophoresis, 543
Mass Spectrometry, 547

Four Perspectives on Proteins, 551
Perspective 1: Protein Domains and Motifs: Modular Nature of Proteins, 552
Added Complexity of Multidomain Proteins, 557
Protein Patterns: Motifs or Fingerprints Characteristic of Proteins, 557
Perspective 2: Physical Properties of Proteins, 559
Accuracy of Prediction Programs, 561
Proteomic Approaches to Phosphorylation, 563
Proteomic Approaches to Transmembrane Regions, 565
Introduction to Perspectives 3 and 4: Gene Ontology Consortium, 567
Perspective 3: Protein Localization, 568
Perspective 4: Protein Function, 570
Perspective, 575
Pitfalls, 575
Advice for Students, 575
Web Resources, 576
Discussion Questions, 578
Problems/Computer Lab, 578
Self-Test Quiz, 579
Suggested Reading, 580
References, 580

## 13 Protein Structure, 589

Overview of Protein Structure, 589
Protein Sequence and Structure, 590
Biological Questions Addressed by Structural Biology: Globins, 591
Principles of Protein Structure, 591
Primary Structure, 591
Secondary Structure, 594
Tertiary Protein Structure: Protein-Folding Problem, 598
Structural Genomics, the Protein Structure Initiative, and Target Selection, 600
Protein Data Bank, 602
Accessing PDB Entries at NCBI Website, 606
Integrated Views of Universe of Protein Folds, 609
Taxonomic System for Protein Structures: SCOP Database, 610
CATH Database, 613
Dali Domain Dictionary, 615
Comparison of Resources, 617
Protein Structure Prediction, 617
Homology Modeling (Comparative Modeling), 618
Fold Recognition (Threading), 619
<i>Ab Initio</i> Prediction (Template-Free Modeling), 621
A Competition to Assess Progress in Structure Prediction, 621
Intrinsically Disordered Proteins, 622
Protein Structure and Disease, 622
Perspective, 625

- Pitfalls, 625
- Advice for Students, 625
  - Discussion Questions, 625
  - Problems/Computer Lab, 626
  - Self-Test Quiz, 627
  - Suggested Reading, 628
  - References, 628

## 14 Functional Genomics, 635

- Introduction to Functional Genomics, 635
  - The Relationship Between Genotype and Phenotype, 637
- Eight-Model Organisms For Functional Genomics, 638
  - 1. The Bacterium *Escherichia coli*, 639
  - 2. The Yeast *Saccharomyces cerevisiae*, 640
  - 3. The Plant *Arabidopsis thaliana*, 643
  - 4. The Nematode *Caenorhabditis elegans*, 643
  - 5. The Fruit Fly *Drosophila melanogaster*, 645
  - 6. The Zebrafish *Danio rerio*, 645
  - 7. The Mouse *Mus musculus*, 646
  - 8. *Homo sapiens*: Variation in Humans, 647
- Functional Genomics Using Reverse and Forward Genetics, 648
  - Reverse Genetics: Mouse Knockouts and the  $\beta$ -Globin Gene, 650
  - Reverse Genetics: Knocking Out Genes in Yeast Using Molecular Barcodes, 653
  - Reverse Genetics: Random Insertional Mutagenesis (Gene Trapping), 657
  - Reverse Genetics: Insertional Mutagenesis in Yeast, 660
  - Reverse Genetics: Gene Silencing by Disrupting RNA, 662
  - Forward Genetics: Chemical Mutagenesis, 665
  - Comparison of Reverse and Forward Genetics, 665
- Functional Genomics and the Central Dogma, 666
  - Approaches to Function and Definitions of Function, 646
  - Functional Genomics and DNA: Integrating Information, 668
  - Functional Genomics and RNA, 668
  - Functional Genomics and Protein, 670
- Proteomics Approaches to Functional Genomics, 670
  - Functional Genomics and Protein: Critical Assessment of Protein Function Annotation, 672
  - Protein–Protein Interactions, 672
    - Yeast Two-Hybrid System, 673
    - Protein Complexes: Affinity Chromatography and Mass Spectrometry, 675
  - Protein–Protein Interaction Databases, 676
  - From Pairwise Interactions to Protein Networks, 678
    - Assessment of Accuracy, 680
    - Choice of Data, 680

Experimental Organism, 680
Variation in Pathways, 681
Categories of Maps, 681
Pathways, Networks, and Integration: Bioinformatics Resources, 682
Perspective, 685
Pitfalls, 686
Advice for Students, 686
Web Resources, 686
Discussion Questions, 686
Problems/Computer Lab, 686
Self-Test Quiz, 687
Suggested Reading, 688
References, 688

## PART III GENOME ANALYSIS

### 15 Genomes Across the Tree of Life, 699

Introduction, 700
Five Perspectives on Genomics, 701
Brief History of Systematics, 701
History of Life on Earth, 705
Molecular Sequences as the Basis of the Tree of Life, 705
Role of Bioinformatics in Taxonomy, 709
Prominent Web Resources, 710
Ensembl Genomes, 710
NCBI Genome, 710
Genome Portal of DOE JGI and the Integrated Microbial Genomes, 710
Genomes On Line Database (GOLD), 710
UCSC, 710
Genome-Sequencing Projects: Chronology, 711
Brief Chronology, 711
1976–1978: First Bacteriophage and Viral Genomes, 711
1981: First Eukaryotic Organellar Genome, 712
1986: First Chloroplast Genomes, 714
1992: First Eukaryotic Chromosome, 715
1995: Complete Genome of Free-Living Organism, 715
1996: First Eukaryotic Genome, 715
1997: <i>Escherichia coli</i> , 715
1998: First Genome of Multicellular Organism, 716
1999: Human Chromosome, 716
2000: Fly, Plant, and Human Chromosome 21, 716
2001: Draft Sequences of Human Genome, 716
2002: Continuing Rise in Completed Genomes, 717
2003: HapMap, 717
2004: Chicken, Rat, and Finished Human Sequences, 717

- 2005: Chimpanzee, Dog, Phase I HapMap, 718
- 2006: Sea Urchin, Honeybee, dbGaP, 718
- 2007: Rhesus Macaque, First Individual Human Genome, ENCODE Pilot, 718
- 2008: Platypus, First Cancer Genome, First Personal Genome Using NGS, 718
- 2009: Bovine, First Human Methlyome Map, 718
- 2010: 1000 Genomes Pilot, Neandertal , Exome Sequencing to Find Disease Genes, 719
- 2011: A Vision for the Future of Genomics, 719
- 2012: Denisovan Genome, Bonobo, and 1000 Genomes Project, 719
- 2013: The Simplest Animal and a 700,000-Year-Old Horse, 719
- 2014: Mouse ENCODE, Primates, Plants, and Ancient Hominids, 719
- 2015: Diversity in Africa, 720
- Genome Analysis Projects: Introduction, 720
  - Large-Scale Genomics Projects, 721
  - Criteria for Selection of Genomes for Sequencing, 722
    - Genome Size, 722
    - Cost, 722
    - Relevance to Human Disease, 723
    - Relevance to Basic Biological Questions, 724
    - Relevance to Agriculture, 724
    - Sequencing of One Versus Many Individuals from a Species, 724
  - Role of Comparative Genomics, 724
  - Resequencing Projects, 725
  - Ancient DNA Projects, 725
  - Metagenomics Projects, 725
- Genome Analysis Projects: Sequencing, 728
  - Genome-Sequencing Centers, 728
  - Trace Archive: Repository for Genome Sequence Data, 728
  - HTGS Archive: Repository for Unfinished Genome Sequence Data, 730
- Genome Analysis Projects: Assembly, 730
  - Four Approaches to Genome Assembly, 730
  - Genome Assembly: From FASTQ to Contigs with Velvet, 733
  - Comparative Genome Assembly: Mapping Contigs to Known Genomes, 734
  - Finishing: When Has a Genome Been Fully Sequenced?, 735
  - Genome Assembly: Measures of Success, 735
  - Genome Assembly: Challenges, 735
- Genome Analysis Projects: Annotation, 737
  - Annotation of Genes in Eukaryotes: Ensembl Pipeline, 738
    - Annotation of Genes in Eukaryotes: NCBI Pipeline, 739
    - Core Eukaryotic Genes Mapping Approach (CEGMA), 739
    - Assemblies from the Genome Reference Consortium, 741
    - Assembly Hubs and Transfers at UCSC, Ensembl, and NCBI, 741
    - Annotation of Genes in Bacteria and Archaea, 741
    - Genome Annotation Standards, 741
  - Perspective, 742

- Pitfalls, 742
- Advice for Students, 743
  - Discussion Questions, 743
  - Problems/Computer Lab, 743
  - Self-Test Quiz, 745
  - Suggested Reading, 743
  - References, 745

## 16 Completed Genomes: Viruses, 755

- Introduction, 755
  - International Committee on Taxonomy of Viruses (ICTV) and Virus Species, 756
- Classification of Viruses, 758
  - Classification of Viruses Based on Morphology, 758
  - Classification of Viruses Based on Nucleic Acid Composition, 758
  - Classification of Viruses Based on Genome Size, 758
  - Classification of Viruses Based on Disease Relevance, 760
  - Diversity and Evolution of Viruses, 762
  - Metagenomics and Virus Diversity, 764
- Bioinformatics Approaches to Problems in Virology, 765
- Human Immunodeficiency Virus (HIV), 766
  - NCBI and LANL resources for HIV-1, 766
- Influenza Virus, 771
- Measles Virus, 774
- Ebola Virus, 775
- Herpesvirus: From Phylogeny to Gene Expression, 776
  - The Pairwise Sequence Comparison (PASC) Tool, 780
- Giant Viruses, 782
  - Comparing genomes with MUMmer, 783
- Perspectives, 785
- Pitfalls, 786
- Advice for Students, 786
- Web Resources, 786
  - Discussion Questions, 787
  - Problems/Computer Lab, 787
  - Self-Test Quiz, 788
  - Suggested Reading, 789
  - References, 789

## 17 Completed Genomes: Bacteria and Archaea, 797

- Introduction, 797
- Classification of Bacteria and Archaea, 798
  - Classification of Bacteria by Morphological Criteria, 800
  - Classification of Bacteria and Archaea Based on Genome Size and Geometry, 801

Classification of Bacteria and Archaea Based on Lifestyle, 805
Classification of Bacteria Based on Human Disease Relevance, 808
Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences, 809
Classification of Bacteria and Archaea Based on Other Molecular Sequences, 810
The Human Microbiome, 811
Analysis of Bacterial and Archaeal Genomes, 814
Nucleotide Composition, 817
Finding Genes, 819
Interpolated Context Model (ICM), 822
GLIMMER3, 824
Challenges of Bacterial and Archaeal Gene Prediction, 825
Gene Annotation, 825
Lateral Gene Transfer, 827
Comparison of Bacterial Genomes, 830
TaxPlot, 830
MUMmer, 833
Perspective, 834
Pitfalls, 835
Advice for Students, 835
Web Resources, 835
Discussion Questions, 836
Problems/Computer Lab, 836
Self-Test Quiz, 836
Suggested Reading, 837
References, 837

## 18 Eukaryotic Genomes: Fungi, 847

Introduction, 847
Description and Classification of Fungi, 848
Introduction to Budding Yeast <i>Saccharomyces Cerevisiae</i> , 849
Sequencing Yeast Genome, 851
Features of Budding Yeast Genome, 851
Exploring Typical Yeast Chromosome, 854
Web Resources for Analyzing a Chromosome, 854
Exploring Variation in a Chromosome with Command-Line Tools, 857
Finding Genes in a Chromosome with Command-Line Tools, 858
Properties of Yeast Chromosome XII, 860
Gene Duplication and Genome Duplication of <i>S. cerevisiae</i> , 860
Comparative Analyses of Hemiascomycetes, 865
Comparative Analyses of Whole-Genome Duplication, 866
Identification of Functional Elements, 868
Analysis of Fungal Genomes, 869
Fungi in the Human Microbiome, 870

- Aspergillus, 871  
Candida albicans, 871  
*Cryptococcus neoformans*: model fungal pathogen, 872  
Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*, 873  
Neurospora crassa, 873  
First Basidiomycete: *Phanerochaete chrysosporium*, 875  
Fission Yeast *Schizosaccharomyces pombe*, 875  
Other Fungal Genomes, 876  
Ten Leading Fungal Plant Pathogens, 876  
Perspective, 876  
Pitfalls, 877  
Advice for Students, 877  
Web Resources, 877  
    Discussion Questions, 877  
    Problems/Computer Lab, 878  
    Self-Test Quiz, 879  
    Suggested Reading, 880  
    References, 880

## 19 Eukaryotic Genomes: From Parasites to Primates, 887

- Introduction, 887  
Protozoans at Base of Tree Lacking Mitochondria, 890  
    *Trichomonas*, 890  
    *Giardia lamblia*: A Human Intestinal Parasite, 891  
Genomes of Unicellular Pathogens: Trypanosomes and *Leishmania*, 890  
    Trypanosomes, 892  
    *Leishmania*, 894  
The Chromalveolates, 895  
    Malaria Parasite *Plasmodium falciparum*, 895  
    More Apicomplexans, 898  
    Astonishing Ciliophora: *Paramecium* and *Tetrahymena*, 899  
    Nucleomorphs, 902  
    Kingdom Stramenopila, 904  
Plant Genomes, 906  
    Overview, 906  
    Green Algae (*Chlorophyta*), 908  
    *Arabidopsis thaliana* Genome, 910  
    The Second Plant Genome: Rice, 913  
    Third Plant: Poplar, 914  
    Fourth Plant: Grapevine, 915  
    Giant and Tiny Plant Genomes, 915  
    Hundreds More Land Plant Genomes, 915  
    Moss, 916  
Slime and Fruiting Bodies at the Feet of Metazoans, 916  
    Social Slime Mold *Dictyostelium discoideum*, 916

- Metazoans, 917  
    Introduction to Metazoans, 917  
    900 MYA: the Simple Animal *Caenorhabditis elegans*, 918  
    900 MYA: *Drosophila melanogaster* (First Insect Genome), 919  
    900 MYA: *Anopheles gambiae* (Second Insect Genome), 921  
    900 MYA: Silkworm and Butterflies, 922  
    900 MYA: Honeybee, 923  
    900 MYA: A Swarm of Insect Genomes, 923  
    840 MYA: A Sea Urchin on the Path to Chordates, 924  
    800 MYA: *Ciona intestinalis* and the Path to Vertebrates, 925  
    450 MYA: Vertebrate Genomes of Fish, 926  
    350 MYA: Frogs, 929  
    320 MYA: Reptiles (Birds, Snakes, Turtles, Crocodiles), 929  
    180 MYA: The Platypus and Opposum Genomes, 931  
    100 MYA: Mammalian Radiation from Dog to Cow, 933  
    80 MYA: The Mouse and Rat, 934  
    5–50 MYA: Primate Genomes, 937  
Perspective, 940  
Pitfalls, 941  
Advice for Students, 941  
Web Resources, 942  
    Discussion Questions, 942  
    Problems/Computer Lab, 942  
    Self-Test Quiz, 943  
    Suggested Reading, 944  
    References, 944

## 20 Human Genome, 957

- Introduction, 957  
    Main Conclusions of Human Genome Project, 958  
    Gateways to Access the Human Genome, 959  
        NCBI, 959  
        Ensembl, 959  
        University of California at Santa Cruz Human Genome Browser, 961  
        NHGRI, 961  
        Wellcome Trust Sanger Institute, 964  
    Human Genome Project, 964  
        Background of Human Genome Project, 964  
        Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence, 966  
        Human Genome Assemblies, 966  
        Broad Genomic Landscape, 968  
            Long-Range Variation in GC Content, 969  
            CpG Islands, 969  
            Comparison of Genetic and Physical Distance, 970

Repeat Content of Human Genome, 971
Transposon-Derived Repeats, 972
Simple Sequence Repeats, 973
Segmental Duplications, 973
Gene Content of Human Genome, 974
Noncoding RNAs, 975
Protein-Coding Genes, 975
Comparative Proteome Analysis, 975
Complexity of Human Proteome, 978
25 Human Chromosomes, 979
Group A (Chromosomes 1–3), 981
Group B (Chromosomes 4, 5), 982
Group C (Chromosomes 6–12, X), 983
Group D (Chromosomes 13–15), 983
Group E (Chromosomes 16–18), 984
Group F (Chromosomes 19, 20), 984
Group G (Chromosomes 21, 22, Y), 984
Mitochondrial Genome, 985
Human Genome Variation, 986
SNPs, Haplotypes, and HapMap, 986
Viewing and Analyzing SNPs and Haplotypes, 988
HaploView, 988
HapMap Browser, 988
Integrative Genomics Browser (IGV), 988
NCBI dbSNP, 988
PLINK, 992
SNPduo, 990
Major Conclusions of HapMap Project, 994
The 1000 Genomes Project, 995
Variation: Sequencing Individual Genomes, 998
Perspective, 999
Pitfalls, 1000
Advice for Students, 1001
Discussion Questions, 1001
Problems/Computer Lab, 1001
Self-Test Quiz, 1003
Suggested Reading, 1004
References, 1004

## 21 Human Disease, 1011

Human Genetic Disease: A Consequence of DNA Variation, 1011
A Bioinformatics Perspective on Human Disease, 1012
Garrod's View of Disease, 1014
Classification of Disease, 1015
NIH Disease Classification: MeSH Terms, 1017

- Categories of Disease, 1020  
    Allele Frequencies and Effect Sizes, 1020  
    Monogenic Disorders, 1021  
    Complex Disorders, 1024  
    Genomic Disorders, 1025  
    Environmentally Caused Disease, 1029  
    Disease and Genetic Background, 1030  
    Mitochondrial Disease, 1030  
    Somatic Mosaic Disease, 1032  
    Cancer: A Somatic Mosaic Disease, 1033
- Disease Databases, 1036  
    OMIM: Central Bioinformatics Resource for Human Disease, 1036  
    Human Gene Mutation Database (HGMD), 1039  
    ClinVar and Databases of Clinically Relevant Variants, 1040  
    GeneCards, 1041  
    Integration of Disease Database Information at the UCSC Genome Browser, 1041  
    Locus-Specific Mutation Databases and LOVD, 1041  
    The PhenCode Project, 1044  
    Limitations of Disease Databases: The Growing Interpretive Gap, 1045  
    Human Disease Genes and Amino Acid Substitutions, 1045
- Approaches to Identifying Disease-Associated Genes and Loci, 1046  
    Linkage Analysis, 1047  
    Genome-Wide Association Studies, 1047  
    Identification of Chromosomal Abnormalities, 1050  
    Human Genome Sequencing, 1051  
        Genome Sequencing to Identify Monogenic Disorders, 1051  
        Genome Sequencing to Solve Complex Disorders, 1051  
        Research Versus Clinical Sequencing and Incidental Findings, 1052  
        Disease-causing Variants in Apparently Normal Individuals, 1054
- Human Disease Genes in Model Organisms, 1055  
    Human Disease Orthologs in Nonvertebrate Species, 1056  
    Human Disease Orthologs in Rodents, 1058  
    Human Disease Orthologs in Primates, 1059
- Functional Classification of Disease Genes, 1060  
Perspective, 1063  
Pitfalls, 1063  
Advice for Students, 1063  
    Discussion Questions, 1064  
    Problems/Computer Lab, 1062

Self-Test Quiz, 1065  
Suggested Reading, 1066  
References, 1066

**GLOSSARY, 1075**

**SELF-TEST QUIZ: SOLUTIONS, 1103**

**AUTHOR INDEX, 1105**

**SUBJECT INDEX, 1109**

# Preface to the Third Edition

When the first edition of this textbook was published in 2003, the Human Genome Project had just been completed at a cost of nearly US\$ 3 billion. When the second edition came into print in 2009, the first genome sequence of an individual (J. Craig Venter) had recently been published at an estimated cost of US\$ 80 million.

Let me tell you a remarkable story. It is now 2015 and it costs just several thousand dollars to obtain the complete genome sequence of an individual. Sturge-Weber syndrome is a rare neurocutaneous disorder (affecting the brain and skin) that is sometimes debilitating: some patients must have a hemispherectomy (removal of half the brain) to alleviate the severe seizures. We obtained paired samples from just three individuals with Sturge-Weber syndrome: biopsies were from affected parts of the body (such as port-wine stains that occur on the face, neck, or shoulder) or from presumably unaffected regions. We purified DNA and sequenced these six whole genomes, compared the matched pairs, and identified a single base pair mutation in the *GNAQ* gene as responsible for Sturge-Weber syndrome. (The mutation is somatic, mosaic, and activating: somatic in that it occurs during development but is not transmitted from the parents; mosaic in that it affects just part of the body; and activating because *GNAQ* encodes a protein that in the mutated form turns on a signaling cascade.) We found that mutations in this gene also cause port-wine stain birthmarks (which affect 1 in 300 people or about 23 million people worldwide). Matt Shirley, then a graduate student in my lab, performed the bioinformatics analyses that led to this discovery. He analyzed about 700 billion bases of DNA. After finding the mutation he confirmed it by re-sequencing dozens of samples, typically at over 10,000-fold depth of coverage. We reported these findings in the *New England Journal of Medicine* in 2013.

This story illustrates several aspects of the fields of bioinformatics and genomics. First, we are in a time period when there is an explosive growth in the availability of DNA sequence. This is enabling us to address biological questions in unprecedented ways. Second, while it is inexpensive to acquire DNA sequences, it is essential to know how to analyze them. One goal of this book is to introduce sequence analysis. Third, bioinformatics serves biology: we can only interpret the significance of DNA sequence variation in the context of some biological process (such as a disease state). In the case of the *GNAQ* mutation, that gene encodes a protein (called G $\alpha$ q) that we can study in tremendous detail using the tools of bioinformatics; we can evaluate its three-dimensional structure, the proteins and chemical messengers it interacts with, and the cellular pathways it participates in. Fourth, bioinformatics and genomics offer us hope. For Sturge-Weber syndrome patients and those with port-wine stain birthmarks, we are hopeful that a molecular understanding of these conditions will lead to treatments.

This book is written by a biologist who has used the tools of bioinformatics to help understand biomedical research questions. I introduce concepts in the context of biological problem-solving. Compared to earlier editions, this new text emphasizes command-line software on the Linux (or Mac) platform, complemented by web-based approaches.

In an era of “Big Data” there is a great divide between those whose intellectual core is centered in biomedical science and those whose focus involves computer science. I hope this book helps to bridge the divide between these two cultures.

Writing a book like this is a wonderful and constant learning experience. I thank past and present members of my lab who taught me including Shruthi Bandyadka (for advice on R), Christopher Bouton, Carlo Colantuoni, Donald Freed (for extensive advice on next-generation sequencing or NGS), Laurence Frelin, Mari Kondo, Sarah McClymont, Nathaniel Miller, Alicia Rizzo, Eli Roberson, Matt Shirley (who also provided extensive NGS advice), Eric Stevens, and Jamie Wangen. For advice on specific chapters, I thank: Ben Busby of the National Center for Biotechnology Information (NCBI) for advice regarding Chapters 1, 2, and 5 and detailed comments on Chapters 9 and 10; Eric Sayers and Jonathan Kans of NCBI for advice on EDirect in Chapter 2; Heiko Schmidt for advice on TREE-PUZZLE and MrBayes in Chapter 7; Joel Benington for detailed comments on Chapters 8 and 15–19 and helpful discussions about teaching; Harold Lehmann for guidance on various fields of informatics; and N. Varg for helpful comments on all chapters. I thank many colleagues who participated in teaching bioinformatics and genomics courses over the years. I've learned from all these teachers, including Dimitri Avramopoulos, Jef Boeke, Kyle Cunningham, Garry Cutting, George Dimopoulos, Egert Hoiczyk, Rafael Irizarry, Akhilesh Pandey, Sean Prigge, Ingo Ruczinski, Alan Scott, Alan F. Scott, Kirby D. Smith, David Sullivan, David Valle, and Sarah Wheelan. I am grateful to faculty members with whom I taught genomics workshops including Elana Fertig, Luigi Marchionni, John McGready, Loris Mulroni, Frederick Tan, and Sarah Wheelan. This book includes several thousand literature references, but I apologize to the many more colleagues whose work I did not cite. I also cite 900 websites and again apologize to the developers of the many I did not include.

I also acknowledge the support of Dr Gary W. Goldstein, President and CEO of the Kennedy Krieger Institute where I work. Kennedy Krieger Institute sees 22,000 patients a year, mostly children with neurodevelopmental disorders from common conditions (such as autism spectrum disorder and intellectual disability) to rare genetic diseases. I am motivated to try to apply the tools of bioinformatics and genomics to help these children. This perspective has guided my writing of this book, which emphasizes the relevance of all the topics in bioinformatics and genomics to human disease in general. We are hopeful that genomics will lead to an understanding of the molecular bases of so many devastating conditions, and this in turn may one day lead to better diagnosis, prevention, treatment, and perhaps even cures.

It is my pleasure to thank my editors at Wiley-Blackwell – Laura Bell, Celia Carden, Beth Dufour, Elaine Rowan, Fiona Seymour, Audrie Tan, and Rachel Wade – for generous support throughout this project. I appreciate all their dedication to the quality of the book.

On a personal note I thank my wife Barbara for her love and support throughout the very long process of writing this textbook. Finally, to my girls Ava and Lillian: I hope you'll always be inspired to be curious and full of wonder about the world around us.

# About the Companion Website

This book is accompanied by a companion website:

**[www.wiley.com/go/pevsnerbioinformatics](http://www.wiley.com/go/pevsnerbioinformatics)**

Readers can visit this website for supplemental information, such as PowerPoint files of all the figures and tables from the book, solutions to the Self-Test Quizzes and Problems found at the end of each chapter.

The author also maintains a comprehensive website for the book:

**[www.bioinfbook.org](http://www.bioinfbook.org)**

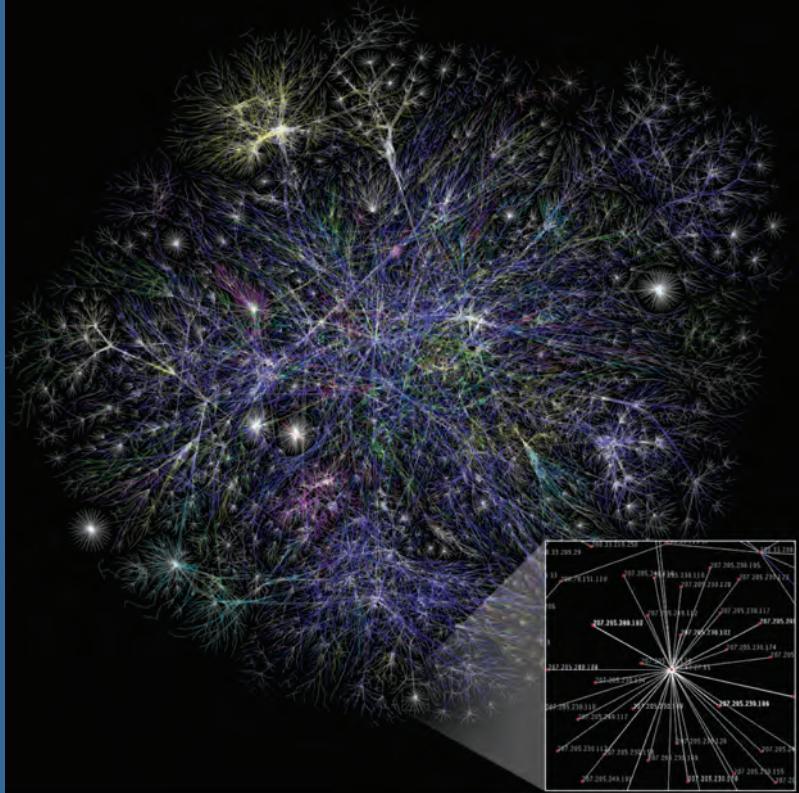
This site features lecture files (in PowerPoint and audiovisual format), over 900 Web Links and over 130 Web Documents that are referred to throughout the book as well as videocasts of how to perform many basic operations.



# Analyzing DNA, RNA, and Protein Sequences

PART

|



The first third of this book covers essential topics in bioinformatics. Chapter 1 provides an overview of the approaches we take, including the use of web-based and command-line software. We describe how to access sequences (Chapter 2). We then align them in a pairwise fashion (Chapter 3) or compare them to members of a database using BLAST (Chapter 4), including specialized searches of protein or DNA databases (Chapter 5). We next perform multiple sequence alignment (Chapter 6) and visualize these alignments as phylogenetic trees with an evolutionary perspective (Chapter 7).

The upper image shows the connectivity of the internet (from the Wikipedia entry for “internet”), while the lower image shows a map of human protein interactions (from the Wikipedia entry for “Protein–protein interaction”). We seek to understand biological principles on a genome-wide scale using the tools of bioinformatics.

Sources: Upper: Dcrjsr, 2002. Licensed under the Creative Commons Attribution 3.0 Unported license. Lower: The Opte Project, 2006. Licensed under the Creative Commons Attribution 2.5 Generic license.

# Introduction

# CHAPTER 1

*Penetrating so many secrets, we cease to believe in the unknowable. But there it sits nevertheless, calmly licking its chops.*

—H.L. Mencken

## LEARNING OBJECTIVES

After reading this chapter you should be able to:

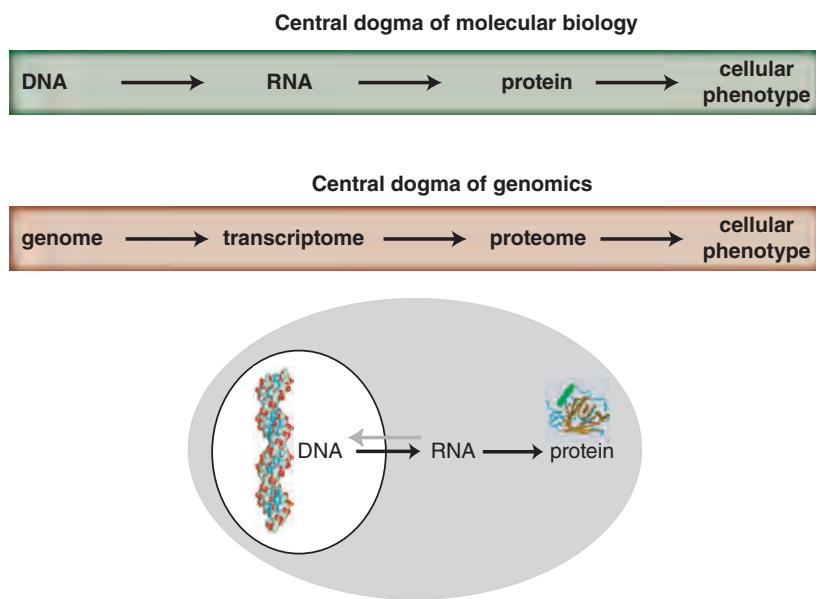
- define the terms bioinformatics;
- explain the scope of bioinformatics;
- explain why globins are a useful example to illustrate this discipline; and
- describe web-based versus command-line approaches to bioinformatics.

Bioinformatics represents a new field at the interface of the ongoing revolutions in molecular biology and computers. I define bioinformatics as the use of computer databases and computer algorithms to analyze proteins, genes, and the complete collection of deoxyribonucleic acid (DNA) that comprises an organism (the genome). A major challenge in biology is to make sense of the enormous quantities of sequence data and structural data that are generated by genome-sequencing projects, proteomics, and other large-scale molecular biology efforts. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms underlying biological problems related to the structure and function of macromolecules, biochemical pathways, disease processes, and evolution.

According to a National Institutes of Health (NIH) definition, bioinformatics is “research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, analyze, or visualize such data.” The related discipline of computational biology is “the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems.” Another definition from the National Human Genome Research Institute (NHGRI) is that “Bioinformatics is the branch of biology that is concerned with the acquisition, storage, display, and analysis of the information found in nucleic acid and protein sequence data.”

Russ Altman (1998) and Altman and Dugan (2003) offer two definitions of bioinformatics. The first involves information flow following the central dogma of molecular biology (**Fig. 1.1**). The second definition involves information flow that is transferred based

The NIH Bioinformatics Definition Committee findings are reported at <http://www.bisti.nih.gov/docs/CompuBioDef.pdf> (WebLink 1.1 at <http://bioinfbook.org/>). The NHGRI definition is available at <http://www.genome.gov/19519278> (WebLink 1.2).



**FIGURE 1.1** A first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, the Sequence Read Archive, and the DNA Database of Japan (DDBJ) serve as repositories for quadrillions ( $10^{15}$ ) of nucleotides of DNA sequence data (see Chapter 2). Corresponding databases of expressed genes (RNA) and protein have been established. A main focus of the field of bioinformatics is to study molecular sequence data to gain insight into a broad range of biological problems.

on scientific methods. This second definition includes problems such as designing, validating, and sharing software; storing and sharing data; performing reproducible research workflows; and interpreting experiments.

While the discipline of bioinformatics focuses on the analysis of molecular sequences, genomics and functional genomics are two closely related disciplines. The goal of genomics is to determine and analyze the complete DNA sequence of an organism, that is, its genome. The DNA encodes genes can be expressed as ribonucleic acid (RNA) transcripts and then, in many cases, further translated into protein. Functional genomics describes the use of genome-wide assays to study gene and protein function. For humans and other species, it is now possible to characterize an individual's genome, collection of RNA (transcriptome), proteome and even the collections of metabolites and epigenetic changes, and the catalog of organisms inhabiting the body (the microbiome) (Topol, 2014).

The aim of this book is to explain both the theory and practice of bioinformatics and genomics. The book is especially designed to help the biology student use computer programs and databases to solve biological problems related to proteins, genes, and genomes. Bioinformatics is an integrative discipline, and our focus on individual proteins and genes is part of a larger effort to understand broad issues in biology such as the relationship of structure to function, development, and disease. For the computer scientist, this book explains the motivations for creating and using algorithms and databases.

## ORGANIZATION OF THE BOOK

There are three main sections of the book. Part I (Chapters 2–7) explains how to access biological sequence data, particularly DNA and protein sequences (Chapter 2). Once sequences are obtained, we show how to compare two sequences (pairwise alignment;

Chapter 3) and how to compare multiple sequences (primarily by the Basic Local Alignment Search Tool or BLAST; Chapters 4 and 5). We introduce multiple sequence alignment (Chapter 6) and show how multiply aligned proteins or nucleotides can be visualized in phylogenetic trees (Chapter 7). Chapter 7 therefore introduces the subject of molecular evolution.

Part II describes functional genomics approaches to DNA, RNA, and protein and the determination of gene function (Chapters 8–14). The central dogma of biology states that DNA is transcribed into RNA then translated into protein. Chapter 8 introduces chromosomes and DNA, while Chapter 9 describes next-generation sequencing technology (emphasizing practical data analysis). We next examine bioinformatic approaches to RNA (Chapter 10), including both noncoding and coding RNAs. We then describe the measurement of mRNA (i.e., gene expression profiling) using microarrays and RNA-seq. Again we focus on practical data analysis (Chapter 11). From RNA we turn to consider proteins from the perspective of protein families, and the analysis of individual proteins (Chapter 12) and protein structure (Chapter 13). We conclude the second part of the book with an overview of the rapidly developing field of functional genomics (Chapter 14), which integrates contemporary approaches to characterizing the genome, transcriptome, and proteome.

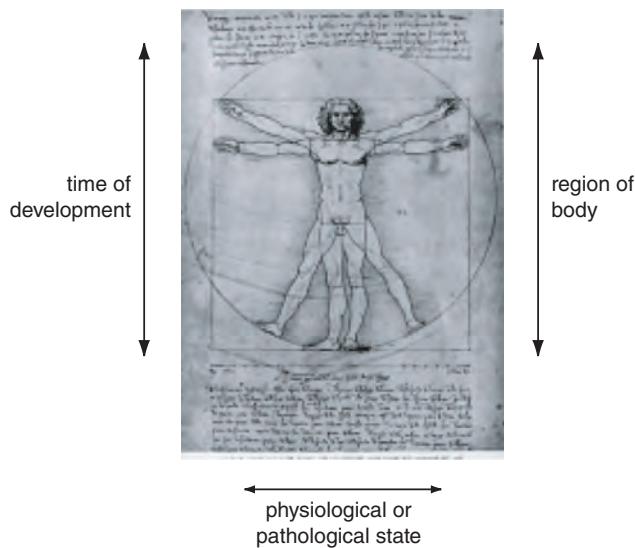
Part III covers genome analysis across the tree of life (Chapters 15–21). Since 1995, the genomes have been sequenced for several thousand viruses, bacteria, and archaea as well as eukaryotes such as fungi, animals, and plants. Chapter 15 provides an overview of the study of completed genomes. We describe bioinformatics resources for the study of viruses (Chapter 16) and bacteria and archaea (Chapter 17; these are two of the three main branches of life). Next we explore the genomes of a variety of eukaryotes including fungi (Chapter 18), organisms from parasites to primates (Chapter 19) and then the human genome (Chapter 20). Finally, we explore bioinformatic approaches to human disease (Chapter 21).

The third part of the book, spanning the tree of life from the perspective of genomics, depends strongly on the tools of bioinformatics from the first two parts of the book. I felt that this book would be incomplete if it introduced bioinformatics without also applying its tools and principles to the genomes of all life.

## BIOINFORMATICS: THE BIG PICTURE

We can summarize the fields of bioinformatics and genomics with three perspectives. The first perspective on bioinformatics is the cell (**Fig. 1.1**). Here we follow the central dogma. A focus of the field of bioinformatics is the collection of DNA (the genome), RNA (the transcriptome), and protein sequences (the proteome) that have been amassed. These millions–quadrillions of molecular sequences present both great opportunities and great challenges. A bioinformatics approach to molecular sequence data involves the application of computer algorithms and computer databases to molecular and cellular biology. Such an approach is sometimes referred to as functional genomics. This typifies the essential nature of bioinformatics: biological questions can be approached from levels ranging from single genes and proteins to cellular pathways and networks or even whole-genomic responses. Our goals are to understand how to study both individual genes and proteins and collections of thousands of genes/proteins.

From the cell we can focus on individual organisms, which represents a second perspective of the field of bioinformatics (**Fig. 1.2**). Each organism changes across different stages of development and (for multicellular organisms) across different regions of the body. For example, while we may sometimes think of genes as static entities that specify features such as eye color or height, they are in fact dynamically regulated across time and region and in response to physiological state. Gene expression varies in disease states or



**FIGURE 1.2** A second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual’s genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. For an individual organism, bioinformatics tools can therefore be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.

in response to a variety of signals, both intrinsic and environmental. Many bioinformatics tools are available to study the broad biological questions relevant to the individual: there are many databases of expressed genes and proteins derived from different tissues and conditions. One of the most powerful applications of functional genomics is the use of DNA microarrays or RNA-seq to measure the expression of thousands of genes in biological samples.

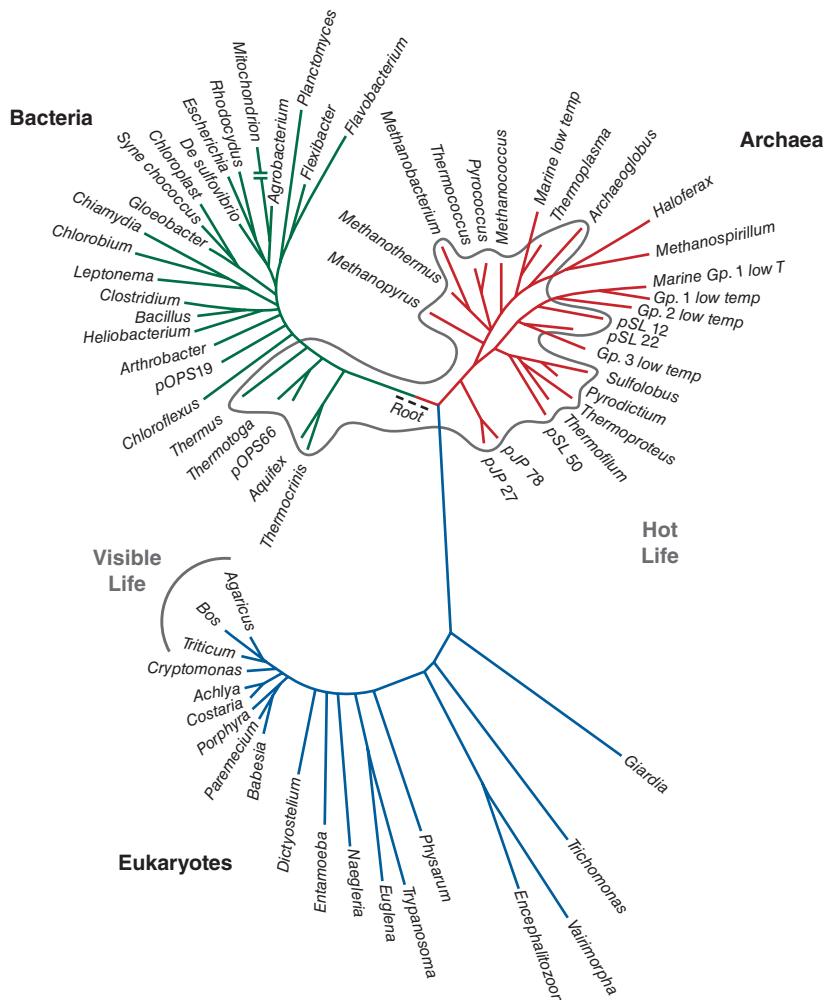
At the largest scale is the tree of life (Fig. 1.3; see also Chapter 15). There are many millions of species alive today, and they can be grouped into the three major branches of bacteria, archaea, and eukaryotes. Molecular sequence databases currently hold DNA sequence from ~300,000 different species. The complete genome sequences of thousands of organisms are now available. One of the main lessons we are learning is the fundamental unity of life at the molecular level. We are also coming to appreciate the power of comparative genomics, in which genomes are compared. Through DNA sequence analysis we are learning how chromosomes evolve and are sculpted through processes such as chromosomal duplications, deletions, and rearrangements, and through whole-genome duplications (Chapters 8 and 18–19).

Figure 1.4 depicts the contents of this book in the context of these three perspectives of bioinformatics.

### A Consistent Example: Globins

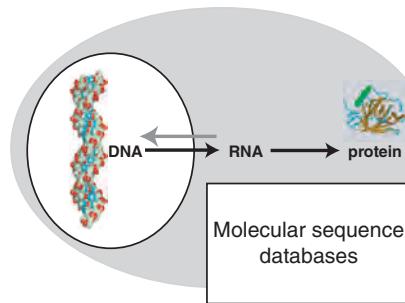
Throughout this book, we will focus on the globin gene family to provide a consistent example of bioinformatics and genomics concepts. The globin family is one of the best characterized in biology.

- Historically, hemoglobin is one of the first proteins to be studied, having been described in the 1830s and 1840s by Gerardus Johannes Mulder, Justus Liebig, and others.
- Myoglobin, a globin that binds oxygen in the muscle tissue, was the first protein to have its structure resolved by X-ray crystallography (Chapter 13).



**FIGURE 1.3** A third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are not depicted here. For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome). We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth. Adapted from Barns *et al.* (1996), Hugenholtz and Pace (1996), and Pace (1997).

- Hemoglobin, a tetramer of four globin subunits (principally  $\alpha_2\beta_2$  in adults), is the main oxygen carrier in the blood of vertebrates. Its structure was also one of the earliest to be described. The comparison of myoglobin, alpha globin, and beta globin protein sequences represents one of the earliest applications of multiple sequence alignment (Chapter 6), and led to the development of amino acid substitution matrices used to score protein relatedness (Chapter 3).
- As DNA sequencing technology emerged in the 1980s, the globin loci on human chromosomes 16 (for  $\alpha$  globin) and 11 (for  $\beta$  globin) were among the first to be sequenced and analyzed. The globin genes are exquisitely regulated across time (switching from embryonic to fetal to adult forms) and with tissue-specific gene expression. We will discuss these loci in the description of the control of gene expression (Chapters 10 and 14).
- While hemoglobin and myoglobin remain the best-characterized globins, the family of homologous proteins extends to separate classes of plant globins, invertebrate



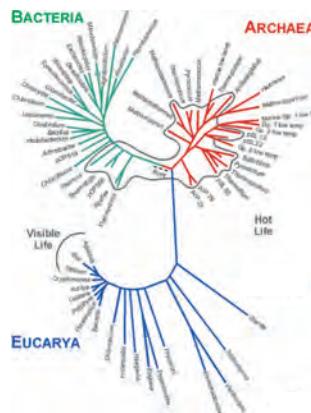
#### Part I: Bioinformatics: analyzing DNA, RNA, and protein

- Chapter 1: Introduction
- Chapter 2: How to obtain sequences
- Chapter 3: How to compare two sequences
- Chapters 4 and 5: How to compare a sequence across databases
- Chapter 6: How to multiply align sequences
- Chapter 7: How to view multiply aligned sequences as phylogenetic trees



#### Part II: Functional genomics: from DNA to RNA to protein

- Chapter 8: DNA: The eukaryotic chromosome
- Chapter 9: DNA analysis: next-generation sequencing
- Chapter 10: Bioinformatics approaches to RNA
- Chapter 11: Microarray and RNA-seq data analysis
- Chapter 12: Protein analysis and protein families
- Chapter 13: Protein structure
- Chapter 14: Functional genomics



#### Part III: Genomics

- Chapter 15: The tree of life
- Chapter 16: Viruses
- Chapter 17: Bacteria and archaea
- Chapter 18: Fungi
- Chapter 19: Eukaryotes from parasites to plants to primates
- Chapter 20: The human genome
- Chapter 21: Human disease

**FIGURE 1.4** Overview of the chapters in this book.

hemoglobins (some of which contain multiple globin domains within one protein molecule), bacterial homodimeric hemoglobins (consisting of two globin subunits), and flavohemoglobins that occur in bacteria, archaea, and fungi. The globin family is therefore useful as we survey the tree of life (Chapters 15–21).

## ORGANIZATION OF THE CHAPTERS

The chapters of this book are intended to provide both the theory of bioinformatics subjects as well as a practical guide to using computer databases and algorithms. Web resources are provided throughout each chapter. Chapters end with brief sections called Perspective, Pitfalls, and Advice for Students. The perspective feature describes the rate of growth of the subject matter in each chapter. For example, a perspective on Chapter 2 (access to sequence information) is that the amount of DNA sequence data deposited in repositories is undergoing an explosive rate of growth. In contrast, an area such as

pairwise sequence alignment, which is fundamental to the entire field of bioinformatics (Chapter 3), was firmly established in the 1970s and 1980s. Even for fundamental operations such as multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7), dozens of novel, ever-improving approaches are being introduced at a rapid rate. For example, hidden Markov models and Bayesian approaches are being applied to a wide range of bioinformatics problems.

The pitfalls section of each chapter describes some common difficulties encountered by biologists using bioinformatics tools. Some errors might seem trivial, such as searching a DNA database with a protein sequence. Other pitfalls are more subtle, such as artifacts caused by multiple sequence alignment programs depending upon the type of parameters that are selected. Indeed, while the field of bioinformatics depends substantially on analyzing sequence data, it is important to recognize that there are many categories of errors associated with data generation, collection, storage, and analysis. We address the problems of false positive and false negative results in a variety of searches and analyses.

Each chapter includes multiple-choice quizzes to test your understanding of the chapter materials. There are also problems that require you to apply the concepts presented in each chapter. These problems may form the basis of a computer laboratory for a bioinformatics course.

The reference list at the end of each chapter is preceded by a discussion of recommended articles. This “Suggested Reading” section includes classic papers that show how the principles described in each chapter were discovered. Particularly helpful review articles and research papers are highlighted.

## SUGGESTIONS FOR STUDENTS AND TEACHERS: EXERCISES, FIND-A-GENE, AND CHARACTERIZE-A-GENOME

This is a textbook for two separate courses: the first course is an introduction to bioinformatics (Parts I and II, i.e., Chapters 1–14), and the second is an introduction to genomics (Part III, i.e., Chapters 15–21). In a sense, the discipline of bioinformatics serves biology, facilitating ways of posing and then answering questions about proteins, genes, and genomes. Part III of this book surveys the tree of life from the perspective of genes and genomes, and could not progress without the bioinformatics tools described in Parts I and II of the book.

Students often have a particular research area of interest such as a gene, a physiological process, a disease, or a genome. It is hoped that, in the process of studying globins and other specific proteins and genes throughout this book, students can simultaneously apply the principles of bioinformatics to their own research questions.

The websites described in this book are posted on the home page for this book (<http://www.bioinfbook.org>) as “WebLinks.” That site contains 900 URLs, organized by chapter. Each chapter also refers to web documents posted on the site. For example, if you see a figure of a phylogenetic tree or a sequence alignment, you can easily retrieve the raw data and make the figure yourself.

Another feature of a Johns Hopkins bioinformatics course is that each student is required to discover a novel gene by the last day of the course. The student must begin with any protein sequence of interest and perform database searches to identify genomic DNA that encodes a protein no one has described before. This problem is described in detail in Chapter 4 (and summarized in Web Document 4.5 at <http://www.bioinfbook.org/chapter4>). The student therefore chooses the name of the gene and its corresponding protein, and describes information about the organism and evidence that the gene has not been described before. The student then creates a multiple sequence alignment of the new protein (or gene) and creates a phylogenetic tree showing its relation to other known sequences.

Each year, some beginning students are slightly apprehensive about accomplishing this exercise; in the end, all of them succeed. A benefit of this exercise is that it requires a student to actively use the principles of bioinformatics. Many students choose a gene (or protein) relevant to their own research area.

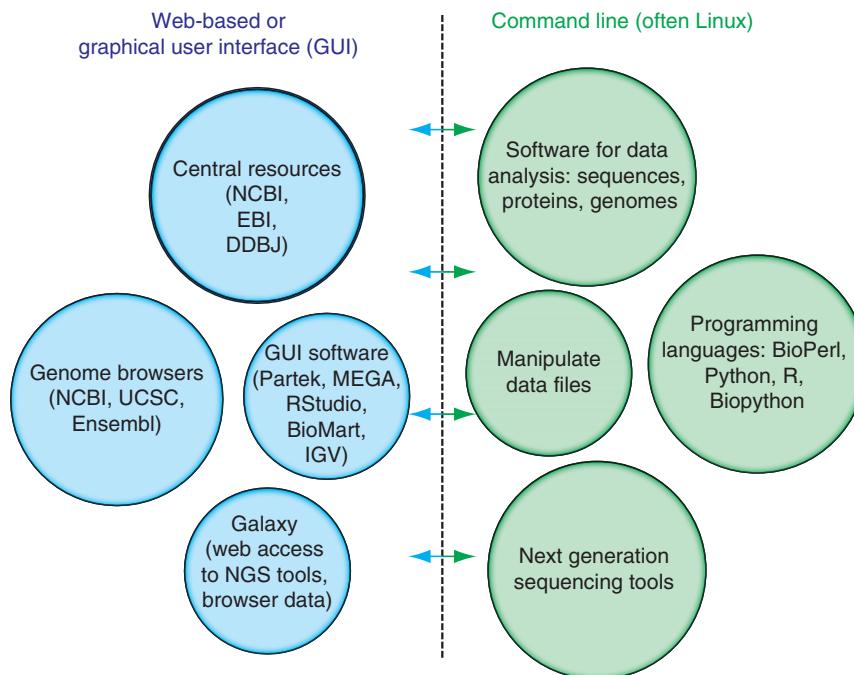
For a genomics course, students select a genome of interest and describe five aspects in depth (described at the start of Chapter 15):

1. The basic characteristics of the genome, such as its size, number of chromosomes, and other features, are described.
2. A comparative genomic analysis is performed to study the relation of the species to its neighbors.
3. The student describes biological principles that are learned through genome analysis.
4. The human disease relevance is described.
5. Bioinformatics aspects are described, such as key databases or algorithms used for genome analysis.

Teaching bioinformatics and genomics is notable for the diversity of students learning these new disciplines. Each chapter provides background on the subject matter. For more advanced students, key research papers are listed at the end of each chapter. These papers are technical, and reading them along with the chapters will provide a deeper understanding of the material.

## BIOINFORMATICS SOFTWARE: TWO CULTURES

There are two dramatically different approaches to bioinformatics: using web-based and command-line tools (Fig. 1.5). Web-based tools, sometimes called “point-and-click,” do not require knowledge of programming and are immediately accessible.



**FIGURE 1.5** Bioinformatics resources. Web-based or “point-and-click” resources are shown to the left, including the major portals (National Center for Biotechnology Information, European Bioinformatics Institute), major genome browsers (Ensembl, UCSC), databases, and specialized websites. Command-line resources are shown to the right. These include programming languages (such as Biopython, BioPerl, and the R language) and command-line software (typically accessed using the Linux operating system).

Command-line tools may have a steeper learning curve, but almost always offer more options for executing programs. They are more appropriate for analyzing large-scale datasets that are now routinely encountered in bioinformatics. Even for smaller datasets, command-line approaches can offer more flexibility and precision in accomplishing your tasks and more reproducible research since you can document your analysis steps.

## Web-Based Software

The field of bioinformatics relies heavily on the Internet as a place to access sequence data, to access software that is useful to analyze molecular data, and as a place to integrate different kinds of resources and information relevant to biology. We will describe a variety of websites. Initially, we will focus on the main publicly accessible databases that serve as repositories for DNA and protein data. These include:

1. the National Center for Biotechnology Information (NCBI), which hosts GenBank and other resources;
2. the European Bioinformatics Institute (EBI);
3. Ensembl, which includes a genome browser and resources to study dozens of genomes; and
4. the University of California at Santa Cruz (UCSC) Genome bioinformatics site, including a web browser and table browser for a variety of species.

Throughout the chapters of this book we introduce almost 1000 additional websites that are relevant to bioinformatics. The main advantages offered by websites are easy access, rapid updates, good visibility to the community, and ease of use (since in general programming skills, command line skills, and the use of Linux-type operating systems are not required).

The URLs for these sites are NCBI, <http://ncbi.nlm.nih.gov> (WebLink 1.3); EBI, <http://www.ebi.ac.uk> (WebLink 1.4); Ensembl, <http://www.ensembl.org> (WebLink 1.5); and UCSC, <http://genome.ucsc.edu> (WebLink 1.6). For information on vast numbers of available databases, see the annual January issue of the journal *Nucleic Acids Research*, <http://nar.oxfordjournals.org/> (WebLink 1.7).

## Command-Line Software

Command-line tools offer distinct, critical advantages. High-throughput approaches to biology result in the creation of both large and small datasets which require sophisticated analyses. We can think about command-line software in several ways.

1. The operating system is often Linux (a Unix-like environment). The Mac O/S is compatible with Linux as well (and is POSIX-compliant). However, while Windows-type operating systems are popular, they are not appropriate for the majority of command-line programs. In this book I assume the reader has no background in Unix. Beginning in Chapter 2, I provide basic instructions for becoming acquainted with Linux by providing examples of commands for a variety of software.
2. Programming languages are commonly used in bioinformatics. Examples are Perl (or its relative BioPerl; Stajich, 2007), Python (as well as Biopython), and R to manipulate data. Learning such languages is important as it is extremely useful to be able to write scripts and thus accomplish a broad range of tasks. Modules are available for hundreds of bioinformatics applications. For example, the BioConductor project currently includes > 1,000 packages that are useful for solving many tasks. Acquiring knowledge of R is a steep learning curve, and I provide suggestions of books, articles, and websites you can use to achieve this aim. It is also possible to use an R package without being an R “power user,” however. For example, in Chapter 8 we use the R package `Biostrings` to extract information about the features on chromosomes, and in Chapter 11 we use R packages to analyze gene expression datasets from microarrays and next-generation sequencing. Once you learn to use a few packages, you will be in a position to learn many more.

POSIX is an acronym for Portable Operating System Interface. It offers standards for maintaining compatibility between operating systems.

See <http://bioinfbook.org/chapter1> for links to resources for learning Unix.

3. The command line of Unix systems offers Bash, a default shell for Linux and Mac OS X operating systems. We introduce a variety of Bash scripts in this book. Bash includes a series of utilities that can accomplish tasks such as sorting a table of data, transposing it, counting the numbers of rows and columns, merging data, or working with regular expressions. We'll see examples of Bash commands in Box 2.3 and in Chapter 9 on next-generation sequencing, for example.

Bio-Linux 8 (released July 2014) is available at <http://environmentalomics.org/bio-linux/> (WebLink 1.8). Cygwin is available at <http://www.cygwin.com> (WebLink 1.9). PuTTY is at <http://www.putty.org> (WebLink 1.10).

Which operating system should you use? Linux is essential for many bioinformatics experts, often because it is used to access very large datasets (e.g., terabytes of data) with large amounts of RAM. For example, I recommend installing Bio-Linux on a laptop or a virtual machine. For many students approaching bioinformatics for the first time, the Macintosh O/S works well because it offers a Unix-like terminal. For Windows users, Cygwin provides a Unix-like environment. If you have access to a Linux server you can access it from a Windows or Mac environment using software such as PuTTY.

We may further distinguish between using command-line software and using a programming language. Learning Perl, Python, or other languages offers tremendous benefits (Dudley and Butte, 2009). However, even if you do not program, you still should learn basic information about how to acquire, store, manipulate, and explore large files. Many files used in bioinformatics and genomics are simply too large to be handled efficiently (if at all) by web-based or GUI-based software. Many files that are generated by software tools require some level of restructuring to be further studied (e.g., to be analyzed by additional software tools). For many students, it has become essential to learn techniques to manipulate files on the command line.

## Bridging the Two Cultures

Many bioinformatics resources are available to bridge the cultures of web-based and command-line software. This book introduces you to both (Table 1.1). For example, NCBI offers the web-based Entrez database that lets you type a query and obtain information. NCBI also provides EDirect, a set of command-line programs to access databases (see Chapter 2). Similarly, Ensembl offers programmatic access using Perl application programming interfaces (APIs). As another example, Galaxy hosts a broad range of web-based tools that are otherwise available as command-line software run on the Linux environment.

What is your best approach? Each person engaged in bioinformatics work should decide what problem he or she wants to solve, then choose the appropriate tool(s). If you are working with next-generation sequence data, it will be essential to learn how to use software tools in the Linux operating system. If that is new to you, you could use the more accessible Galaxy tools to start becoming familiar with the types of data and algorithms you will encounter as you transition to Linux-based tools. If you are doing phylogeny you can also start with MEGA software to learn a variety of approaches before complementing your analyses with command-line software to perform Bayesian analyses (see Chapter 7).

In this book we will use examples to try to help bridge these cultures. In Chapter 8 we will encounter both BioMart (an Ensembl web-based resource that interconnects hundreds of databases) and `biomaRt` (an R package that performs BioMart queries).

We will also see that the bioinformatics community is continuously improving existing software and developing new methods. There are often “competitions” in which organizers of an event obtain evidence of the gold standard “truth” for some problem, such as solving a protein structure or assembling a genome. Members of the community are then invited to compete to solve the answer within some time frame. By comparing the various results it is possible to assess the performance of each software

**TABLE 1.1** Overview of some web-based (or graphical user interface (GUI)) and command-line software used in various chapters of this book.

Part: Chapter	Topic	Web-based or GUI software	Command-line software
I: 2	Access to information	BioMart Genome Workbench	EDirect
I: 3	Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
I: 4	BLAST	BLAST	BLAST+
I: 5	Database searching	DELTA-BLAST Megablast	HMMER
I: 6	Multiple alignment	Pfam, MUSCLE	MAFFT
I: 7	Phylogeny	MEGA	MrBayes
II: 8	Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
II: 9	Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
II: 10	RNA	RNAfam, tRNAscan	
II: 11	RNAseq	Galaxy	affy (R package), RSEM
II: 12	Proteomics	ExPASy	pepstats (EMBOSS)
II: 13	Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
II: 14	Functional genomics	FLink, Cytoscape	
III: 15	Tree of life		Velvet (assembly)
III: 16	Viruses		MUMmer (alignment)
III: 17	Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
III: 18	Fungi	YGOB	Ensembl (variants)
III: 19	Eukaryotic genomes		
III: 20	Human genome		PLINK
III: 21	Human disease	OMIM, BioMart	EDirect, MitoSeek

(i.e., true and false positives, true and false negatives); by defining the sensitivity and specificity of software we learn which tools to use. Examples of critical assessments are given in **Table 1.2**.

## New Paradigms for Learning Programming for Bioinformatics

It is an excellent idea to learn a programming language to facilitate your bioinformatics work. You may want to run programs that are written in a language such as R or Python (as we do in this book), or you may want to write your own code and manipulate data to solve some task. In addition to available books and courses, many websites offer online training in the forms of tutorials or courses. David Searls (2012a, 2014) has reviewed many such online resources. These include Massive Open Online Courses (MOOCs) that tens of thousand of students may register for. Searls (2012b) also suggests ten rules for online learning. Briefly, these include: make a plan; be selective; organize your learning environment; do the readings; do the exercises; do the assessments; exploit the advantages (e.g., convenience); reach out to others; document your achievements; and be realistic

Excellent websites that guide you to learn a language include Code School (<https://www.codeschool.com>, WebLink 1.11), Code Academy (<http://www.codecademy.com>, WebLink 1.12), Data Camp (<https://www.datacamp.com>, WebLink 1.13), and Software Carpentry (<http://software-carpentry.org>, WebLink 1.14). Rosalind offers bioinformatics instruction through problem solving (<http://rosalind.info/problems/locations/>, WebLink 1.15).

**TABLE 1.2 Critical assessment competitions in bioinformatics.**

Name/Acronym	Competition	Chapter
Alignathon	Compare whole-genome alignment methods	6
EGASP	ENCODE Genome Annotation Assessment Project	8
Assemblathon	Compare the performance of genome assemblers	9
GAGE	Genome Assembly Gold-standard Evaluations	9
ABRF	Association of Biomolecular Resource Facilities (ABRF) assessment of phosphorylation	12
CASP	Critical Assessment of Structure Prediction	13
CAFA	Critical Assessment of Protein Function	14
CAGI	Critical Assessment of Genome Interpretation	14

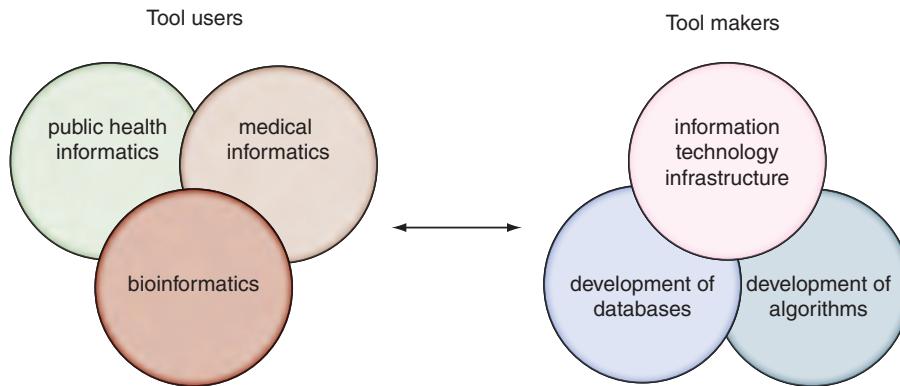
about your expectations for what you can learn. These rules also apply to reading a textbook such as this one.

### Reproducible Research in Bioinformatics

Science by its nature is cumulative and progressive. Whether you use web-based or command-line tools, research should be conducted in a way that is reproducible by the investigator and by others. This facilitates the cumulative, progressive nature of your work. In the realm of bioinformatics this means the following.

- A workflow should be well documented. This may include keeping text documents on your computer in which you can copy and paste complex commands, URLs, or other forms of data. Many people choose to maintain a traditional lab notebook, written by hand, but increasingly this must be accompanied by some form of electronic notebook.
- To facilitate your work, information stored on a computer should be well organized. In Box 2.3 we introduce a paper by Noble (2009), offering guidance on how to organize your files.
- Data should be made available to others. Repositories are available to store high-throughput data in particular. Examples are Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI and ArrayExpress and European Nucleotide Archive (ENA) at EBI.
- Metadata can be equally as crucial as data. Metadata refers to information about datasets. For a bacterial genome that has been sequenced, the metadata may include the location from which the bacterium was isolated, the culture conditions, and whether it is pathogenic. For a study of gene expression in human brain, the metadata may include the post-mortem interval, the gender, the disease phenotype, and the method of RNA isolation. Metadata provide key information for statistical analyses, allowing the investigator to explore the effects of various parameters on the outcome measure.
- Databases that are used should be documented. Since the contents of databases change over time, it is important to document the version number and the date(s) of access.
- Software should be documented. For established packages, the version number should be provided. Further documenting the specific steps you use allows others to independently repeat your analyses. In an effort to share software, many researchers use repositories such as GitHub.

Git is the most popular distributed version control system for software development. It allows scientists to access software having specific versions. Github hosts both open and private projects. It is available online at <https://github.com> (WebLink 1.16). As of early 2015, it has almost 20 million repositories and 8 million users.



**FIGURE 1.6** Tool users and tool makers. The term “informatics” has been applied to an increasing number of disciplines in recent years including bioinformatics, public health informatics, medical informatics, and library informatics. Each of these disciplines is concerned with systematizing and analyzing increasingly large datasets. The focus of bioinformatics and genomics is on proteins, genes, and genomes in particular.

## BIOINFORMATICS AND OTHER INFORMATICS DISCIPLINES

In recent years there has been a proliferation of other informatics fields including medical informatics, health care informatics, nursing informatics, and library informatics (Fig. 1.6). Bioinformatics has some overlap with these disciplines but is distinguished by its emphasis on DNA and other biomolecules. We may also distinguish tool users (e.g., biologists using bioinformatics software to study gene function, or medical informaticists using electronic health records) from tool makers (e.g., those who build databases, create information technology infrastructure, or write computer software). In bioinformatics, more than in other informatics disciplines, the tool users are also increasingly adept at being tool makers.

## ADVICE FOR STUDENTS

The fields of bioinformatics and genomics are extremely broad. You should decide what range of problems you want to study, and what techniques are best suited to tackling those problems. Looking at Figure 1.5, you can see a broad range of available tools and approaches. As we move through the chapters it will likely become clear which is right for you. I encourage you to approach this textbook as actively as possible. When we discuss a website or a software package, take it as an opportunity to explore it in depth.

There are many ways to get help. Try using Biostars, an online forum in which you can post questions, get answers from the community, explore tutorials, and more (Parnell *et al.*, 2011). By the year 2015, over 16,000 registered users have created >125,000 posts. Try joining Biostars or other bioinformatics forums to find others who have questions similar to yours.

Biostars was started in 2009 by Istvan Albert of Penn State University. Visit Biostars at <http://www.biostars.org> (WebLink 1.17).

## SUGGESTED READING

Dudley and Butte (2009) provide an excellent guide to developing effective bioinformatics programming skills (including the use of open source software and Unix). There have been relatively few general overviews of the field of bioinformatics in the past five years, perhaps because of its broadening scope. Thousands of reviews cover specialized topics. For all of Chapters 2–21 I provide sets of recent review articles.

Visit the NAR database issue at  
✉ <http://nar.oxfordjournals.org/>  
(WebLink 1.18).

In 2011 Eric Green, Mark Guyer and colleagues at the National Human Genome Research Institute published the highly recommended article: “Charting a course for genomic medicine from base pairs to bedside” (Green *et al.*, 2011). This paper describes achievements in genomics and prospects for the coming decade.

Each January the journal Nucleic Acids Research offers a Database Issue that describes many central bioinformatics resources (Fernández-Suárez *et al.*, 2014). That journal provides access to a vast number of papers via its website.

## REFERENCES

- Altman, R.B. 1998. Bioinformatics in support of molecular medicine. *Proceedings of AMIA Symposium 1998*, 53–61. PMID: 9929182.
- Altman, R.B., Dugan, J.M. 2003. Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis* **44**, 3–14. PMID: 12647379.
- Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences, USA* **93**(17), 9188–9193. PMID: 8799176.
- Dudley, J.T., Butte, A.J. 2009. A quick guide for developing effective bioinformatics programming skills. *PLoS Computational Biology* **5**(12), e1000589. PMID: 20041221.
- Fernández-Suárez, X.M., Rigden, D.J., Galperin, M.Y. 2014. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Research* **42**(1), D1–6. PMID: 24316579.
- Green, E.D., Guyer, M.S. 2011. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204–213. PMID: 21307933.
- Hugenholtz, P., Pace, N.R. 1996. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology* **14**, 190–197. PMID: 8663938.
- Noble, W.S. 2009. A quick guide to organizing computational biology projects. *PLoS Computational Biology* **5**(7), e1000424. PMID: 19649301.
- Pace, N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740. PMID: 9115194.
- Parnell, L.D., Lindenbaum, P., Shameer, K. *et al.* 2011. BioStar: an online question & answer resource for the bioinformatics community. *PLoS Computational Biology* **7**(10), e1002216. PMID: 22046109.
- Searls, D.B. 2012a. An online bioinformatics curriculum. *PLoS Computational Biology* **8**(9), e1002632. PMID: 23028269.
- Searls, D.B. 2012b. Ten simple rules for online learning. *PLoS Computational Biology* **8**(9), e1002631. PMID: 23028268.
- Searls, D.B. 2014. A new online computational biology curriculum. *PLoS Computational Biology* **10**(6), e1003662. PMID: 24921255.
- Stajich, J.E. 2007. An Introduction to BioPerl. *Methods in Molecular Biology* **406**, 535–548. PMID: 18287711.
- Topol, E.J. 2014. Individualized medicine from prewomb to tomb. *Cell* **157**(1), 241–253. PMID: 24679539.



admodum secernantur, exponam. Res est parvi laboris.  Farina sumitur ex optimo tritico, modice trita, ne cibrum furfures subeant; oportet enim ab his esse quam expurgatissimam, ut omnis miltura tollatur suspicio. Tum aqua purissima permiscetur, ac subigitur. Quod reliquum est operis, lotura absolvit. Aqua enim partes omnes, quascumque potest solvere, secum avehit; alias intactas relinquit.

Porro haec, quas aqua relinquit, concrectata manibus, pressaque sub aqua reliqua, paullatim in massam coguntur mollem, & supra, quam credi potest, tenacem: egregium glutinis genus, & ad opificia multa aptissimum; in quo illud notatum dignum est, quod aqua permisceri se amplius non sinit. Illa aliae, quas aqua secum avehit, aliquandiu innatant, & aquam lacteam reddunt; post paullatim deferuntur ad fundum, & subsidunt; nec admodum inter se coherent; sed quasi pulvis vel levissimo concussum redeunt. Nihil his affinius est amylo; vel potius ipsa verissimum sunt amylum. Atque haec scilicet duo sunt illa partium genera, quae sibi Beccarius proposuit ad chymicum opus faciendum, queque ut suis nominibus distingueret, glutinosum alterum appellare solebat, alterum amylaceum. 

Tanta est autem horum generum diversitas, ut si utrumque vel digestione, vel destillatione resolvias, & principia, unde conuant, chymicorum more, elicias, non ex una ac simplici, sed ex duabus longissimeque inter se diversis rebus producibile videantur; cum enim amyacea pars suum prae se genus ferat, eaque principia ostendat, quae a vegetabili natura duci solent; glutinosa originem quasi detrectat suam, ac se per omnia sic praebet, quasi esset ab animante quopiam profecta. Quod ut melius intelligatur, generatim primum scire convenit, quam dissimiliter vegetabilia atque animalia in digestionibus destillationibusque se praestent.

In digestionibus, quas lenis & diurnus calor facit, animalium partes numquam ad veram absolutamque fermentationem perducuntur; sed putreficiunt tetricime semper. Vegetabilia quasi sua sponte fermentantur, neque putreficiunt, nisi ars adiuvet; eaque inter fermentandum manifesta acoris indicia praebent, quae nulla sunt in animalibus, dum putreficiunt. Fermentatione autem concocta, vinosum aut acetosum liquorem vegetabilia largiuntur; animalia, si putreficiunt, urinose.



*to many uses; and what is especially worthy of note, it cannot any longer be mixed with water. The other particles, which water carries away with itself, for some time float and render the water milky; but after a while they are carried to the bottom and sink; nor in any way do they adhere to each other; but like powder they return upward on the lightest contact. Nothing is more like this than starch, or rather this truly is starch. And these are manifestly the two sorts of bodies which Beccari displayed through having done the work of a chemist and he distinguished them by their names, one being appropriately called glutinous (see open arrowhead) and the other amyaceous.*

In addition to purifying gluten, Beccari identified it as an “animal substance” in contrast to starch, a “vegetable substance,” based on differences on how they decomposed with heat or distillation. A century later Jons Jakob Berzelius proposed the word protein; he also posited that plants form “animal materials” that are eaten by herbivorous animals.

Source: Zanotti (1745).

Chapter 2 introduces ways to access molecular data, including information about DNA and proteins. One of the first scientists to study proteins was Iacopo Bartolomeo Beccari (1682–1776), an Italian philosopher and physician who discovered protein as a component of vegetables. This image is from page 123 of the Bologna Commentaries, written by a secretary on the basis of a 1728 lecture given by Beccari (Zanotti, 1745). Beccari separated gluten (plant proteins) from wheaten flour. The passage beginning *Res est parvi laboris* (“it is a thing of little labor”; see closed arrowhead) is translated as follows (Beach, 1961, p. 362):

*It is a thing of little labor. Flour is taken of the best wheat, moderately ground, the bran not passing though the sieve, for it is necessary that this be fully purged away, so that all traces of a mixture have been removed. Then it is mixed with pure water and kneaded. What is left by this procedure, washing clarifies. Water carries off with itself all it is able to dissolve, the rest remains untouched. After this, what the water leaves is worked with the hands, and pressed upon in the water that has stayed. Slowly it is drawn together in a doughy mass, and beyond what is possible to be believed, tenacious, a remarkable sort of glue, and suited*

# Access to Sequence Data and Related Information

# CHAPTER

# 2

*The body of data available in protein sequences is something fundamentally new in biology and biochemistry, unprecedented in quantity, in concentrated information content and in conceptual simplicity ... For the past four years we have published an annual Atlas of Protein Sequence and Structure, the latest volume of which contains nearly 500 sequences or partial sequences established by several hundred workers in various laboratories.*

—Margaret Dayhoff (1969), p. 87

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

- define the types of molecular databases;
- define accession numbers and the significance of RefSeq identifiers;
- describe the main genome browsers and use them to study features of a genomic region; and
- use resources to study information about both individual genes (or proteins) and large sets of genes/proteins.

## INTRODUCTION TO BIOLOGICAL DATABASES

All living organisms are characterized by the capacity to reproduce and evolve. The genome of an organism is defined as the collection of DNA within that organism, including the set of genes that encode RNA molecules and proteins. In 1995 the complete genome of a free-living organism was sequenced for the first time, the bacterium *Haemophilus influenzae* (Fleischmann *et al.*, 1995; Chapters 15 and 17). In the years since then the genomes of thousands of organisms have been completely sequenced, ushering in a new era of biological data acquisition and information accessibility. Publicly available databases now contain quadrillions ( $>10^{15}$ ) of nucleotides of DNA sequence data, soon to be quintillions ( $>10^{18}$  bases). These have been collected from over 300,000 different species of organisms (Benson *et al.*, 2015). The goal of this chapter is to introduce the databases that store these data and strategies to extract information from them.

There are two main technologies for DNA sequencing (we will discuss these in detail in Chapter 9). Beginning in the 1970s dideoxynucleotide sequencing (“Sanger sequencing”) was the principal method. Since 2005 next-generation sequencing (NGS) technology has emerged, allowing orders of magnitude more sequence data to be generated. The availability of vastly more sequence data (at a relatively low cost per base) has impacted most areas of bioinformatics and genomics. There are new challenges in acquiring,

analyzing, storing, and distributing such data. It is no longer unusual for researchers to analyze datasets that are many terabytes in size.

In this chapter (and in this book) we will introduce two ways of thinking about accessing data. The first is in terms of individual genes, proteins, or related molecules. Taking human beta globin as an example, there is a locus (on chromosome 11) harboring the beta globin gene (*HBB*) and associated genomic elements such as a promoter and introns. There is tremendous variation between people (variants include single-nucleotide variants, differences in repetitive DNA elements, and differences in chromosomal copy number). This gene can be transcribed to beta globin mRNA which is expressed in particular tissues (and particular times of development) and may be translated into beta globin protein. This protein is a subunit of the hemoglobin protein, a tetramer that has various functions in health and diseases. All this information about the beta globin gene, RNA, and protein is accessible through the databases and resources introduced in this chapter.

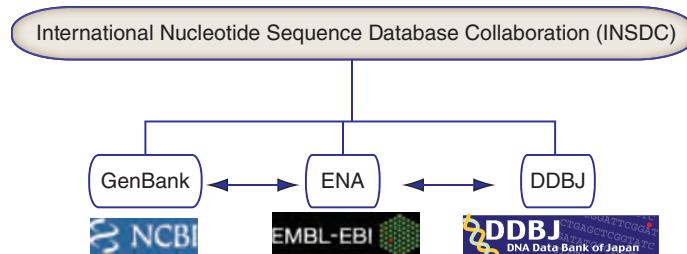
We write gene names using the official gene notation, for example *HBB*. For human genes, this is given by the HUGO Gene Nomenclature Committee (HGNC) (<http://www.genenames.org>) (WebLink 2.1) (Gray *et al.*, 2013). The website also provides guidelines for when to use upper case and italics for human gene symbols. For other species (e.g., yeast, Chapter 18) the conventions vary.

A second perspective is on large datasets related to a problem of interest. Here are three examples:

1. We might want to study all the variants that have been identified across all human globin genes.
2. In patients having mutations in a gene we might want to study the collection of all of the tens of thousands of RNA transcripts in a given cell type in order to assess the functional consequences of that variation. After performing a microarray or RNAseq experiment (see Chapter 11), it might be of interest to identify a set of regulated transcripts and assign their protein products to some cellular pathways.
3. Perhaps we want to sequence the DNA corresponding to a set of 100 genes implicated in hemoglobin function. Databases and resources such as Entrez, BioMart, and Galaxy (introduced below) facilitate the manipulation of larger datasets. You can acquire, store, and analyze datasets involving some set of molecules that have been previously characterized (e.g., all known protein-coding genes on human chromosome 11) or are novel (e.g., data you obtain experimentally that you can annotate and compare to known data).

## CENTRALIZED DATABASES STORE DNA SEQUENCES

How much DNA sequence is stored in public databases? Where are the data stored? We begin with three main sites that have been responsible for storing nucleotide sequence data from 1982 to the present (Fig. 2.1). These are: (1) GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda (NCBI Resource Coordinators, 2014; Benson *et al.*, 2015); (2) the European Molecular Biology Laboratory (EMBL)-Bank Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) in Hinxton, England

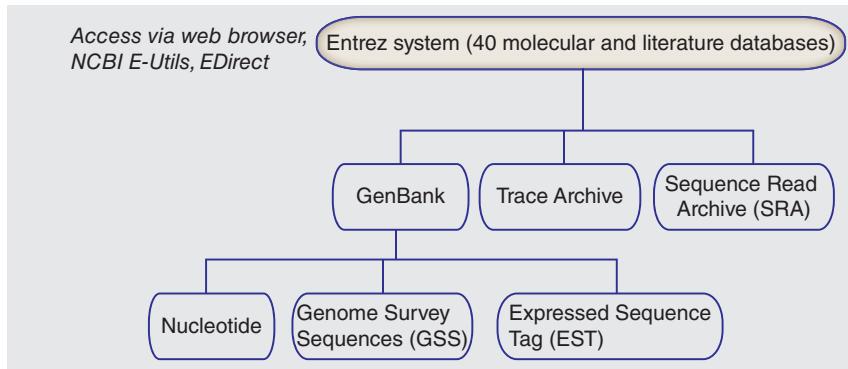


**FIGURE 2.1** The nucleotide collections of GenBank at NCBI, EMBL-Bank at the European Bioinformatics Institute, and DDBJ at the DNA Data Bank of Japan are all coordinated by the International Nucleotide Sequence Database Collaboration (INSDC).

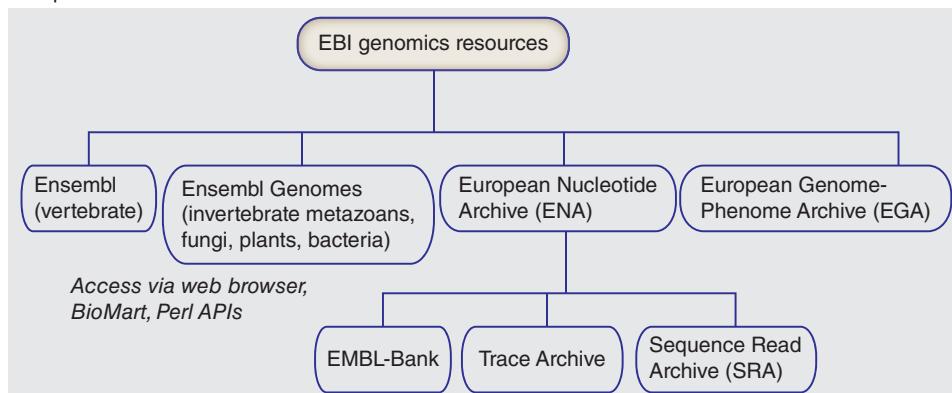
(Pakseresht *et al.*, 2014; Brooksbank *et al.*, 2014); and (3) the DNA Database of Japan (DDBJ) at the National Institute of Genetics in Mishima (Ogasawara *et al.*, 2013; Kosuge *et al.*, 2014). All three are coordinated by the International Nucleotide Sequence Database Collaboration (INSDC) (Nakamura *et al.*, 2013; Fig. 2.1), and they share their data daily. GenBank, EMBL-Bank, and DDBJ are organized as databases within NCBI, EBI, and DDBJ which offer many dozens of other resources for the study of sequence data (see Fig. 2.2).

NCBI is at <http://www.ncbi.nlm.nih.gov/> and GenBank is at <http://www.ncbi.nlm.nih.gov/Genbank>; DDBJ is at <http://www.ddbj.nig.ac.jp/>; and EMBL-Bank is at <http://www.ebi.ac.uk/>. You can visit the INSDC at <http://www.insdc.org/>. You can access these URLs by visiting this book's website (<http://bioinfbook.org>) and using Chapter 2 WebLinks 2.2 to 2.6.

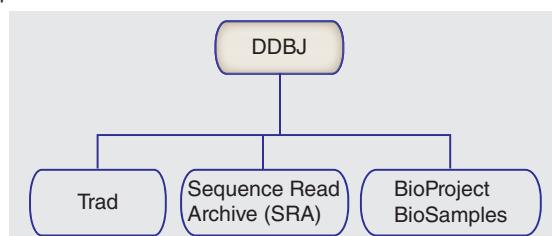
(a) National Center for Biotechnology Information



(b) European Bioinformatics Institute



(c) DNA Database of Japan



**FIGURE 2.2** DNA sequences are shared by three major repositories. (a) The National Center for Biotechnology Information (NCBI) houses GenBank as part of its Entrez system of 40 molecular and literature databases. The Trace Archive stores sequence traces, and the Sequence Read Archive (SRA) stores next-generation sequence data. GenBank includes separate divisions for nucleotides, genome survey sequences, and expressed sequence tags. (b) The European Bioinformatics Institute resources include Ensembl (with a focus on vertebrate genomes), Ensembl Genomes (centralizing data on broader groups of species), the European Nucleotide Archive (ENA), and the European Genome-Phenome Archive (EGA). Within ENA, EMBL-Bank includes the same raw sequence data as GenBank at NCBI. Similar data are also housed in the Trace Archive and SRA. (c) The DNA Database of Japan (DDBJ) also includes a SRA. Its traditional (Trad) division shares the same raw sequence data with GenBank and EMBL-Bank on a daily basis. All these various databases can be accessed by web browsing or via programs such as EDirect (for command-line access to Entrez databases).

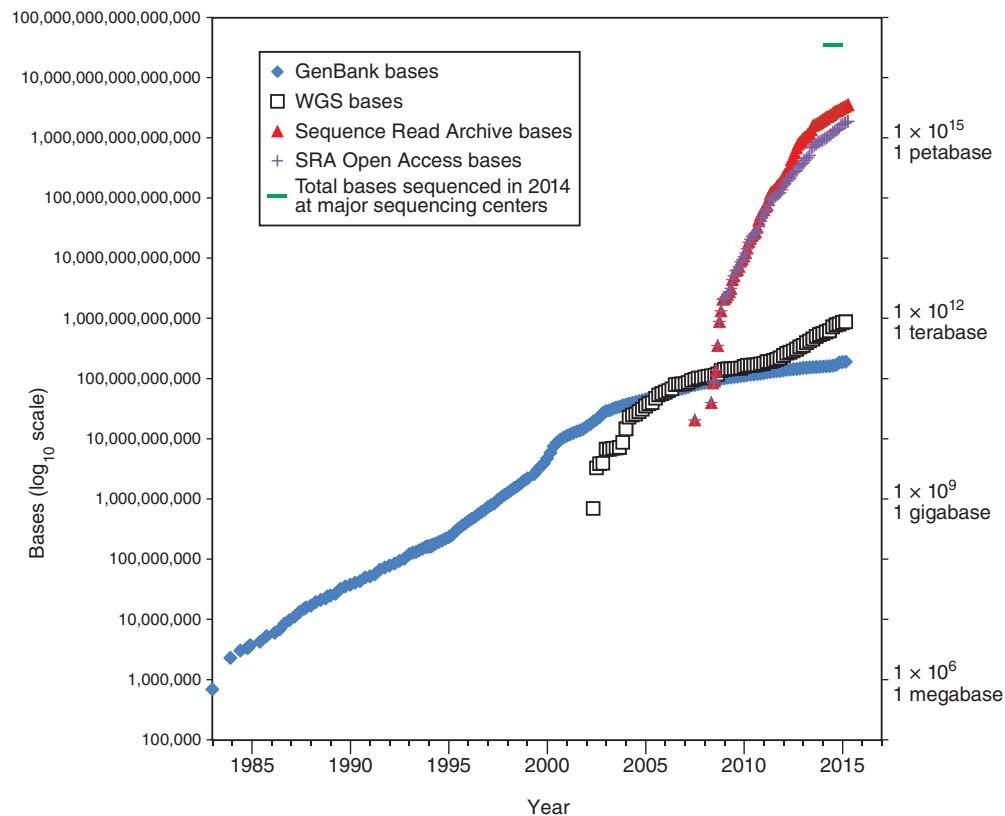
Members of the research community can submit records directly to sequence repositories at NCBI, EBI, and DDBJ. Quality control is assured through guidelines enforced at the time of submission, and through projects such as RefSeq that reconcile differences between submitted entries. For GenBank, NCBI offers the command-line tool `tbl2asn` to automate the creation of sequence records.

The growth of DNA in repositories is shown in **Figure 2.3**. GenBank (representative of the holdings of EMBL-Bank and DDBJ) has received submissions since 1982, including sequences from thousands of individual submitters. Over the past 30 years the number of bases in GenBank has doubled approximately every 18 months.

GenBank, EMBL-Bank, and DDBJ accept sequence data that consist of complete or incomplete genomes (or chromosomes) analyzed by a whole-genome shotgun (WGS) strategy. The WGS division consists of sequences generated by high-throughput sequencing efforts. WGS sequences have been available since 2002, but they are not considered part of the GenBank/EMBLBank/DDBJ releases. As indicated in **Figure 2.3**, the number of base pairs of DNA included among WGS sequences now exceeds the holdings of GenBank.

You can access `tbl2asn` at <http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/> (WebLink 2.7).

In 2015 the number of bases in GenBank reached 188 billion (contained in ~181 million sequences). To see statistics on the growth of EMBL-Bank visit <http://www.ebi.ac.uk/ena/about/statistics> (WebLink 2.8).



**FIGURE 2.3** Growth of DNA sequence in repositories. Data are shown for GenBank (blue diamonds) from release 3 (December 1982) to release 206 (February 2015). Additional DNA sequences from the whole-genome shotgun sequencing projects, begun in 2002, are shown (open black squares). SRA data from NCBI are plotted including total bases (red triangles) and the subset of open-access bases (purple + symbols). Data plotted from the GenBank release notes at <http://www.ncbi.nlm.nih.gov/Genbank> and SRA notes at <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?>. The total number of DNA bases sequenced at major sequencing centers in 2014 is shown (green bar; ~40 petabases). This estimate is extrapolated from the output of the Broad Institute for 2014 which is ~9% of the output of the set of major centers described in **Figure 15.10**. Consideration of additional output from sources such as companies involved in high-throughput sequencing would greatly increase this estimate. According to NCBI, for SRA the  $3.5 \times 10^{15}$  bases in the current release (March 2015) correspond to  $2.3 \times 10^{15}$  bytes of data.

**TABLE 2.1 Scales of DNA base pairs.**

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	Size of a typical coding region of a gene
1,000,000	1 megabase pair	1 Mb	Size of a typical bacterial genome
$10^9$	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
$10^{12}$	1 terabase pair	1 Tb	
$10^{15}$	1 petabase pair	1 Pb	

Inspection of **Figure 2.3** reveals that the recently developed Sequence Read Archive (SRA) contains vastly more sequence data than the sum of GenBank and WGS; in fact, SRA currently holds 3000 times more bases of DNA. Each sequence read in SRA is relatively short (typically 50–400 base pairs), reflecting next-generation sequencing technology (described in Chapter 9). Most of the SRA data are publicly available (such as sequences from various organisms across the tree of life); these are shown as open-access bases in **Figure 2.3**. Some of the data are derived from humans, and can potentially lead to the identification of particular clinical subjects or research participants. Access to those data is therefore restricted, requiring application to a committee from qualified researchers who agree to adhere to ethical guidelines. **Figure 2.3** shows data from SRA at NCBI, including total data and open access data.

To make sense of such large numbers of bases of DNA we can look at several specific examples (**Table 2.1**). The first eukaryotic genome to be completed (*Saccharomyces cerevisiae*; Chapter 19) is ~13 million base pairs (Mb) in size. An average-sized human chromosome is ~150 Mb, and a single human genome consists of >3 billion base pairs (3 Gb). For next-generation sequencing (Chapter 9), short sequence reads (typically 100–300 base pairs in length) are obtained in vast quantities that allow each single base pair to be represented (“covered”) by some average number of independent reads such as 30. There is therefore  $30 \times$  depth of coverage. For a recent study in my lab, we obtained paired affected/unaffected samples from three individuals with a disease, performed whole-genome sequencing, and obtained 700 billion ( $7 \times 10^{11}$ ) bases of DNA sequence. For a large-scale cancer study involving 20,000 tumor/normal comparisons, a massive  $10^{16}$  bases of DNA can be generated. Even larger studies involving 200,000 tumor/normal comparisons are being planned. Other experimental approaches such as whole-exome sequencing (involving sequencing the collection of exons in a genome that are thought to be functionally most important) and sequencing of the transcriptome (RNASeq; Chapter 11) also generate large amounts of data.

We can also consider amounts of sequence data in terms of terabytes. A byte is a unit of computer storage information, consisting of 8 bits and encoding one character. **Table 2.2** shows

We will discuss WGS in Chapter 15. To learn more about it, visit <http://www.ncbi.nlm.nih.gov/genbank/wgs> (WebLink 2.9). By February 2015 there were ~873 billion bases in WGS at NCBI (release 206).

In addition to SRA, next-generation sequence data are stored and can be obtained from the European Nucleotide Archive (ENA, <http://www.ebi.ac.uk/ena>, WebLink 2.10) and the DDBJ Read Archive (DRA, [http://trace.ddbj.nig.ac.jp/dra/index\\_e.html](http://trace.ddbj.nig.ac.jp/dra/index_e.html), WebLink 2.11).

**TABLE 2.2 Range of files sizes and typical examples.**

Size	Abbreviation	No. bytes	Examples
Bytes	–	1	1 byte is typically 8 bits, used to encode a single character of text
Kilobytes	1 kb	$10^3$	Size of a text file with up to 1000 characters
Megabytes	1 MB	$10^6$	Size of a text file with 1 million characters
Gigabytes	1 GB	$10^9$	600 GB: size of GenBank (uncompressed flat files) <a href="ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt">ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt</a> (WebLink 2.84)
Terabytes	1 TB	$10^{12}$	385 TB: United States Library of Congress web archive ( <a href="http://www.loc.gov/webarchiving/faq.html">http://www.loc.gov/webarchiving/faq.html</a> ) (WebLink 2.85) 464 TB: Data generated by the 1000 Genomes Project ( <a href="http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project">http://www.1000genomes.org/faq/how-much-disk-space-used-1000-genomes-project</a> ) (WebLink 2.86)

(Continued)

**TABLE 2.2** (continued)

Size	Abbreviation	No. bytes	Examples
Petabytes	1 PB	$10^{15}$	1 PB: size of dataset available from The Cancer Genome Atlas (TCGA) 5 PB: size of SRA data available for download from NCBI 15 PB: amount of data produced each year at the physics facility CERN (near Geneva) ( <a href="http://home.web.cern.ch/about/computing">http://home.web.cern.ch/about/computing</a> ) (WebLink 2.87)
Exabytes	1 EB	$10^{18}$	2.5 exabytes of data are produced worldwide (Lampitt, 2014)

A megabase is one million ( $10^6$ ) bases of DNA. A gigabase is one billion ( $10^9$ ) bases. A terabase is one trillion ( $10^{12}$ ) bases.

some typical sizes for various files and projects. A typical desktop might have 500 gigabytes (Gb) of storage. The uncompressed flatfiles of the current release of GenBank (introduced below) are ~600 Gb. One thousand gigabytes is equivalent to one terabyte (1000 Gb = 1 Tb), which is the amount of storage some researchers use to study a single whole human genome. One thousand terabytes is equivalent to one petabyte (1000 Tb = 1 Pb). Large-scale sequencing projects, for example one that involves whole-genome sequences of 10,000 individuals, require several Pb of storage.

## CONTENTS OF DNA, RNA, AND PROTEIN DATABASES

While the sequence information underlying DDBJ, EMBL-Bank, and GenBank are equivalent, we begin our discussion with GenBank. GenBank is a database consisting of most known public DNA and protein sequences (Benson *et al.*, 2015), excluding next-generation sequence data. In addition to storing these sequences, GenBank contains bibliographic and biological annotation. Its data are available free of charge from NCBI.

### Organisms in GenBank/EMBL-Bank/DDBJ

Over 310,000 different species are represented in GenBank, with over 1000 new species added per month (Benson *et al.*, 2015). The number of organisms represented in GenBank is shown in **Table 2.3**. We define the bacteria, archaea, and eukaryotes in detail in Chapters 15–19. Briefly, eukaryotes have a nucleus and are often multicellular, while bacteria do not have a nucleus. Archaea are single-celled organisms, distinct from eukaryotes and bacteria, and constitute a third major branch of life. Viruses, which contain nucleic acids (DNA or RNA) but can only replicate in a host cell, exist at the borderline of the definition of living organisms.

**TABLE 2.3** Taxa represented in GenBank.

Ranks	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	143	140	525	0	808
Bacteria	1,370	2,611	13,331	819	18,131
Eukaryota	20,443	67,606	297,207	22,608	407,864
Fungi	1,550	4,620	29,450	1,128	36,748
Metazoa	14,670	45,517	145,044	11,428	216,659
Viriplantae	2,622	14,680	113,529	9,789	140,620
Viruses	618	442	2,349	0	3,409
All taxa	22,603	70,806	313,443	23,427	430,279

Source: GenBank, NCBI, <http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi>.

**TABLE 2.4** Ten most sequenced organisms in GenBank.

Entries	Bases	Species	Common name
20,614,460	17,575,474,103	<i>Homo sapiens</i>	Human
9,724,856	9,993,232,725	<i>Mus musculus</i>	Mouse
2,193,460	6,525,559,108	<i>Rattus norvegicus</i>	Rat
2,203,159	5,391,699,711	<i>Bos taurus</i>	Cow
3,967,977	5,079,812,801	<i>Zea mays</i>	Maize
3,296,476	4,894,315,374	<i>Sus scrofa</i>	Pig
1,727,319	3,128,000,237	<i>Danio rerio</i>	Zebrafish
1,796,154	1,925,428,081	<i>Triticum aestivum</i>	Bread wheat
744,380	1,764,995,265	<i>Solanum lycopersicum</i>	Tomato
1,332,169	1,617,554,059	<i>Hordeum vulgare</i> subsp. <i>vulgare</i>	Barley

Source: GenBank, NCBI, <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt> (GenBank release 194.0).

We have seen so far that GenBank is very large and growing rapidly. From **Table 2.3** we see that the organisms in GenBank consist mostly of eukaryotes. Of the microbes, there are about 25 times more bacterial than archaeal species represented in GenBank.

The number of entries and bases of DNA/RNA for the 10 most sequenced organisms in GenBank is provided in **Table 2.4** (excluding chloroplast and mitochondrial sequences). This list includes some of the most common model organisms that are studied in biology. Notably, the scientific community is studying a series of mammals (e.g., human, mouse, cow), other vertebrates (chicken, frog), and plants (corn, rice, bread wheat, wine grape). Different species are useful for a variety of different studies. Bacteria, archaea, fungi, and viruses are absent from the list in **Table 2.4** because they have relatively small genomes.

To help organize the available information, each sequence name in a GenBank record is followed by its data file division and primary accession number. (We will define accession numbers below.) The following codes are used to designate the data file divisions:

1. PRI: primate sequences
2. ROD: rodent sequences
3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences
10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: expressed sequence tags
13. PAT: patent sequences
14. STS: sequence-tagged sites
15. GSS: genome survey sequences
16. HTG: high-throughput genomic sequences
17. HTC: high-throughput cDNA sequences
18. ENV: environmental sampling sequences
19. CON: constricted sequences
20. TSA: transcriptome shotgun assembly sequences.

We will discuss how genomes of various organisms are selected for complete sequencing in Chapter 15.

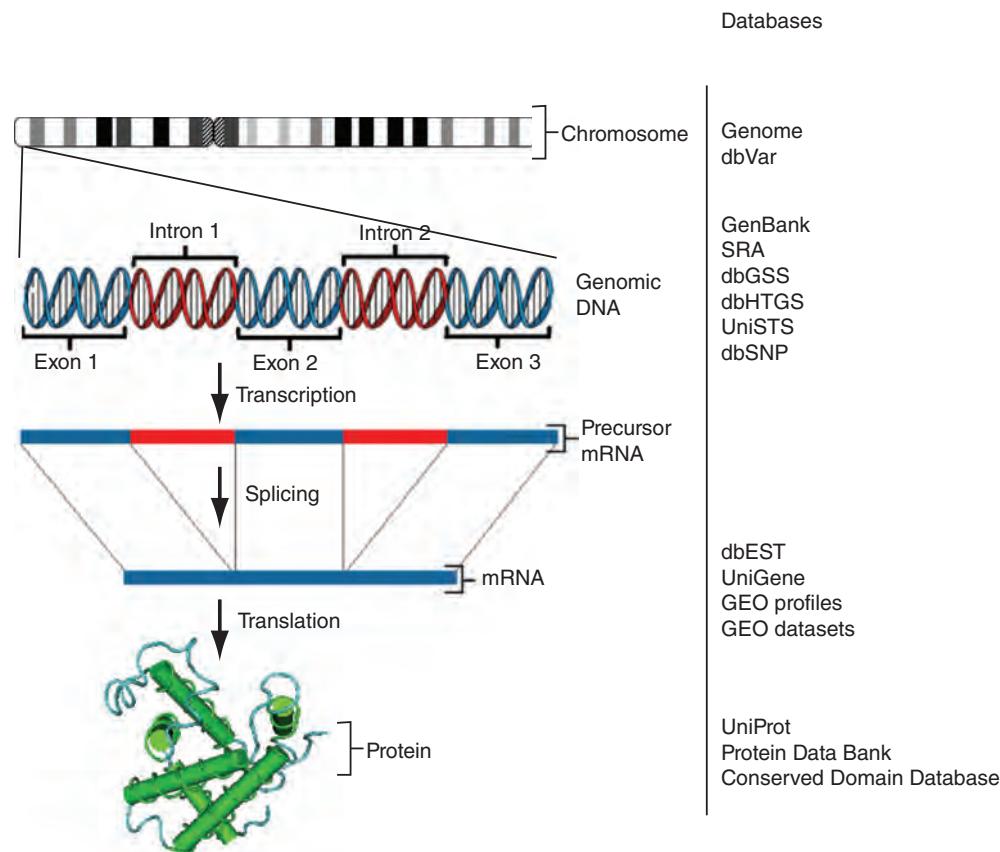
The International Human Genome Sequencing Consortium adopted the Bermuda Principles in 1996, calling for the rapid release of raw genomic sequence data. You can read about recent versions of these principles at <http://www.genome.gov/10506376> (WebLink 2.12).

## Types of Data in GenBank/EMBL-Bank/DDBJ

There are enormous numbers of molecular sequences in the DDBJ, EMBL-Bank, and GenBank databases. We will next look at some of the basic kinds of data present in GenBank. We then address strategies to extract the data you want from GenBank.

We start with an example. We want to find out the sequence of human beta globin. A fundamental distinction is that both DNA, RNA-based, and protein sequences are stored in discrete databases. Furthermore, within each database sequence data are represented in a variety of forms. For example, beta globin may be described at the DNA level (e.g., as a gene), at the RNA level (as a messenger RNA or mRNA transcript), and at the protein level (see Fig. 2.4). Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety of databases contain cDNA sequences corresponding to RNA transcripts.

Beginning with the DNA, a first task is to learn the official name and symbol of a gene (and its gene products, including the protein). Beta globin has the official name of “hemoglobin, beta” and the symbol *HBB*. (From one point of view there is no such thing as a “hemoglobin gene” because globin genes encode globin proteins, and the combination of these globins with heme forms the various types of hemoglobin. Perhaps “globin, beta” might be a more appropriate official name.) For humans and many other species, the RNA or cDNA is generally given the same name (e.g., *HBB*), while the protein name may



**FIGURE 2.4** The types of data stored in various databases (right column) can be conceptualized in terms of the central dogma of biology in which genomic DNA (organized in chromosomes; top rows) includes protein-coding genes that are transcribed to precursor messenger RNA (mRNA), processed to mature mRNA, and translated to protein. The protein structure is from accession 1HBS (see Cn3D software, Chapter 13). To learn more about these various databases, search the alphabetical list of resources from the NCBI homepage.

Source: NCBI (<http://www.ncbi.nlm.nih.gov/>).

differ and is not italicized. Often, multiple investigators study the same gene or protein and assign different names. The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins.

For our example of beta globin, the various forms are described in the following sections.

See <http://www.genenames.org> (WebLink 2.1).

## Genomic DNA Databases

A gene is localized to a chromosome. The gene is the functional unit of heredity (further defined in Chapter 8) and is a DNA sequence that typically consists of regulatory regions, protein-coding exons, and introns. Often, human genes are 10–100 kb in size. In the case of human *HBB* this gene is situated on chromosome 11 (see Chapter 8 on the eukaryotic chromosome). The beta globin gene may be part of a large fragment of DNA such as a cosmid, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC) that may contain several genes. A BAC is a large segment of DNA (typically up to 200,000 base pairs or 200 kb) that is cloned into bacteria. Similarly, YACs are used to clone large amount of DNA into yeast. BACs and YACs are useful vectors with which to sequence large portions of genomes.

Human chromosome 11, which is a mid-sized chromosome, contains about 1800 genes and is about  $134 \times 10^6$  base pairs (134 Mb) in length.

### DNA-Level Data: Sequence-Tagged Sites (STSs)

The Probe database at NCBI includes STSs, which are short (typically 500 base pairs long) genomic landmark sequences for which both DNA sequence data and mapping data are available (Olson *et al.*, 1989). STSs have been obtained from several hundred organisms, including primates and rodents. Because they are sometimes polymorphic, containing short sequence repeats (Chapter 8), STSs can be useful for mapping studies.

Visit the Probe database at <http://www.ncbi.nlm.nih.gov/probe> (WebLink 2.13). Search for STSs within this database with the qualifier “unists”[Properties]. As of February 2015 there are 300,000 human STSs.

### DNA-Level Data: Genome Survey Sequences (GSSs)

All searches of the NCBI Nucleotide database provide results that are divided into three sections: GSS, ESTs, and “CoreNucleotide” (i.e., the remaining nucleotide sequences; Fig. 2.2a). The GSS division of GenBank consists of sequences that are genomic in origin (in contrast to entries in the EST division which are derived from cDNA [mRNA]). The GSS division contains the following types of data (see Chapters 8 and 15):

- random “single-pass read” genome survey sequences;
- cosmid/BAC/YAC end sequences;
- exon-trapped genomic sequences; or
- the *Alu* polymerase chain reaction (PCR) sequences.

There are currently 38 million GSS entries from over 1000 organisms (February 2015). The top four organisms account for about one-third of all entries (these are the mouse *Mus musculus*, a marine metagenome collection, the maize *Zea mays*, and human). This database is accessed via <http://www.ncbi.nlm.nih.gov/nucgss> (WebLink 2.14).

### DNA-Level Data: High-Throughput Genomic Sequence (HTGS)

The HTGS division was created to make “unfinished” genomic sequence data rapidly available to the scientific community. It was set up from a coordinated effort between the three international nucleotide sequence databases: DDBJ, EMBL, and GenBank. The HTGS division contains unfinished DNA sequences generated by the high-throughput sequencing centers.

## RNA data

We have described some of the basic kinds of DNA sequence data in GenBank, EMBL-Bank, and DDBJ. We next consider RNA-level data.

The HTGS home page is <http://www.ncbi.nlm.nih.gov/HTGS/> (WebLink 2.15) and its sequences can be searched via BLAST (see Chapters 4 and 5).

### RNA-Level Data: cDNA Databases Corresponding to Expressed Genes

Protein-coding genes, pseudogenes, and noncoding genes are all transcribed from DNA to RNA (see Chapters 8 and 10). Genes are expressed from particular regions of the

In DNA databases, the convention is to use the four DNA nucleotides (guanine, adenine, thymidine, cytosine; G, A, T, C) when referring to DNA derived from RNA. The RNA base uridine (U) corresponding to T is not used.

In February 2015 GenBank had about 76,000,000 ESTs. We will discuss ESTs further in Chapter 10.

To find the entry for beta globin, go to <http://www.ncbi.nlm.nih.gov>, select All Databases then click UniGene, select human, then enter beta globin or HBB. The UniGene accession number is Hs.523443; note that Hs refers to *Homo sapiens*. The HBB entry in UniGene is at <http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?UGID=914190&TAXID=9606&SEARCH=%20globin> (WebLink 2.16). To see the DNA sequence of a typical EST, click on an EST accession number from the UniGene page (e.g., AA970968.1), then follow the link to the GenBank entry in NCBI Nucleotide (<http://www.ncbi.nlm.nih.gov/nucleotide/3146258>; WebLink 2.17).

body and times of development. If one obtains a tissue such as liver, purifies RNA, then converts the RNA to the more stable form of complementary DNA (cDNA), some of the cDNA clones contained in that cDNA are likely to encode beta globin. Beta globin RNA is therefore represented in databases as an expressed sequence tag (EST), that is, a cDNA sequence derived from a particular cDNA library.

#### **RNA-Level Data: Expressed Sequence Tags (ESTs)**

The database of expressed sequence tags (dbEST) is a division of GenBank that contains sequence data and other information on “single-pass” cDNA sequences from a number of organisms (Boguski *et al.*, 1993). An EST is a partial DNA sequence of a cDNA clone. All cDNA clones, and therefore all ESTs, are derived from some specific RNA source such as human brain or rat liver. The RNA is converted into a more stable form, cDNA, which may then be packaged into a cDNA library (refer to Fig. 2.4). Typically ESTs are randomly selected cDNA clones that are sequenced on one strand (and therefore may have a relatively high sequencing error rate). ESTs are often 300–800 base pairs in length. The earliest efforts to sequence ESTs resulted in the identification of many hundreds of genes that were novel at the time (Adams *et al.*, 1991).

Currently, GenBank divides ESTs into three major categories: human, mouse, and other. Table 2.5 shows the 10 organisms from which the greatest number of ESTs has been sequenced. Assuming that there are 20,300 human protein-coding genes (see Chapter 20) and given that there are about 8.7 million human ESTs, there is currently an average of over 400 ESTs corresponding to each human protein-coding gene.

#### **RNA-Level Data: UniGene**

The goal of the UniGene (unique gene) project is to create gene-oriented clusters by automatically partitioning ESTs into nonredundant sets. Ultimately there should be one UniGene cluster assigned to each gene of an organism. There may be as few as one EST in a cluster, reflecting a gene that is rarely expressed, to tens of thousands of ESTs associated with a highly expressed gene. We discuss UniGene clusters further in Chapter 10 (on gene expression). The 19 phyla containing 142 organisms currently represented in UniGene are listed in Table 2.6.

For human beta globin, there is only a single UniGene entry. This entry currently has ~2400 human ESTs that match the beta globin gene. This large number of ESTs reflects how abundantly the beta globin gene has been expressed in cDNA libraries that have

**TABLE 2.5 Top ten organisms for which ESTs have been sequenced. Many thousands of cDNA libraries have been generated from a variety of organisms, and the total number of public entries is currently over 41 million.**

Organism	Common name	Number of ESTs
<i>Homo sapiens</i>	Human	8,704,790
<i>Mus musculus + domesticus</i>	Mouse	4,853,570
<i>Zea mays</i>	Maize	2,019,137
<i>Sus scrofa</i>	Pig	1,669,337
<i>Bos taurus</i>	Cattle	1,559,495
<i>Arabidopsis thaliana</i>	Thale Cress	1,529,700
<i>Danio rerio</i>	Zebrafish	1,488,275
<i>Glycine max</i>	Soybean	1,461,722
<i>Triticum aestivum</i>	Wheat	1,286,372
<i>Xenopus (Silurana) tropicalis</i>	Western clawed frog	1,271,480

Source: NCBI, [http://www.ncbi.nlm.nih.gov/dbEST/dbEST\\_summary.html](http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) (dbEST release 130101).

**TABLE 2.6 19 Phyla and 142 organisms represented in UniGene.**

Phylum	Number of species	Example
Chordata	42	<i>Equus caballus</i> (horse)
Echinodermata	2	<i>Strongylocentrotus purpuratus</i> (purple sea urchin)
Arthropoda	19	<i>Apis mellifera</i> (honey bee)
Mollusca	2	<i>Aplysia californica</i> (California sea hare)
Annelida	2	<i>Alvinella pompejana</i>
Nematoda	2	<i>Caenorhabditis elegans</i> (nematode)
Platyhelminthes	3	<i>Schistosoma mansoni</i>
Porifera	1	<i>Amphimedon queenslandica</i>
Cnidaria	3	<i>Nematostella vectensis</i> (starlet sea anemone)
Ascomycota	5	<i>Neurospora crassa</i>
Basidiomycota	1	<i>Filobasidiella neoformans</i>
Codonosigidae	1	<i>Monosiga ovata</i>
Streptophyta	50	<i>Zea mays</i> (maize)
Chlorophyta	2	<i>Chlamydomonas reinhardtii</i>
Apicomplexa	1	<i>Toxoplasma gondii</i>
Bacillariophyta	1	<i>Phaeodactylum tricornutum</i>
Oomycetes	2	<i>Phytophthora infestans</i> (potato late blight agent)
Dictyosteliida	1	<i>Dictyostelium discoideum</i> (slime mold)
Ciliophora	2	<i>Paramecium tetraurelia</i>

Source: UniGene, NCBI (accessed April 2013).

been sequenced. A UniGene cluster is a database entry for a gene containing a group of corresponding ESTs (**Fig. 2.5**).

There are now thought to be approximately 20,300 human protein-coding genes (see Chapter 20). One might expect an equal number of UniGene clusters. However, there are far more human UniGene clusters (currently 130,000) than there are genes. This discrepancy could occur for three reasons.

1. Much of the genome is transcribed at low levels (see the description of the ENCODE project in Chapters 8 and 10). Currently (UniGene build 235), 64,000 human UniGene clusters consist of a single EST and ~100,000 UniGene clusters consist of just 1–4 ESTs. These could reflect rare transcription events of unknown biological relevance.
2. Some DNA may be transcribed during the creation of a cDNA library without corresponding to an authentic transcript; it is therefore a cloning artifact. We discuss the criteria for defining a eukaryotic gene in Chapter 8. Alternative splicing (Chapter 10) may introduce apparently new clusters of genes because the spliced exon has no homology to the rest of the sequence.
3. Clusters of ESTs could correspond to distinct regions of one gene. In that case there would be two (or more) UniGene entries corresponding to a single gene (see **Fig. 2.5**). As a genome sequence becomes finished, it may become apparent that the two UniGene clusters should properly cluster into one. The number of UniGene clusters may therefore collapse over time.

We are using beta globin as a specific example. If you want to type “globin” as a query, you will simply get more results from any database; in UniGene, you will find almost 200 entries corresponding to a variety of globin genes in various species.

## Access to Information: Protein Databases

In many cases you are interested in obtaining protein sequences. The Protein database at NCBI consists of translated coding regions from GenBank as well as sequences from external databases such as UniProt (UniProt Consortium 2012), The Protein Information

The UniGene project has become extremely important in the effort to identify protein-coding genes in newly sequenced genomes. We discuss this in Chapter 15.

(a)

UGID:914190 UniGene Hs.523443 *Homosapiens* (human) HBB Order cDNA clone, Links

### Hemoglobin, beta (HBB)

Human protein-coding gene HBB. Represented by 2363 ESTs from 234 cDNA libraries. Corresponds to reference sequence NM\_000518.4. [UniGene 914190 - Hs.523443]

#### SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

	Best Hits and Hits from model organisms	Species	Id(%)	Len(aa)
<a href="#">XP_508242.1</a>	PREDICTED: hemoglobin subunit beta isoform 2	<i>P. troglodytes</i>	100.0	146
<a href="#">NP_000509.1</a>	HBB gene product	<i>H. sapiens</i>	100.0	146
<a href="#">NP_001188320.1</a>	hemoglobin subunit beta-1-like	<i>M. musculus</i>	83.7	146
<a href="#">NP_001091375.1</a>	uncharacterized protein LOC100037217	<i>X. laevis</i>	61.9	146
<a href="#">NP_571095.1</a>	ba1 gene product	<i>D. rerio</i>	52.7	147
Other hits (2 of 21) [Show all]		Species	Id(%)	Len(aa)
<a href="#">NP_001157900.1</a>	HBB gene product	<i>M. mulatta</i>	95.9	146
<a href="#">NP_001162318.1</a>	HBB gene product	<i>P. anubis</i>	95.2	146

#### GENE EXPRESSION

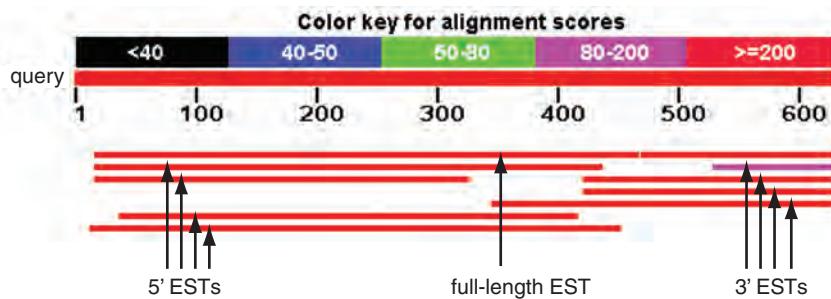
Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

**EST Profile:** Approximate expression patterns inferred from EST sources. [Show more entries with profiles like this]

**GEO Profiles:** Experimental gene expression data (Gene Expression Omnibus).

**cDNA Sources:** blood; mixed; muscle; placenta; bone marrow; lung; brain; spleen; pancreas; connective tissue; pharynx; eye; ovary; uterus; liver; bone; heart; prostate; mammary gland; kidney; uncharacterized tissue; skin; adipose tissue; intestine; stomach; umbilical cord; adrenal gland; nerve; vascular; thymus; testis; embryonic tissue; pituitary gland; parathyroid; ganglia; thyroid; lymph node; pineal gland; ear

(b)



**FIGURE 2.5** The UniGene database includes clusters of expressed sequence tags (ESTs) from human and a large variety of other eukaryotes. (a) The UniGene entry for human HBB indicates that 2363 ESTs have been identified from 234 different cDNA libraries. UniGene reports selected protein similarities, and summarizes gene expression including profiles of regional and temporal expression of HBB. (b) ESTs are mapped to a particular gene and to each other. The number of ESTs that constitute a UniGene cluster ranges from 1 to over 1000; on average there are 100 ESTs per cluster. Sometimes, separate UniGene clusters correspond to distinct regions of a gene (particularly for large genes). Here human beta globin (HBB) mRNA (NM\_000518.4) was used as a query with BLAST (Chapter 4) and searched against nine ESTs selected from among >2000 available ESTs. Four of them are 5' ESTs, four are 3' ESTs (including a poly(A)+ tail), and one is a full-length EST. The accession numbers are AA985606.1, AA910627.1, AI089557.1, AI150946.1, R25417.1, R27238.1, R27242.1, R27252.1, R31622.1, R32259.1.

EBI offers access to over a dozen different protein databases, listed at  
<http://www.ebi.ac.uk/services/proteins/> (WebLink 2.18).

Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), and the Protein Data Bank (PDB) (Rose *et al.*, 2013). The EBI similarly provides information on proteins via these major databases. We will next explore ways to obtain protein data through UniProt, an authoritative and comprehensive protein database.

### *UniProt*

The Universal Protein Resource (UniProt) is the most comprehensive, centralized protein sequence catalog (Magrane and UniProt Consortium, 2011). Formed as a collaborative effort in 2002, it consists of a combination of three key databases:

1. Swiss-Prot is considered the best-annotated protein database, with descriptions of protein structure and function added by expert curators.
2. The translated EMBL (TrEMBL) Nucleotide Sequence Database Library provides automated (rather than manual) annotations of proteins not in Swiss-Prot. It was created because of the vast number of protein sequences that have become available through genome sequencing projects.
3. PIR maintains the Protein Sequence Database, another protein database curated by experts.

UniProt is organized in three database layers.

1. The UniProt Knowledgebase (UniProtKB) is the central database that is divided into the manually annotated UniProtKB/Swiss-Prot and the computationally annotated UniProtKB/TrEMBL.
2. The UniProt Reference Clusters (UniRef) offer nonredundant reference clusters based on UniProtKB. UniRef clusters are available with members sharing at least 50%, 90%, or 100% identity.
3. The UniProt Archive, UniParc, consists of a stable, nonredundant archive of protein sequences from a wide variety of sources (including model organism databases, patent offices, RefSeq, and Ensembl).

You can access UniProt directly from its website, or from EBI or ExPASy. A search for beta globin yields dozens of results. At present RefSeq accessions are not displayed, so for a given query it may be unclear which sequence is the prototype.

## CENTRAL BIOINFORMATICS RESOURCES: NCBI AND EBI

We have looked at the amount of DNA in centralized databases, and the types of DNA, RNA, and protein entries. We next visit two of the main centralized bioinformatics hubs: the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). The relation of DNA repositories in NCBI, EBI, and DDBJ is outlined in Figure 2.2.

### Introduction to NCBI

The NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information (Sayers *et al.*, 2012; NCBI Resource Coordinators, 2014). Prominent resources include the following:

- PubMed is the search service from the National Library of Medicine (NLM) that provides access to over 24 million citations in MEDLINE (Medical Literature, Analysis, and Retrieval System Online) and other related databases, with links to participating online journals.
- Entrez integrates the scientific literature, DNA, and protein sequence databases, three-dimensional protein structure data, population study datasets, and assemblies of complete genomes into a tightly coupled system. PubMed is the literature component of Entrez. For tips on searching Entrez databases see Box 2.1.
- BLAST (Basic Local Alignment Search Tool) is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases (Altschul *et al.*, 1990, 1997). BLAST is a set of similarity search programs designed to explore all of

The European Bioinformatics Institute (EBI) in Hinxton and the Swiss Institute of Bioinformatics (SIB) in Geneva created Swiss-Prot and TrEMBL. PIR is a division of the National Biomedical Research Foundation (<http://pir.georgetown.edu/>, WebLink 2.19) in Washington, DC. PIR was founded by Margaret Dayhoff, whose work is described in Chapter 3. The UniProt web site is <http://www.uniprot.org> (WebLink 2.20).

To access UniProt from EBI, visit <http://www.ebi.ac.uk/uniprot/> (WebLink 2.21). To access UniProt from the major proteomics resource ExPASy, visit [http://web.expasy.org/docs/swiss-prot\\_guideline.html](http://web.expasy.org/docs/swiss-prot_guideline.html) (WebLink 2.22). For release 2014\_09 (September 2014) UniProtKB contains 84 million sequence entries, comprising ~27 billion amino acids. Additional statistics are available at <ftp://ftp.uniprot.org/pub/databases/uniprot/relnotes.txt> (WebLink 2.23).

Extremely useful tutorials are available for Entrez, PubMed, and other NCBI resources at an NCBI education site (<http://www.ncbi.nlm.nih.gov/Education/>) (WebLink 2.24) as well as the PubMed home page (<http://www.ncbi.nlm.nih.gov/pubmed>, WebLink 2.25). You can also access this from the education link on the NCBI home page (<http://www.ncbi.nlm.nih.gov>).

## BOX 2.1 TIPS FOR USING ENTREZ DATABASES

- The Boolean operators AND, OR, and NOT must be capitalized. By default, AND is assumed to connect two terms; subject terms are automatically combined.
- Perform a search of a specific phrase by adding quotation marks. This may potentially restrict the output, so it is a good idea to repeat a search with and without quotation marks.
- Boolean operators are processed from left to right. If you add parentheses, the enclosed terms will be processed as a unit rather than sequentially. A search of NCBI Gene with the query “globin AND promoter OR enhancer” yields 31,000 results; however, by adding parentheses, the query “globin AND (promoter OR enhancer)” yields just 66 results.
- If interested in obtaining results from a particular organism (or from any taxonomic group such as the primates or viruses), try beginning with TaxBrowser to select the organism first. Adding the search term human[ORGN] will restrict the output to human. Alternatively, you can use the taxonomy identifier for human, 9606: txid9606[Organism:exp]
- A variety of limiters can be added. In NCBI Protein, the search 500000:999999[Molecular weight] will return proteins having a molecular weight from 500,000 to 1 million daltons. To view proteins between 10,000 and 50,000 daltons that I have worked on, enter 010000:050000[Molecular weight] pevsner j (or, equivalently, 010000[MOLWT] : 050000[MOLWT] AND pevsner j[Author]).
- By truncating a query with an asterisk, you can search for all records that begin with a particular text string. For example, a search of NCBI Nucleotide with the query “globin” returns 6777 results; querying with “glob\*” returns 490,358 results. These include entries with the species *Chaetomium globosum* or the word global.
- Keep in mind that any Entrez query can be applied to a BLAST search to restrict its output (Chapter 4).

the available sequence databases, regardless of whether the query is protein or DNA. We explore BLAST in Chapters 3–5.

- Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. It was created by Victor McKusick and his colleagues and developed for the World Wide Web by NCBI (Amberger *et al.*, 2011). The database contains detailed reference information. It also contains links to PubMed articles and sequence information. We describe OMIM in Chapter 21 (on human disease).
- Books: NCBI offers about 200 books online. These books are searchable, and are linked to PubMed. See recommended reading (at the end of this chapter) for several relevant bioinformatics titles.
- Taxonomy: the NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms (archaea, bacteria, eukaryota, and viruses) (Fig 2.6). The site features taxonomy information such as genetic codes and taxonomy resources and additional information such as molecular data on extinct organisms and recent changes to classification schemes. We visit this site in Chapters 7 (on evolution) and 15–19 (on genomes and the tree of life).
- Structure: the NCBI structure site maintains the Molecular Modelling Database (MMDB), a database of macromolecular three-dimensional structures, as well as tools for their visualization and comparative analysis. MMDB contains experimentally determined biopolymer structures obtained from the Protein Data Bank (PDB). Structure resources at NCBI include PDBeast (a taxonomy site within MMDB), Cn3D (a three-dimensional structure viewer), and a vector alignment search tool (VAST) which allows comparison of structures (see Chapter 13 on protein structure.)

The Protein Data Bank (<http://www.rcsb.org/pdb/>, WebLink 2.26) is the single worldwide repository for the processing and distribution of biological macromolecular structure data. We explore PDB in Chapter 13.

## The European Bioinformatics Institute (EBI)

The EBI website is comparable to NCBI in its scope and mission, and it represents a complementary, independent resource. EBI features six core molecular databases (Brooksbank *et al.*, 2014): (1) EMBL-Bank is the repository of DNA and RNA sequences that is complementary to GenBank and DDBJ (Brooksbank *et al.*, 2014); (2) Swiss-Prot and (3) TrEMBL are two protein databases that are further described in Chapter 12; (4) MSD is a protein structure database (see Chapter 13); (5) Ensembl is one

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there is a search bar with "Homo sapiens" entered, options to search as a complete name or lock the results, and buttons for "Go" and "Clear". Below the search bar, it says "Display 0 levels using filter: none". The main content area is titled "Homo sapiens" and contains the following information:

- Taxonomy ID: 9606
- Genbank common name: **human**
- Inherited blast name: **primates**
- Rank: species
- Genetic code: [Translation table 1 \(Standard\)](#)
- Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)
- Other names:

  - common name: **man**
  - authority: **Homo sapiens Linnaeus, 1758**

Lineage (full)

[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homininae](#); [Homo](#)

To the right of the main content is a table titled "Entrez records" showing the number of records for various data types:

Database name	Subtree links	Direct links
Nucleotide	10,217,570	10,217,541
Nucleotide EST	8,704,803	8,704,803
Nucleotide GSS	1,729,196	1,727,870
Protein	696,378	696,243
Structure	20,041	20,041
Genome	1	1
Popset	22,687	22,687
SNP	63,228,028	63,228,028
Domains	12	12
GEO Datasets	475,213	475,213
UniGene	130,045	130,045
UniSTS	328,844	328,844
PubMed Central	11,154	11,148
Gene	43,470	43,433
HomoloGene	18,473	18,473
SRA Experiments	53,471	53,469
Probe	24,258,933	24,258,933
Assembly	25	25
Bio Project	13,443	13,442
Bio Sample	812,246	812,243
Bio Systems	2,518	2,518
dbVar	2,517,546	2,517,546
Epigenomics	4,186	4,186
GEO Profiles	27,034,750	27,034,750
Protein Clusters	13	13
Taxonomy	3	1

**FIGURE 2.6** The entry for *Homo sapiens* at the NCBI Taxonomy Browser displays information about the genus and species as well as a variety of links to Entrez records. By following these links, a list of proteins, genes, DNA sequences, structures, or other data types that are restricted to this organism can be obtained. This can be a useful strategy to find a protein or gene from a particular organism (e.g., a species or subspecies of interest), excluding data from all other species.

Source: Taxonomy Browser, NCBI.

of the main genome browsers (described below); and (6) ArrayExpress is one of the two main worldwide repositories for gene expression data, along with the Gene Expression Omnibus at NCBI; both are described in Chapter 10.

Throughout this book we will focus on both the NCBI and EBI websites. In many cases those sites begin with similar raw data and then provide distinct ways of organizing, analyzing, and displaying data across a broad range of bioinformatics applications. When

You can access EBI at <http://www.ebi.ac.uk/> (WebLink 2.5).

Ensembl is a joint project of the EBI and WTSI (<http://www.ensembl.org>, WebLink 2.27). Related Ensembl projects include Metazoans (<http://metazoa.ensembl.org/>, WebLink 2.28), plants (<http://plants.ensembl.org/>, WebLink 2.29), fungi (<http://fungi.ensembl.org/>, WebLink 2.30), protists (<http://protists.ensembl.org/>, WebLink 2.31), and bacteria (<http://bacteria.ensembl.org/>, WebLink 2.32).

working on a problem, such as studying the structure or function of a particular gene, it is often helpful to explore the wealth of resources in both these sites. For example, each offers expert functional annotation of particular sequences and expert curation of databases. The NCBI and EBI websites increasingly offer an integration of their database resources so that information between the two sites can be easily linked.

## Ensembl

Founded in 1999 to annotate the human genome, the Ensembl project now spans over 70 vertebrate species. Related Ensembl projects include hundreds of other species from insects to bacteria.

## ACCESS TO INFORMATION: ACCESSION NUMBERS TO LABEL AND IDENTIFY SEQUENCES

If studying a problem that involves any gene or protein, it is likely that you will need to find information about some database entries. You can begin your research problem with information obtained from the literature, or you may have the name of a specific sequence of interest. Perhaps you have raw amino acid and/or nucleotide sequence data; we will explore how to analyze these in Chapters 3–5. The problem we will address now is how to extract information about your gene or protein of interest from databases.

An essential feature of DNA and protein sequence records is that they are tagged with accession numbers. An accession number is a string of about 4–12 numbers and/or alphabetic characters that are associated with a molecular sequence record (some are much longer). An accession number may also label other entries, such as protein structures or the results of a gene expression experiment (Chapters 10 and 11). Accession numbers from molecules in different databases have characteristic formats (Box 2.2). These formats vary because each database employs its own system. As you explore databases

### BOX 2.2 TYPES OF ACCESSION NUMBERS

Type of Record	Sample Accession Format
GenBank/EMBL/DDBJ nucleotide sequence records	One letter followed by five digits (e.g., X02775); two letters followed by six digits (e.g., AF025334).
GenPept sequence records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them)	Three letters and five digits (e.g., AAA12345).
Protein sequence records from SwissProt and PIR	Usually one letter and five digits (e.g., P12345). SwissProt numbers may also be a mixture of numbers and letters.
Protein sequence records from the Protein Research Foundation	A series of digits (often six or seven) followed by a letter (e.g., 1901178A).
RefSeq nucleotide sequence records	Two letters, an underscore bar, and six or more digits (e.g., mRNA records (NM_*): NM_006744; genomic DNA contigs (NT_*): NT_008769).
RefSeq protein sequence records	Two letters (NP), an underscore bar, and six or more digits (e.g., NP_006735).
Protein structure records	PDB accessions generally contain one digit followed by three letters (e.g., 1TUP). They may contain other mixtures of numbers and letters (or numbers only). MMDB ID numbers generally contain four digits (e.g., 3973.)

Many accession numbers include a suffix (e.g., .1 in NP\_006735.1), indicating a version number.



The screenshot shows a search interface with a search bar containing 'beta globin'. Below the search bar, it says 'About 75,478 search results for "beta globin"'. The results are organized into several sections: Literature, Genes, Health, Proteins, Genomes, and Chemicals. Each section lists various databases and their counts.

Literature			Genes		
Books	339	books and reports	EST	2,042	expressed sequence tag sequences
MeSH	4	ontology used for PubMed indexing	Gene	113	collected information about gene loci
NLM Catalog	10	books, journals and more in the NLM Collections	GEO DataSets	148	functional genomics studies
PubMed	8,827	scientific & medical abstracts/citations	GEO Profiles	3,828	gene expression and molecular abundance profiles
PubMed Central	18,185	full-text journal articles	HomoloGene	4	homologous gene sets for selected organisms
<b>Health</b>			PopSet	59	sequence sets from phylogenetic and population studies
ClinVar	163	human variations of clinical significance	UniGene	41	clusters of expressed transcripts
dbGaP	1,368	genotype/phenotype interaction studies	<b>Proteins</b>		
GTR	18	genetic testing registry	Conserved Domains	8	conserved protein domains
MedGen	13	medical genetics literature and links	Protein	2,316	protein sequences
OMIM	119	online mendelian inheritance in man	Protein Clusters	0	sequence similarity-based protein clusters
PubMed Health	21	clinical effectiveness, disease and drug reports	Structure	404	experimentally-determined biomolecular structures
<b>Genomes</b>			<b>Chemicals</b>		
Assembly	0	genomic assembly information	BioSystems	283	molecular pathways with links to genes, proteins and chemicals
BioProject	19	biological projects providing data to NCBI	PubChem BioAssay	45	bioactivity screening studies
BioSample	21	descriptions of biological source materials	PubChem Compound	0	chemical information with structures, information and links
Clone	32,086	genomic and cDNA clones	PubChem Substance	186	deposited substance and chemical information
dbVar	214	genome structural variation studies			
Epigenomics	24	epigenomic studies and display tools			
Genome	351	genome sequencing projects by organism			
GSS	3	genome survey sequences			
Nucleotide	3,276	DNA and RNA sequences			
Probe	125	sequence-based probes and primers			
SNP	789	short genetic variations			
SRA	13	high-throughput DNA and RNA sequence read archive			
Taxonomy	0	taxonomic classification and nomenclature catalog			

**FIGURE 2.7** The Entrez search engine (accessed from the home page of NCBI) provides links to results from 40 different NCBI databases. For many genes and proteins there are thousands of accession numbers. The RefSeq project is particularly important in trying to provide the best representative sequence of each normal (nonmutated) transcript produced by a gene and of each distinct, wildtype protein sequence.

Source: Entrez search engine, NCBI.

from which you extract DNA and protein data, try to become familiar with the different formats for accession numbers. Some of the various databases (Fig. 2.2) employ accession numbers that tell you whether the entry contains nucleotide or protein data.

For a typical molecule such as beta globin there are thousands of accession numbers (Fig. 2.7). Many of these correspond to ESTs and other fragments of DNA that match beta globin. How can you assess the quality of sequence or protein data? Some sequences are full-length, while others are partial. Some reflect naturally occurring variants such as single-nucleotide polymorphisms (SNPs; Chapter 8) or alternatively spliced transcripts (Chapter 10). Many of the sequence entries contain errors, particularly in the ends of EST reads. When we compare beta globin sequence derived from mRNA and from genomic DNA we may expect them to match perfectly (or nearly so) but, as we will see, there are often discrepancies (Chapter 10).

Using Sanger sequencing, DNA is usually sequenced on both strands. However, ESTs are often sequenced on one strand only, and therefore have a high error rate. We discuss sequencing error rates in Chapter 9.

For an NCBI page discussing GI numbers see <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html> (WebLink 2.33).

To see and compare the three myoglobin RefSeq entries at the DNA and the protein levels, visit <http://www.bioinfbook.org/chapter2> and select Web Document 2.1. As another example, the human alpha 1 globin and alpha 2 globin genes (*HBA1* and *HBA2*) are physically separate genes that encode proteins with identical sequences. The encoded alpha 1 globin and alpha 2 globin proteins are assigned the RefSeq identifiers NP\_000549.1 and NP\_000508.1.

Allelic variants, such as single base mutations in a gene, are not assigned different RefSeq accession numbers. However, OMIM and dbSNP (Chapters 8 and 21) do catalog allelic variants.

In addition to accession numbers, NCBI also assigns unique sequence identification numbers that apply to the individual sequences within a record. GenInfo (GI) numbers are assigned consecutively to each sequence that is processed. For example, the human beta globin DNA sequence associated with the accession number NM\_000518.4 has a gene identifier GI:28302128. The suffix .4 on the accession number refers to a version number; NM\_000518.3 has a different gene identifier, GI: 13788565.

### The Reference Sequence (RefSeq) Project

One of the most important developments in the management of molecular sequences is RefSeq. The goal of RefSeq is to provide the best representative sequence for each normal (i.e., nonmutated) transcript produced by a gene and for each normal protein product (Pruitt *et al.*, 2014). There may be hundreds of GenBank accession numbers corresponding to a gene, since GenBank is an archival database that is often highly redundant. However, there will be only one RefSeq entry corresponding to a given gene or gene product, or several RefSeq entries if there are splice variants or distinct loci.

Consider human myoglobin as an example. There are three RefSeq entries (NM\_005368.2, NM\_203377.1, and NM\_203378.1), each corresponding to a distinct splice variant. Each splice variant involves the transcription of different exons from a single-gene locus. In this example, all three transcripts happen to encode an identical protein having the same amino acid sequence. The source of the transcript distinctly varies, and may be regulated and expressed under different physiological conditions. It therefore makes sense that each protein sequence, although having an identical string of amino acid residues, is assigned its own protein accession number (NP\_005359.1, NP\_976311.1, and NP\_976312.1, respectively).

RefSeq entries are curated by the staff at NCBI and are nearly nonredundant (Pruitt *et al.*, 2014). RefSeq entries have different status levels (predicted, provisional, and reviewed), but in each case the RefSeq entry is intended to unify the sequence records. You can recognize a RefSeq accession by its format, such as NP\_000509 (P stands for beta globin protein) or NM\_006744 (for beta globin mRNA). The corresponding XP\_12345 and XM\_12345 formats imply that the sequences are not based on experimental evidence. A variety of RefSeq formats are shown in Table 2.7 and identifiers corresponding to human beta globin are shown in Table 2.8.

A GenBank or RefSeq accession number refers to the most recent version of a given sequence. For example, NM\_000558.3 is currently a RefSeq identifier for human

**TABLE 2.7 Formats of accession numbers for RefSeq entries. There are currently 22 different RefSeq accession formats. The methods include expert manual curation, automated curation, or a combination. Abbreviations: BAC, bacterial artificial chromosome; WGS, whole-genome shotgun (see Chapter 15). Adapted from <http://www.ncbi.nlm.nih.gov/refseq/about/>.**

Molecule	Accession format	Genome
Complete genome	NC_123456	Complete genomic molecules, including genomes, chromosomes, organelles, and plasmids
Genomic DNA	NW_123456 or NW_123456789	Intermediate genomic assemblies
Genomic DNA	NZ_ABCD12345678	Collection of whole-genome shotgun sequence data
Genomic DNA	NT_123456	Intermediate genomic assemblies (BAC and/or WGS sequence data)
mRNA	NM_123456 or NM_123456789	Transcript products; mature mRNA protein-coding transcripts
Protein	NP_123456 or NM_123456789	Protein products (primarily full-length)
RNA	NR_123456	Noncoding transcripts (e.g., structural RNAs, transcribed pseudogenes)

**TABLE 2.8 RefSeq accession numbers corresponding to human beta globin. Adapted from <http://www.ncbi.nlm.nih.gov/refseq/about/>.**

Category	Accession	Size	Description
DNA	NC_000011.9	135,006,516 bp	Genomic contig
DNA	NM_000518.4	626 bp	DNA corresponding to mRNA
DNA	NG_000007.3	81,706 bp	Genomic reference
protein	NP_000509.1	147 amino acids	Protein

alpha 1 hemoglobin. We mentioned above that a suffix such as “.3” is the version number. By default, if you do not specify a version number then the most recent version is provided.

### RefSeqGene and the Locus Reference Genomic Project

While the RefSeq project has a critical role in defining reference sequences, it has several limitations. The changing version numbers of some sequences can lead to ambiguity when scientists report RefSeq accession numbers without their version numbers. For example, a patient may have a variant at a specific nucleotide position in the beta globin gene corresponding to NM\_000518.3 but (as often happens) the version number is not given. Once the record is subsequently updated to NM\_000518.4, anyone studying this variant might be unsure of the correct position of the variant since it depends on which sequence version was used.

To address these concerns about gene variant reporting, the Locus Reference Genomic (LRG) sequence format was introduced (Dagleish *et al.*, 2010). The goal of this project is to define genomic sequences that can be used as reference standards for genes, representing a standard allele. No version numbers are used and sequence records are stable and designed to be independent of updates to reference genome assemblies. In a related response to this issue, the RefSeq project was expanded to include RefSeq Gene.

### The Consensus Coding Sequence CCDS Project

The Consensus Coding Sequence (CCDS) project was established to identify a core set of protein coding sequences that provide a basis for a standard set of gene annotations (Farrell *et al.*, 2014). The CCDS project is a collaboration between four groups (EBI, NCBI, the Wellcome Trust Sanger Institute and the University of California, Santa Cruz or UCSC). Currently, the CCDS project has been applied to the human and mouse genomes; its scope is considerably more limited than RefSeq. Its strength is that it offers a “gold standard” of best supported gene and protein annotations with extensive manual annotation by experts, enhancing the quality of the database (Harte *et al.*, 2012).

### The Vertebrate Genome Annotation (VEGA) Project

It is essential to correctly annotate each genome; in particular, we need to define gene loci and all their features. The Vertebrate Genome Annotation (VEGA) database offers high-quality, manual (expert) annotation of the human and mouse genomes, as well as selected other vertebrate genomes (Harrow *et al.*, 2014).

Performing a search for HBB at the VEGA website, there is one human entry. This includes two main displays: (1) a transcript view which provides information such as cDNA and coding sequences and protein domain information; and (2) a gene view which includes data on orthologs and alternative alleles.

Carry out a NCBI nucleotide search for NM\_000558.1 and learn about the revision history of that accession number. In Chapter 3 we will learn how to compare two sequences; you can BLAST NM\_000558.1 against NM\_000558.3 to see the differences, or view the results in Web Document 2.2 at <http://www.bioinfbook.org/chapter2>. If you do not specify a version number for BLAST searches then the most recent version is used by default.

LRG is pronounced “large.” You can access this project at <http://www.lrg-sequence.org> (WebLink 2.34). You can access RefSeqGene at <http://www.ncbi.nlm.nih.gov/refseq/rsg/> (WebLink 2.35).

You can learn about the CCDS project at <http://www.ncbi.nlm.nih.gov/projects/CCDS/> (WebLink 2.36). As of October 2014 there are 18,800 human gene IDs (and over 30,000 CCDS IDs) for this project.

VEGA is a project of the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Wellcome Trust Sanger Institute. There are three main portals to access HAVANA annotation: Ensembl, UCSC, and VEGA. You can access Vega at <http://vega.sanger.ac.uk/> (WebLink 2.37). The HAVANA website is <http://www.sanger.ac.uk/research/projects/vertebratogenome/havana/> (WebLink 2.38). At NCBI, Vega annotations are available in the Gene resource.

We discuss the definition of a gene and complex features such as alternative splice sites, pseudogenes, polyadenylation sites, other regulatory sites, and the structure of exons and introns in Chapter 8.

You can view the VEGA page for HBB at [http://vega.sanger.ac.uk/Homo\\_sapiens/Gene/Summary?g=OTTHUMG00000066678;r=11:5246694-5250625](http://vega.sanger.ac.uk/Homo_sapiens/Gene/Summary?g=OTTHUMG00000066678;r=11:5246694-5250625) (WebLink 2.39). The NCBI Gene entry for HBB also contains a link to the VEGA result.

## ACCESS TO INFORMATION VIA GENE RESOURCE AT NCBI

How can one navigate through the bewildering number of protein and DNA sequences in the various databases? An emerging feature is that databases are increasingly interconnected, providing a variety of convenient links to each other and to algorithms that are useful for DNA, RNA, and protein analysis. NCBI's Gene resource (formerly called Entrez Gene, and LocusLink before that) is particularly useful as a major portal. It is a curated database containing descriptive information about genetic loci (Maglott *et al.*, 2007). You can obtain information on official nomenclature, aliases, sequence accessions, phenotypes, Enzyme Commission (EC) numbers, OMIM numbers, UniGene clusters, HomoloGene (a database that reports eukaryotic orthologs), map locations, and related websites.

To illustrate the use of NCBI Gene we search for human beta globin. The result of entering an NCBI Gene search is shown in **Figure 2.8**. Note that in performing this search, it can be convenient to restrict the search to a particular organism of interest. (This can be done using the “limits” tab on the NCBI Gene page.) The “Links” button

Name/Gene ID	Description	Location	Aliases
<a href="#">HBB</a> ID: 3043	hemoglobin, beta [ <i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, <b>beta-globin</b>
<a href="#">hbq1</a> ID: 394453	hemoglobin, gamma A [ <i>Xenopus (Silurana)</i> <i>tropicalis</i> (western clawed frog)]	NW_004668244.1 (60116737..60118249)	<b>beta-globin</b> , hbb1, hbga, hbgr, hsggl1
<a href="#">hbq1</a> ID: 734881	hemoglobin, gamma A [ <i>Xenopus laevis</i> (African clawed frog)]		<b>beta-globin</b> , hbb1, hbga, hbgr, hsggl1
<a href="#">Hbb-bh1</a> ID: 15132	hemoglobin Z, beta-like embryonic chain [ <i>Mus</i> <i>musculus</i> (house mouse)]	Chromosome 7, NC_000073.6 (103841638..103843162, complement)	<b>betaH1</b>
<a href="#">HBG2</a> ID: 396485	hemoglobin, gamma G [ <i>Gallus</i> <i>gallus</i> (chicken)]	Chromosome 1, NC_006088.3 (193724299..193725801)	HBB, HBD, HBE1

**FIGURE 2.8** Result of a search for “beta globin” in NCBI Gene (via an Entrez search). Information is provided for a variety of organisms including *Homo sapiens*, *Mus musculus*, and several frog species. Links provides access to information on beta globin from a variety of other databases.

Source: NCBI Gene.

NCBI Resources How To

Gene Gene Limits Advanced Search Help

Display Settings:  Full Report Send to:

### HBB hemoglobin, beta [ *Homo sapiens* (human) ]

Gene ID: 3043, updated on 16-Apr-2013

**Summary**

Official Symbol HBB provided by HGNC  
 Official Full Name hemoglobin, beta provided by HGNC  
 Primary source HGNC:4827  
 See related Ensembl:ENSG00000244734; HPRD:00786; MIM:141900; Vega:OTTHUMG00000066678  
 Gene type protein coding  
 RefSeq status REVIEWED  
 Organism *Homo sapiens*  
 Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Earchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo  
 Also known as CD113t-C; beta-globin  
 Summary The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta-3'. [provided by RefSeq, Jul 2008]

**Genomic context**

Location: 11p15.5 See HBB in Epigenomics, MapViewer  
 Sequence: Chromosome: 11; NC\_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC\_000011.9

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Interactions
- Pathways
- General gene information
- Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- Reference sequences
- Related sequences
- Additional links

Related information

- Order cDNA clone
- 3D structures
- BioAssay
- BioAssay, by Protein Target
- BioProjects
- BioSystems
- Books
- CCDS
- ClinVar
- Conserved Domains

**FIGURE 2.9** Portion of the NCBI Gene entry for human beta globin. Information is provided on the gene structure and chromosomal location, as well as a summary of the protein's function. RefSeq accession numbers are also provided (not shown); access these by clicking "Reference sequences" in the table of contents (top right). The menu (right sidebar) provides extensive links to additional databases including PubMed, OMIM (Chapter 21), UniGene (Chapter 10), a variation database (dbSNP; Chapter 20), HomoloGene (with information on homologs; Chapter 6), a gene ontology database (Chapter 12), and Ensembl viewers at EBI (Chapter 8).

Source: NCBI Gene entry.

(Fig. 2.8, top right) provides access to various other databases entries on beta globin. Clicking on the main link to the human beta globin entry results in the following information (Fig. 2.9):

- At the top right, there is a table of contents for the NCBI Gene beta globin entry. Below it are further links to beta globin entries in NCBI databases (e.g., protein and nucleotide databases and PubMed), as well as external databases (e.g., Ensembl and UCSC; see below and Chapter 8).
- Gene provides the official symbol (*HBB*) and name for human beta globin.
- A schematic overview of the gene structure is provided, hyperlinked to the Map Viewer (see "The Map Viewer at NCBI" below).
- There is a brief description of the function of beta globin, defining it as a carrier protein of the globin family.
- The Reference Sequence (RefSeq) and GenBank accession numbers are provided.

Gene is accessed from the main NCBI web page (by clicking All Databases). Currently (2014), Gene encompasses about 12,000 taxa and 15 million genes. We explore many of the resources within NCBI's Gene in later chapters such as its links to information on genes (Chapter 8), expression data such as RNA-seq data as available within its browser (Chapter 11), proteins (Chapter 12), links to pathway data (Chapter 14), and disease relevance (Chapter 21).

Figure 2.10 shows the standard, default form of a typical NCBI Protein record (for beta globin). It is simple to obtain a variety of formats by changing the display options. By clicking a tab (Fig. 2.10a) the commonly used FASTA format for protein (or DNA) sequences can be obtained, as shown in Figure 2.11. Note also that by clicking the CDS

**Display Settings:** GenPept

**hemoglobin subunit beta [Homo sapiens]**

NCBI Reference Sequence: NP\_000509.1

[FASTA](#) [Graphics](#)

**Go to:** [NP\\_000509](#)

LOCUS	NP_000509	147 aa	linear	PRI 17-APR-2013
DEFINITION	hemoglobin subunit beta [Homo sapiens].			
ACCESSION	NP_000509			
VERSION	NP_000509.1	GI:4504349		
DBSOURCE	REFSEQ: accession <a href="#">NM_000518.4</a>			
KEYWORDS	.			
SOURCE	Homo sapiens (human)			
ORGANISM	<a href="#">Homo sapiens</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.			
REFERENCE	1 (residues 1 to 147)			
AUTHORS	Lacerza,G., Prezioso,R., Musollino,G., Piluso,G., Mastrullo,L. and De Angioletti,M.			
TITLE	Identification and molecular characterization of a novel 55-kb deletion recurrent in southern Italy: the Italian (G) gamma((A) gammadelta(beta)) degrees -thalassemia			
JOURNAL	<a href="#">Eur. J. Haematol.</a> 90 (3), 214-219 (2013)			
PUBMED	<a href="#">23281611</a>			

**CDS**

```

1..147
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C"
/coded_by="NM_000518.4:51..494"
/db_xref="CCDS:CCDS7753.1"
/db_xref="GeneID:3043"
/db_xref="HGNC:4827"
/db_xref="HPRD:00786"
/db_xref="MIM:141900"
```

**ORIGIN**

```

1 mvhltpeeks avtalwgkvn vdevggealq r11vvypwtq rffesfgdls tpdavmgnpk
61 vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
121 keftppvqaa yqkvvagvan alahkyh
//
```

**FIGURE 2.10** Display of an NCBI Protein record for human beta globin. This is a typical entry for any protein. (Above) Top portion of the record. Key information includes the length of the protein (147 amino acids), the division (PRI, or primate), the accession number (NP\_000509.1), the organism (*H. sapiens*), literature references, comments on the function of globins, and links to other databases (right side). At the top of the page, the display option allows this record to be obtained in a variety of formats, such as FASTA (Fig. 2.11). (Below) Bottom portion of the record, which includes features such as the coding sequence (CDS). The amino acid sequence is provided at the bottom in the single-letter amino acid code (although here not in the FASTA format).

Source: NCBI Protein entry.

The screenshot shows the NCBI Protein search interface. At the top, there are links for "NCBI Resources" and "How To". Below that, a search bar has "Protein" selected. Underneath the search bar are "Limits" and "Advanced" buttons. A "Display Settings" dropdown is set to "FASTA". The main content area displays the protein header "hemoglobin subunit beta [Homo sapiens]" and the NCBI Reference Sequence "NP\_000509.1". Below the header are links for "GenPept" and "Graphics". The FASTA sequence is shown as follows:

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAQKVVAAGVAN
ALAHKYH
```

**FIGURE 2.11** Protein entries can be displayed in the FASTA format. This includes a header row (beginning with the > symbol) containing a single line of text, then a single line break and the sequence (whether protein in the single-letter amino acid code or DNA in the GATC format). The FASTA format is used in a variety of software programs that we will use involving topics such as pairwise alignment (Chapter 3), BLAST (Chapter 4), next-generation sequencing (Chapter 9), and proteomics (Chapter 12).

Source: NCBI.

(coding sequence) link of an NCBI Protein or NCBI Nucleotide record (shown in Fig. 2.10b at the upper left), the nucleotides that encode a particular protein, typically beginning with a start methionine (ATG) and ending with a stop codon (TAG, TAA, or TGA), can be obtained. This can be useful for a variety of applications including multiple sequence alignment (Chapter 6) and molecular phylogeny (Chapter 7).

### Relationship Between NCBI Gene, Nucleotide, and Protein Resources

If interested in obtaining information about a particular DNA or protein sequence, it is reasonable to visit NCBI Nucleotide or NCBI Protein and perform a search. A variety of search strategies are available, such as limiting the output to a particular organism or taxonomic group of interest, or limiting the output to RefSeq entries.

There are also many advantages to beginning your search through NCBI Gene. The official gene name can be identified there, and you can be assured of the chromosomal location of the gene. Furthermore, each Gene entry includes a section of reference sequences that provides all the DNA and protein variants that are assigned RefSeq accession numbers.

FASTA is both an alignment program (described in Chapter 3) and a commonly used sequence format (further described in Chapter 4 and used in web documents throughout this book). It is related to FASTQ and FASTG (formats used in next-generation sequence analysis; see Chapter 9).

### Comparison of NCBI's Gene and UniGene

As described above, the UniGene project assigns one cluster of sequences to one gene. For example, for *HBB* there is one UniGene entry with the UniGene accession number Hs.523443. This UniGene entry includes a list of all the GenBank entries, including ESTs, that correspond to the *HBB* gene. The UniGene entry also includes mapping information, homologies, and expression information (i.e., a list of the tissues from which cDNA libraries were generated that contain ESTs corresponding to the RBP gene).

NCBI Gene now has >200,000 human entries (as of 2015). These include gene predictions, pseudogenes, and mapped phenotypes.

HomoloGene is available by clicking All Databases from the NCBI home page, or at <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene> (WebLink 2.40). Release 68 (2014) has >230,000 groups (including 19,000 human groups). We define homologs in Chapter 3.

UniGene and NCBI Gene have features in common, such as links to OMIM, homologs, and mapping information. They both show RefSeq accession numbers. There are four main differences between UniGene and NCBI Gene:

1. UniGene has detailed expression information; the regional distributions of cDNA libraries from which particular ESTs have been sequenced are listed.
2. UniGene lists ESTs corresponding to a gene, allowing them to be studied in detail.
3. Gene may provide a more stable description of a particular gene; as described above, UniGene entries may be collapsed as genome-sequencing efforts proceed.
4. Gene has fewer entries than UniGene, but these entries are more richly curated.

## NCBI's Gene and HomoloGene

The HomoloGene database provides groups of annotated proteins from a set of completely sequenced eukaryotic genomes. Proteins are compared (by BLASTP; see Chapter 4), placed in groups of homologs, and the protein alignments are then matched to the corresponding DNA sequences. You can find a HomoloGene entry for a gene/protein of interest by following a link on the NCBI Gene page.

A search of HomoloGene with the term hemoglobin results in dozens of matches for myoglobin, alpha globin, and beta globin. By clicking on the beta globin group, access can be gained to a list of proteins with RefSeq accession numbers from human, chimpanzee, dog, mouse, and chicken. The pairwise alignment scores (see Chapter 3) are summarized and linked, and the sequences can be downloaded (in genomic DNA, mRNA, and protein formats) and displayed as a protein multiple sequence alignment (Chapter 6).

## COMMAND-LINE ACCESS TO DATA AT NCBI

The websites of NCBI, EBI, Ensembl, and other bioinformatics sites offer convenient access to resources through a web browser; an alternative is to use command-line tools. We now introduce command-line use and describe Entrez Direct (EDirect), which allows command-line access to Entrez databases.

### Using Command-Line Software

Many bioinformatics software packages were designed for command-line usage. We use such software such for a variety of applications such as BLAST (Chapter 4), sequence alignment (Chapter 6), phylogeny (Chapter 7), DNA analysis (Chapter 8), next-generation sequence analysis (Chapter 9), RNA-seq (Chapter 11), genome comparisons (Chapter 16), and genome annotation (Chapter 17).

The three most popular operating systems are Windows, Mac OS, and Unix. Each operating system manages resources on a computer, executes tasks, and provides the user interface. Linux is a flavor of Unix that offers several advantages, especially for those manipulating datasets and software programs for bioinformatics:

- It is a free operating system.
- It has been developed by thousands of programmers and now features applications and interfaces that can provide an experience closer to the Windows and Mac OS environments that are more familiar to many students.
- It is highly customizable and flexible.
- For bioinformatics applications, it is well suited to process large datasets such as tables with millions of rows, or smaller data matrices that require sophisticated manipulation.
- Microsoft Excel limits the number of rows a spreadsheet can have and, more importantly, as a default it automatically changes some names and numbers. Tables in a Unix environment are unrestricted in size (limited only by available disk space) and are not automatically reformatted.

A user types commands via a command processor. Bash is a Unix shell that is the default command processor for Linux and Mac OS X.

You can access a computer running Linux on a laptop or desktop, or by accessing a Linux server. For example, you can work on Microsoft Windows and access a Linux machine with a Secure Shell (SSH) client such as PuTTY. This is a free, open-source terminal emulator that enables one machine to communicate with another. PuTTY implements the client end of a session, opening a window on a PC that lets you type commands and receive results obtained from a remote Linux machine.

Mac OS offers a terminal (visit Applications > Utilities > Terminal). This provides a Unix-based shell (called Portable Operating System Interface or POSIX-compliant). For many bioinformatics researchers, the availability of a terminal with access to a vast number of Unix-based tools and resources makes Mac OS preferable to a PC.

For PC users, Cygwin offers a Unix-like environment and command-line interface on Microsoft Windows. We demonstrate some PC-based command-line tools in this book, but in most cases we rely on Linux or Mac OS.

Box 2.3 introduces several basic command-line tasks and operations. Open a terminal and try them. You will see other basic commands as we use command-line tools throughout this book.

Bash stands for Bourne-again shell.

Cygwin is available at <http://www.cygwin.com/> (WebLink 2.41).

## BOX 2.3 LINUX COMMANDS

We can explore the command-line environment with six topics. A hash (#) symbol indicates a comment; any commented text is ignored. (If the # appears at the beginning of the line, the entire line is ignored; if # appears in the middle of a line, the commands that precede it are executed.) A \$ symbol indicates a Unix command prompt whether you are working with Linux or Mac OS; some operating systems use other command prompts.

1. *Finding where you are and moving around.*

```
$ pwd # print working directory
/home/pevsner # this is your beginning working directory
$ cd /home/pevsner/mysubdirectory # change directory
# This results in a new command prompt; enter pwd to confirm that you have moved down
# into a subdirectory.
$ cd .. # The current directory is represented by a single dot (..). Using two dots
# (...) we change to the parent directory
$ cd ~ # Use this from any location to return to the home directory, e.g., /home/
pevsner
```

To find out what files are stored within a directory, use the following code.

```
$ ls # list contents in a directory
$ ls -l # list files in the "long" format including file sizes and permissions
$ ls -lh # list files including file sizes (in human readable format) and permissions
```

2. *Getting help.* Try the manual (`man`) for usage of many utilities (or try `info` on some Mac OS terminals). The `man` page can have so much information that it is difficult to know the best way to begin using some function of interest. Many people therefore rely heavily on searches with their favorite search engine (typically Google) for help on accomplishing some task. Many other people have had questions similar to yours! There are also excellent forums such as Biostars (<http://www.biostars.org>) where you can read others' questions and answers.

```
$ man pwd # type q to exit any man entry
$ man cd
$ man ls
```

## BOX 2.3 (CONTINUED)

3. *Permissions.* When you use `ls -l` to view your files, permissions are shown with the first 10 characters. For example:

```
$ ls -lh
total 20K
-rw-rw-r--. 1 pevsner pevsner 1.5K Sep 24 2013 9globins.txt
drwxrwxr-x. 2 pevsner pevsner 43 Oct 17 09:09 ch01_intro
drwxrwxr-x. 3 pevsner pevsner 103 Apr 19 15:35 ch04_blast
```

The first character is usually either `d` (for directory) or `-` (a regular file, and not a directory); in the example above there are two directories and one file, then three sets of three characters: `rwx` (read, write, executable). These three groups are (a) the owner of the file; (b) members of the group; and (c) all other users. These permissions settings specify who can read files, write to them, or execute them. Users routinely need to examine (and update) permissions.

```
$ sudo chmod ugo+rwx path/to/file
```

`sudo` should be used carefully by new users. It allows some users to execute a command as the “superuser,” for example setting permissions. `sudo` requires an administrator’s password.

`chmod` refers to “change file mode bits” and changes the permissions for a file or directory, for example making it accessible to other users. The `ugo+rwx` option makes the file and/or folder readable, writable, and executable by the user (`u`), group (`g`), and others (`o`).

4. *Making a directory.*

```
$ mkdir myproject
```

You can organize your data in many different ways. William Noble (2009) has written an excellent guide suggesting that you create subfolders such as `doc` (to store documents), `data` (to store fixed datasets such as sequence records or alignment files), `results` (to track experiments you perform on your data), `src` (for source code), and `bin` (for compiled binaries or scripts). A goal is to make it possible for someone unfamiliar with your work to examine your files and understand what you did and why.

5. *Making a text file.* There are several excellent editors. `nano` is perhaps the easiest to learn if you are just beginning; it offers helpful prompts to facilitate editing and saving files. Here we use `vim`.

```
$ man vim # get information on vim usage
$ vim mydocument.txt # we create a text file called mydocument.txt
# In the vim text editor,
# press :h for a main help file
# press i to insert text
# press Esc (escape key) to leave insert mode
# press :wq to write changes and quit
```

6. *Importing a file.* Go to a web browser and visit NCBI > Downloads > FTP:RefSeq > Mitochondrion > <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.protein.faa.gz>. To grab a URL, be sure to “Copy Link Location,” which you can subsequently paste.

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/mitochondrion.1.protein.faa.gz
# Your file will be downloaded into your directory! On a Mac try curl in place of wget.
```

The EDirect documentation also lists some basic Unix filters for sorting text documents (`sort`), removing repeated lines (`uniq`), matching patterns (`grep`), and more.

## BOX 2.4 USING NCBI'S EDIRECT: COMMAND-LINE ACCESS TO ENTREZ DATABASES

The Entrez system currently includes 40 databases, including those we will encounter for nucleotide and protein records (this chapter), multiple alignments (HomoloGene and Conserved Domain Database, Chapter 6), gene expression (Gene Expression Omnibus, Chapter 9), proteins (Chapter 12), and protein structure (Chapter 13). An easy way to access these databases is by web searches.

In many cases it is essential to use a structured interface to perform large-scale queries. For example, suppose you obtain a list of 100 genes of interest (perhaps they are significantly regulated in a gene expression study, or they have variants of interest from a whole-genome sequence). NCBI offers two main options. (1) The Entrez Programming Utilities (E-utils) allow you to search and retrieve information from Entrez databases. You use software that posts an E-util URL to NCBI using a fixed URL recognized by E-util servers at NCBI. We can employ Biopython, Perl, or other languages for this purpose. (2) EDirect allows command-line access to the Entrez databases. It is convenient, versatile, and far easier to use than E-utils.

The programs accessed by EDirect (and the E-utils) are as follows:

1. Einfo: database statistics. This provides the number of records available in each field of a database. For example, you can determine how many records are in PubMed. Einfo also describes which other Entrez databases link to the given database you are interrogating.
2. Esearch: text searches. When you provide a text query (such as “globin”) this returns a list of UIDs. These UIDs can later be used in Esummary, Efetch, or Elink.
3. Epost: UID uploads. You may have a list of UIDs, such as PMIDs for a favorite query. You can upload these UIDs and store them on a History Server.
4. Esummary: document summary downloads. When you provide a list of UIDs, Esummary returns the corresponding document summaries.
5. Efetch: data record downloads. Note that Esearch and Efetch can be combined for more efficient searching.
6. Elink: Entrez links.
7. EGQuery: global query. Given a text query, this utility reports the number of records in each Entrez database. Similarly, when you enter a text query into the main page of NCBI you can see various database matches.
8. Espell: spelling suggestions.

Try EDirect. Start by installing it; directions are available at the NCBI website, along with sample queries. Repeat the examples given in this chapter. When you do any Entrez search using the NCBI website, see if you can repeat it using EDirect! To get started, copy the following commands from the EDirect website (also available at the Chapter 2 page for <http://bioinfbook.org/>). This will download scripts into a folder called `edirect` in your home directory.

```
cd ~
perl -MNet::FTP -e \
    '$ftp = new Net::FTP("ftp.ncbi.nlm.nih.gov", Passive => 1); $ftp->login;
     $ftp->binary; $ftp->get("/entrez/entrezdirect/edirect.zip");'
unzip -u -q edirect.zip
rm edirect.zip
export PATH=$PATH:$HOME/edirect
./edirect/setup.sh
```

### Accessing NCBI Databases with EDirect

EDirect is a suite of Perl scripts that allows queries in the Unix environment, including users of Linux and Macintosh OSX computers. (It also works with the Cygwin Unix-emulation environment on Windows computers.) EDirect allows you to access information in the various Entrez databases using command-line arguments (from a terminal window). Installation is simple (see Box 2.4) and produces a folder called `edirect` in your home directory. On a Linux machine, open a terminal window where you typically begin in your home directory. The # sign below indicates a comment that is not implemented as a command.

```
$ cd edirect # navigate to the folder with edirect scripts
$ ls # ls is a utility that lists entries within a directory
README      edirutil    einfo      epost      esummary
econtact    efetch      elink      eproxy     nquery
edirect.pl   efilter     enotify    esearch    xtract
```

Entrez Direct can be downloaded by file transfer protocol (FTP) at <ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/> (WebLink 2.42). EDirect documentation is provided at <http://www.ncbi.nlm.nih.gov/books/NBK179288> (WebLink 2.43). NCBI developed EDirect to provide simplified access to NCBI's Entrez Programming Utilities (E-utilities), which are a set of server-side programs that use a fixed URL syntax to provide a stable interface into the Entrez databases. EDirect can accomplish practically any E-utilities task on the command line, but without the need for programming experience. Visit <http://www.ncbi.nlm.nih.gov/books/NBK25500/> (WebLink 2.44) to learn more about the E-utilities and obtain a deeper understanding of what EDirect can accomplish.

When you download EDirect as described in Box 2.4, its scripts can be used when you are working in any directory. If you need to move the `edirect` folder to another location, you should also edit the `.bash_profile` configuration file, updating the statement that sets the PATH environment variable. The general pattern for this statement is as follows:

```
export PATH=$HOME/
subdirectory_
with_edirect_
scripts:$PATH:.
```

These are the various scripts available in EDirect.

EDirect has functions that facilitate your ability to navigate Entrez databases (`esearch`, `elink`, `efilter`), retrieval functions (`esummary`, `efetch`), extracting fields from XML results (`xtract`), and assorted other functions such as `epost` to upload unique identifiers or accession numbers. We next provide several specific examples, adapted from the EDirect online documentation at NCBI.

### *EDirect Example 1*

Search PubMed for articles by the author J. Pevsner including the term GNAQ, fetch the results in the form of summaries, and send the results first to the screen and then to a file called `example1.out`. The `$` sign indicates the start of a Unix (or Linux or Mac OSX) command.

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetech -format docsum
1: Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, Cohen B, North PE, Marchuk DA, Comi AM, Pevsner J. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N Engl J Med. 2013 May 23;368(21):1971-9. doi: 10.1056/NEJMoa1213507. Epub 2013 May 8. PubMed PMID: 23656586; PubMed Central PMCID: PMC3749068.
```

Here we used the pipe symbol (`|`) to send our results from the `esearch` utility, `efetch`. That allowed us to select a particular output format called `docsum` for document summary. We can also use `>` to send the result to a file (called `example1.txt`):

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetech -format docsum >
example1.txt
```

You can view your query results on the screen, send them to a file, or view just part of the output. The `less` utility displays the output one page at a time; use the space bar to advance a page. In Linux you can enter `$ man less` to use the manual (`man`) utility for more information about `less` (or any other function). (Or try `$ info less` on a Mac.) Use `head` without an argument to display just the first 10 lines of the file.

### *EDirect Example 2*

Perform a PubMed search without piping the results to `efetch`. Instead we will pipe the results to `less`. This will display on the screen how many results there are for various queries.

```
$ esearch -db pubmed -query "pevsner j" | less
<ENTREZ_DIRECT>
<Db>pubmed</Db>

<WebEnv>NCID_1_142748046_130.14.18.34_9001_1391877213_1550387237</WebEnv>
<QueryKey>1</QueryKey>
<Count>99</Count>
<Step>1</Step>
</ENTREZ_DIRECT>
( END )
```

This command searches PubMed for articles by J. Pevsner and shows that there are 99. Similar searches show the number of articles for the query hemoglobin (~155,000), bioinformatics (~131,000), or BLAST (~23,000). Instead of using the pipe `|` to send the results to `less`, we could also send the results to a file with an argument such as `> myoutput.txt`.

### *EDirect Example 3*

Search PubMed to find which authors have published the most in the area of bioinformatics software. EDirect includes a useful function called `sort-uniq-count-rank`. Unix is a good environment for tasks such as sorting a large list and counting items.

Some Unix commands are used frequently and can be combined to simplify tasks. The `sort-uniq-count-rank` function will read lines of text, sort them alphabetically, count the number of occurrences of each unique line, and then resort by the line count.

We are now ready to search PubMed for a topic. Here we will use the major topic “bioinformatics” in the Medical Subjects Headings browser (MeSH, introduced in “Example of PubMed Search” below) and “software” in the title/abstract (the [TIAB] indexed field). We use `esearch` to search PubMed, then we send (“pipe” or `|`) the output to the `efetch` program that formats the results in Extensible Markup Language (XML). We further use `xtract` to obtain the authors’ last names and first initials, then `sort-uniq-count-rank` to list the results.

```
$ esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" |
efetch -format xml | xtract -pattern PubmedArticle -block Author -sep " "
-tab "\n" -element LastName,Initials | sort-uniq-count-rank
29 Aebersold R
27 Wang Y
22 Deutsch EW
22 Zhang J
21 Chen Y
21 Martens L
20 Wang J
19 Zhang Y
18 Smith RD
17 Hermjakob H
17 Wang X
15 Li X
15 Zhang X
14 Chen L
14 Li C
14 Li L
14 Yates JR
13 Durbin R
13 Liu J
13 Salzberg SL
13 Sun H
13 Zhang L
```

The authors who have published the most articles on bioinformatics software (according to the particular search criteria we chose) include: Ruedi Aebersold (a pioneer in proteomics), Eric Deutsch (Institute for Systems Biology); Lennart Martens (proteomics and systems biology); Henning Hermjakob (European Bioinformatics Institute); Richard Durbin (Wellcome Trust Sanger Institute); and Steven Salzberg (Johns Hopkins).

#### *EDirect Example 4*

Perform a search of the Protein database for entries matching the query term “hemoglobin”, and pipe the results in the FASTA format to `head` to see the first 6 lines of the output.

```
$ esearch -db protein -query "hemoglobin" | efetch -format fasta | head -6
# the -6 argument specifies that we want to see the first 6 lines of
# output; the default setting is 10 lines
>gi|582086208|gb|EVU02130.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 52-G]
MIIVTNTAKITKGNGHKLIDRFNKGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
AFKAAHSHQGGMPDYILDNKISYYDVKVVRMPMAAAQ

>gi|582080234|gb|EVT96395.1| heme-degrading monooxygenase IsdG [Bacillus
anthracis 9080-G]
MIIVTNTAKITKGNGHKLIDRFNKGQVETMPGFLGLEVLLTQNTVDYDEVTISTRWNAKEDFQGWTKSP
```

Although we searched the protein database, note that you can search any of the dozens of Entrez databases.

***EDirect Example 5***

Find PubMed articles related to the query “hemoglobin”, use elink to find related articles, then use elink again to find proteins.

```
esearch -db pubmed -query "hemoglobin" | \
elink -related | \
elink -target protein
```

This example shows how commands can be entered on separate lines with the \ symbol.

***EDirect Example 6***

List the genes on human chromosome 16 including their start and stop positions.

```
$ esearch -db gene -query "16[chr] AND human[orgn] AND alive[prop]" \
| esummary | xtract -pattern DocumentSummary -element Id -block \
LocationHistType -match "AssemblyAccVer:GCF_000001405.25" -pxf "\n" \
-element AnnotationRelease,ChrAccVer,ChrStart,ChrStop > example6.out
```

The results are stored in the file example6.out (you can select any name). We use head -5 to view the first five lines of the output.

```
$ head -5 example6.out
999
105 NC_000016.9    68771127    68869444
4313
105 NC_000016.9    55513080    55540585
64127
```

This example shows a complex command that can be used (by copying and pasting from the EDirect website documentation into a terminal prompt) without programming experience.

***EDirect Example 7***

Find the taxonomic family name and BLAST division for a set of organisms. In Chapter 14 we explore eight model organisms. First make a text file listing these organisms (you can use a text editor to create a file by typing vim organisms.txt or nano organisms.txt, and you can find this resulting file at <http://bioinfbook.org>). Let’s use cat (catalog) to display the contents of this file.

```
$ cat organisms.txt
Escherichia coli
Saccharomyces cerevisiae
Arabidopsis thaliana
Caenorhabditis elegans
Drosophila melanogaster
Danio rerio
Mus musculus
Homo sapiens
```

Next write a shell script called taxonomy.sh (it is provided at the EDirect website at NCBI and also available on this book’s website).

```
$ cat taxonomy.sh
#!/bin/bash
#EDirect script
while read org
do
    esearch -db taxonomy -query "$org [LNGE] AND family [RANK]" < /dev/null |
        efetch -format docsum |
            xtract -pattern DocumentSummary -lbl "$org" -element ScientificName
Division
done
```

To execute this script we need appropriate permissions (see Box 2.3). We first use `ls -lh` (list the directory contents in the long format) to check the permissions on this file, then after changing the permissions it becomes executable.

```
$ ls -lh taxonomy.sh
-rw-rw-r--. 1 pevsner pevsner 244 Oct 17 17:00 taxonomy.sh
$ chmod ugo+rwx taxonomy.sh
$ ls -lh taxonomy.sh
-rwxr-xr-x. 1 pevsner pevsner 244 Oct 17 17:00 taxonomy.sh
```

The `x` (in the read/write/execute groups) indicates this is executable. We can now print the list of organisms (with the `cat` command), and pipe (`|`) the results to our shell script.

```
$ cat organisms.txt | ./taxonomy.sh
Escherichia coli Enterobacteriaceae enterobacteria
Saccharomyces cerevisiae Saccharomycetaceae ascomycetes
Arabidopsis thaliana Brassicaceae eudicots
Caenorhabditis elegans Rhabditidae nematodes
Drosophila melanogaster Drosophilidae flies
Danio rerio Cyprinidae bony fishes
Mus musculus Muridae rodents
Homo sapiens Hominidae primates
```

## ACCESS TO INFORMATION: GENOME BROWSERS

Genome browsers are databases with a graphical interface that presents a representation of sequence information and other data as a function of position across the chromosomes. We focus on viral, bacterial, archaeal, and eukaryotic chromosomes in Chapters 16–20. Genome browsers have emerged as essential tools for organizing information about genomes. We now briefly introduce three principal genome browsers (Ensembl, UCSC, and NCBI) and describe how they may be used to acquire information about a gene or protein of interest.

### Genome Builds

On using the UCSC, Ensembl, or other genome browsers there is a corresponding “genome build” for any organism being studied. A genome build refers to an assembly in which DNA sequence is collected and arranged to reflect the sequence along each chromosome. For a given organism’s genome, a build is released only occasionally (typically every few years). This build includes annotation, that is, the assignment of information such as the start and stop position of genes, exons, repetitive DNA elements, or other features. When you use a browser you should explore available genome builds. In some cases it is best to use the most recent available build. It is however common for earlier builds to have richer annotation, and very different categories of information are presented in different builds.

The Genome Reference Consortium (GRC) maintains the reference genomes for human, mouse, and zebrafish. The most recent human genome build is GRCh38 (sometimes called hg38), released in 2013. Previous builds were GRCh37 (also called hg19) in 2009 and GRCh36 (also called hg18) in 2006. Issues that must be addressed for any genome build include the following:

- What are the coordinates (start and end position) of each chromosome? For the human *HBB* gene spanning 1606 base pairs on chromosome 11, the start and end positions are given as chr11:5,246,696–5,248,301 in the GRCh37 build of February 2009, and chr11:5,203,272–5,204,877 in the previous build (NCBI36/hg18 of March 2006).
- How many gaps are there in the genome sequence, and can they be closed? Some regions such as the short arms of acrocentric chromosomes, telomeres, and

We discuss genome assembly in more detail in Chapters 9 and 15. NCBI describes the eukaryotic genome annotation process at [http://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/process/](http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) (WebLink 2.45).

The GRC website is <http://www.genomereference.org> (WebLink 2.46).

The MHC in humans is present on chromosome 6 from ~29.6 to 33.1 megabases of GRCh37/hg19.

The UCSC genome browser is available from the UCSC bioinformatics site at <http://genome.ucsc.edu> (WebLink 2.47). You can see examples of it in Figures 5.16 and 6.10. We encounter specialized versions such as browsers for Ebola virus (Chapter 16) and cancer (Chapter 21).

Ensembl (<http://www.ensembl.org>, WebLink 2.27) is supported by the Wellcome Trust Sanger Institute (WTSI; <http://www.sanger.ac.uk/>, WebLink 2.48) and the EBI (<http://www.ebi.ac.uk/>, WebLink 2.49). Ensembl focuses on vertebrate genomes, although its genome browser format is being adopted for the analysis of many additional eukaryotic genomes.

centromeres are so highly repetitive that it is extremely challenging to obtain an accurate sequence (see Chapter 8).

- How are structurally variant genomic loci represented? How are polymorphisms in nonfunctional sites (such as pseudogenes) represented? We define structural variants in Chapter 8.
- How many erroneous bases are present in a genome build, and how can they be identified and corrected? If a reference genome assembly is accurate to an error rate of 1 in 100,000 bases, then for 3 billion base pairs of sequence there are 30,000 expected errors. As reference genomes continue to be sequenced deeply (as described in Chapter 9) this error rate is expected to decline.

Some loci are challenging to represent in a genome build. An example is the major histocompatibility complex (MHC) which is so diverse in humans that there is no single consensus. Primary and alternate loci are defined, with some genes (such as *HLA-DRB3*) appearing only on the alternate locus. Patches are released (such as patch 10 abbreviated GRCh37.p10) which correct errors, represent alternative loci that occur due to allelic diversity, and also involve as few changes as possible to chromosomal coordinates.

## The University of California, Santa Cruz (UCSC) Genome Browser

The UCSC browser currently supports the analysis of three dozen vertebrate and invertebrate genomes, and is perhaps the most widely used genome browser for human and other prominent organisms such as mouse. The Genome Browser provides graphical views of chromosomal locations at various levels of resolution (from several base pairs up to hundreds of millions of base pairs spanning an entire chromosome). Each chromosomal view is accompanied by horizontally oriented annotation tracks. There are hundreds of available user-selected tracks in categories such as mapping and sequencing, phenotype and disease associations, genes, expression, comparative genomics, and genomic variation. These annotation tracks offer the Genome Browser tremendous depth and flexibility. Literature on the UCSC Genome Browser includes an overview of its function (Pevsner, 2009; Karolchik *et al.*, 2014), its resources for analyzing variation (Thomas *et al.*, 2007), its Table Browser (Karolchik *et al.*, 2004), and BLAT (Kent, 2002) (Chapter 5).

As an example of how to use the browser, go the UCSC bioinformatics site, click Genome Browser, set the clade (group) to Vertebrate, the genome to human, the assembly to March 2009 (or any other build date), and under “position or search term” type hbb (Fig. 2.12a). Click submit and you will see a list of known genes and a RefSeq gene entry for beta globin on chromosome 11 (Fig. 2.12b). By following this RefSeq link you can view the beta globin gene (spanning about 1600 base pairs) on chromosome 11, and can perform detailed analyses of the beta globin gene (including neighboring regulatory elements), the messenger RNA (see Chapter 8), and the protein (Fig. 2.12c).

## The Ensembl Genome Browser

The Ensembl project offers a series of comprehensive websites emphasizing a variety of eukaryotic organisms (Flicek *et al.*, 2014). To many users, it is comparable in scope and importance to the UCSC Genome Browser, and it is often useful for new users to visit both sites. The Ensembl project’s goals are to automatically analyze and annotate genome data (see Chapter 15) and to present genomic data via its web browser.

We can begin to explore Ensembl from its home page by selecting *Homo sapiens* and performing a text search for “hbb,” the gene symbol for beta globin. This yields a link to the beta globin protein and gene; we will return to the Ensembl resource in later chapters. This entry contains a large number of features relevant to HBB, including identifiers, the

(a) Specifying the genome, assembly, and gene (or region or feature)

group: Mammal; genome: Human; assembly: Feb. 2009 (GRCh37/hg19); position: chr21:33,031,597-33,041,570; search term: hbb. Buttons include submit, track search, add custom tracks, track hubs, and configure tracks and display.

[Click here to reset](#) the browser user interface settings to their defaults.

(b) Selecting a gene

### UCSC Genes

HBB (uc001mae.1) at chr11:5246696-5248301 - Homo sapiens hemoglobin, beta (HBB), mRNA.  
 HBD (uc001maf.1) at chr11:5254059-5255858 - Homo sapiens hemoglobin, delta (HBD), mRNA.  
 RBM17 (uc010qav.2) at chr10:6131309-6159422 - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 2, mRNA.  
 RBM17 (uc001jzb.3) at chr10:6130949-6159422 - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 1, mRNA.  
 HBA1 (uc002cfx.1) at chr16:226679-227520 - Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.  
 HBA2 (uc002cfv.4) at chr16:222846-223709 - Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA.  
 HBBP1 (uc001mag.3) at chr11:5263185-5264822 - Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding RNA.  
 TMEM158 (uc011baf.2) at chr3:45265956-45267814 - Homo sapiens transmembrane protein 158 (gene/pseudogene) (TMEM158), mRNA.

### RefSeq Genes

HBB at chr11:5246696-5248301 - (NM\_000518) hemoglobin subunit beta  
 HBBP1 at chr11:5263185-5264822 - (NR\_001589)

(c) Genome browser

### UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

Position: chr11:5,246,696-5,248,301 1,606 bp. Search bar: enter position, gene symbol or search terms. Go button.

Tracks displayed:

- Scale: chr11: 5,247,888 | hg19: 5,247,500
- UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
- RefSeq Genes
- SNPs
- Human mRNAs
- Mammal Cons
- Rhesus
- Mouse
- Dog
- Elephant
- Opossum
- Chicken
- X\_tropicalis
- Zebrafish
- Common SNPs(137)
- RepeatMasker

Annotations include:

- 500 bases
- Publications: Sequences in scientific articles
- Human mRNAs from GenBank
- Placental Mammal Baseewise Conservation by PhyloP
- Multiz Alignments of 46 Vertebrates
- Simple Nucleotide Polymorphisms (dbSNP 137) Found in > 1% of Samples
- Repeating Elements by RepeatMasker

Control buttons: move start, move end, track search, default tracks, default order, hide all, add custom tracks, track hubs, configure, reverse, resize, refresh.

**FIGURE 2.12** Using the UCSC Genome Browser. (a) Select from dozens of organisms (mostly vertebrates) and assemblies, then enter a query such as “beta globin” (shown here) or an accession number or chromosomal position. (b) By clicking submit, a list of known genes as well as RefSeq genes is displayed. (c) Following the link to the RefSeq gene for beta globin, a browser window is opened showing 1606 base pairs on human chromosome 11. A series of horizontal tracks are displayed including a list of RefSeq genes and Ensembl gene predictions; exons are displayed as thick bars, and arrows indicate the direction of transcription (from right to left, toward the telomere or end of the short arm of chromosome 11).

Source: UCSC Genome Browser (<http://genome.ucsc.edu>). Courtesy of UCSC.

**TABLE 2.9 Ensembl stable identifiers. For human entries the prefix is ENS, while other common species prefixes include ENSBTA (cow *Bos taurus*), ENSMUS (mouse *Mus musculus*), ENSRNO (rat *Rattus norvegicus*) and FB (fruit fly *Drosophila melanogaster*).**

Feature prefix	Definition	Human beta globin example
E	exon	ENSE00001829867
FM	protein family	ENSM00250000000136
G	gene	ENSG00000244734
GT	gene tree	ENSGT00650000093060
P	protein	ENSP00000333994
R	regulatory feature	ENSR00000557622
T	transcript	ENST00000335295

Source: Ensembl Release 76; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

DNA sequence, and convenient links to many other database resources. Ensembl offers a set of stable identifiers (**Table 2.9**).

### The Map Viewer at NCBI

The Map Viewer is accessed from the main page of NCBI or via <http://www.ncbi.nlm.nih.gov/mapview/> (WebLink 2.50). Records in NCBI Gene, Nucleotide, and Protein also provide direct links to the Map Viewer.

The NCBI Map Viewer includes chromosomal maps (both physical maps and genetic maps; see Chapter 20) for a variety of organisms including metazoans (animals), fungi, and plants. Map Viewer allows text-based queries (e.g., “beta globin”) or sequence-based queries (e.g., BLAST; see Chapter 4). For each genome, four levels of detail are available: (1) the home page of an organism; (2) the genome view, showing ideograms (representations of the chromosomes); (3) the map view, allowing you to view regions at various levels of resolution; and (4) the sequence view, displaying sequence data as well as annotation of interest such as the location of genes.

Entries in NCBI’s Gene resource include access to the graphical viewer. We will return to this browser in later chapters. Visit the *HBB* entry of NCBI Gene (Fig. 2.9.), scroll to the viewer, and try the Tools and Configure pull-downs to begin exploring its features.

## EXAMPLES OF HOW TO ACCESS SEQUENCE DATA: INDIVIDUAL GENES/PROTEINS

We next explore two practical problems in accessing data: human histones and the Human Immunodeficiency Virus-1 (HIV-1) pol protein. Each presents distinct challenges.

### Histones

By viewing the search details on an NCBI Protein query, you can see that the command is interpreted as “txid9606[Organism:exp] AND histone[All Fields].” The Boolean operator AND is included between search terms by default.

The biological complexity of proteins can be astonishing, and accessing information about some proteins can be extraordinarily challenging. Histones are among the most familiar proteins by name. They are small proteins (12–20 kilodaltons) that are localized to the nucleus where they interact with DNA. There are five major histone subtypes as well as additional variant forms; the major forms serve as core histones (the H2A, H2B, H3, and H4 families), which ~147 base pairs of DNA wrap around, and linker histones (the H1 family). Suppose you want to inspect a typical human histone for the purpose of understanding the properties of a representative gene and its corresponding protein; the challenge is that there are currently 470,000 histone entries in NCBI Protein (April 2015).

The output can be restricted to a species or other taxonomic group of interest from the NCBI Protein site or from the Taxonomy Browser. Each organism or group in GenBank

(e.g., kingdom, phylum, order, genus, species) is assigned a unique taxonomy identifier. Following the link to *Homo sapiens*, the identifier 9606, the lineage, and a summary of available Entrez records can be found (**Fig. 2.6**).

Using the NCBI Protein search string (“txid9606[Organism:exp] histone”) there are currently over 8000 human histone proteins of which >2000 have RefSeq accession numbers. Some of these are histone deacetylases and histone acetyltransferases; by expanding the query to “txid9606[Organism:exp] AND histone[All Fields] NOT deacetylase NOT acetyltransferase” there are over 1700 proteins with RefSeq accession numbers.

How can the search be further pursued?

1. The NCBI Gene entry for any histone offers a brief summary of the family, provided by RefSeq. We saw an example for globins in **Figure 2.9**.
2. You could select a histone at random and study it, although you may not know whether it is representative.
3. There are specialized, expert-curated databases available online for many genes, proteins, diseases, and other molecular features of interest. The Histone Sequence Database (Mariño-Ramírez *et al.*, 2011) shows that the human genome has about 113 histone genes, including a cluster of 56 adjacent genes on chromosome 6p. This information is useful to understand the scope of the family.
4. There are databases of protein families, including Pfam and InterPro. We introduce these in Chapter 6 (multiple sequence alignment) and Chapter 12 (proteomics). Such databases offer succinct descriptions of protein and gene families and can orient you toward identifying representative members.

## HIV-1 pol

Consider reverse transcriptase, the RNA-dependent DNA polymerase of HIV-1 (Frankel and Young, 1998). The gene encoding reverse transcriptase is called *pol* (for polymerase). How do you obtain its DNA and protein sequence?

From the home page of NCBI enter “hiv-1” (do not use quotation marks; the use of capital letters is optional). All Entrez databases are searched. Under the Nucleotide category, there are over half a million entries. Click Nucleotide to see these entries. Over 3000 entries have RefSeq identifiers; while this narrows the search considerably, there are still too many matches to easily find HIV-1 *pol*. One reason for the large number of entries in NCBI Nucleotide is that the HIV-1 genome has been re-sequenced thousands of times in efforts to identify variants. Another reason for the many hits is that entries for a variety of organisms, including mouse and human, refer to HIV-1 and are therefore listed in the output.

We can again use the species filter and restrict the output to HIV. There is now only one RefSeq entry (NC\_001802.1). This entry refers to the 9181 bases that constitute HIV-1, encoding just nine genes including *gag-pol*. Given the thousands of HIV-1 *pol* variants that exist this example highlights the usefulness of the RefSeq project, allowing the research community to have a common reference sequence to explore.

As alternative strategies, from the Entrez results for HIV-1 select the genome, assembly, or taxonomy page to link to the single NCBI Genome record for HIV-1 and, through the genome annotation report, find a table of the nine genes (and nine proteins) encoded by the genome. Each of these nine NCBI Genome records contains detailed information on the genes; in the case of *gag-pol*, there are seven separate RefSeq entries, including one for the *gag-pol* precursor (NP\_057849.4, 1435 amino acids in length) and one for the mature HIV-1 *pol* protein (NP\_789740.1, 995 amino acids).

Note that other NCBI databases are not appropriate for finding the sequence of a viral reverse transcriptase: UniGene does not incorporate viral records, while OMIM is

The Histone Sequence Database is available at <http://research.nhgri.nih.gov/histones/> (WebLink 2.51). It was created by David Landsman, Andy Baxevanis, and colleagues at the National Human Genome Research Institute.

You can find links to a large collection of specialized databases at <http://www.expasy.org/links.html> (WebLink 2.52), the Life Science Directory at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB).

We explore bioinformatics approaches to HIV-1 in detail in Chapter 16 on viruses.

As of October 2014 there are over 600,000 entries in NCBI Nucleotide for the query “hiv-1.”

We will see that BLAST searches (Chapter 4) can be limited by any Entrez query; you can enter the taxonomy identifier into a BLAST search to restrict the output to any organism or taxonomic group of interest.

From the NCBI Genome or other Entrez pages, try exploring the various options. For example, for the NCBI Genome entry for NC\_001802.1 you can display a convenient protein table; from NCBI Nucleotide or Protein you can select Graph to obtain a schematic view of the HIV-1 genome and the genes and proteins it encodes. The table of nine proteins is available at [http://www.ncbi.nlm.nih.gov/genome/proteins/10319?project\\_id=15476](http://www.ncbi.nlm.nih.gov/genome/proteins/10319?project_id=15476) (WebLink 2.53).

A 2011 issue of the journal *Database* is dedicated to BioMart. See [http://www.oxfordjournals.org/our\\_journals/database/biomart\\_virtual\\_issue.html](http://www.oxfordjournals.org/our_journals/database/biomart_virtual_issue.html) (WebLink 2.54).

A “relational schema” refers to the use of a relational database. Ensembl stores its data in a popular relational database called MySQL (<http://www.mysql.com>, WebLink 2.55). Web Document 2.3 shows a schema of the tables used at Ensembl (from [http://useast.ensembl.org/info/docs/api/core/core\\_schema.html](http://useast.ensembl.org/info/docs/api/core/core_schema.html), WebLink 2.56).

The UCSC Table Browser can be reached via <http://genome.ucsc.edu/cgi-bin/hgTables> (WebLink 2.57).

limited to human entries (e.g., human genes implicated in susceptibility to HIV infection). UniGene and OMIM do however have links to genes that are related to HIV, such as eukaryotic reverse transcriptases.

## HOW TO ACCESS SETS OF DATA: LARGE-SCALE QUERIES OF REGIONS AND FEATURES

### Thinking About One Gene (or Element) Versus Many Genes (Elements)

In many cases we are interested in a single gene. Throughout this book we focus on the beta globin gene (*HBB*) and the hemoglobin protein as a prototypical example of a gene and an associated protein product.

In many other cases we want to know about large collections of genes, proteins, or indeed any other element.

- What is the complete set of human globin genes?
- To which chromosomes are they assigned?
- How many exons are on chromosome 11, and how many repeat elements occur in each exon?

It would be tedious, inefficient, and error-prone to collect information one gene at a time. There are many bioinformatics tools that allow us to collect genome-wide information. We will focus on two sources: the Ensembl database (including the BioMart resource); and the UCSC Genome Browser (and Table Browser). These are complementary, equally useful resources that offer powerful search options. They differ significantly in format, and offer access to large datasets that are closely related but not exactly the same. Each can be accessed via Galaxy, also introduced below.

### The BioMart Project

The BioMart offers easy access to a vast amount of information in multiple databases. This project is based on two principles (Kasprzyk, 2011). The first is its “data agnostic modeling”: very large numbers of datasets are imported from assorted domains (including third-party databases), and a relational schema is employed to access data. This relational schema allows a query (such as a gene name or chromosomal locus) to be connected to associated information (such as annotation of gene structure), even if the information originated in projects that modeled the data in different ways. The second principle is data federation: many distributed databases are organized into a single, integrated, virtual database. When you use BioMart it is therefore possible to search information relevant to hundreds of resources (including topics we have described in this chapter such as RefSeq, Ensembl, HGNC, LRG, UniProt, and CCDS) while BioMart functions as a single database resource.

We will explore two different ways to extract information from BioMart in Computer Lab problems 2.4–2.6 below. Later we approach BioMart through the R package `biomaRt` (Chapter 8).

### Using the UCSC Table Browser

The UCSC Table Browser is equally important and useful as the corresponding Genome Browser (Karolchik *et al.*, 2014). The Table Browser enables accurate, complete tabular descriptions of the same data that can be visualized in the Genome Browser. These tables can be downloaded, viewed, and queried. For example, set the genome to human (clade: Mammal; genome: Human; assembly: GRCh37/hg19; **Fig. 2.13a**, arrow 1), and choose a

(a)

**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal   genome: Human   assembly: Feb. 2009 (GRCh37/hg19) ← 1

group: Genes and Gene Prediction Tracks   track: RefSeq Genes   add custom tracks   track hubs

table: refGene   describe table schema

region:  genome  ENCODE Pilot regions  position chr11:5240001-5300000   lookup   define regions ← 2

identifiers (names/accessions): paste list   upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table   Send output to  Galaxy  GREAT ← 4

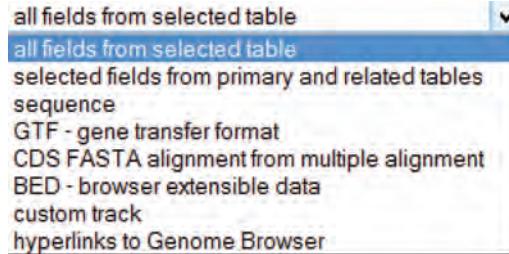
output file: (leave blank to keep output in browser)

file type returned:  plain text  gzip compressed

get output   summary/statistics ← 5

To reset **all** user cart settings (including custom tracks), [click here](#).

(b)



(c)

chr11	5246695	5248301	NM_000518	0	-	5246827	5248251	0	3	261,223,142,	0,1111,1464,
chr11	5254058	5255858	NM_000519	0	-	5254193	5255663	0	3	264,223,287,	0,1162,1513,
chr11	5263184	5264822	NR_001589	0	-	5264822	5264822	0	3	293,223,143,	0,1151,1495,
chr11	5269501	5271087	NM_000559	0	-	5269588	5271034	0	3	216,223,145,	0,1096,1441,
chr11	5274420	5276011	NM_000184	0	-	5274506	5275958	0	3	215,223,145,	0,1101,1446,
chr11	5289579	5291373	NM_005330	0	-	5289698	5291120	0	3	248,223,345,	0,1104,1449,

**FIGURE 2.13** The University of California, Santa Cruz (UCSC) Genome Browser offers a complementary Table Browser that is equally useful. (a) The Table Browser includes options to select the clade, genome, and assembly (arrow 1), for example GRCh37 (also called hg19). (We discuss human genome assemblies in Chapter 20.) Groups (e.g., genes) and tracks (e.g., RefSeq genes) and a region of interest (arrow 2) can be selected. Note that in the position box (arrow 3) you can enter a gene name (e.g., hbb), click “lookup”, and those genomic coordinates will be entered. Next, choose the output format (arrow 4). Click “summary statistics” (arrow 5) for a summary of how many elements occur in your query, or click “get output” for full results. (b) Examples of available output formats. These typically lead to a further webpage offering additional options (e.g., sequence can include DNA or protein; a BED file can include a whole gene, coding exons, or other options). (c) Example of a BED file output. Such files are versatile and can be used for many further analyses, for example using next-generation sequencing software (described in Chapter 9).

Source: UCSC Genome Browser (<http://genome.ucsc.edu>). Courtesy of UCSC.

track such as RefSeq genes. A region of interest such as the entire human genome, the ENCODE region (introduced in Chapter 8), or a user-selected genomic region (**Fig. 2.13a**, arrow 2) can be defined. In the position box (arrow 3) you can also type the name of a gene of interest. The output format can be set to BED (browser extensible data; see below) or several other formats (**Fig. 2.13b**). Note that by checking the Galaxy or Great links (arrow 4) you can send the results to other programs. **Fig. 2.13c** shows the output for this particular query in the BED format. For any Table Browser query, you can get a summary of the size of the output (arrow 5) or click “get output” to return the results to a plain text html or (if you prefer) to a compressed file.

### Custom Tracks: Versatility of the BED File

Genome browsers display many categories of information about chromosomal features, including genes, regulatory regions, variation, and conservation. There are two main reasons we might want to customize this information: either to obtain selected types of information (e.g., all microRNA genes within a particular distance from a set of exons), or to upload information that we are interested in (e.g., results from a microarray experiment showing which RNA transcripts are regulated in our experiment, or many other types of data we acquire experimentally).

We will also encounter BED files as we analyze next-generation sequence data (Chapter 9). BED files include information from DNA sequencing experiments (as well as RNA sequencing or RNA-seq studies). We explore BEDTools software that analyzes BED files in a variety of ways, for example showing regions of overlap.

There are many file formats for custom tracks. The BED file (shown in **Fig 2.13c** as a Table Browser output) is one of the most popular. It can be uploaded to UCSC for visualization in the Genome Browser and/or for analysis in the Table Browser. It includes three required fields (columns): chromosome, start position, and end position. Additional, optional fields are as follows:

- Column 4: name. In our example the RefSeq identifiers are given. (One way you could learn the corresponding gene names is to input that list into BioMart.)
- Column 5: score. This ranges from 0 to 1000, with higher scores displayed as increasing shades of gray.
- Column 6: strand. These are all the minus strand (–) in our example.
- Columns 7, 8: thickStart and thickEnd. It is sometimes useful to display subportions of an entry with thick lines, such as coding regions within genes.
- Column 9: itemRgb. The Red Blue Green (RGB) value (such as 0, 255, 0) specifies the color of the output.
- Columns 10–12: blockCount, blockSizes, blockStarts. These display the number of blocks (e.g., exons) in each row, the block sizes, and the block start positions.

Details of the BED format are provided at <http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED> (WebLink 2.58).

For examples of file formats visit <http://bioinfbook.org> and see Web Document 2.4. UCSC lists many publicly available custom tracks at <http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html> (WebLink 2.59). Extensive help on custom tracks is given at <http://genome.ucsc.edu/goldenPath/help/customTrack.html> (WebLink 2.60).

Many custom file formats are supported by Ensembl and UCSC (**Table 2.10**). For each, we provide a web document allowing you to further explore it.

There are several caveats to using custom files. First, be careful to check whether the chromosome should be specified as a number (e.g., 11 for chromosome 11) or with the prefix chr (e.g., chr11 as in **Fig. 2.13.c**). Second, be careful to check whether the counting is zero-based or one-based (0-based or 1-based; **Table 2.11**). We explain these counting schemes in Box 2.5. For the UCSC Genome Browser, which uses 1-based counting, the first nucleotide of the *HBB* gene begins on chromosome 11 at nucleotide position 5,246,696; however, using the UCSC Table Brower the starting position is 5,246,695. This is not an error, but exemplifies how two different counting schemes are commonly employed. Of course, a one-nucleotide difference can be crucially important when you are analyzing genomic variants.

**TABLE 2.10 File formats for custom tracks used at Ensembl and/or UCSC. Two definitions of GTF (from Ensembl and UCSC) are given.**

File Format	Definition	Typical file size
BAM		Any size; often millions of rows
BED	Browser extensible data	Any size; often dozens to thousands or millions of rows
BedGraph		Any size
bigBed		
GFF/GTF	General feature format, General transfer format Gene transfer format	Any size
MAF		
PSL		Any size
WIG	Wiggle	Any size
BAM	Binary alignment/map	Very large
BigWig		Very large
VCF	Variant call format	Very large

## Galaxy: Reproducible, Web-Based, High-Throughput Research

Galaxy is a web-based analysis platform that accepts input from a variety of sources including BioMart and the UCSC Table Browser. Visit the Galaxy site and note that there are three panels: tools (at left), display (at center), and history (at right). The main advantages of Galaxy are:

1. it provides a large, integrated collection of software tools to import a variety of data types (particularly large, high-throughput datasets) and analyze them;
2. it is web-based, providing access to many software packages that are otherwise available only in the command-line environment (for those learning about these tools it provides ready access to at least a simple version of the software); and
3. it fosters reproducible research because the analysis steps you follow may be documented, stored, and shared with others.

The Galaxy Team has written articles on how to use Galaxy (Blankenberg *et al.*, 2011; Goecks *et al.*, 2010, 2013; Hillman-Jackson *et al.*, 2012), including its use in next-generation sequence analysis (Goecks *et al.*, 2012) and its Tool Shed and Tool Factory (Lazarus *et al.*, 2012).

**TABLE 2.11 One-based and zero-based counting.**

Resource	System	WebLink
Python	0-based	
UCSC browser in BED or other format	0-based	
UCSC data returned in BED or other format	0-based	
BAM files (Chapter 9)	0-based	<a href="http://samtools.sourceforge.net/SAM1.pdf">http://samtools.sourceforge.net/SAM1.pdf</a> (WebLink 2.88)
Ensembl	1-based	<a href="http://www.ensembl.org/Help/Faq?id=286">http://www.ensembl.org/Help/Faq?id=286</a> (WebLink 2.89)
UCSC browser in coordinate format	1-based	<a href="http://genome.ucsc.edu/FAQ/FAQtracks.html">http://genome.ucsc.edu/FAQ/FAQtracks.html</a> (WebLink 2.90)
BLAST (Chapter 4)	1-based	
GFF files (Chapter 9)	1-based	
VCF files (Chapter 9)	1-based	<a href="http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41">http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41</a> (WebLink 2.91)

Source: <http://alternateallele.blogspot.com/2012/03/genome-coordinate-cheat-sheet.html> (WebLink 2.92).

## BOX 2.5. 0-BASED AND 1-BASED COUNTING

Counting nucleotide positions is surprisingly complicated. If we enter HBB into the UCSC Genome Browser (GRCh37/hg19 build), we can see that this gene spans 1606 base pairs at coordinates chr11:5,246,696–5,248,301. But if you then link to the UCSC Table Browser, choose this position (chr11:5,246,696–5,248,301) under the Region option, and select the BED (browser extensible data) output format, the result is chr11:5,246,695–5,248,301. Try it for any gene or locus! The first position now ends in a 5 rather than a 6, indicating a discrepancy of one base pair. Why?

There are two different ways to count coordinate positions. The first is one-based (or 1-based) counting, in which the first base has position 1. Let's use the example of the (hypothetical) nucleotide string GATCG at the beginning of chromosome 1. This would have the position chr1:1–5. The interval length is end – begin + 1 (here 5 – 1 + 1 = 5). The nucleotides TCG occur at positions 3–5. Such straightforward 1-based counting is used in the Ensembl and UCSC Genome Browsers as well as GFF, GTF, and VCF files that we describe in Chapter 9 (these provide information about variants in a genome). BLAST (Chapters 3–5) uses 1-based counting, as does the R programming language. The advantage of 1-based counting is that it is intuitive and most of us are used to it. The disadvantage is that if you want to know the length of the interval, subtracting the lowest value (1) from the highest value (5) yields a length of 4, which is not correct.

An alternate way to count is zero-based (or 0-based) counting. This is implemented in BED files that are part of the UCSC Genome browser, as well as other formats in which genomic data are presented. BAM/SAM files (Chapter 9) which represent nucleotide sequences aligned to a genome reference are 0-based, as is Python. For our simple example, 0-based coordinates of GATCG would be chr1:0–4. The end is at position 5, so the interval length is end – begin (here 5 – 0 = 5). Subtracting the value 0 from 5 yields the correct result of length 5 for this string.

**Table 2.11** lists several resources that use either 0-based or 1-based counting. The 0-based BED format is also “half-open.” This means that the start position is inclusive, but the end position is not. For the region of five nucleotides that spans positions 1:5 in a 1-based format, in the 0-based BED the start position is position 0 while the end position is 5.

Visit Galaxy at <http://usegalaxy.org> (WebLink 2.61).

To try Galaxy, select “Get Data” from the list of tools then choose data from the UCSC Table Browser, which becomes available in the central Galaxy panel. Select beta globin (hbb), set the format to sequence, choose protein sequence, and send the output back to Galaxy. There the sequence will appear in the history panel at right; by clicking the eye icon you can display it. Then you can select from hundreds of tools to further analyze it.

We will encounter Galaxy in several contexts:

- We can extract protein sequences (e.g., from UCSC) and perform pairwise alignment (problem 3.3 in Chapter 3).
- It is useful to explore genomic DNA alignments (Chapter 6).
- In exploring chromosomes we extract human microsatellites; we will create a table including their genomic coordinates and sort the results to find which is longest (Chapter 8, problem 8.1).
- In analyzing next-generation sequence data (Chapter 9) we can import FASTQ files (and assess their quality using the FASTQC program within Galaxy), perform alignments, and analyze BAM and VCF files (introduced in Chapter 9).
- Galaxy is popular for its suite of RNA-seq analysis tools; command-line software such as Bowtie and BWA that we introduce in Chapter 11 is also available in Galaxy.

The NLM website is <http://www.nlm.nih.gov/> (WebLink 2.61), and PubMed is at <http://www.ncbi.nlm.nih.gov/pubmed/> (WebLink 2.63). Over 2.5 billion MEDLINE/PubMed searches were performed in 2013 (see [http://www.nlm.nih.gov/bsd/bsd\\_key.html](http://www.nlm.nih.gov/bsd/bsd_key.html), WebLink 2.64).

A PubMed tutorial is offered at [http://www.nlm.nih.gov/bsd/pubmed\\_tutorial/m1001.html](http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html) (WebLink 2.65).

## ACCESS TO BIOMEDICAL LITERATURE

The National Library of Medicine (NLM) is the world’s largest medical library. In 1971 the NLM created MEDLINE (Medical Literature, Analysis, and Retrieval System Online), a bibliographic database. MEDLINE currently contains over 24 million references to journal articles in the life sciences with citations from over 5600 biomedical journals. Free access to MEDLINE is provided through PubMed, which is developed by NCBI. While MEDLINE and PubMed both provide bibliographic citations, PubMed also contains links to online full-text journal articles. PubMed also provides access and links to the integrated molecular biology databases maintained by NCBI. These databases contain DNA and protein sequences, genome-mapping data, and three-dimensional protein structures.

## Example of PubMed Search

A search of PubMed for information about “beta globin” (in quotation marks) yields ~6700 entries. Box 2.6 describes the basics of using Boolean operators in PubMed. There are many additional ways to limit this search. Use filters (on the left sidebar) and try applying features such as restricting the output to articles that are freely available through PubMed Central.

The Medical Subject Headings (MeSH) browser provides a convenient way to focus or expand a search. MeSH is a controlled vocabulary thesaurus containing over 26,000 descriptors (headings). From PubMed (or from the main NCBI homepage), select MeSH and enter “beta globin.” The result suggests a series of possibly related topics including one for “beta-Globins.” By adding MeSH terms, a search can be focused and structured according to the specific information you seek. Lewitter (1998) and Fielding and Powell (2002) discuss strategies for effective MEDLINE searches, such as avoiding inconsistencies in MeSH terminology and finding a balance between sensitivity (i.e., finding relevant articles) and specificity (i.e., excluding irrelevant citations). For example, for a subject that is not well indexed, it is helpful to combine a text keyword with a MeSH term. It can also be helpful to use truncations; for example, the search “therap\*” introduces a wildcard that will retrieve variations such as therapy, therapist, and therapeutic.

The growth of MEDLINE is described at [http://www.nlm.nih.gov/bsd/index\\_stats\\_comp.html](http://www.nlm.nih.gov/bsd/index_stats_comp.html) (WebLink 2.66). Despite the multinational contributions to MEDLINE, the percentage of articles written in English has risen from 59% at its inception in 1966 to 93% in the year 2014  
[http://www.nlm.nih.gov/bsd/medline\\_lang\\_distr.html](http://www.nlm.nih.gov/bsd/medline_lang_distr.html) (WebLink 2.67).

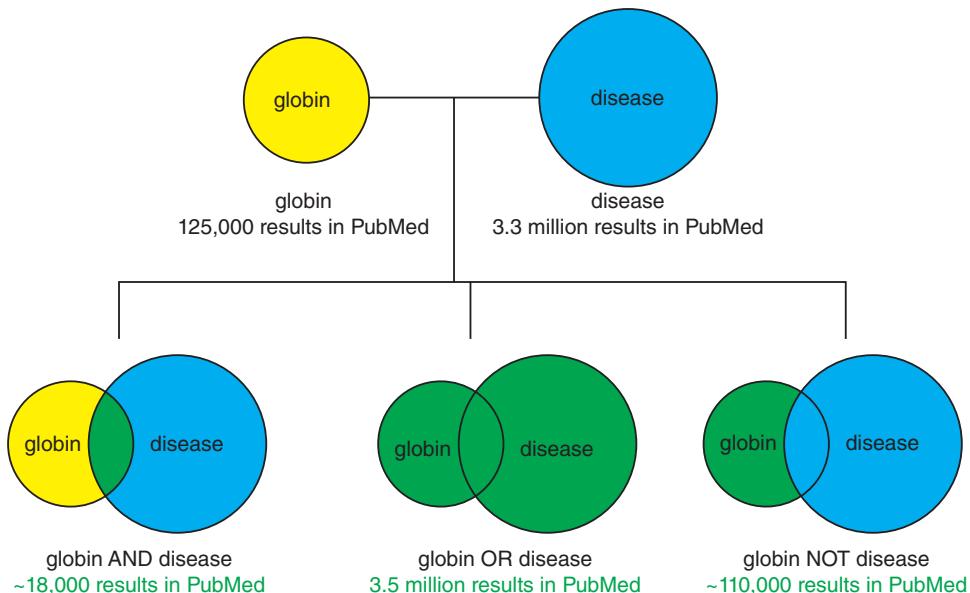
The MeSH website is at <http://www.ncbi.nlm.nih.gov/mesh> (WebLink 2.68); you can also access MeSH via the NCBI website including its PubMed page.

## PERSPECTIVE

Bioinformatics is an emerging field whose defining feature is the accumulation of biological information in databases. The three major traditional DNA databases – GenBank, EMBL-Bank, and DDBJ – are adding several million new sequences each year as well as billions of nucleotides. At the same time, next-generation sequencing technology is

### BOX 2.6 VENN DIAGRAMS OF BOOLEAN OPERATORS AND, OR, AND NOT FOR HYPOTHETICAL SEARCH TERMS 1 AND 2

The AND command restricts the search to entries that are both present in a query. The OR command allows either one or both of the terms to be present. The NOT command excludes query results. The green areas represent search queries that are retrieved. Examples are provided for the queries “globin” or “disease” in PubMed. The Boolean operators affect the searches as indicated.



producing vastly greater amounts of DNA. A single lab that is sequencing ten human genomes might generate a trillion base pairs of DNA sequences (a terabase) within a month.

In this chapter, we have described ways to find information on the DNA and/or protein sequences of individual genes (using beta globin as an example) as well as sets of genes. Many other databases and resources are available, some as websites and some (such as R packages or NCBI E-Utilities) via programming languages. Increasingly, there is no single correct way to find information; many approaches are possible. Moreover, resources such as those described in this chapter (e.g., NCBI, ExPASy, EBI/EMBL, and Ensembl) are closely interrelated, providing links between the databases.

## PITFALLS

There are many pitfalls associated with the acquisition of both sequence and literature information. In any search, the most important first step is to define your goal: for example, decide whether you want protein or DNA sequence data. A common difficulty that is encountered in database searches is receiving too much information; this problem can be addressed by learning how to generate specific searches with appropriate limits.

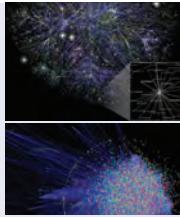
It is surprising how often students begin studying the wrong gene. It is a good idea to visit the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) website (<http://www.genenames.org>, WebLink 2.69). This shows the official gene symbol for human genes, with links to key resources such as Ensembl and NCBI. Given a list of gene symbols of interest, you can upload them in a text file to BioMart to confirm all symbols are correct.

## ADVICE FOR STUDENTS

I recommend that you visit the major bioinformatics websites (EBI, NCBI, Ensembl, UCSC) and spend many hours exploring each one. Some students have a favorite protein, gene, pathway, disease, organism, or other topic. If so, learn all you can about your favorite topic; within reason you should know all that can be known about it. If you don't have a particular topic, keep focused on our example of beta globin, a famous gene/protein that is well characterized. Try to practice studying one gene at a time versus a group of genes (or proteins or other molecules). When we mention performing batch queries on BioMart, try it yourself. Later we will work with high-throughput datasets that contain thousands or even many millions of rows of data, and it can be just as easy to query 100 objects (such as accession numbers) as a million. When you have questions, try Biostars (<http://www.biostars.org>; WebLink 2.70) to see if others have posed similar questions, or sign up and post your own.

## WEB RESOURCES

You can visit the website for this book (<http://www.bioinfbook.org>) to find WebLinks; Web Documents; PowerPoint, PDF, and audiovisual files of lectures; and additional URLs. Major sites often offer portals that are rich in information such as training and site overviews. These include sites within Ensembl (<http://www.ensembl.org/info/>, WebLink 2.71), EBI (<http://www.ebi.ac.uk/training/>, WebLink 2.72), NCBI (<http://www.ncbi.nlm.nih.gov/guide/training-tutorials/>, WebLink 2.73), and UCSC Genome Bioinformatics (<http://genome.ucsc.edu/training.html>, WebLink 2.74). For literature searches, the National Library of Medicine offers a PubMed tutorial (<http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/>, WebLink 2.75) and excellent online training resources (<http://www.nlm.nih.gov/bsd/disted/pubmed.html>, WebLink 2.76).



## Discussion Questions

**[2-1]** What categories of errors occur in databases? How are these errors assessed?

**[2-2]** How is quality control maintained in GenBank, given that thousands of individual investigators submit data?

### PROBLEMS/COMPUTER LAB

**[2-1]** The purpose of this problem is to introduce you to using Entrez and related NCBI resources. How many human proteins are bigger than 300,000 daltons? What is the longest human protein? There are several different ways to solve these questions.

- (1) From the home page of NCBI select the alphabetical list of resources or the pull-down menu, find Protein, and use the filter on the left sidebar to limit entries to human.
- (2) Enter a command in the format `xxxxxx:yyyyyy[molwt]` to restrict the output to a certain number of daltons; for example, `002000:010000[molwt]` will select proteins of molecular weight 2000–10,000.
- (3) As a different approach, search `30000:50000[Sequence Length]`
- (4) You can read more about titin (NP\_596869.4), the longest human protein, at NCBI Gene (<http://www.ncbi.nlm.nih.gov/gene/7273>, WebLink 2.77). While the average protein has a length of several hundred amino acids, incredibly titin is 34,423 amino acids in length.
- (5) Explore additional ways to limit Entrez searches by using an NCBI Handbook chapter (<http://www.ncbi.nlm.nih.gov/books/NBK44864/>, WebLink 2.78).

**[2-2]** The purpose of this problem is to obtain information from the NCBI website. The RefSeq accession number of human beta globin protein is NP\_000509. Go to NCBI (<http://www.ncbi.nlm.nih.gov/>). What is the RefSeq accession number of beta globin protein from the chimpanzee (*Pan troglodytes*)?

- (1) There are several different ways to solve this. Try typing chimpanzee globin into the home page of NCBI; or use the species limiter of NCBI Protein, or use the Taxonomy Browser to find chimpanzee NCBI Gene entries.
- (2) HomoloGene (<http://www.ncbi.nlm.nih.gov/homologene>, WebLink 2.38) is a great resource to learn about sets of related eukaryotic proteins. Use HomoloGene to find a set of beta globins including chimpanzee.

**[2-3]** The purpose of this exercise is to become familiar with the EBI website and how to use it to access information.

- (1) Visit the site (<http://www.ebi.ac.uk/>, WebLink 2.5). Enter hemoglobin beta in the main query box (alternatively, use the query human hemoglobin beta).
- (2) Inspect the results. Explore the various links to information about pathways, genomes, nucleotide and protein sequences, structures, protein families, and more.

**[2-4]** Accessing information from BioMart: the beta globin locus.

- (1) Go to <http://www.ensembl.org> and follow the link to BioMart.
- (2) First choose a database; we will select Ensembl Genes 71.
- (3) Choose a dataset: Homo sapiens genes (GRCh37.p10). Note the other available datasets.
- (4) Choose a filter. Here the options include region, gene, transcript event, expression, multispecies comparisons, protein domains, and variation. Select “region”, chromosome 11, and enter 5240000 for the Gene Start (base pairs) and 5300000 for the Gene End. (Note that this region spans 60 kilobases and corresponds to chr11:5,240,001–5,300,000.)
- (5) Choose attributes. Select the following features. Under “Gene” select Ensembl Gene ID and %GC content; under “External” select the external references CCDS ID, HGNC symbol (this is the official gene symbol), and HGNC ID(s).
- (6) At the top left select “Count.” Currently there are 8 genes matching these criteria.
- (7) To view these results select “Results.” Note that you can export your results in several formats (including a comma separated values or CSV file) that can be further manipulated (e.g., converted to a BED file).

**[2-5]** BioMart: working with lists. The goal of this exercise is to access information in BioMart by uploading a text file listing gene identifiers of interest. Follow the steps from problem (2.4), but for the filter set choose Gene (instead of Region), select ID list limit and adjust the pulldown menu to HGNC symbol, then browse for a text file having a list of gene symbols. See Web Document 2.5 for a text file listing official HGNC symbols for 13 human globin genes (*CYGB, HBA1, HBA2, HBB, HBD, HBE1, HBG1, HBG2, HBM, HBQ1, HBZ, MB, NGB*). You could also enter these

gene symbols manually. For attributes choose any set of features that is different from that in problem (2.4), so that you can further explore BioMart resources.

**[2-6]** Accessing information from Ensembl.

- (1) Visit the Ensembl resource for humans (<http://www.ensembl.org/human>).
- (2) In the main search box enter 11:5,240,001–5,300,000. The resulting page displays several panels. At the top, all of chromosome 11 is shown. Where on the chromosome is the region we have selected? In what chromosomal band does this region reside?
- (3) The next panel shows the region in detail. What is the size of the displayed region, in base pairs? In general, genes encoding olfactory receptors are gamed OR followed by a string of numbers and letters (e.g., *OR51F1*). Approximately how many olfactory receptor genes flank the 60 kb region we have selected? Can you determine exactly how many ORs are in that region?
- (4) Next we see the region we selected (11:5240001–5300000). Note that there are horizontal tracks (similar to the UCSC Genome Browser).

**[2-7]** Accessing information from UCSC. Hemoglobin is a tetramer composed primarily of two alpha globin subunits and two beta globin subunits. Consider alpha globin. There are two related human genes (official gene symbols *HBA1* and *HBA2*). Use the UCSC Genome Browser (<http://genome.ucsc.edu/>) to determine the length of the intergenic region between the *HBA1* and *HBA2* genes.

**[2-8]** Accessing information from UCSC. What types of repetitive DNA elements occur in the human beta globin gene? The purpose of this exercise is for you to gain familiarity with the UCSC Genome Browser. As a user, you choose which tracks to display. Visit and explore as many as possible. Try to get a sense for the main categories of information offered at the Genome Browser. As you work in the genome browser you may want to switch between builds GRCh37 and GRCh38. To do so, go the the “View” pull-down menu and use “In other genomes (convert).” Carry out the following steps.

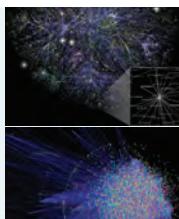
- (1) Go to <http://genome.ucsc.edu/cgi-bin/hgGateway>. Make sure the clade is Mammal, genome is Human, assembly is NCBI37/hg19, and in the “gene” box enter hbb for beta globin. Click Submit. Note that HBB is the official gene symbol for beta globin, but you can use the lowercase hbb for this search. Use NCBI Gene (or <http://www.genenames.org> for the HGNC site) to find the official gene symbol of your favorite gene.

- (2) Click the “default tracks” button. Note the position you have reached (chromosome 11, spanning 1606 base pairs close to the beginning of the short or “p” arm of the chromosome). Note the appearance of over a dozen graphical tracks that are horizontally oriented.
- (3) One of the tracks is “Repeating Elements by Repeatmasker.” There are two black blocks. Right click on the block and select “Full.” Alternatively, scroll down to the section entitled “Variation and Repeats,” locate “RepeatMasker,” and change the pull-down menu setting from “dense” to “full.” Note also that by clicking the blue heading “RepeatMasker” you visit a page describing the RepeatMasker program and its use at the UCSC Genome Browser.
- (4) View the RepeatMasker output. Choose one answer.
  - (a) There are no repetitive elements.
  - (b) There is one SINE element and one LINE element.
  - (c) There is one LTR and one satellite.
  - (d) There is one LINE element and one low-complexity element.
  - (e) There are well over a dozen repetitive elements.

**[2-9]** Accessing information from the UCSC Table Browser. How many SNPs span the human beta globin gene? To solve this problem, use the UCSC Table Browser. The Table Browser is as equally useful as the Genome Browser. Instead of offering visual output, it offers tabular output. Often it is not practical (or accurate) to visually count elements from the Genome Browser. We often want quantitative information about genomic features in some chromosomal region or across the whole genome. This problem asks about single-nucleotide polymorphisms (SNPs), which are positions that vary (i.e., exhibit polymorphism) across individuals in a population. Carry out the following steps.

- (1) Start at the HBB region of the UCSC Genome Browser and click the “Tables” tab along the top. Alternatively, you can go to the UCSC website (<http://genome.ucsc.edu>), and click Tables. Set the clade to Mammal, genome to Human, assembly to GRCh37/hg19, group to Variation, track to AllSNPs(142), table to snp142, and region to position chr11:5246696–5248301. Note that if the position is not already set, you can type hbb into the position box, click “lookup” and the correct position will be entered.
- (2) To see the answer to this problem, click “summary/statistics.” The item count tells you how many SNPs there are.
- (3) To see the answer as a table, set the output format to “all fields from selected table,” make sure the “Send

- output to Galaxy/GREAT” boxes are not checked, and click the “get output” box. The SNPs are shown as a table including chromosome, start, and stop position.
- (4) Try the various output options, such as a bed file or a custom track. Note that you can output the information as a file saved to your computer.
- [2-10]** Accessing information from Galaxy. How big is the largest RefSeq gene on human chromosome 21? Solve this problem by using Galaxy.
- (1) First go to Galaxy (<http://usegalaxy.org>). Optionally, you can register (under the “User” tab).
  - (2) On the left sidebar, choose “Get Data” then “UCSC Main Table Browser.”
  - (3) Set the clade (Mammal), genome (Human), assembly (GRCh37, or try GRCh38), group (Genes and Gene Prediction Tracks), track (RefSeq Genes), table (RefGene), region (click position then enter “chr21” without the quotation marks) then click “lookup” right next to the position. Under output format choose “BED-browser extensible data” and click the box “Send output to Galaxy.”
  - (4) Optionally, click “summary/statistics” to get a quick look at how many proteins are assigned to chromosome 21. (That answer is currently 636.)
  - (5) At the lower left part of the page, click “get output.” Note that you now have a variety of output options; choose BED and click “Send query to Galaxy.”
- (6) Galaxy’s central panel informs you that the job is added to the queue.
  - (7) Your dataset is available in the history panel to the right. Click the dataset header (1: UCSC Main on Human: refGene (chr21:1-46944323)) to see the number of regions and to see the column headers. Click the “eye” icon to see your data in the central panel.
  - (8) Next figure out the size of the genes. First, add a new column. On the left Galaxy panel click “Text Manipulation” then “Compute an expression on every row.” Add the expression  $c3-c2$  to take the end position of each gene and subtract the beginning. For “Round result?” choose “Yes.” Click “Execute.”
  - (9) A new dataset is created, called “Compute on data 1.” There is a new column 13 with the sizes of all the genes. Go to the left sidebar of Galaxy, click “Filter and Sort,” click “Sort data in ascending or descending order” and choose the query; the column (c13); the flavor (numerical sort); the order (descending); and click Execute.
  - (10) A new dataset is created. Click the eye icon to see your spreadsheet in the main Galaxy panel. Your answer is there on the first (top) row. Alternatively, go to “Text Manipulation,” select “Cut columns from a table,” and Cut columns (c5, c6, c7, c8, c9, c10, c11, c12). This will clean up your table, making it easier to see column 13 with the gene lengths.



## Self-Test Quiz

**[2-1]** Which one of the following does not have the proper format of an accession number? (Note: To answer the question, you do not need to look up the particular entries corresponding to each of these accession numbers.)

- (a) rs41341344;
- (b) J03093;
- (c) 1PBO;
- (d) NT\_030059; or
- (e) all of these have proper formats.

**[2-2]** KEY: Accession number NM\_005368.2 corresponds to a human gene that is located on which chromosome? Suggestion: try following the link to NCBI Gene. Choose one answer.

- (a) 11p15.5;
- (b) 2q13.1;
- (c) Xq28;
- (d) 21q12; or
- (e) 22q13.1.

**[2-3]** Approximately how many human clusters are currently in UniGene?

- (a) About 8000;
- (b) About 20,000;
- (c) About 140,000; or
- (d) About 400,000.

**[2-4]** You have a favorite gene, and you want to determine in what tissues it is expressed. Which one of the following resources is likely the most direct route to this information?

- (a) UniGene;
- (b) Entrez;
- (c) PubMed; or
- (d) PCR.

**[2-5]** Is it possible for a single gene to have more than one UniGene cluster?

- (a) Yes; or
- (b) No.

**[2-6]** Which of the following databases is derived from mRNA information?

- (a) dbEST;
- (b) PBD;
- (c) OMIM; or
- (d) HTGS.

**[2-7]** Which of the following databases can be used to access text information about human diseases?

- (a) EST;
- (b) PBD;
- (c) OMIM; or
- (d) HTGS.

**[2-8]** What is the difference between RefSeq and GenBank?

- (a) RefSeq includes publicly available DNA sequences submitted from individual laboratories and sequencing projects.
- (b) GenBank provides nonredundant curated data.
- (c) GenBank sequences are derived from RefSeq.
- (d) RefSeq sequences are derived from GenBank and provide nonredundant curated data.

**[2-9]** If you want literature information, what is the best website to visit?

- (a) OMIM;
- (b) Entrez;
- (c) PubMed; or
- (d) PROSITE.

You can access the annual NAR database issue at <http://nar.oxfordjournals.org/> (WebLink 2.79). The NCBI Help Manual “Entrez Sequences Help” is online at <http://www.ncbi.nlm.nih.gov/books/NBK44864/> (WebLink 2.80). Other titles include “MyNCBI Help” (<http://www.ncbi.nlm.nih.gov/books/NBK3843/>, WebLink 2.81) and PubMed Help (<http://www.ncbi.nlm.nih.gov/books/NBK3830/>, WebLink 2.82). The NCBI Handbook is available at <http://www.ncbi.nlm.nih.gov/books/NBK143764/> (WebLink 2.83).

## SUGGESTED READING

Bioinformatics databases are evolving extremely rapidly. Each January, the first issue of the journal *Nucleic Acids Research* includes nearly 100 brief articles on databases. These include descriptions of NCBI (NCBI Resource Coordinators, 2014), GenBank (Benson *et al.*, 2015), and EMBL (Cochrane *et al.*, 2008). Gretchen Gibney and Andreas Baxevanis (2011) wrote an excellent tutorial, “Searching NCBI Databases Using Entrez”. The NCBI website offers extensive online documentation such as Entrez Sequences Help.

## REFERENCES

- Adams, M. D., Kelley, J.M., Gocayne, J.D. *et al.* 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252**, 1651–1656. PMID: 2047873.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410. PMID: 2231712.
- Altschul, S. F., Madden, T.L., Schäffer, A.A. *et al.* 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402. PMID: 9254694.
- Amberger, J., Bocchini, C., Hamosh, A. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutations* **32**(5), 564–567. PMID: 21472891.
- Beach, E.F. 1961. Beccari of Bologna. The discoverer of vegetable protein. *Journal of the History of Medicine* **16**, 354–373.
- Benson, D.A., Clark, K., Karsch-Mizrachi, I. *et al.* 2015. GenBank. *Nucleic Acids Research* **43**(Database issue), D30–35. PMID: 25414350.

- Blankenberg, D., Coraor, N., Von Kuster, G. *et al.* 2011. Integrating diverse databases into an unified analysis framework: a Galaxy approach. *Database (Oxford)* **2011**, bar011. PMID: 21531983.
- Boguski, M. S., Lowe, T. M., Tolstoshev, C. M. 1993. dbEST: database for “expressed sequence tags.” *Nature Genetics* **4**, 332–333. PMID: 8401577.
- Brooksbank, C., Bergman, M.T., Apweiler, R., Birney, E., Thornton, J. 2014. The European Bioinformatics Institute’s data resources 2014. *Nucleic Acids Research* **42**(1), D18–25. PMID: 24271396.
- Cochrane, G., Akhtar, R., Aldebert, P. *et al.* 2008. Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research* **36**, D5–12. PMID: 18039715.
- Dalgleish, R., Fllice, P., Cunningham, F. *et al.* 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Medicine* **2**(4), 24. PMID: 20398331.
- Farrell, C.M., O’Leary, N.A., Harte, R.A., *et al.* 2014. Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Research* **42**(1), D865–872. PMID: 24217909.
- Fielding, A. M., Powell, A. 2002. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Annals of Clinical Biochemistry* **39**, 345–350. PMID: 12117438.
- Fleischmann, R. D., Adams, M.D., White, O. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus nsemble* Rd. *Science* **269**, 496–512. PMID: 7542800.
- Fllice, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Frankel, A. D., Young, J. A. 1998. HIV-1: Fifteen proteins and an RNA. *Annual Reviews of Biochemistry* **67**, 1–25. PMID: 9759480.
- Gibney, G., Baxevanis, A.D. 2011. Searching NCBI databases using Entrez. *Current Protocols in Bioinformatics Chapter* 1, Unit 1.3. PMID: 21633942.
- Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**(8), R86. PMID: 20738864.
- Goecks, J., Coraor, N., Galaxy Team, Nekrutenko, A., Taylor, J. 2012. NGS analyses by visualization with Trackster. *Nature Biotechnology* **30**(11), 1036–1039. PMID: 23138293.
- Goecks, J., Eberhard, C., Too, T. *et al.* 2013. Web-based visual analysis for high-throughput genomics. *BMC Genomics* **14**, 397. PMID: 23758618.
- Gray, K.A., Daugherty, L.C., Gordon, S.M. *et al.* 2013. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Research* **41**(Database issue), D545–552. PMID: 23161694.
- Harrow, J.L., Steward, C.A., Frankish, A. *et al.* 2014. The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Research* **42**(1), D771–779. PMID: 24316575.
- Harte, R.A., Farrell, C.M., Loveland, J.E. *et al.* 2012. Tracking and coordinating an international curation effort for the CCDS Project. *Database* **2012**, bas008. PMID: 22434842.
- Hillman-Jackson, J., Clements, D., Blankenberg, D. *et al.* 2012. Using Galaxy to perform large-scale interactive data analyses. *Current Protocols in Bioinformatics Chapter* 10, Unit10.5. PMID: 22700312.
- Karolchik, D., Hinrichs, A.S., Furey, T.S. *et al.* 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32**(Database issue), D493–496. PMID: 14681465.
- Karolchik, D., Barber, G.P., Casper, J. *et al.* 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research* **42**(1), D764–770. PMID: 24270787.
- Kasprzyk, A. 2011. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011**, bar049. PMID: 22083790.
- Kent, W.J. 2002. BLAT: the BLAST-like alignment tool. *Genome Research* **12**(4), 656–664. PMID: 11932250.
- Kosuge, T., Mashima, J., Kodama, Y. *et al.* 2014. DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Research* **42**(1), D44–49. PMID: 24194602.
- Lampitt, A. 2014. Hadoop: A platform for the big data era. Deep Dive Series, Infoworld.com. Available at: [http://www.infoworld.com/d/big-data/download-the-hadoop-deep-dive-210169?idlg=ifwsite\\_na\\_General\\_Deep%20Dive\\_na\\_lgna\\_na\\_na\\_wpl](http://www.infoworld.com/d/big-data/download-the-hadoop-deep-dive-210169?idlg=ifwsite_na_General_Deep%20Dive_na_lgna_na_na_wpl) (accessed 30 January 2014).

- Lazarus, R., Kaspi, A., Ziemann, M., Galaxy Team. 2012. Creating reusable tools from scripts: the Galaxy Tool Factory. *Bioinformatics* **28**(23), 3139–3140. PMID: 23024011.
- Lewitter, F. 1998. Text-based database searching. *Bioinformatics: A Trends Guide* **1998**, 3–5.
- Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* **35**, D26–31. PMID: 17148475.
- Magrane, M., UniProt Consortium. 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009. PMID: 21447597.
- Mariño-Ramírez, L., Levine, K.M., Morales, M. *et al.* 2011. The Histone Database: an integrated resource for histones and histone fold-containing proteins. *Database* **2011**, article ID bar048, doi:10.1093/database/bar048.
- Nakamura, Y., Cochrane, G., Karsch-Mizrachi, I., International Nucleotide Sequence Database Collaboration. 2013. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research* **41**(Database issue), D21–24. PMID: 23180798.
- NCBI Resource Coordinators. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **42**(Database issue), D7–17. PMID: 24259429.
- Noble, W.S. 2009. A quick guide to organizing computational biology projects. *PLoS Computational Biology* **5**(7), e1000424. PMID: 19649301.
- Ogasawara, O., Mashima, J., Kodama, Y. *et al.* 2013. DDBJ new system and service refactoring. *Nucleic Acids Research* **41**(Database issue), D25–29. PMID: 23180790.
- Olson, M., Hood, L., Cantor, C., Botstein, D. 1989. A common language for physical mapping of the human genome. *Science* **245**, 1434–1435.
- Pakseresht, N., Alako, B., Amid, C. *et al.* 2014. Assembly information services in the European Nucleotide Archive. *Nucleic Acids Research* **42**(1), D38–43. PMID: 24214989.
- Pevsner, J. 2009. Analysis of genomic DNA with the UCSC genome browser. *Methods in Molecular Biology* **537**, 277–301. PMID: 19378150.
- Pruitt, K.D., Brown, G.R., Hiatt, S.M. *et al.* 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research* **42**(1), D756–763. PMID: 24259432.
- Rose, P.W., Bi, C., Bluhm,W.F. *et al.* 2013. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research* **41**(Database issue), D475–482. PMID: 23193259.
- Sayers, E.W., Barrett, T., Benson, D.A. *et al.* 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **40**(Database issue), D13–25. PMID: 22140104.
- Thomas, D.J., Trumbower, H., Kern, A.D. *et al.* 2007. Variation resources at UC Santa Cruz. *Nucleic Acids Research* **35**(Database issue), D716–720. PMID: 17151077.
- UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **40**, D71–D75.
- Zanotti, F.M. 1745. De Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii. Bononiae, Bologna.



### Adrenocorticotropin (ACTH)

The complete amino acid sequences are known for corticotropins isolated from the anterior pituitary glands of three different species, pig, beef, and sheep. The structure of sheep ACTH was discussed in the last chapter, and the sequences shown in Table 9 include only those areas of the three molecules where differences are to be found. Although some difference between the content of amide nitrogen groups has been reported for the three species, these are not included in the figure since it has not been possible to rule out, with certainty, the possibility that these variations are due, in part, to the rigors of the isolation and purification techniques employed.

**TABLE 9**  
Variations in Amino Acid Sequences Among Different Preparations of ACTH

Preparation	Species	Residue No.								
		25	26	27	28	29	30	31	32	33
$\beta$ -Corticotropin	sheep }									
Corticotropin A	beef <sup>a</sup> } pig									
		Ala.Gly.Glu.Asp.Asp.Glu						Ala.Ser.Glu.NH <sub>2</sub>		
		Asp.Gly.Ala.Glu.Asp.Glu						Leu.Ala.Glu		

<sup>a</sup> Identity with sheep hormone not absolutely certain but very probable as judged from the nearly complete sequence analysis by J. S. Dixon and C. H. Li (personal communication to the author).

Two points are of particular interest in regard to the sequences shown. First, the corticotropins of sheep and beef are identical and differ from that of the pig. This finding is consonant with the closer phylogenetic relationship of sheep and cows to each other than of either to pigs. Second, chemical differences are found only in that portion of the ACTH molecule which has been shown to be unessential for hormonal activity. Genetic mutations leading to such differences might, therefore, not be expected to impose significant disadvantages in terms of survival, and these genes could become established in the gene pools of the species.

### Melanotropin (MSH)

Melanotropin, like the other hormones considered in this chapter, is a typically chordate polypeptide. Indeed, the demonstration of melanocyte-stimulating activity in extracts of tunicates constitutes an

Pairwise alignment involves matching two protein or DNA sequences. The first proteins that were sequenced include insulin (by Frederick Sanger and colleagues; see Fig. 7.3) and globins. This figure is from *The Molecular Basis of Evolution* by the Nobel laureate Christian Anfinsen (1959, p. 152). It shows the results of a pairwise alignment of a portion of adrenocortotropic hormone (ACTH) from sheep or cow (top) with that of pig (below). Such alignments, performed manually, led to the realization that amino acid sequences of proteins reflect the phylogenetic relatedness of different species. Furthermore, pairwise alignments reveal the portions of a protein that may be important for its biological function.

Source: Anfinsen (1959).

# Pairwise Sequence Alignment

# CHAPTER 3

*An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently.*

—Margaret Dayhoff (1978, p. 345)

## LEARNING OBJECTIVES

Upon completion of this chapter, you should be able to:

- define homology as well as orthologs and paralogs;
- explain how PAM (accepted point mutation) matrices are derived;
- contrast the utility of PAM and BLOSUM scoring matrices;
- define dynamic programming and explain how global (Needleman–Wunsch) and local (Smith–Waterman) pairwise alignments are performed; and
- perform pairwise alignment of protein or DNA sequences at the NCBI website.

## INTRODUCTION

One of the most basic questions about a gene or protein is whether it is related to any other gene or protein. Relatedness of two proteins at the sequence level suggests that they are homologous. Relatedness also suggests that they may have common functions. By analyzing many DNA and protein sequences, it is possible to identify domains or motifs that are shared among a group of molecules. These analyses of the relatedness of proteins and genes are accomplished by aligning sequences. As we complete the sequencing of the genomes of many organisms, the task of finding out how proteins are related within an organism and between organisms becomes increasingly fundamental to our understanding of life.

In this chapter we introduce pairwise sequence alignment. We adopt an evolutionary perspective in our description of how amino acids (or nucleotides) in two sequences can be aligned and compared. We then describe algorithms and programs for pairwise alignment.

Two genes (or proteins) are homologous if they have evolved from a common ancestor.

To see an example of this use human beta globin protein (NP\_000509.1) in a DELTA-BLAST query against plant RefSeq proteins; we learn how to do this in Chapter 5. There are many dozens of significant matches. Perform a BLASTN search with the coding region of the corresponding DNA (NM\_000518.4); there are no significant matches. When BLASTN is used to query DNA from organisms that last shared a common ancestor with humans more recently, such as fish, there are many significant matches.

The website <http://timetree.org> (WebLink 3.1) of Sudhir Kumar and colleagues provides estimates of the divergence times of species across the tree of life (Hedges *et al.*, 2006).

Some researchers use the term *analogous* to refer to proteins that are not homologous but share some similarity by chance. Such proteins are presumed not to have descended from a common ancestor.

You can see the protein sequences used to generate Figures 3.2 and 3.3 in Web Documents 3.1 and 3.2 and <http://www.bioinfbook.org/chapter3>.

## Protein Alignment: Often More Informative than DNA Alignment

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is often more informative to compare protein sequence. There are several reasons for this. Many changes in a DNA sequence (particularly at the third position of a codon) do not change the amino acid that is specified. Furthermore, many amino acids share related biophysical properties (e.g., lysine and arginine are both basic amino acids). The important relationships between related (but mismatched) amino acids in an alignment can be accounted for using scoring systems (described in this chapter). DNA sequences are less informative in this regard. Protein sequence comparisons can identify homologous sequences while the corresponding DNA sequence comparisons cannot (Pearson, 1996).

When a nucleotide coding sequence is analyzed, it is often preferable to study its translated protein. In Chapter 4 (on BLAST searching), we see that we can move easily between the worlds of DNA and protein. For example, the TBLASTN tool from the NCBI BLAST website allows related proteins derived from a DNA database to be searched for with a protein sequence. This query option is accomplished by translating each DNA sequence into all of the six proteins that it potentially encodes.

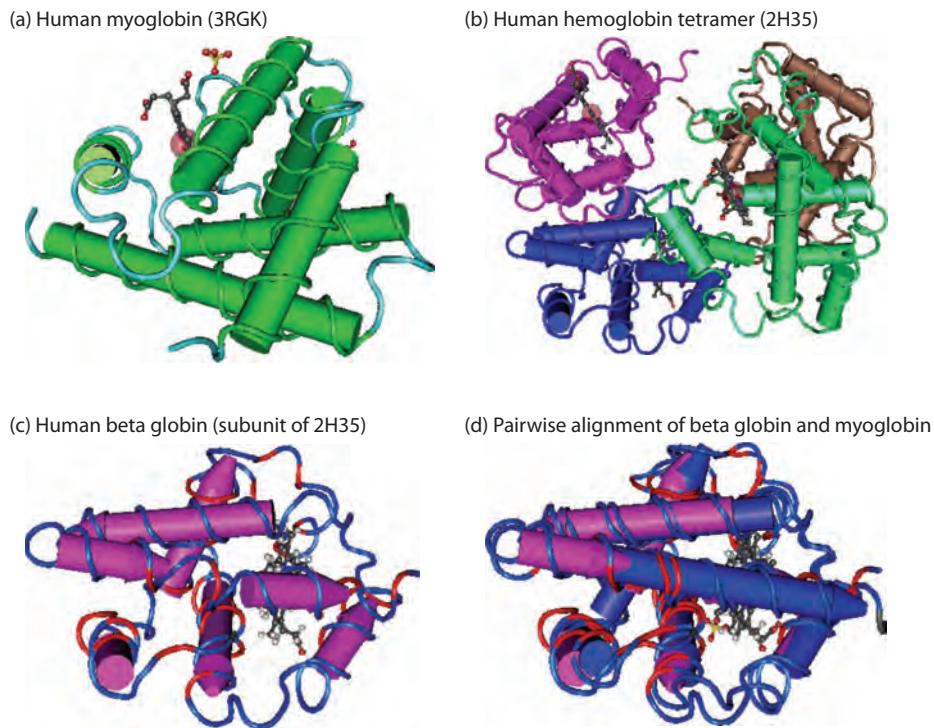
Nevertheless, in many cases it is appropriate to compare nucleotide sequences. This comparison can be important in confirming the identity of a DNA sequence in a database search, in searching for polymorphisms, in analyzing the identity of a cloned cDNA fragment, in comparing regulatory regions, or in many other applications.

## Definitions: Homology, Similarity, Identity

Let us consider the globin family of proteins. We begin with human myoglobin (accession number NP\_005359.1) and beta globin (accession number NP\_000509.1) as two proteins that are distantly but significantly related. The accession numbers are obtained from Gene at NCBI (Chapter 2). Myoglobin and the hemoglobin chains (alpha, beta, and other) are thought to have diverged some 450 million years ago, near the time that human and cartilagenous fish lineages diverged (Fig. 19.22).

Two sequences are *homologous* if they share a common evolutionary ancestry. There are no degrees of homology; sequences are either homologous or not (Reeck *et al.*, 1987; Tautz, 1998). Homologous proteins almost always share a significantly related three-dimensional structure. Myoglobin and beta globin have very similar structures, as determined by X-ray crystallography (Fig. 3.1). When two sequences are homologous, their amino acid or nucleotide sequences usually share significant identity. While homology is a qualitative inference (sequences are homologous or not), identity and similarity are quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid (or nucleotide) identity. In the globin family for example, all the members are homologous but some have sequences that have diverged so greatly that they share no recognizable sequence identity (e.g., human beta globin and human neuroglobin share only 22% amino acid identity). Perutz and colleagues demonstrated that individual globin chains share the same overall shape as myoglobin, even though the myoglobin and alpha globin proteins share only about 26% amino acid identity. In general, three-dimensional structures diverge much more slowly than amino acid sequence identity between two proteins (Chothia and Lesk, 1986). Recognizing this type of homology is an especially challenging bioinformatics problem.

Proteins that are homologous may be orthologous or paralogous. *Orthologs* are homologous sequences in different species that arose from a common ancestral gene during speciation. Figure 3.2 shows a tree of myoglobin orthologs. There is a human myoglobin gene and a rat gene. Humans and rodents diverged about 90 million years ago (MYA) (see Chapter 19), at which time a single ancestral myoglobin gene diverged by



**FIGURE 3.1** Three-dimensional structures of: (a) myoglobin (accession 3RGK); (b) the tetrameric hemoglobin protein (2H35); (c) the beta globin subunit of hemoglobin; and (d) myoglobin and beta globin superimposed. The images were generated with the program Cn3D (see Chapter 13). These proteins are homologous (descended from a common ancestor) and share very similar three-dimensional structures. However, pairwise alignment of the amino acid sequences of these proteins reveals that the proteins share very limited amino acid identity.

Source: Cn3D, NCBI.

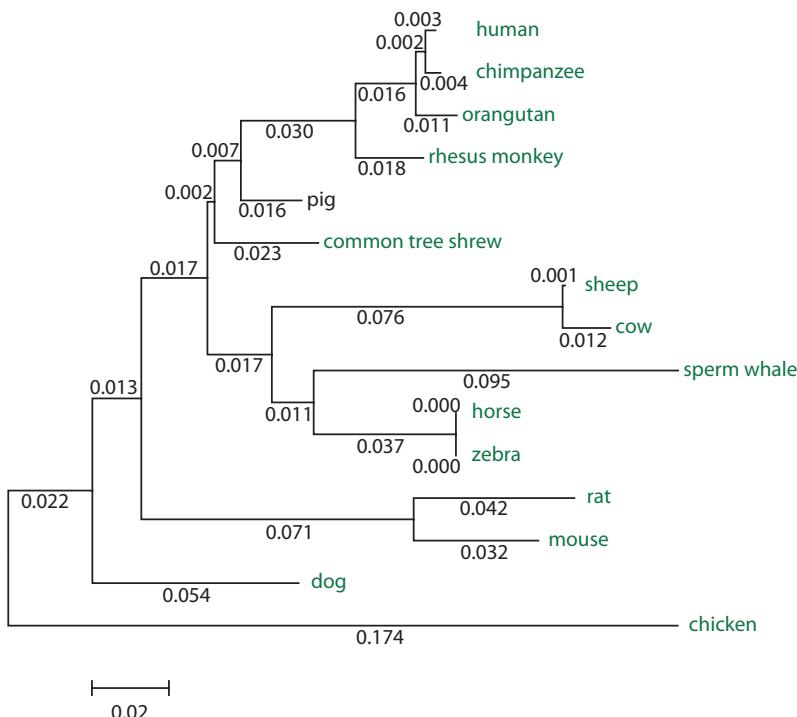
speciation. Orthologs are presumed to have similar biological functions; in this example, human and rat myoglobins both transport oxygen in muscle cells. *Paralogs* are homologous sequences that arose by a mechanism such as gene duplication. For example, human alpha 1 globin (NP\_000549.1) is paralogous to alpha 2 globin (NP\_000508.1); indeed, these two proteins share 100% amino acid identity. Human alpha 1 globin and beta globin are also paralogs (as are all the proteins shown in Fig. 3.3). All of the globins have distinct properties, including regional distribution in the body, developmental timing of gene expression, and abundance. They are all thought to have distinct but related functions as oxygen carrier proteins.

The concept of homology has a rich history dating back to the nineteenth century (Box 3.1). Walter M. Fitch (1970, p. 113) provided our current definitions of these terms. He wrote that “there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example,  $\alpha$  and  $\beta$  hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example  $\alpha$  hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact).”

Notably, orthologs and paralogs do not necessarily have the same function. We provide various definitions of gene and protein function in Chapters 8–14. Later in the book, we explore genomes across the tree of life (Chapters 15–20). In all genome sequencing projects, orthologs and paralogs are identified based on database searches. Two DNA (or protein) sequences are defined as homologous based on achieving significant alignment

In general, when we consider other paralogous families they are presumed to share common functions. Consider the lipocalins: all are about 20 kilodalton proteins that have a hydrophobic binding pocket that is thought to be used to transport a hydrophobic ligand. Members include retinol binding protein (a retinol transporter), apolipoprotein D (a cholesterol transporter), and odorant-binding protein (an odorant transporter secreted from the lateral nasal gland).

We therefore define homologous genes within the same organism as paralogous. But consider further the case of globins. Human  $\alpha$ -globin and  $\beta$ -globin are paralogs, as are mouse  $\alpha$ -globin and mouse  $\beta$ -globin. Human  $\alpha$ -globin and mouse  $\alpha$ -globin are orthologs. What is the relation of human  $\alpha$ -globin to mouse  $\beta$ -globin? These could be considered paralogs, because  $\alpha$ -globin and  $\beta$ -globin originate from a gene duplication event rather than from a speciation event. However, they are not paralogs because they do not occur in the same species. It may therefore be more appropriate to simply call them “homologs,” reflecting their descent from a common ancestor. Fitch (1970, p. 113) notes that phylogenies require the study of orthologs (see also Chapter 7).



**FIGURE 3.2** A group of myoglobin orthologs, visualized by multiply aligning the sequences (Chapter 6) then creating a phylogenetic tree by neighbor-joining (Chapter 7). The accession numbers and species names are as follows: human, NP\_005359 (*Homo sapiens*); chimpanzee, XP\_001156591 (*Pan troglodytes*); orangutan, P02148 (*Pongo pygmaeus*); rhesus monkey, XP\_001082347 (*Macaca mulatta*); pig, NP\_999401 (*Sus scrofa*); common tree shrew, P02165 (*Tupaia glis*); horse, P68082 (*Equus caballus*); zebra, P68083 (*Equus burchellii*); dog, XP\_850735 (*Canis familiaris*); sperm whale, P02185 (*Physeter catodon*); sheep, P02190 (*Ovis aries*); rat, NP\_067599 (*Rattus norvegicus*); mouse, NP\_038621 (*Mus musculus*); cow, NP\_776306 (*Bos taurus*); chicken\_XP\_416292 (*Gallus gallus*). The sequences are shown in Web Document 3.1 (<http://www.bioinfbook.org/chapter3>). In this tree, sequences that are more closely related to each other are grouped closer together. Note that as entire genomes continue to be sequenced (Chapters 15–20), the number of known orthologs will grow rapidly for most families of orthologous proteins.

scores, as discussed below and in Chapter 4. However, some homologous proteins have entirely distinct functions.

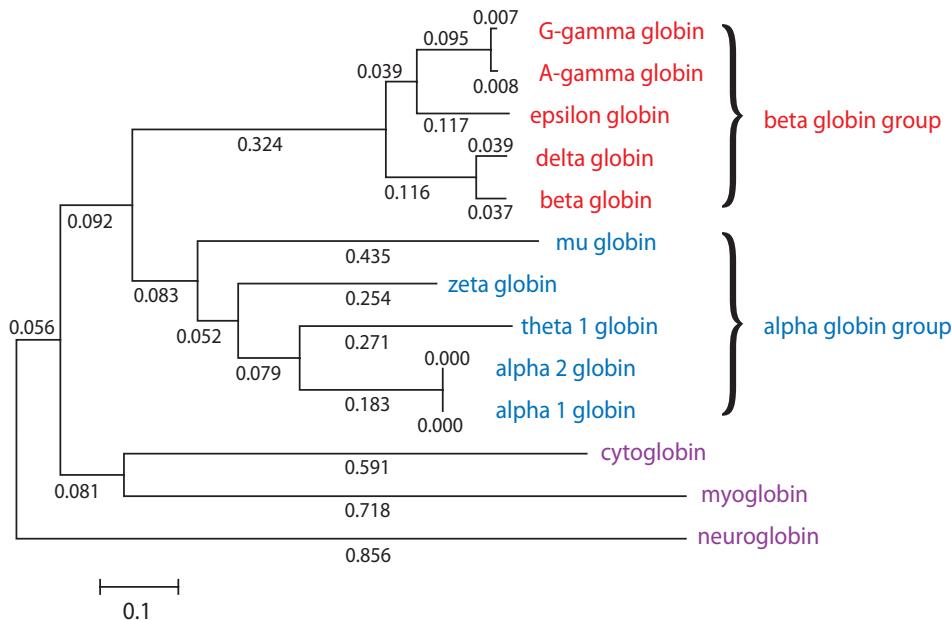
We can assess the relatedness of any two proteins by performing a *pairwise alignment*. In this procedure, we place the two sequences directly next to each other. One

### BOX 3.1 A HISTORY OF HOMOLOGY

Richard Owen (1804–1892) was one of the first biologists to use the term homology. He defined homology as “the same organ in different animals under every variety of form and function” (Owen, 1843, p. 379). Charles Darwin (1809–1882) also discussed homology in the 6th edition of *The Origin of Species or The Preservation of Favoured Races in the Struggle for Life* (1872). He wrote:

That relation between parts which results from their development from corresponding embryonic parts, either in different animals, as in the case of the arm of man, the foreleg of a quadruped, and the wing of a bird; or in the same individual, as in the case of the fore and hind legs in quadrupeds, and the segments or rings and their appendages of which the body of a worm, a centipede, &c., is composed. The latter is called serial homology. The parts which stand in such a relation to each other are said to be homologous, and one such part or organ is called the homologue of the other. In different plants the parts of the flower are homologous, and in general these parts are regarded as homologous with leaves.

For a review of the history of the concept of homology see Hossfeld and Olsson (2005).



**FIGURE 3.3** Paralogous human globins: Each of these proteins is human, and each is a member of the globin family. This unrooted tree was generated using the neighbor-joining algorithm in MEGA (see Chapter 7). The proteins and their RefSeq accession numbers (also shown in Web Document 3.2) are delta globin (NP\_000510), G-gamma globin (NP\_000175), beta globin (NP\_000509), A-gamma globin (NP\_000550), epsilon globin (NP\_005321), zeta globin (NP\_005323), alpha 1 globin (NP\_000549), alpha 2 globin (NP\_000508), theta 1 globin (NP\_005322), hemoglobin mu chain (NP\_001003938), cytoglobin (NP\_599030), myoglobin (NP\_005359), and neuroglobin (NP\_067080). A Poisson correction model was used (see Chapter 7).

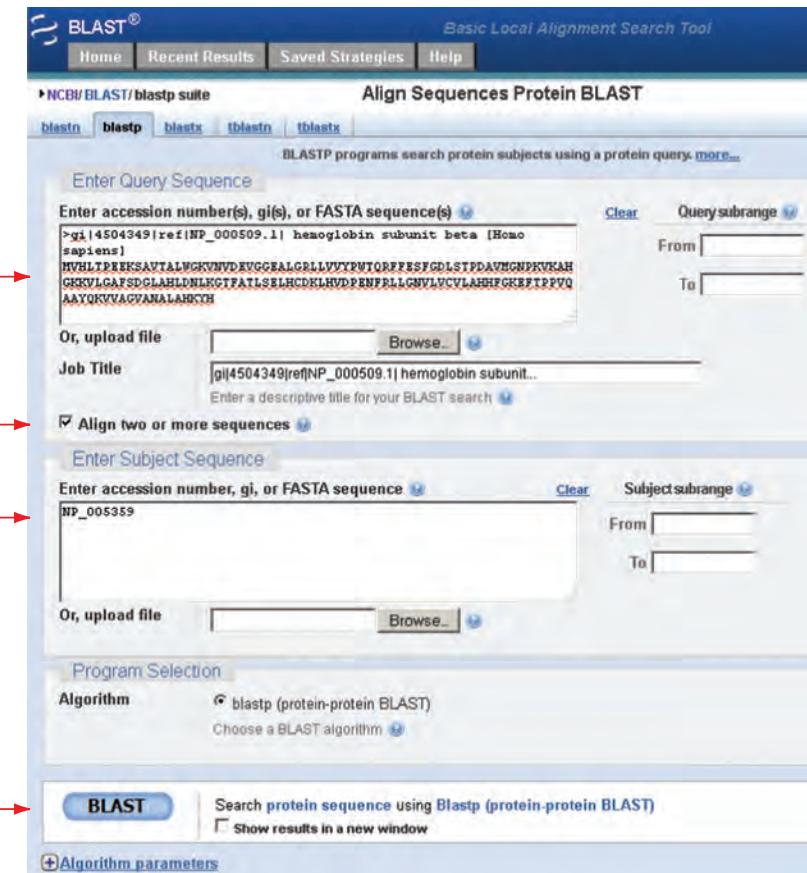
practical way to do this is through the NCBI BLASTP tool (for proteins) or BLASTN (for nucleotides) (Tatusova and Madden, 1999; **Fig. 3.4**). Perform the following steps:

1. Choose the program BLASTP (for “BLAST proteins”) for our comparison of two proteins. Check the box “Align two or more sequences.”
2. Enter the sequences or their accession numbers. Here we use the sequence of human beta globin in the FASTA format, and for myoglobin we use the accession number (**Fig. 3.4**).
3. Select any optional parameters.
  - You can choose from eight scoring matrices: BLOSUM90, BLOSUM80, BLOSUM62, BLOSUM50, BLOSUM45, PAM250, PAM70, PAM30. Select PAM250.
  - You can change the gap creation penalty and gap extension penalty.
  - For BLASTN searches you can change reward and penalty values.
  - There are other parameters you can change, such as word size, expect value, filtering, and dropoff values. We discuss these in more detail in Chapter 4.
4. Click “align.” The output includes a pairwise alignment using the single-letter amino acid code (**Fig. 3.5a**).

Note that the FASTA format uses the single-letter amino acid code; those abbreviations are shown in Box 3.2.

It is impractical to align proteins by visual inspection. Also, if we allow gaps in the alignment to account for deletions or insertions in the two sequences, the number of possible alignments rises exponentially. Clearly, we need a computer algorithm to perform

The BLAST suite of programs is available at the NCBI site, <http://www.ncbi.nlm.nih.gov/BLAST/> (WebLink 3.2). We discuss various options for using the Basic Local Alignment Search Tool (BLAST) in Chapter 4.



**FIGURE 3.4** The BLAST tools at the NCBI website allow the comparison of two DNA or protein sequences. Here the program is set to BLASTP for the comparison of two proteins (arrow 2). Human beta globin (NP\_000509) is input in the FASTA format (arrow 1), while human myoglobin (NP\_005359) is input as an accession number (arrow 3). Click BLAST to start the search (arrow 4), and note the option at bottom left to view and adjust the algorithm parameters.

Source: BLAST, NCBI.

We discuss global and local alignments in the section “Alignment Algorithms: Global and Local”.

an alignment (see Box 3.3). In the pairwise alignments shown in **Figure 3.5a**, beta globin is on top (on the line labeled “query”) and myoglobin is below (on the subject line). An intermediate row indicates the presence of *identical* amino acids in the alignment. For example, note that near the beginning of the alignment the residues WGKV are identical between the two proteins. We can count the total number of identical residues; in this case, the two proteins share 25% identity (37 of 145 aligned residues). Identity is the extent to which two amino acid (or nucleotide) sequences are invariant. Note that this particular alignment is called *local* because only a subset of the two proteins is aligned: the first and last few amino acid residues of each protein are not displayed. A global pairwise alignment includes all residues of both sequences.

Another aspect of this pairwise alignment is that some of the aligned residues are similar but not identical; they are related to each other because they share similar biochemical properties. *Similar* pairs of residues are structurally or functionally related. For example, on the first row of the alignment we can find threonine and serine (T and S connected by a + sign in **Fig. 3.5a**); nearby we can see a leucine and a valine residue that are aligned. These are *conservative substitutions*. Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydroxylated amino acids (S, T), and hydrophobic amino acids (W, F, Y, L, I, V, M, A). Later in this chapter we will see how scores are assigned to aligned amino acid residues. In the pairwise alignment

(a)

Score = 43.9 bits (102), Expect = 1e-09, Method: Composition-based stats.  
 Identities = 37/145 (25%), Positives = 57/145 (39%), Gaps = 2/145 (1%)

Query	4	LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
	→ L+ E V +WGKV D G E L RL +P T F+ F L + D + + +		
Sbjct	3	LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHKLKSEDEMKAEDL	62
Query	62	KAHGKKVLGAFSDGLAHLNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGK	121
	→ K HG VL A L + + L++ H K + + + ++ VL		
Sbjct	63	KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG	122
Query	122	EFTPPVQAAYQKVVAGVANALAHKY	146
	→ +F Q A K + +A Y		
Sbjct	123	DFGADAQGAMNKALELFRKDMASNY	147

(b)

Score = 18.1 bits (35), Expect = 0.015, Method: Composition-based stats.  
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query	12	VITALWGKVNVD--EVGGEALGRLL	33
		V +WGKV D G E L RL	
Sbjct	11	VLNVWGKVEADIPGHGQEVLIRLF	34
match	4 11 5 6 6 5 4 5 4	sum of matches: +60 (round up to +61)	
mismatch	-1 1 0 -2 -2 -4 0	sum of mismatches: -13	
gap open		sum of gap penalties: -13	
gap extend			total raw score: 61 - 13 - 13 = 35

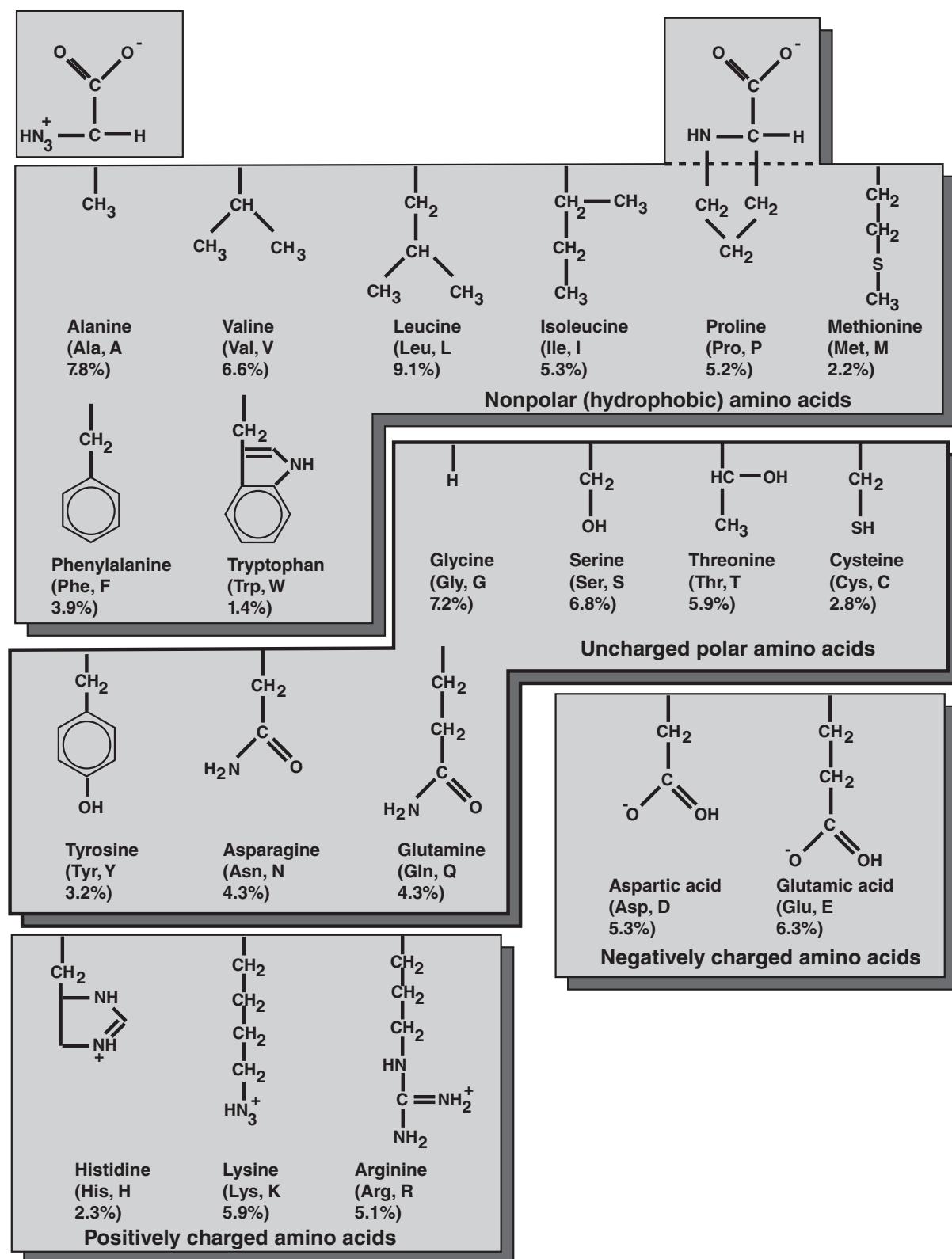
**FIGURE 3.5** Pairwise alignment of human beta globin (the “query”) and myoglobin (the “subject”). (a) The alignment from the search shown in **Figure 3.4**. Note that this alignment is local (i.e., the entire lengths of each protein are not compared), and there are many positions of identity between the two sequences (indicated with amino acids intervening between the query and subject lines; see rows with arrows). The alignment contains an internal gap (indicated by two dashes). (b) Illustration of how raw scores are calculated, using the result of a separate search with just amino acids 12–33 of HBB (corresponding to the region with green shaded letters between the arrowheads in (a). The raw score is 35, rounded up to 36; this represents the sum of the match scores (from a BLOSUM62 matrix in this case), the mismatch scores, the gap opening penalty (set to -11 for this search), and the gap extension penalty (set to -1). Raw scores are subsequently converted to bit scores.

of a segment of HBB and myoglobin, you can see that each pair of residues is assigned a score that is relatively high for matches and often negative for mismatches.

The *percent similarity* of two protein sequences is the sum of both identical and similar matches. In **Figure 3.5a**, there are 57 aligned amino acid residues which are similar. In general, it is more useful to consider the identity shared by two protein sequences rather than the similarity, because the similarity measure may be based upon a variety of definitions of how related (similar) two amino acid residues are to each other.

In summary, pairwise alignment is the process of lining up two sequences to achieve maximal levels of identity (and maximal levels of conservation in the case of amino acid alignments). The purpose of a pairwise alignment is to assess the degree of similarity and the possibility of homology between two molecules. For example, we may say that two proteins share 25% amino acid identity and 39% similarity. If the amount of sequence identity is sufficient, then the two sequences are probably homologous. It is never correct to say that two proteins share a certain percent homology, because they are either homologous or not. Similarly, it is not appropriate to describe two sequences as “highly homologous;” instead, it can be said that they share a high degree of similarity.

## BOX 3.2 STRUCTURES AND ONE- AND THREE-LETTER ABBREVIATIONS OF 20 COMMON AMINO ACIDS



It is very helpful to memorize these abbreviations and to become familiar with the physical properties of the amino acids. The percentages refer to the relative abundance of each amino acid in proteins.

### BOX 3.3 ALGORITHMS AND PROGRAMS

An *algorithm* is a procedure that is structured in a computer program (Sedgewick, 1988). For example, there are many algorithms used for pairwise alignment. A computer *program* is a set of instructions that uses an algorithm (or multiple algorithms) to solve a task. For example, the BLAST program (Chapters 3–5) uses a set of algorithms to perform sequence alignments. Other programs that we will introduce in Chapter 7 use algorithms to generate phylogenetic trees.

Computer programs are essential to solve a variety of bioinformatics problems because millions of operations may need to be performed. The algorithm used by a program provides the means by which the operations of the program are automated. Throughout this book, note how many hundreds of programs have been developed using many hundreds of different algorithms. Each program and algorithm is designed to solve a specific task. An algorithm that is useful to compare one protein sequence to another may not work in a comparison of one sequence to a database of 10 million protein sequences.

Why might an algorithm that is useful for comparing two sequences be less useful to compare millions of sequences? Some problems are so inherently complex that an exhaustive analysis would require a computer with enormous memory or the problem would take an unacceptably long time to complete. A *heuristic algorithm* is one that makes approximations of the best solution without exhaustively considering every possible outcome. The 13 proteins in **Figure 3.2** can be arranged in a tree over a billion distinct ways (see Chapter 7); finding the optimal tree is a problem that a heuristic algorithm can solve in a second.

See the section “The Statistical Significance of Pairwise Alignments” for further discussion, including the use of expect values to assess whether an alignment of two sequences is likely to have occurred by chance (Chapter 4). Such analyses provide evidence to assess the hypothesis that two proteins are homologous. Ultimately, the strongest evidence to determine whether two proteins are homologous comes from structural studies in combination with evolutionary analyses.

Two proteins could have similar structures due to convergent evolution. Molecular evolutionary studies are essential (based on sequence analyses) to assess this possibility.

### BOX 3.4 DAYHOFF'S PROTEIN SUPERFAMILIES

Dayhoff (1978, p. 3) studied 34 protein “superfamilies” grouped into 71 phylogenetic trees. These proteins ranged from some that are very well conserved (e.g., histones and glutamate dehydrogenase; see **Fig. 3.10**) to others that have a high rate of mutation acceptance (e.g., immunoglobulin (Ig) chains and kappa casein; see **Fig. 3.11**). Protein families were aligned; then they counted how often any one amino acid in the alignment was replaced by another. Here is a partial list of the proteins they studied, including the rates of mutation acceptance. For a more detailed list, see **Table 7.1**. There is a range of almost 400-fold between the families that evolve fastest and slowest, but within a given family the rate of evolution (measured in PAMs per unit time) varies only two- to three-fold between species. Used with permission.

PROTEIN	PAMS PER 100 MILLION YEARS
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome c	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

## Gaps

Pairwise alignment is useful as a way to identify mutations that have occurred during evolution and have caused divergence of the sequences of the two proteins we are studying. The most common mutations are *substitutions*, *insertions*, and *deletions*. In protein sequences, substitutions occur when a mutation results in the codon for one amino acid being changed into that for another. This results in the alignment of two nonidentical amino acids, such as serine and threonine. Insertions and deletions occur when residues are added or removed and are typically represented by dashes that are added to one or the other sequence. Insertions or deletions (even those just one character long) are referred to as *gaps* in the alignment.

In our alignment of human beta globin and myoglobin there is one gap (Fig. 3.5a, between the arrowheads). Gaps can occur at the ends of the proteins or in the middle. Note that one of the effects of adding gaps is to make the overall length of each alignment exactly the same. The addition of gaps can help to create an alignment that models evolutionary changes that have occurred.

In a typical scoring scheme there are two gap penalties called affine gap costs. One is a score  $-a$  for creating a gap ( $-11$  in the example of Fig. 3.5b). A second penalty is  $-b$  for each residue that a gap extends. If a gap extends for  $k$  residues it is assigned a penalty of  $-(a + bk)$ . For a gap of length 1, the score is  $-(a + b)$ .

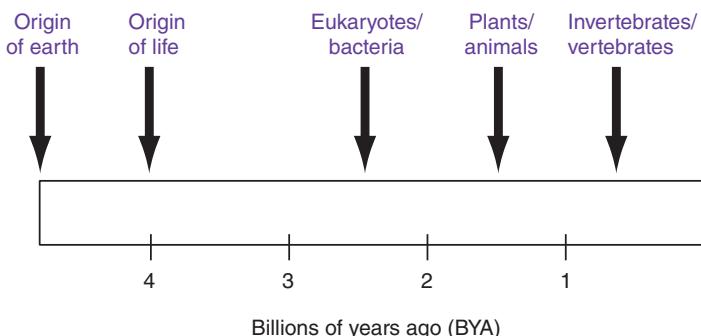
For a description of affine gap penalties at NCBI see [http://www.ncbi.nlm.nih.gov/blast/html/sub\\_matrix.html](http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html) (WebLink 3.3).

It is possible to infer the sequence of the common ancestor (see Chapter 7).

## Pairwise Alignment, Homology, and Evolution of Life

If two proteins are homologous, they share a common ancestor. Generally, we observe the sequence of proteins (and genes) from organisms that are extant. We can compare myoglobins from species such as human, horse, and chicken, and see that the sequences are homologous (Fig. 3.2). This implies that an ancestral organism had a myoglobin gene and lived sometime before the divergences of the lineages that gave rise to human and chicken ~310 MYA (see Chapter 19). Descendants of that ancestral organism include many vertebrate species. The study of homologous protein (or DNA) sequences by pairwise alignment involves an investigation of the evolutionary history of that protein (or gene).

For a brief overview of the time scale of life on Earth, see Figure 3.6 (refer to Chapter 15 for a more detailed discussion). The divergence of different species is established through the use of many sources of data, especially the fossil record. Fossils of bacteria have been discovered in rocks 3.5 billion years old or even older (Schopf, 2002). Fossils of methane-producing archaea, representative of a second domain of life, are found in rocks over 3 billion years old. The other main domain of life, the eukaryotes, emerged at a similar time. In the case of globins, in addition to the vertebrate proteins represented in Figure 3.2 there



**FIGURE 3.6** Overview of the history of life on Earth. See Chapters 15 and 19 for details. Gene/protein sequences are analyzed in the context of evolution. Which organisms have orthologous genes? When did these organisms evolve? How related are human and bacterial globins?

are plant globins that must have shared a common ancestor with the metazoan (animal) globins some 1.5 billion years ago. There are also many bacterial and archaeal globins, suggesting that the globin family arose earlier than two billion years ago.

## SCORING MATRICES

When two proteins are aligned, what scores should they be assigned? For the alignment of beta globin and myoglobin in **Figure 3.5a** there were specific scores for matches and mismatches; how were they derived? Margaret Dayhoff (1966, 1978) provided a model of the rules by which evolutionary change occurs in proteins. We now examine the Dayhoff model in seven steps (following the article from Dayhoff, 1978). This provides the basis of a quantitative scoring system for pairwise alignments between any proteins, whether they are closely or distantly related. We then describe the BLOSUM matrices of Steven Henikoff and Jorja G. Henikoff. Most database searching methods such as BLAST and HMMER (Chapters 4 and 5) depend in some form upon the evolutionary insights of the Dayhoff model.

### Dayhoff Model Step 1 (of 7): Accepted Point Mutations

Dayhoff and colleagues considered the problem of how to assign scores to aligned amino acid residues. Their approach was to catalog hundreds of proteins and compare the sequences of closely related proteins in many families. They considered the question of which specific amino acid substitutions are observed to occur when two homologous protein sequences are aligned. They defined an *accepted point mutation* as a replacement of one amino acid in a protein by another residue that has been accepted by natural selection. Accepted point mutation is abbreviated PAM (which is easier to pronounce than APM). An amino acid change that is accepted by natural selection occurs when: (1) a gene undergoes a DNA mutation such that it encodes a different amino acid; and (2) the entire species adopts that change as the predominant form of the protein.

Which point mutations are accepted in protein evolution? Intuitively, conservative replacements such as serine for threonine would be most readily accepted. In order to determine all possible changes, Dayhoff and colleagues examined 1572 changes in 71 groups of closely related proteins (Box 3.4). Their definition of “accepted” mutations was therefore based on empirically observed amino acid substitutions. Their approach involved a phylogenetic analysis: rather than comparing two amino acid residues directly, they compared them to the inferred common ancestor of those sequences (**Fig. 3.7**; Box 3.5).

The empirical results of observed substitutions are shown in **Figure 3.8**, which describes the frequency with which any amino acid pairs  $i, j$  are aligned. Inspection of this table reveals which substitutions are unlikely to occur (for example, cysteine and tryptophan have noticeably few substitutions), while others such as asparagine and serine tolerate replacements quite commonly. Today, we could generate a table like this with vastly more data (refer to **Fig. 2.3** and the explosive growth of DNA sequence repositories). Several groups have produced updated versions of the PAM matrices (Gonnet *et al.*, 1992; Jones *et al.*, 1992). Nonetheless, the findings from 1978 are essentially correct. The largest inaccuracies in **Figure 3.8** occur for pairs of rarely substituted residues such as cys and asp, for which zero substitutions were observed in the 1978 dataset (35 of 190 total possible exchanges were never observed).

### Dayhoff Model Step 2 (of 7): Frequency of Amino Acids

To model the probability that one aligned amino acid in a protein changes to another, we need to know the frequencies of occurrence of each amino acid. **Table 3.1** shows the frequency with which each amino acid is found ( $f_i$ ).

The Dayhoff (1978) reference is to the *Atlas of Protein Sequence and Structure*, a book with 25 chapters (and various co-authors) describing protein families. The 1966 version of the Atlas described the sequences of just several dozen proteins (cytochromes c, other respiratory proteins, globins, some enzymes such as lysozyme and ribonucleases, virus coat proteins, peptide hormones, kinins, and fibrinopeptides). The 1978 edition included about 800 protein sequences.

Dayhoff *et al.* focused on proteins sharing 85% or more identity; they could therefore construct their alignments with a high degree of confidence. In the section “Global Sequence Alignment: Algorithm of Needleman and Wunsch” below, we will see how the Needleman and Wunsch algorithm (1970) permits the optimal alignment of protein sequences.

Look up a recent estimate of the frequency of occurrence of each amino acid at the SwissProt website <http://www.expasy.org/sprot/relnotes/relistat.html> (WebLink 3.4). From the UniProtKB/Swiss-Prot protein knowledgebase (release 51.7), the amino acid composition of all proteins is shown in Web Document 3.3 (<http://www.bioinfbook.org/chapter3>).



**FIGURE 3.7** Dayhoff's approach to determining amino acid substitutions. (a) Partial multiple sequence alignment of human alpha 1 globin, beta globin, delta globin, and myoglobin. Four columns in which alpha 1 globin and myoglobin have different amino acid residues are indicated in red. For example, A is aligned with G (arrow). (b) Phylogenetic tree that shows the four extant sequences (labeled 1–4) as well as two internal nodes that represent the ancestral sequences (labeled 5 and 6). The inferred ancestral sequences were identified by maximum parsimony analysis using the software PAUP (Chapter 7), and are displayed in (a). From this analysis it is apparent that at each of the columns labeled in red, there was no direct interchange of two amino acids between alpha 1 globin and myoglobin. Instead, an ancestral residue diverged. For example, the arrow in (a) indicates an ancestral glutamate that evolved to become alanine or glycine, but it would not be correct to suggest that alanine had been converted directly to glycine.

### Dayhoff Model Step 3 (of 7): Relative Mutability of Amino Acids

Dayhoff *et al.* calculated the relative mutabilities of the amino acids (Table 3.2). This simply describes how often each amino acid is likely to change over a short evolutionary period. (We note that the evolutionary period in question is short because this analysis involves protein sequences that are closely related to each other.) To calculate relative mutability, they divided the number of times each amino acid was observed to mutate ( $m_i$ ) by the overall frequency of occurrence of that amino acid ( $f_i$ ).

Why are some amino acids more mutable than others? The less mutable residues probably have important structural or functional roles in proteins, such that the consequence of replacing them with any other residue could be harmful to the organism.

### BOX 3.5 A PHYLOGENETIC APPROACH TO ALIGNING AMINO ACIDS

Dayhoff and colleagues did not compare the probability of one residue mutating directly into another. Instead, they constructed phylogenetic trees using parsimony analysis (see Chapter 7). They then described the probability that two aligned residues derived from a common ancestral residue. With this approach, they could minimize the confounding effects of multiple substitutions occurring in an aligned pair of residues. As an example, consider an alignment of the four human proteins alpha 1 globin, beta globin, delta globin, and myoglobin. A direct comparison of alpha 1 globin would suggest several amino acid replacements such as ala↔gly, asn↔leu, lys↔leu, and ala↔val (Fig. 3.7a). However, a phylogenetic analysis of these four proteins results in the estimation of internal nodes that represent ancestral sequences. In Figure 3.7b the external nodes (corresponding to the four existing proteins) are labeled, as are internal nodes 5 and 6 that correspond to inferred ancestral sequences. In the four cases that are highlighted in Figure 3.7a, the ancestral sequences suggest that a glu residue changed to ala and gly in alpha 1 globin and myoglobin, but ala and gly never directly interchanged (Fig. 3.7a, arrow). The Dayhoff approach was therefore more accurate by taking an evolutionary perspective.

In a further effort to avoid the complicating factor of multiple substitutions occurring in alignments of protein families, Dayhoff et al. also focused on using multiple sequence alignments of closely related proteins. For example, their analysis of globins considered the alpha globins and beta globins separately.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	Y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	17	20	90	167	0	17								
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

**FIGURE 3.8** Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences. Amino acids are presented alphabetically according to the three-letter code. Notice that some substitutions (green shaded boxes) are very commonly accepted (such as V and I or S and T). Other amino acids, such as C and W, are rarely substituted by any other residue (orange shaded boxes).

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

(We will see in Chapter 21 that many human diseases, from cystic fibrosis to the autism-related Rett syndrome to hemoglobinopathies, can be caused by a single amino acid substitution in a protein.) Conversely, the most mutable amino acids – asparagine, serine, aspartic acid, and glutamic acid – have functions in proteins that are easily assumed by other residues. The most common substitutions seen in Figure 3.8 are glutamic acid for aspartic acid (both are acidic), serine for alanine, serine for threonine (both are hydroxylated), and isoleucine for valine (both are hydrophobic and of a similar size).

**TABLE 3.1 Normalized frequencies of amino acid. These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence.**

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

**TABLE 3.2 Relative mutabilities of amino acids. The value of alanine is arbitrarily set to 100.**

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation. Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

The substitutions that occur in proteins can also be understood with reference to the genetic code (Box 3.6). Observe how common amino acid substitutions tend to require only a single-nucleotide change. For example, aspartic acid is encoded by GAU or GAC, and changing the third position to either A or G causes the codon to encode a glutamic acid. Also note that four of the five least mutable amino acids (tryptophan, cysteine, phenylalanine, and tyrosine) are specified by only one or two codons. A mutation of any of the three bases of the W codon is guaranteed to change that amino acid. The low mutability of this amino acid suggests that substitutions are not tolerated by natural selection. Of the eight least mutable amino acids (Table 3.2), only one (leucine) is specified by six codons. Dayhoff *et al.* also noted that a fairly large number (20%) of the interchanges observed in Figure 3.8 required two nucleotide changes. In other cases such as gly and trp, only a single-nucleotide change would be required for the substitution; this was never empirically observed however, presumably because such a change has been rejected by natural selection.

#### Dayhoff Model Step 4 (of 7): Mutation Probability Matrix for the Evolutionary Distance of 1 PAM

Dayhoff and colleagues next used the data on accepted mutations (Fig. 3.8) and the probabilities of occurrence of each amino acid to generate a *mutation probability matrix M* (Fig. 3.9). Each element of the matrix  $M_{ij}$  shows the probability that an original amino acid  $j$  (see the columns) will be replaced by another amino acid  $i$  (see the rows) over a defined evolutionary interval. In the case of Figure 3.9 the interval is one PAM, which is defined as the unit of evolutionary divergence in which 1% of the amino acids have been changed between the two protein sequences. Note that the evolutionary interval of this PAM matrix is defined in terms of percent amino acid divergence and not in units of years. 1% divergence of protein sequence may occur over vastly different time frames for protein families that undergo substitutions at different rates (see Fig. 7.5 in which we introduce the molecular clock).

Examination of Figure 3.9 reveals several important features. The highest scores are distributed in a diagonal from top left to bottom right. The values in each column sum to 100%. The value 98.7 at the top left indicates that, when the original sequence consists of an alanine, there is a 98.7% likelihood that the replacement amino acid will also be an alanine over an evolutionary distance of one PAM. There is a 0.3% chance that it will be changed to serine. The most mutable amino acid (from Table 3.2), asparagine, has only a

### BOX 3.6. THE STANDARD GENETIC CODE

In this table, the 64 possible codons are depicted along with the frequency of codon utilization and the single-letter code of the amino acid that is specified. There are four bases (A, C, G, U) and three bases per codon, so there are  $4^3 = 64$  codons.

		Second nucleotide				Third nucleotide
		T	C	A	G	
First nucleotide	T	TTT Phe 171 TTC Phe 203 TTA Leu 73 TTG Leu 125	TCT Ser 147 TCC Ser 172 TCA Ser 118 TCG Ser 45	TAT Tyr 124 TAC Tyr 158 TAA Ter 0 TAG Ter 0	TGT Cys 99 TGC Cys 119 TGA Ter 0 TGG Trp 122	T C A G
	C	CTT Leu 127 CTC Leu 187 CTA Leu 69 CTG Leu 392	CCT Pro 175 CCC Pro 197 CCA Pro 170 CCG Pro 69	CAT His 104 CAC His 147 CAA Gln 121 CAG Gln 343	CGT Arg 47 CGC Arg 107 CGA Arg 63 CGG Arg 115	T C A G
	A	ATT Ile 165 ATC Ile 218 ATA Ile 71 ATG Met 221	ACT Thr 131 ACC Thr 192 ACA Thr 150 ACG Thr 63	AAT Asn 174 AAC Asn 199 AAA Lys 248 AAG Lys 331	AGT Ser 121 AGC Ser 191 AGA Arg 113 AGG Arg 110	T C A G
	G	GTT Val 111 GTC Val 146 GTA Val 72 GTG Val 288	GCT Ala 185 GCC Ala 282 GCA Ala 160 GCG Ala 74	GAT Asp 230 GAC Asp 262 GAA Glu 301 GAG Glu 404	GGT Gly 112 GGC Gly 230 GGA Gly 168 GGG Gly 160	T C A G

Adapted from the International Human Genome Sequencing Consortium (2001), figure 34. Used with permission.

Several features of the genetic code should be noted. Amino acids may be specified by one codon (M, W), two codons (C, D, E, F, H, K, N, Q, Y), three codons (I), four codons (A, G, P, T, V), or six codons (L, R, S). UGA is rarely read as a selenocysteine (abbreviated sec, and the assigned single-letter abbreviation is U).

For each block of four codons that are grouped together, one is often used dramatically less frequently. For example, for F, L, I, M, and V (i.e., codons with a U in the middle, occupying the first column of the genetic code), adenine is used relatively infrequently in the third-codon position. For codons with a cytosine in the middle position, guanine is strongly under-represented in the third position.

Also note that in many cases mutations cause a conservative change (or no change at all) in the amino acid. Consider threonine (ACX). Any mutation in the third position causes no change in the specified amino acid, because of “wobble.” If the first nucleotide of any threonine codon is mutated from A to U, the conservative replacement to a serine occurs. If the second nucleotide C is mutated to a G, a serine replacement occurs. Similar patterns of conservative substitution can be seen along the entire first column of the genetic code, where all of the residues are hydrophobic, and also for the charged residues D, E and K, R.

Codon usage varies between organisms and between genes within organisms. Note also that while this is the standard genetic code, some organisms use alternate genetic codes. A group of two dozen alternate genetic codes are listed at the NCBI Taxonomy website, <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/> (WebLink 3.20). As an example of a nonstandard code, vertebrate mitochondrial genomes use AGA and AGG to specify termination (rather than arg in the standard code), ATA to specify met (rather than ile), and TGA to specify trp (rather than termination).

98.22% chance of remaining unchanged; the least mutable amino acid, tryptophan, has a 99.76% chance of remaining the same.

The nondiagonal elements of this matrix have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}} \quad (3.1)$$

		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid																					
A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2	
R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0
N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0
D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0
H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3	0.0
L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2
K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0	0.0
P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0	0.0
S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0	0.0
T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1	0.0
W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0	0.0
Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0	0.0
V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	99.0	0.0

**FIGURE 3.9** The PAM1 mutation probability matrix. The original amino acid  $j$  is arranged in columns (across the top), while the replacement amino acid  $i$  is arranged in rows. Dayhoff et al. multiplied values by 10,000 (offering added precision) while here we multiply by 100 so that, for example, the first cell's value of 98.7 corresponds to 98.7% occurrence of alanine remaining alanine over this evolutionary interval.

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

where  $M_{ij}$  refers to the probability that an original amino acid  $j$  will be replaced by an amino acid from row  $i$ .  $A_{ij}$  is an element of the accepted point mutation matrix of **Figure 3.8**, such as the value corresponding to the original alanine being substituted by an arginine.  $\lambda$  is a proportionality constant (discussed below) and  $m_j$  is the mutability of the  $j$ th amino acid (from **Table 3.2**). We can further consider the diagonal elements of **Figure 3.9** which have the values:

$$M_{jj} = 1 - \lambda m_j \quad (3.2)$$

where  $M_{jj}$  is the probability that original amino acid  $j$  will remain  $j$  without undergoing a substitution to another amino acid. Let's understand these two equations by inspecting the first column of the mutation probability matrix in which the original amino acid is alanine. The total probability (the sum of all elements) is 1 or, considering the elements as percentages, the sum of the column is 100%. It is intuitively reasonable that the probability of observing a change to the amino acid – equivalent to the sum of all the elements other than alanine remaining itself,  $M_{ji}$  – is proportional to the mutability of alanine.

For each original amino acid, it is easy to observe the amino acids that are most likely to replace it if a change should occur. These data are very relevant to pairwise sequence alignment because they will form the basis of a scoring system (described below in Dayhoff Model Steps 5–7) in which reasonable amino acid substitutions in an alignment are rewarded while unlikely substitutions are penalized.

Almost all molecular sequence data are obtained from extant organisms. We can infer ancestral sequences, as described in Box 3.5 and Chapter 7. In general however, for an aligned pair of residues  $i, j$  we do not know which mutated into the other. Dayhoff and colleagues used the assumption that accepted amino acid mutations are undirected, that is, they are equally likely in either direction. In the PAM1 matrix, the close relationship of the proteins makes it unlikely that the ancestral residue is entirely different from both of the observed, aligned residues.

### Dayhoff Model Step 5 (of 7): PAM250 and Other PAM Matrices

The PAM1 matrix was based upon the alignment of closely related protein sequences, having an average of 1% change. To ensure that the multiple alignments were valid,

<u>NP_002037.2</u>	164	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAQNI	207
<u>XP_001162057.1</u>	164	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAQNI	207
<u>NP_001003142.1</u>	162	IHDHFGIVEGLMTTVHAITATQRTVDGPGSKMWRDGRGAAQNI	205
<u>XP_893121.1</u>	168	IHDNPGIMEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	211
<u>XP_576394.1</u>	162	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	205
<u>NP_058704.1</u>	162	IHDNPGTVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	205
<u>XP_001070653.1</u>	162	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	205
<u>XP_001062726.1</u>	162	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	205
<u>NP_989636.1</u>	162	IHDNPGIVEGLMTTVHAITATQRTVDGPGSKLWRDGRGAAQNI	205
<u>NP_525091.1</u>	161	INDNPEIVEGLMTTVHATTATQRTVDGPGSKLWRDGRGAAQNI	204
<u>XP_318655.2</u>	161	INDNFGILEGLMTTVHATTATQRTVDGPGSKLWRDGRGAAQNI	204
<u>NP_508535.1</u>	170	INDNFGIIEEGLMTTVHATTATQRTVDGPGSKLWRDGRGAGQNI	213
<u>NP_595236.1</u>	164	INDTPGIEEGLMTTVHATTATQRTVDGPGSKDKWRGGRASANII	207
<u>NP_011708.1</u>	162	INDAPGIEEGLMTTVHSILTATQRTVDGPGSHKDWRGGRTASGNII	205
<u>XP_456022.1</u>	161	INDEFGIDEALMTTVHSITATQRTVDGPGSHKDWRGGRTASGNII	204
<u>NP_001060897.1</u>	166	IHDNFGIIEGLMTTVHAITATQRTVDGPGSSKDWRGGRAASFNI	205

**FIGURE 3.10** Multiple sequence alignment of a portion of the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein from 13 organisms: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Canis lupus* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat; three variants), *Gallus gallus* (chicken), *Drosophila melanogaster* (fruit fly), *Anopheles gambiae* (mosquito), *Caenorhabditis elegans* (worm), *Schizosaccharomyces pombe* (fission yeast), *Saccharomyces cerevisiae* (baker's yeast), *Kluyveromyces lactis* (a fungus), and *Oryza sativa* (rice). Columns in the alignment having even a single amino acid change are indicated with arrowheads. The accession numbers are given in the figure. The alignment was created by searching HomoloGene at NCBI with the term gapdh.

proteins within a family were at least 85% identical. We are often interested in exploring the relationships of proteins that share far less than 99% amino acid identity. We can accomplish this by constructing probability matrices for proteins that share any degree of amino acid identity. Consider closely related proteins, such as the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) proteins shown in Figure 3.10. A mutation from one residue to another is a relatively rare event, and a scoring system used to align two such closely related proteins should reflect this. (In the PAM1 mutation probability matrix of Fig. 3.9, some substitutions such as tryptophan to threonine are so rare that they were never observed in Dayhoff's dataset.)

Orthologous kappa caseins from various species provide an example of a less well-conserved family (Fig. 3.11). Some columns of residues in this alignment are perfectly conserved among the selected species but most are not, and many gaps need to be introduced. Several positions at which four or even five different residues occur in an aligned column are indicated.

Here, substitutions are likely to be very common. PAM matrices such as PAM100 or PAM250 were generated to reflect the kinds of amino acid substitutions that occur in distantly related proteins.

How are PAM matrices other than PAM1 derived? The proportionality constant  $\lambda$  of Equations (3.1) and (3.2) applies to all columns of the mutation probability matrix of Figure 3.9. In that matrix,  $\lambda$  is chosen to correspond to an evolutionary distance of 1 PAM. As we make  $\lambda$  larger, we model a greater evolutionary distance. We could for example make a PAM2, PAM3, or PAM4 matrix by multiplying  $\lambda$ . This approach will fail for greater evolutionary distances (such as PAM250, in which 250 changes occur in two aligned sequences of length 100); the problem is that adjusting  $\lambda$  does not account for multiple substitutions. Dayhoff *et al.* instead used matrix multiplication: they multiplied the PAM1 matrix by itself, up to hundreds of times, to obtain other PAM matrices (see Box 3.7), therefore extrapolating from the PAM1 matrix. Today this approach is considered valid, although it depends on the accuracy of the PAM1 matrix to avoid propagating errors.

Databases such as Pfam (Chapter 6) summarize the phylogenetic distribution of gene/protein families across the tree of life.

The GAPDH sequences used to generate Figure 3.10 and the kappa casein sequences used to generate Figure 3.11 are shown in Web Documents 3.4 and 3.5 at <http://www.bioinfbook.org/chapter3>.

	▼	▼	▼	▼	▼	▼	▼
mouse	AIPNPSI	FLAMPTNENQDNTA	IPTIDP	PIVST--PVPTM	-----ESIVNTVANPEAST		
rabbit	S--HPFI	MAILPNKMQDKAVT	PTTNTIAAVEPT--PIPTT	-----EPVVSTEVIAEASP			
sheep	PHPHLSFMAI	IPPKKDQDKTEI	PAINTIASAEPTVHSTPTT	-----EAVVNAVDNPEASS			
cattle	PHPHLSFMAI	IPPKKNQDKTEI	PTINTIASGEPT--STPTT	-----EAVESTVATLEDSP			
pig	PRPHASFI	AIIPPKKNQDKTA	PAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS				
human	PNLHPSFI	AIIPPKKIQDKII	PTINTIATVEPT--PAPAT	-----EPTVDSVVTPEAFS			
horse	PCPHPSFI	AIIPPKKLQEITV	IPKINTIATVEPT--PIPTP	-----EPTVNNAVIPDASS			
.	: * : * :	.. : * :	* .. : ..	* : * : * :	*	.. : ..	:

**FIGURE 3.11** Multiple sequence alignment of seven kappa caseins, representing a protein family that is relatively poorly conserved. Only a portion of the entire alignment is shown. Note that just eight columns of residues are perfectly conserved (indicated with asterisks), and gaps of varying length form part of the alignment. In several columns, there are four different aligned amino acids (arrowheads); in two instances there are five different residues (double arrowheads). The sequences were aligned with MUSCLE 3.6 (see Chapter 6) and were human (*NP\_005203*), equine (*Equus caballus*; *NP\_001075353*), pig (*Sus scrofa* *NP\_001004026*), ovine (*Ovis aries* *NP\_001009378*), rabbit (*Oryctolagus cuniculus* *P33618*), bovine (*Bos taurus* *NP\_776719*) and mouse (*Mus musculus* *NP\_031812*).

To make sense of what different PAM matrices mean, consider the extreme cases. When PAM equals zero, the matrix is a unit diagonal (Fig. 3.12, upper panel) because no amino acids have changed. PAM can be extremely large (e.g., PAM greater than 2000, or the matrix can even be multiplied by itself an infinite number of times). In the resulting  $\text{PAM}_\infty$  matrix there is an equal likelihood of any amino acid being present and all the values consist of rows of probabilities that approximate the background probability for the frequency occurrence of each amino acid (Fig. 3.12, lower panel). We described these background frequencies in Table 3.1.

		original amino acid							
		A	R	N	D	C	Q	E	G
replacement amino acid	PAM0	100	0	0	0	0	0	0	0
	A	0	100	0	0	0	0	0	0
	R	0	0	100	0	0	0	0	0
	N	0	0	0	100	0	0	0	0
	D	0	0	0	0	100	0	0	0
	C	0	0	0	0	0	100	0	0
	Q	0	0	0	0	0	0	100	0
	E	0	0	0	0	0	0	0	100
	G	0	0	0	0	0	0	0	100

		original amino acid							
		A	R	N	D	C	Q	E	G
replacement amino acid	PAM $\infty$	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
	A	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
	R	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
	N	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
	D	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	C	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
	Q	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	E	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9
	G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

**FIGURE 3.12** Portion of the matrices for a zero PAM value (PAM0; upper panel) or for an infinite  $\text{PAM}_\infty$  value (lower panel). At  $\text{PAM}_\infty$  (i.e., if the PAM1 matrix is multiplied by itself an infinite number of times), all the entries in each row converge on the normalized frequency of the replacement amino acid (see Table 3.1). A PAM2000 matrix has similar values that tend to converge on these same limits. In a PAM2000 matrix, the proteins being compared are at an extreme of unrelatedness. In contrast, at PAM0 no mutations are tolerated and the residues of the proteins are perfectly conserved.

### BOX 3.7 MATRIX MULTIPLICATION

A matrix is an orderly array of numbers. An example of a matrix with rows  $i$  and columns  $j$  is:

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

In a symmetric matrix, such as the one above,  $a_{ij} = a_{ji}$ . This means that all the corresponding nondiagonal elements are equal. Matrices may be added, subtracted, or manipulated in a variety of ways. Two matrices can be multiplied together providing that the number of columns in the first matrix  $M_1$  equals the number of rows in the second matrix  $M_2$ .

We can view PAM matrices in R. Try working with a PAM1 matrix. Since it is not readily available in R packages or at the NCBI ftp site, we provide the text file pam1.txt at Web Document 3.10 (<http://bioinfbook.org>). Import it into RStudio, look at its properties, and view its first five rows and columns:

```
> dim(pam1) # this shows the dimensions of the matrix
[1] 20 20
> length(pam1) # this displays the length
[1] 20
> str(pam1) # this displays the structure of pam1; just the first several
# lines are shown here
'data.frame': 20 obs. of 20 variables:
$ A: num 0.9867 0.0001 0.0004 0.0006 0.0001 ...
$ R: num 0.0002 0.9913 0.0001 0 0.0001 ...
...
> pam1 # this shows the full matrix (not shown here)
> pam1[1:5,1:5] # this displays the first five rows and columns
      A      R      N      D      C
1 0.9867 0.0002 0.0009 0.0010 0.0003
2 0.0001 0.9913 0.0001 0.0000 0.0001
3 0.0004 0.0001 0.9822 0.0036 0.0000
4 0.0006 0.0000 0.0042 0.9859 0.0000
5 0.0001 0.0001 0.0000 0.0000 0.9973
```

Next, multiply the PAM1 mutation probability matrix by itself 250 times, creating the data frame called pam250, obtaining a PAM250 matrix.

```
> pam250 <- pam1^250 # we multiply the PAM1 matrix by itself 250 times
> pam250[1:5,1:5] # we view the first five rows and columns
      A      R      N      D      C
[1,] 0.03517888 0.0000000 0.00000000 0.00000000 0.0000000
[2,] 0.00000000 0.1125321 0.00000000 0.00000000 0.0000000
[3,] 0.00000000 0.0000000 0.01121973 0.00000000 0.0000000
[4,] 0.00000000 0.0000000 0.00000000 0.02872213 0.0000000
[5,] 0.00000000 0.0000000 0.00000000 0.00000000 0.5086918
```

The PAM250 matrix is of particular interest (Fig. 3.13). It is produced when the PAM1 matrix is multiplied by itself 250 times, and it is one of the common matrices used for BLAST searches of databases (Chapter 4). This matrix applies to an evolutionary distance where proteins share about 20% amino acid identity. Compare this matrix to the PAM1 mutation probability matrix (Fig. 3.9), and note that much of the information content is lost. The diagonal from top left to bottom right tends to contain higher values than elsewhere in the matrix, but not in the dramatic fashion of the PAM1 matrix. As an example of how to read the PAM250 matrix, if the original amino acid is an alanine there is just a 13% chance that the second sequence will also have

		Original amino acid																			
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3	
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2	
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3	
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3	
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7	
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2	
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9	
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13	
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5	
M	1	1	1	1	0	1	1	1	2	3	2	6	2	1	1	1	1	1	2		
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3	
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4	
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6	
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6	
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0	
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2	
V	7	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17		

**FIGURE 3.13** The PAM250 mutation probability matrix. At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to **Figure 3.11**, and the columns sum to 100.

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

an alanine. In fact, there is a nearly equal probability (12%) that the alanine will have been replaced by a glycine. For the least mutable amino acids, tryptophan and cysteine, there is more than a 50% probability that those residues will remain unchanged at this evolutionary distance.

### Dayhoff Model Step 6 (of 7): From a Mutation Probability Matrix to a Relatedness Odds Matrix

Dayhoff *et al.* defined a relatedness odds matrix. For the elements  $M_{ij}$  of any given mutation probability matrix, what is the probability that amino acid  $j$  will change to  $i$  in a homologous sequence?

$$R_{ij} = \frac{M_{ij}}{f_i}. \quad (3.3)$$

Equation (3.3) describes an odds ratio (Box 3.8). For the numerator, Dayhoff *et al.* considered an entire spectrum of models for evolutionary change in determining target frequencies. For the denominator, the normalized frequency  $f_i$  is the probability of amino acid residue  $i$  occurring in the second sequence by chance.

For the relatedness odds matrix, a value for  $R_{ij}$  of 1 means that the substitution (e.g., alanine replaced by asparagine) occurs as often as can be expected by chance. Values greater than 1 indicate that the alignment of two residues occurs more often than expected by chance (e.g., a conservative substitution of serine for threonine). Values less than 1 suggest that the alignment is not favored. For a comparison of two proteins, it is necessary to determine the values for  $R_{ij}$  at each aligned position and then multiply the resulting probabilities to achieve an overall score for an alignment.

### BOX 3.8. STATISTICAL CONCEPT: THE ODDS RATIO

Dayhoff *et al.* (1972) developed their scoring matrix by using odds ratios. The mutation probability matrix has elements  $M_{ij}$  that give the probability that amino acid  $j$  changes to amino acid  $i$  in a given evolutionary interval. The normalized frequency  $f_i$  gives the probability that amino acid  $i$  will occur at that given amino acid position by chance. The relatedness odds matrix in Equation (3.3) may also be expressed  $R_{ij} = M_{ij}/f_i$ , where  $R_{ij}$  is the relatedness odds ratio.

Equation (3.3) may also be written:

$$\text{Probability of an authentic alignment} = \frac{P(\text{aligned} \mid \text{authentic})}{P(\text{aligned} \mid \text{random})}.$$

The right side of this equation can be read: “the probability of an alignment given that it is authentic (i.e., the substitution of amino acid  $j$  with amino acid  $i$ ) divided by the probability that the alignment occurs given that it happened by chance.” An odds ratio can be any positive ratio. The probability that an event will occur is the fraction of times it is expected to be observed over many trials; probabilities have values ranging from 0 to 1. Odds and probability are closely related concepts. A probability of 0 corresponds to an odds of 0; a probability of 0.5 corresponds to an odds of 1.0; a probability of 0.75 corresponds to odds of 75:25 or 3. Odds and probabilities may be converted as follows:

$$\text{odds} = \frac{\text{probability}}{1 - \text{probability}} \quad \text{and probability} = \frac{\text{odds}}{1 + \text{odds}}.$$

### Dayhoff Model Step 7 (of 7): Log-Odds Scoring Matrix

The logarithmic form of the relatedness odds matrix is called a log-odds matrix. The log-odds form is given by:

$$s_{ij} = 10 \times \log_{10} \left( \frac{M_{ij}}{f_i} \right). \quad (3.4)$$

The cells in a log-odds matrix consist of scores ( $s_{ij}$ ) for aligning any two residues (including an amino acid with itself) along the length of a pairwise alignment.  $M_{ij}$  (also written as  $q_{ij}$ ) is the observed frequency of substitution for each pair of amino acids. The values for  $q_{ij}$ , also called the “target frequencies,” are derived from a mutation probability matrix such as those shown in **Figures 3.9** (for PAM1) and 3.13 (for PAM250). These values consist of positive numbers that sum to 1. The background frequency  $f_i$  refers to the independent, background probability of replacement amino acid  $i$  occurring in this position.

The log-odds matrix for PAM250 is shown in **Figure 3.14**. The values have been rounded off to the nearest integer. Using the logarithm here is convenient because it allows us to sum the scores of the aligned residues when we perform an overall alignment of two sequences. (If we did not take the logarithm we would need to multiply the ratios at all the aligned positions, and this is computationally more cumbersome.)

Try using Equation (3.4) to make sure you understand how the mutation probability matrix (**Fig. 3.13**) is converted into the log-odds scoring matrix (**Fig. 3.14**). As an example, to determine the score assigned to a substitution from cysteine to leucine, the PAM250 mutation probability matrix value is 0.02 (**Fig. 3.13**) and the normalized frequency of leucine is 0.085 (**Table 3.1**). We therefore have:

$$s_{(\text{cysteine, leucine})} = 10 \times \log_{10} \left( \frac{0.02}{0.085} \right) = -6.3. \quad (3.5)$$

Note that this log-odds scoring matrix is symmetric, in contrast to the mutation probability matrix in **Figure 3.13**. In a comparison of two sequences it does not matter which is given first. As another example, an original lysine replaced by an arginine (frequency 4.1%) has a mutation probability matrix score of 0.09, and employing Equation (3.4) yields a log-odds score of 3.4 (matching the score of 3 in **Fig. 3.14**). The values in the matrix are rounded off.

**FIGURE 3.14** Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30. Adapted from NCBI, <ftp://ftp.ncbi.nlm.nih.gov/blast/matrices/>.

What do the scores in the PAM250 matrix signify? A score of +17 for tryptophan matching tryptophan indicates that this correspondence is 50 times more frequent than the chance alignment of this residue in a pairwise alignment. From Equation (3.4), let  $s_{i,j} = +17$  and let the probability of replacement  $q_{ij}/p_i = x$ . Then  $+17 = 10 \log_{10} x$ ;  $+1.7 = \log_{10} x$ ; and  $10^{1.7} = x = 50$ .

A score of  $-10$  indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one-tenth as frequent as the chance alignment of these amino acids. A score of zero is neutral. A score of  $+2$  indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance ( $+2 = 10 \log_{10} x; x = 10^{0.2} = 1.6$ ).

The highest values in this particular log-odds scoring matrix (**Fig. 3.14**) are for tryptophan (17 for an identity) and cysteine (12), while the most severe penalties are associated with substitutions for those two residues. When two sequences are aligned and a score is given, that score is simply the sum of the scores for all the aligned residues across the alignment.

The “target frequencies”  $q_{ij}$  are estimated in reference to a particular amount of evolutionary change. For example, in a comparison of human beta globin versus the closely related chimpanzee beta globin, the likelihood of any particular residue matching another in a pairwise alignment is extremely high; in a comparison of human beta globin and a bacterial globin, the likelihood of a match is low. If in a particular comparison of closely related proteins a serine were aligned to a threonine 5% of the time, then that target frequency  $q_{S,T}$  would be 0.05. If in a different comparison of differently related proteins serine were aligned to threonine more often, say 40% of the time, then that target frequency  $q_{S,T}$  would be 0.4.

It is easy to see how different PAM matrices score amino acid substitutions by comparing the PAM250 matrix (**Fig. 3.14**) with a PAM10 matrix (**Fig. 3.15**). In the PAM10 matrix, identical amino acid residue pairs tend to produce a higher score than in the PAM250 matrix; for example, a match of alanine to alanine scores 7 versus 2, respectively.

**FIGURE 3.15** Log-odds matrix for PAM10. Low PAM values such as this are useful for aligning very closely related sequences. Compare this with the PAM250 matrix (Fig. 3.14) and note that there are larger positive scores for identical matches in this PAM10 matrix and larger penalties for mismatches. Adapted from NCBI, <ftp://ftp.ncbi.nlm.nih.gov/blast/matrices/>.

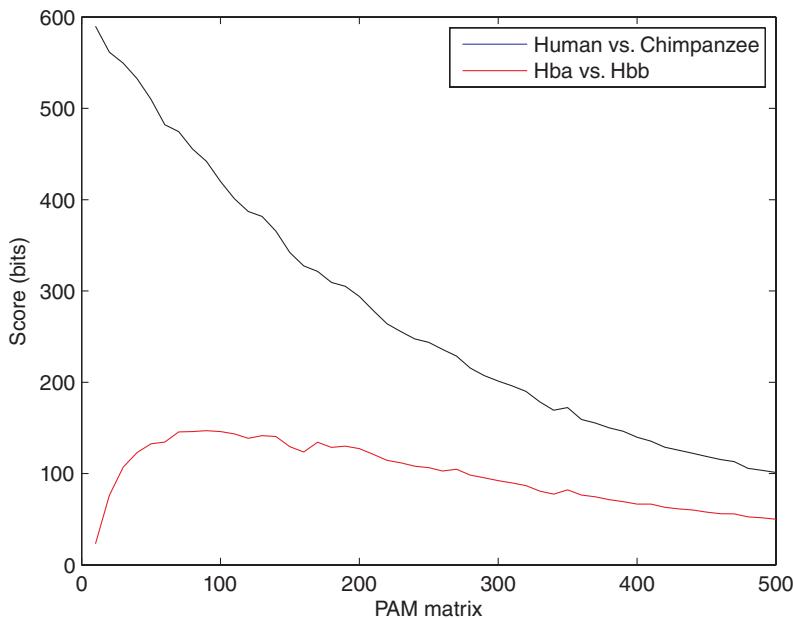
The penalties for mismatches are greater in the PAM10 matrix; for example, a mutation of aspartate to arginine scores -17 (PAM10) versus -1 (PAM250). PAM10 even has negative scores for substitutions (such as glutamate to asparagine: -5) that are scored positively in the PAM250 matrix (+1).

## Practical Usefulness of PAM Matrices in Pairwise Alignment

We can demonstrate the usefulness of PAM matrices by performing a series of global pairwise alignments of both closely related proteins and distantly related proteins. For the closely related proteins we will use human beta globin (NP\_000509.1) and beta globin from the chimpanzee *Pan troglodytes* (XP\_508242.1); these proteins share 100% amino acid identity. The bit scores proceed in a fairly linear, decreasing fashion from about 590 bits using the PAM10 matrix to 200 bits using the PAM250 matrix and 100 bits using the PAM500 matrix (Fig. 3.16, black line). In this pairwise alignment there are no mismatches or gaps, and the high bit scores associated with low PAM matrices (such as PAM10) are accounted for by the higher relative entropy (defined in “Percent Identity and Relative Entropy”). The PAM10 matrix is therefore appropriate for comparisons of closely related proteins. Next consider pairwise alignments of two relatively divergent proteins, human beta globin and alpha globin (NP\_000549.1; Fig. 3.16, red line). The PAM70 matrix yields the highest score. Lower PAM matrices (e.g., PAM10 to PAM60) produce lower bit scores because the sequences share only 42% amino acid identity, and mismatches are assigned large negative scores. We conclude that different scoring matrices vary in their sensitivity to protein sequences (or DNA sequences) of varying relatedness. When comparing two sequences, it may be necessary to repeat the search using several different scoring matrices. Alignment programs cannot be preset to choose the right matrix for each pair of sequences. Instead, they begin with the most broadly useful scoring matrix such as BLOSUM62, which we describe in the following section.

## Important Alternative to PAM: BLOSUM Scoring Matrices

In addition to the PAM matrices, another very common set of scoring matrices is the blocks substitution matrix (BLOSUM) series. Henikoff and Henikoff (1992, 1996) used



**FIGURE 3.16** Global pairwise alignment scores using a series of PAM matrices. Two closely related globins (human and chimpanzee beta globin; black line) were aligned using a series of PAM matrices ( $x$  axis) and the bit scores were measured ( $y$  axis). For two distantly related globins (human alpha versus beta globin; red line) the bit scores are smaller for low PAM matrices (such as PAM1 to PAM20) because mismatches are severely penalized.

Note that the denominator in Equations (3.6) and (3.7) includes  $p_i p_j$ , reflecting the background probabilities of the two aligned amino acids. This is given by Henikoff and Henikoff (1992) and Karlin and Altschul (1990) and others (reviewed by Altschul *et al.*, 2005).

The PAM matrix is given as 10 times the log base 10 of the odds ratio. The BLOSUM matrix is given as 2 times the log base 2 of the odds ratio. BLOSUM scores are therefore not quite as large as they would be if given on the same scale as PAM scores. Practically, this difference in scales is not important because alignment scores are typically converted from raw scores to normalized bit scores (Chapter 4).

the BLOCKS database, which consisted of over 500 groups of local multiple alignments (blocks) of distantly related protein sequences. The Henikoffs therefore focused on conserved regions (blocks) of proteins that are distantly related to each other. The BLOSUM scoring scheme employs a log-odds ratio using the base 2 logarithm:

$$S_{ij} = 2 \times \log_2 \left( \frac{q_{ij}}{p_{ij}} \right). \quad (3.6)$$

Equation (3.6) resembles Equation (3.4) in its format. Karlin and Altschul (1990) and Altschul (1991) have shown that substitution matrices can be described in general in a log-odds form as follows:

$$S_{ij} = \left( \frac{1}{\lambda} \right) \ln \left( \frac{q_{ij}}{p_i p_j} \right) \quad (3.7)$$

where  $S_{ij}$  refers to the score of amino acid  $i$  aligning with  $j$  and  $q_{ij}$  are the positive target frequencies; these sum to 1.  $\lambda$  is a positive parameter that provides a scale for the matrix. We will again encounter  $\lambda$  when we describe the basic statistical measure of a BLAST result (Chapter 4, Equation (4.5)).

The BLOSUM62 matrix is the default scoring matrix for the BLAST protein search programs at NCBI. It merges all proteins in an alignment that has 62% amino acid identity or greater into one sequence. If a block of aligned globin orthologs includes several that have 62, 80, and 95% amino acid identity, these would all be weighted (grouped) as one sequence. Substitution frequencies for the BLOSUM62 matrix are weighted more heavily by blocks of protein sequences having less than 62% identity. (This matrix is therefore useful for scoring proteins that share less than 62% identity.) The BLOSUM62 matrix is shown in Fig. 3.17.

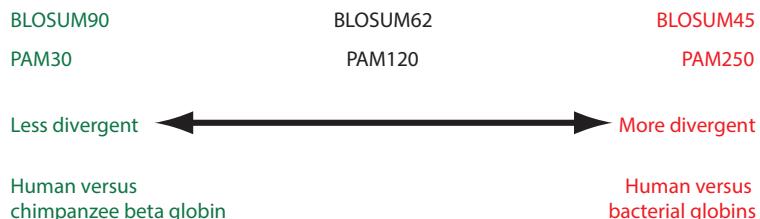
A	4																		
R	-1 5																		
N	-2 0 6																		
D	-2 -2 1 6																		
C	0 -3 -3 -3 9																		
Q	-1 1 0 0 -3 5																		
E	-1 0 0 2 -4 2 5																		
G	0 -2 0 -1 -3 -2 -2 6																		
H	-2 0 1 -1 -3 0 0 -2 8																		
I	-1 -3 -3 -3 -1 -3 -3 -4 -3 4																		
L	-1 -2 -3 -4 -1 -2 -3 -4 -3 2 4																		
K	-1 2 0 -1 -1 1 1 -2 -1 -3 -2 5																		
M	-1 -2 -2 -3 -1 0 -2 -3 -2 1 2 -1 5																		
F	-2 -3 -3 -3 -2 -3 -3 -1 0 0 -3 0 6																		
P	-1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7																		
S	1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4																		
T	0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5																		
W	-3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11																		
Y	-2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7																		
V	0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4																		
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

**FIGURE 3.17** The BLOSUM62 scoring matrix of Henikoff and Henikoff (1992). This matrix merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. BLOSUM62 performs better than alternative BLOSUM matrices or a variety of PAM matrices at detecting distant relationships between proteins. It is therefore the default scoring matrix for most database search programs such as BLAST (Chapter 4).

*Source:* Henikoff & Henikoff (1992). Reproduced with permission from S. Henikoff.

Henikoff and Henikoff (1992) tested the ability of a series of BLOSUM and PAM matrices to detect proteins in BLAST searches of databases. They found that BLOSUM62 performed slightly better than BLOSUM60 or BLOSUM70 and dramatically better than PAM matrices at identifying various proteins. Their matrices were especially useful for identifying weakly scoring alignments. BLOSUM50 and BLOSUM90 are other commonly used scoring matrices in BLAST searches. (For an alignment of two proteins sharing about 50% identity, try using the BLOSUM50 matrix. The FASTA family of sequence comparison programs use BLOSUM50 as a default.)

The relationships of the PAM and BLOSUM matrices are depicted in Figure 3.18. To summarize, BLOSUM and PAM matrices both use log-odds values in their scoring systems. In each case, when performing a pairwise sequence alignment (or when searching a query sequence against a database), specify the exact matrix to use based on the suspected degree of identity between the query and its matches. PAM matrices are based on data from the alignment of closely related protein families, and they involve the assumption that substitution probabilities for highly related proteins (e.g., PAM40) can be extrapolated to probabilities for distantly related proteins (e.g., PAM250). In contrast, the BLOSUM matrices are based on empirical observations of more distantly related protein alignments. Note that a PAM30 matrix, which is available as an option on standard BLASTP searches at NCBI (Chapter 4), may be useful for identifying significant conservation between two closely related proteins. However a BLOSUM matrix with a high value (such as the BLOSUM80 matrix, available from the NCBI BLASTP site) is not necessarily suitable for scoring closely related sequences. This is because the BLOSUM80 matrix is adapted to regions of sequences that share up to 80% identity, but beyond that limited region two proteins may share dramatically less amino acid identity (Pearson and Wood, 2001).



**FIGURE 3.18** Summary of PAM and BLOSUM matrices. High-value BLOSUM matrices and low-value PAM matrices are best suited to study well-conserved proteins such as mouse and rat beta globin. BLOSUM matrices with low numbers (e.g., BLOSUM45) or high PAM numbers are best suited to detect distantly related proteins. Remember that in a BLOSUM45 matrix all members of a protein family with greater than 45% amino acid identity are grouped together, allowing the matrix to focus on proteins with less than 45% identity.

A hit is a change in an amino acid residue that occurs by mutation. We discuss mutations (including multiple hits at a nucleotide position) in Chapter 7 (see Fig. 7.15). We discuss mutations associated with human disease in Chapter 21.

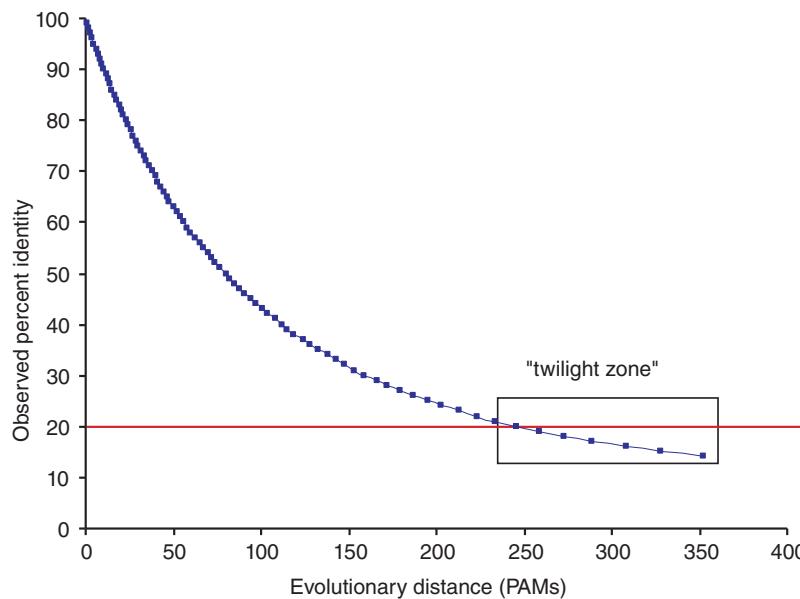
The plot in **Figure 3.19** reaches an asymptote below about 15% amino acid identity. This asymptote would reach about 5% (or the average background frequency of the amino acids) if no gaps were allowed in the comparison between the proteins.

### Pairwise Alignment and Limits of Detection: The “Twilight Zone”

When we compare two protein sequences, how many mutations can occur between them before their differences make them unrecognizable? When we compared glyceraldehyde 3-phosphate dehydrogenase proteins, it was easy to see their relationship (Fig. 3.10). However, when we compared human beta globin and myoglobin, the relationship was much less obvious (Fig. 3.5). Intuitively, at some point two homologous proteins are too divergent for their alignment to be recognized as significant.

The best way to determine the detection limits of pairwise alignments is through statistical tests that assess the likelihood of finding a match by chance. These are described in “The Statistical Significance of Pairwise Alignments” below and in Chapter 4. In particular we will focus on the expect (*E*) value. It can also be helpful to compare the percent identity (and percent divergence) of two sequences versus their evolutionary distance. Consider two protein sequences, each 100 amino acids in length, in which one sequence is fixed and various numbers of mutations are introduced into the other sequence. A plot of the two diverging sequences has the form of a negative exponential (Fig. 3.19) (Dayhoff, 1978; Doolittle, 1987). If the two sequences have 100% amino acid identity, they have zero changes per 100 residues. If they share 50% amino acid identity, they have sustained an average of 80 changes per 100 residues. One might have expected 50 changes per 100 residues in the case of two proteins that share 50% amino acid identity. However, any position can be subject to multiple hits. Percent identity is therefore not an exact indicator of the number of mutations that have occurred across a protein sequence. When a protein sustains about 250 hits per 100 aligned amino acids (as characterized by the PAM250 matrix), it may have about 20% identity with the original protein and can still be recognizable as significantly related. If a protein sustains 360 changes per 100 residues (PAM360), it evolves to a point at which the two proteins share about 15% amino acid identity and are no longer recognizable as significantly related in a direct pairwise comparison.

The PAM250 matrix assumes the occurrence of 250 point mutations per 100 amino acids. As shown in **Figure 3.19**, this corresponds to the “twilight zone.” At this level of divergence, it is usually difficult to assess whether the two proteins are homologous. Other techniques, including multiple sequence alignment (Chapter 6) and structural predictions (Chapter 13), are often very useful to assess homology in these cases. PAM matrices are available from PAM1 to PAM250 or higher, and a specific number of observed amino acid differences per 100 residues is associated with each PAM matrix (Table 3.3; Fig. 3.19). Consider the case of the human alpha globin compared to myoglobin. These proteins are approximately 150 amino acid residues in length, and they may have undergone over



**FIGURE 3.19** Two randomly diverging protein sequences change in a negatively exponential fashion. This plot shows the observed number of amino acid identities per 100 residues of two sequences (y axis) versus the number of changes that must have occurred (the evolutionary distance in PAM units). The twilight zone (Doolittle, 1987) refers to the evolutionary distance corresponding to about 20% identity between two proteins. Proteins with this degree of amino acid sequence identity may be homologous, but such homology is difficult to detect. Data from Dayhoff (1978; see **Table 3.3**).

**TABLE 3.3 Relationship between observed number of amino acid differences per 100 residues of two aligned protein sequences and evolutionary difference. The number of changes that must have occurred, in PAM units.**

Observed differences in 100 residues	Evolutionary distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

Source: Dayhoff (1972). Reproduced with permission from National Biomedical Research Foundation.

There are about  $2^{2n}/\sqrt{\pi n}$  possible global alignments between two sequences of length  $n$  (Durbin *et al.*, 2000; Ewins and Grant, 2001). For two sequences of length 1000, there are about  $10^{600}$  possible alignments. For two proteins of length 200 amino acid residues, the number of possible alignments is over  $6 \times 10^{58}$ .

300 amino acid substitutions since their divergence (Dayhoff *et al.*, 1972, p. 19). Suppose there are 345 changes per 150 amino acids (this corresponds to 230 changes per 100 amino acids). An additional 100 changes would result in only 10 more observable differences (Dayhoff *et al.*, 1972).

## ALIGNMENT ALGORITHMS: GLOBAL AND LOCAL

Our discussion so far has focused on matrices that allow us to score an alignment between two proteins. This involves the generation of scores for identical matches, mismatches, and gaps. We also need an appropriate algorithm to perform the alignment. When two proteins are aligned, there is an enormous number of possible alignments.

There are two main types of alignment: global and local. We explore these approaches next. A *global alignment*, such as the method of Needleman and Wunsch (1970), contains the entire sequence of each protein or DNA molecule. A *local alignment*, such as the method of Smith and Waterman (1981), focuses on the regions of greatest similarity between two sequences. We saw a local alignment of human beta globin and myoglobin in **Figure 3.5** above. For many purposes, a local alignment is preferred, because only a portion of two proteins aligns. (We study the modular nature of proteins in Chapter 12.) Most database search algorithms, such as BLAST (Chapter 4), use local alignments.

Each of these methods is guaranteed to find one or more optimal solutions to the alignment of two protein or DNA sequences. We then describe two rapid-search algorithms, BLAST and FASTA. BLAST represents a simplified form of local alignment that is popular because the algorithm is very fast and easily accessible.

### Global Sequence Alignment: Algorithm of Needleman and Wunsch

One of the first and most important algorithms for aligning two protein sequences was described by Needleman and Wunsch (1970). This algorithm is important because it produces an optimal alignment of protein or DNA sequences, even allowing the introduction of gaps. The result is optimal, but not all possible alignments need to be evaluated. An exhaustive pairwise comparison would be too computationally expensive to perform.

We can describe the Needleman–Wunsch approach to global sequence alignment in three steps: (1) setting up a matrix; (2) scoring the matrix; and (3) identifying the optimal alignment.

#### *Step 1: Setting Up a Matrix*

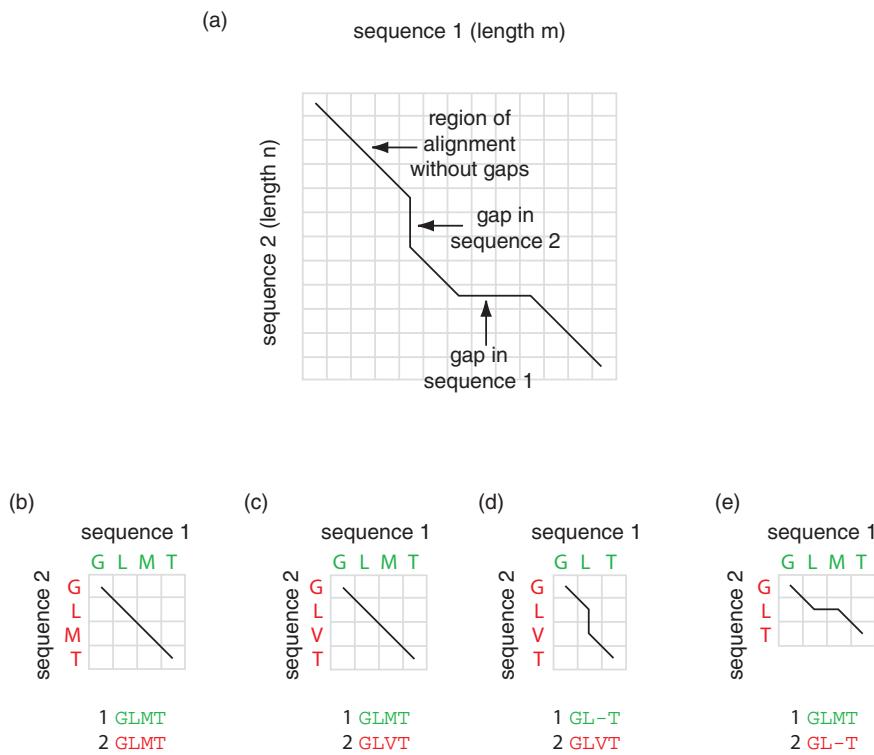
First, we compare two sequences in a two-dimensional matrix (**Fig. 3.20**). The first sequence, of length  $m$ , is listed horizontally along the  $x$  axis so that its amino acid residues correspond to the columns. The second sequence, of length  $n$ , is listed vertically along the  $y$  axis, with its amino acid residues corresponding to rows.

We will describe rules for tracing a diagonal path through this matrix in the following section; the path describes the alignment of the two sequences. A perfect alignment between two identical sequences would simply be represented by a diagonal line extending from the top left to the bottom right (**Fig. 3.20a, b**). Any mismatches between two sequences would still be represented on this diagonal path (**Fig. 3.20c**). However, the score that is assigned might be adjusted according to some scoring system. In the example of **Figure 3.20c**, the mismatch of V and M residues might be assigned a score lower than the perfect match of M and M shown in **Figure 3.20b**.

Gaps are represented in this matrix using horizontal or vertical paths, as shown in **Figure 3.20a, d, e**. Any gap in the top sequence is represented as a vertical line (**Fig. 3.20a, d**),

The Needleman and Wunsch approach is an example of a dynamic programming algorithm. It is called “dynamic” because the alignment is created on a residue-by-residue basis in a search for the optimal alignment. The word “programming” refers to the use of a set of rules to determine the alignment.

This algorithm is also sometimes called the Needleman–Wunsch–Sellers algorithm. Sellers (1974) provided a related alignment algorithm (one that focuses on minimizing differences, rather than on maximizing similarities). Smith *et al.* (1981) showed that the Needleman–Wunsch and Sellers approaches are mathematically equivalent.



**FIGURE 3.20** Pairwise alignment of two amino acid sequences using a dynamic programming algorithm of Needleman and Wunsch (1970) for global alignment. (a) Two sequences can be assigned a diagonal path through the matrix and, when necessary, the path can deviate horizontally or vertically, reflecting gaps that are introduced into the alignment. (b) Two identical sequences form a path on the matrix that fits a diagonal line. (c) If there is a mismatch (or multiple mismatches), the path still follows a diagonal, although a scoring system may penalize the presence of mismatches. If the alignment includes a gap in (d) the first sequence or (e) the second sequence, the path includes a vertical or horizontal line.

while any gap in the bottom sequence is drawn as a horizontal line (Fig. 3.20a, e). These gaps can be of any length. Gaps can be internal or terminal.

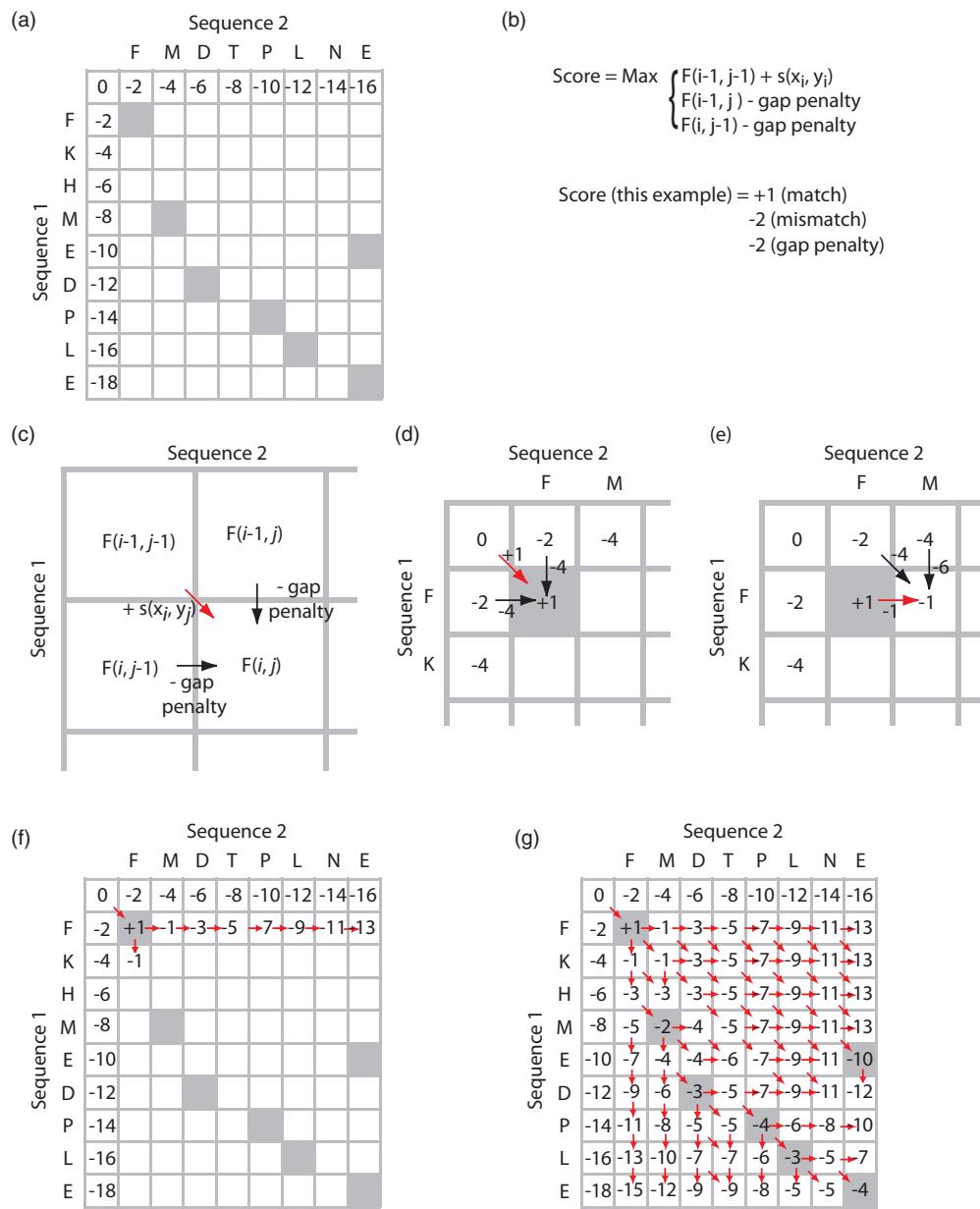
### Step 2: Scoring the Matrix

The goal of this algorithm is to identify an optimal alignment. We set up two matrices: an amino acid identity matrix and then a scoring matrix. We create a matrix of dimensions  $m + 1$  by  $n + 1$  (for the first and second sequences on the  $x$ - and  $y$ -axes respectively; Fig. 3.21a). Gap penalties (here having a value of  $-2$  for each gap position) are placed along the first row and column. This will allow us to introduce a terminal gap of any length. We fill in positions of identity (Fig. 3.21a, gray-filled cells); this is called an identity matrix. For two identical sequences this would include a series of gray-filled cells along the diagonal.

Next, we define a scoring system (Fig. 3.21b). Our goal in finding an optimal alignment is to determine the path through the matrix that maximizes the score. This entails finding a path through as many positions of identity as possible while introducing as few gaps as possible. There are four possible occurrences at each position  $i, j$  (i.e., in each cell in the matrix; Fig. 3.21b):

1. two residues may be perfectly matched (i.e., identical); in this example the score is  $+1$ ;
2. they may be mismatched; here we assign a score of  $-2$ ;

Note that in linear algebra an identity matrix is a special kind of number matrix that has the number 1 from top left to bottom right. For sequence alignments, the amino acid identity matrix is simply a matrix showing all the positions of shared amino acid identity between two sequences, as shown in Fig. 3.20b.



**FIGURE 3.21** Pairwise alignment of two amino acid sequences using the dynamic programming algorithm of Needleman and Wunsch (1970) for global alignment. (a) For sequences of length  $m$  and  $n$  we form a matrix of dimensions  $m + 1$  by  $n + 1$  and add gap penalties in the first row and column. Each gap position receives a score of -2. The cells having identity are shaded gray. (b) The scoring system in this example is +1 for a match, -2 for a mismatch, and -2 for a gap penalty. In each cell, the score is assigned using the recursive algorithm that identifies the highest score from three calculations. (c) In each cell  $F(i, j)$  we calculate the scores derived from following a path from the upper left cell (we add the score of that cell + the score of  $F(i, j)$ ), the cell to the left (including a gap penalty), and the cell directly above (again including a gap penalty). (d) To calculate the score in the cell of the second row and column, we take the maximum of the three scores +1, -4, -4. This best score (+1) follows the path of the red arrow, and we maintain the information of the best path, resulting in each cell's score in order to later reconstruct the pairwise alignment. (e) To calculate the score in the second row, third column we again take the maximum of the three scores -4, -1, -4. The best score follows from the left cell (red arrow). (f) We proceed to fill in scores across the first row of the matrix. (g) The completed matrix includes the overall score of the optimal alignment (-4; see cell at bottom right, corresponding to the carboxy terminus of each protein). Red arrows indicate the path(s) by which the highest score for each cell was obtained.

3. a gap may be introduced from the first sequence, for which we assign a score of  $-2$ ; or
4. a gap may be introduced from the second sequence, also resulting in  $-2$ .

The Needleman and Wunsch algorithm provides a score corresponding to each of these possible outcomes for each position of the aligned sequences. The algorithm also specifies a set of rules describing how we can move through the matrix.

Consider the cell at the lower right-hand corner of **Figure 3.21c**. There are several rules for deciding the optimal score:

- First, both  $i$  and  $j$  must increase. We therefore evaluate scores from three positions (top, left, upper left), moving towards a given cell  $F(i, j)$ . It would not make sense to be able to violate the linear arrangement of amino acids (or nucleotides) in a sequence.
- It is acceptable for a gap to extend an arbitrary number of positions; a scoring system may include separate gap creation and gap extension penalties.
- The particular score that is assigned may come from a scoring matrix such as BLOSUM62.

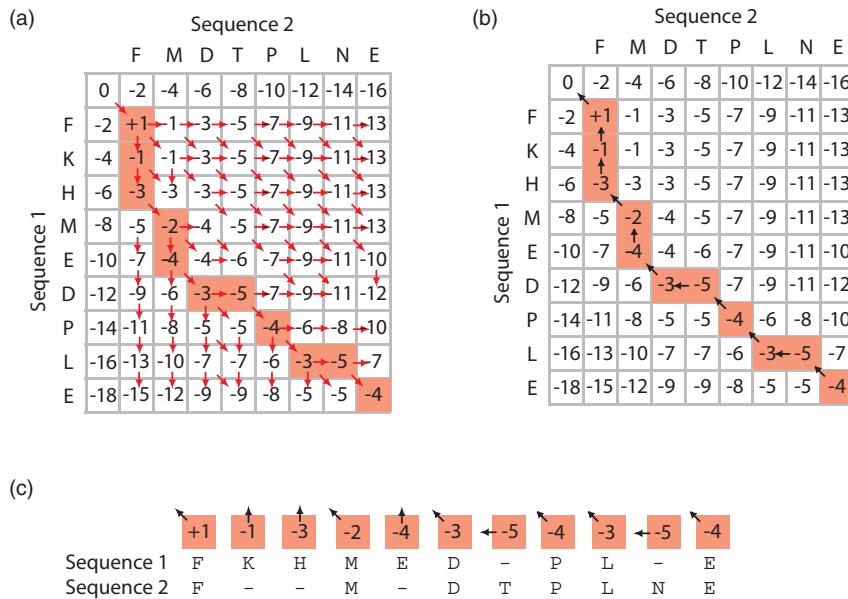
As we begin to align the two sequences in our example we fill in a cell with the value  $+1$  because of the alignment of two F residues (**Fig. 3.21d**). The alternative options of introducing a gap in either sequence would necessitate a gap penalty and a poorer score. We indicate the preferred (highest-scoring) path with a red arrow throughout **Figure 3.21**. We proceed to the next cell to the right, selecting the score of  $-1$  (coming from the left, consisting of  $+1$  (from the previous cell)  $-2$  (for introducing a gap)  $= -1$ ) as better than the alternative scores of  $-4$  and  $-6$  (**Fig. 3.21e**). This process of analyzing possible scores for each cell continues across each row (**Fig. 3.21f**) until the entire matrix is filled in (**Fig. 3.21g**).

### *Step 3: Identifying the Optimal Alignment*

After the matrix is filled, the alignment is determined by a trace-back procedure. Begin with the cell at the lower right of the matrix (carboxy termini of the proteins or 3' end of the nucleic acid sequences). In our example, this has a score of  $-4$  and corresponds to an alignment of two glutamate residues. For this and every cell we can determine from which of the three adjacent cells the best score was derived. This procedure is outlined in **Figure 3.22a**, in which red arrows indicate the paths from which the best scores were obtained for each cell. We therefore define a path (see pink-shaded cells) that will correspond to the actual alignment. In **Figure 3.22b**, we show just the arrows indicating from which cell each best score was derived. This is a different way of defining the optimal path of the pairwise alignment. We build that alignment, including gaps in either sequence, proceeding from the carboxy to the amino terminus. The final alignment (**Fig. 3.22c**) is guaranteed to be optimal, given this scoring system. There may be multiple alignments that share an optimal score, although this rarely occurs when scoring matrices such as BLOSUM62 are implemented.

A variety of programs implement global alignment algorithms (see Web Resources at the end of this chapter). An example is the Needle program from EMBOSS, which can be accessed via Galaxy (Box 3.9). Two bacterial globin family sequences are entered: one from *Streptomyces avermitilis* MA-4680 (NP\_824492.1, 260 amino acids); and another from *Mycobacterium tuberculosis* CDC1551 (NP\_337032.1, 134 amino acids). Penalties are selected for gap creation and extension, and each sequence is pasted into an input box in the FASTA format. The resulting global alignment includes descriptions of the percent identity and similarity shared by the two proteins, the length of the alignment, and the number of gaps introduced (**Fig. 3.23a**).

The Needle program for global pairwise alignment is part of the EMBOSS package available online at the European Bioinformatics Institute (<http://www.ebi.ac.uk/emboss/align/>, WebLink 3.5) or at Galaxy (<http://usegalaxy.org/>, WebLink 3.6). It is further described at the EMBOSS website under applications (<http://emboss.sourceforge.net/>, WebLink 3.7). The *E. coli* and *S. cerevisiae* proteins are available in the FASTA format, as well as globally and locally aligned in Web Document 3.6 (<http://www.bioinfbook.org/chapter3>).



**FIGURE 3.22** Global pairwise alignment of two amino acid sequences using a dynamic programming algorithm: scoring the matrix and using the trace-back procedure to obtain the alignments. (a) The alignment of Figure 3.21(g) is shown. The cells highlighted in pink represent the source of the optimal scores. (b) In an equivalent representation, arrows point back to the source of each cell’s optimal score. (c) This trace-back allows us to determine the sequence of the optimal alignment. Vertical or horizontal arrows correspond to the positions of gap insertions, while diagonal lines correspond to exact matches (or mismatches). Note that the final score ( $-4$ ) equals the sum of matches ( $6 \times 1 = 6$ ), mismatches (none in this example), and gaps ( $5 \times -2 = -10$ ).

The Needleman–Wunsch algorithm is an example of dynamic programming (Sedgewick, 1988). This means that an optimal path (i.e., an optimal alignment) is detected by incrementally extending optimal subpaths, that is, by making a series of decisions at each step of the alignment as to which pair of residues corresponds to the best score. The overall goal is to find the path moving along the diagonal of the matrix that lets us obtain the maximal score. This path specifies the optimal alignment.

### BOX 3.9 EMBOSS

EMBOSS (European Molecular Biology Open Software Suite) is a collection of freely available programs for DNA, RNA, or protein sequence analysis (Rice et al., 2000). There are over 200 available programs in three dozen categories. The home page of EMBOSS (<http://emboss.sourceforge.net/>, WebLink 3.21) describes the various packages. A variety of web servers offer EMBOSS, including Galaxy. You can also visit sites such as <http://emboss.bioinformatics.nl/> (WebLink 3.22) and <http://www.bioinformatics2.wsu.edu/emboss/> (WebLink 3.23).

To perform pairwise sequence alignment using EMBOSS at Galaxy, try the following steps:

1. Visit Galaxy at <https://main.g2.bx.psu.edu/> (WebLink 3.24) and sign in.
2. On the left sidebar (Tools menu) select Get Data and choose UCSC Main. From the human genome (hg19) select the RefGenes table, enter hbb for the position (upon clicking “lookup” the coordinates chr11:5246696-5248301 are added), set the output format to “sequence” and check the box to send to Galaxy. When you click “Get output” select protein and submit.
3. Repeat step (2) to import the HBA2 protein. For both proteins in Galaxy, use Edit Attributes (the pencil icon in the history panel) to rename the sequences hbb and hba2.
4. In the tools panel choose EMBOSS, and scroll to find the water tool for Smith–Waterman local alignment. Alternatively, enter “water” into the Tools search box. Select the two proteins, use default settings, and click Execute. The pairwise alignment is returned.

Once you have entered one or more sequences into Galaxy, explore some of the >100 other EMBOSS tools!

(a)

NP_824492.1	1 MCGDMTVHTVEYIRYRIPEQQSAEFLAAYTRAAQLAAPQCVDYELARC	50
NP_337032.1	1	0
NP_824492.1	51 EEDFEHFVLRITWTSTEDHIEGFRKSELFPDFLAEIRPYISSLIEEMRHYK	100
NP_337032.1	1	0
NP_824492.1	101 PTTVRGRTGAAPVPTLYAWAGGAEEAFARLTTEVFYEKVLKDDVLAPVFEGLMAP .: .....: ...   :.. ...   : .: : : .. :	150
NP_337032.1	1 MEGMDQMPKSFYDAVGGAKTFDAIVSRFYAQVAEDEVLRRVY---P	43
NP_824492.1	151 EH----AAHVALWLGEVFGGPAAAYSETQGGHGHMVAHLGKNITEVQRR  . .....: ..: ...     .  ...: ... : .. .	195
NP_337032.1	44 EDDLAGAERLRLMFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISSLIERD	92
NP_824492.1	196 RWVNLLQDAADDAGLPT-DAEFRSAFLAYAEWGTRLAVYFSGPDAVPPAE .  : ... .....   . .  ... . .  ...  : .. .	244
NP_337032.1	93 AWLRCMHTAVASIDSETLDDEHRRELLDYLEMAAHSLV--NSPF	134
NP_824492.1	245 QPVQPQWSWGAMPPYQP	260
NP_337032.1	135	134

(b)

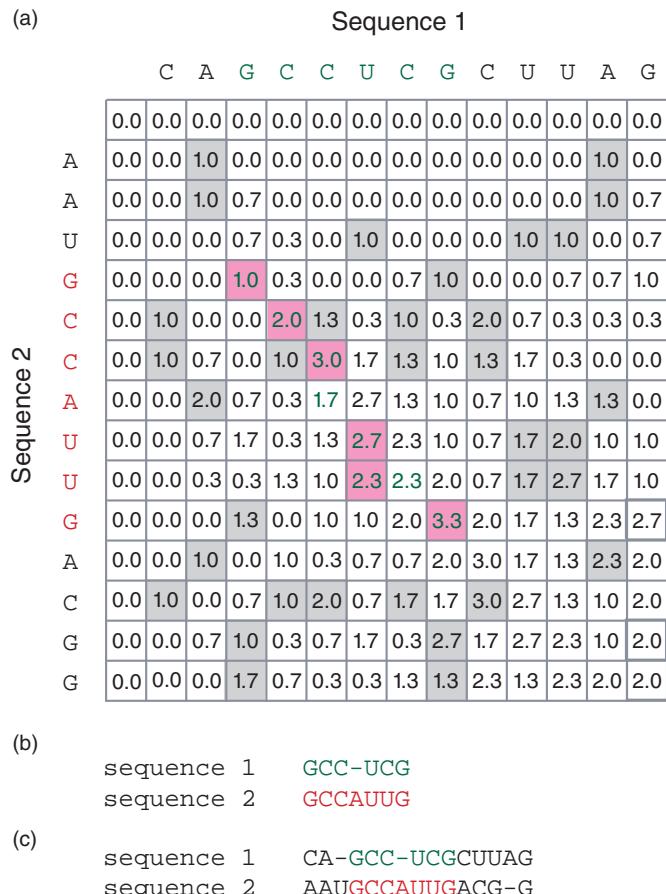
NP_824492.1	113 TLYAWAGGAEEAFARLTTEVFYEKVLKDDVLAPVFEGLMAPEH----AAHVA .: ..... : .. ...   : .: : .. : .. : .. .	157
NP_337032.1	10 SFYDAVGGAKTFDAIVSRFYAQVAEDEVLRRVY---PEDDLAGAERL	55
NP_824492.1	158 LWLGEVFGGPAAAYSETQGGHGHMVAHLGKNITEVQRRRWVNLLQDAADD .  : ..: ...     .  ...: ...: ... : .. : .. .	207
NP_337032.1	56 MFLEQYWGGPRTYSE-QRGHPRLRMRHAPFRISSLIERDAWLRCMHTAVAS	104
NP_824492.1	208 AGLPT-DAEFRSAFLAYAE	225
NP_337032.1	105 IDSETLDDEHRRELLDYLE	123

**FIGURE 3.23** (a) Global pairwise alignment of bacterial proteins containing globin domains from *Streptomyces avermitilis* MA-4680 (NP\_824492) and *Mycobacterium tuberculosis* CDC1551 (NP\_337032). The scoring matrix was BLOSUM62. The aligned proteins share 14.7% identity (39/266 aligned residues), 22.6% similarity (60.266), and 51.9% gaps (138/266). (b) A local pairwise alignment of these two sequences lacks the unpaired amino- and carboxy-terminal extensions and shows 30% identity (35/115 aligned residues). The alignment in (b) corresponds to the shaded region of (a). The arrowheads in (a) indicate aligned residues that were not seen in the local alignment. In performing local alignments (as is done in BLAST, Chapter 4) some authentically aligned regions may therefore be missed.

### Local Sequence Alignment: Smith and Waterman Algorithm

The local alignment algorithm of Smith and Waterman (1981) is the most rigorous method by which subsets of two protein or DNA sequences can be aligned. Local alignment is extremely useful in a variety of applications such as database searching in which we may wish to align domains of proteins (but not the entire sequences). A local sequence alignment algorithm resembles that for global alignment in that two proteins are arranged in a matrix and an optimal path along a diagonal is sought. However, there is no penalty for starting the alignment at some internal position, and the alignment does not necessarily extend to the ends of the two sequences.

For the Smith–Waterman algorithm a matrix is constructed with an extra row along the top and an extra column on the left side. For sequences of lengths  $m$  and  $n$ , the matrix has dimensions  $m + 1$  and  $n + 1$ . The rules for defining the value at each position of the matrix differ slightly from those used in the Needleman–Wunsch algorithm. The score in



**FIGURE 3.24** Local sequence alignment method of Smith and Waterman (1981). (a) In this example, the matrix is formed from two RNA sequences (CAGCCUCGCUUAG and AAUGCCAUUGACGG). While this is not an identity matrix (such as that shown in Fig. 3.21a), positions of nucleotide identity are shaded gray (or shaded pink in the region of local alignment). They scoring system here is +1 for a match, minus one-third for a mismatch, and a gap penalty of the difference between a match and a mismatch (-1.3 for a gap of length one). The matrix is scored based on finding the maximum of four possible non-negative values. The highest value in the matrix (3.3) corresponds to the beginning of the optimal local alignment, and the aligned residues (green font) extend up and to the left until a value of zero is reached. (b) The local alignment derived from this matrix is shown. Note that this alignment includes identities, a mismatch, and a gap. (c) A global alignment of the two sequences is shown for comparison to the local alignment. Note that it encompasses the entirety of both sequences.

Source: Adapted from Smith and Waterman (1981). Reproduced with permissions from Elsevier.

each cell is selected as the maximum of the preceding diagonal or the score obtained from the introduction of a gap. However, the score cannot be negative: a rule introduced by the Smith–Waterman algorithm is that, if all other score options produce a negative value, then a zero must be inserted in the cell. The score  $S(i,j)$  is given as the maximum of four possible values (Fig. 3.24):

1. The score from the cell at position  $i - 1, j - 1$ , that is, the score diagonally up to the left. To this score, add the new score at position  $s[i, j]$ , which consists of either a match or a mismatch.
2.  $S(i, j - 1)$  (i.e., the score one cell to the left) minus a gap penalty.

3.  $S(i-1, j)$  (i.e., the score immediately above the new cell) minus a gap penalty.
4. The number zero.

This condition ensures that there are no negative values in the matrix. In contrast, negative numbers commonly occur in global alignments because of gap or mismatch penalties (note the log-odds matrices in this chapter).

An example of the use of a local alignment algorithm to align two nucleic acid sequences adapted from Smith and Waterman (1981) is shown in **Figure 3.24**. The topmost row and the leftmost column of the matrix are filled with zeros. The maximal alignment can begin and end anywhere in the matrix (within reason; the linear order of the two amino acid sequences cannot be violated). The procedure is to identify the highest value in the matrix (this value is 3.3 in **Fig. 3.24a**). This represents the end (3' end for nucleic acids, or carboxy-terminal portion proteins) of the alignment. This position is not necessarily at the lower right corner as it must be for a global alignment. The trace-back procedure begins with this highest-value position and proceeds diagonally up to the left until a cell is reached with a value of zero. This defines the start of the alignment, and it is not necessarily at the extreme top left of the matrix.

An example of a local alignment of two proteins using the Smith–Waterman algorithm is shown in **Figure 3.23b**. Compare this with the global alignment of **Figure 3.23a** and note that the aligned region is shorter for the local alignment, while the percent identity and similarity are higher. Note also that the local alignment ignores several identically matching residues (**Fig. 3.23a**, arrowheads). Since database searches such as BLAST (Chapter 4) rely on local alignments, there may be conserved regions that are not reported as aligned, depending on the chosen search parameters.

### Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST

While the Smith–Waterman algorithm is guaranteed to find the optimal alignment(s) between two sequences, it suffers from the fact that it is relatively slow. For pairwise alignment, speed is usually not a problem. When a pairwise alignment algorithm is applied to the problem of comparing one sequence (a “query”) to an entire database however, the speed of the algorithm becomes a significant issue and may vary by orders of magnitude.

Most algorithms have a parameter  $N$  that refers to the number of data items to be processed (see Sedgewick, 1988). This parameter can greatly affect the time required for the algorithm to perform a task. If the running time is proportional to  $N$ , then doubling  $N$  doubles the running time. If the running time is quadratic ( $N^2$ ), then for  $N = 1000$  the running time is one million. For both the Needleman–Wunsch and the Smith–Waterman algorithms, both the computer space and the time required to align two sequences is proportional to at least the length of the two query sequences multiplied by each other ( $m \times n$ ). For the search of a database of size  $N$ , this is  $m \times N$ .

Another useful descriptor is  $O$ -notation (called “big-Oh notation”) which provides an approximation of the upper bounds of the running time of an algorithm. The Needleman–Wunsch algorithm requires  $O(mn)$  steps, while the Smith–Waterman algorithm requires  $O(m^2n)$  steps. Subsequently, Gotoh (1982) and Myers and Miller (1988) improved the algorithms so they require less time and space.

Two popular local alignment algorithms have been developed that provide rapid alternatives to Smith–Waterman: FASTA (Pearson and Lipman, 1988) and BLAST (Basic Local Alignment Search Tool; Altschul *et al.*, 1990). Each of these algorithms requires less time to perform an alignment. The time saving occurs because FASTA and BLAST restrict the search by scanning a database for likely matches before performing more

The modified alignment algorithms introduced by Gotoh (1982) and Myers and Miller (1988) require only  $O(nm)$  time and occupy  $O(n)$  in space. Instead of committing the entire matrix to memory, the algorithms ignore scores below a threshold in order to focus on the maximum scores that are achieved during the search.

FASTA stands for FAST-All, referring to its ability to perform a fast alignment of all sequences (i.e., proteins or nucleotides).

The parameter *ktup* refers to multiples such as duplicate, triplicate, or quadruplicate (for  $k=2$ ,  $k=3$ ,  $k=4$ ). The *ktup* values are usually 3–6 for nucleotide sequences and 1–2 for amino acid sequences. A small *ktup* value yields a more sensitive search but requires more time to complete.

William Pearson of the University of Virginia provides FASTA online. Visit [http://fasta.bioch.virginia.edu/fasta\\_www2/fasta\\_list2.shtml](http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml) (WebLink 3.8). Another place to try FASTA is at the European Bioinformatics Institute website, <http://www.ebi.ac.uk/fasta33/> (WebLink 3.9).

Dotlet is a web-based diagonal plot tool available from the Swiss Institute of Bioinformatics (<http://myhits.isb-sib.ch/cgi-bin/dotlet>, WebLink 3.10). It was written by Marco Pagni and Thomas Junier. The website provides examples of the use of Dotlet to visualize repeated domains, conserved domains, exons and introns, terminators, frameshifts, and low-complexity regions.

The accession number of the snail globin is CAJ44466.1, while the accession of human cytoglobin is NP\_599030.1.

rigorous alignments. These are heuristic algorithms (Box 3.3) that sacrifice some sensitivity in exchange for speed; in contrast to Smith–Waterman, they are not guaranteed to find optimal alignments.

The FASTA search algorithm introduced by Pearson and Lipman (1988) proceeds in four steps.

1. A lookup table is generated consisting of short stretches of amino acids or nucleotides from a database. The size of these stretches is determined from the *ktup* parameter. If *ktup* = 3 for a protein search, then the query sequence is examined in blocks of three amino acids against matches of three amino acids found in the lookup table. The FASTA program identifies the 10 highest scoring segments that align for a given *ktup*.
2. These 10 aligned regions are rescored, allowing for conservative replacements, using a scoring matrix such as PAM250.
3. High-scoring regions are joined together if they are part of the same proteins.
4. FASTA then performs a global (Needleman–Wunsch) or local (Smith–Waterman) alignment on the highest scoring sequences, thus optimizing the alignments of the query sequence with the best database matches.

Dynamic programming is therefore applied to the database search in a limited fashion, allowing FASTA to return its results very rapidly because it evaluates only a portion of the potential alignments.

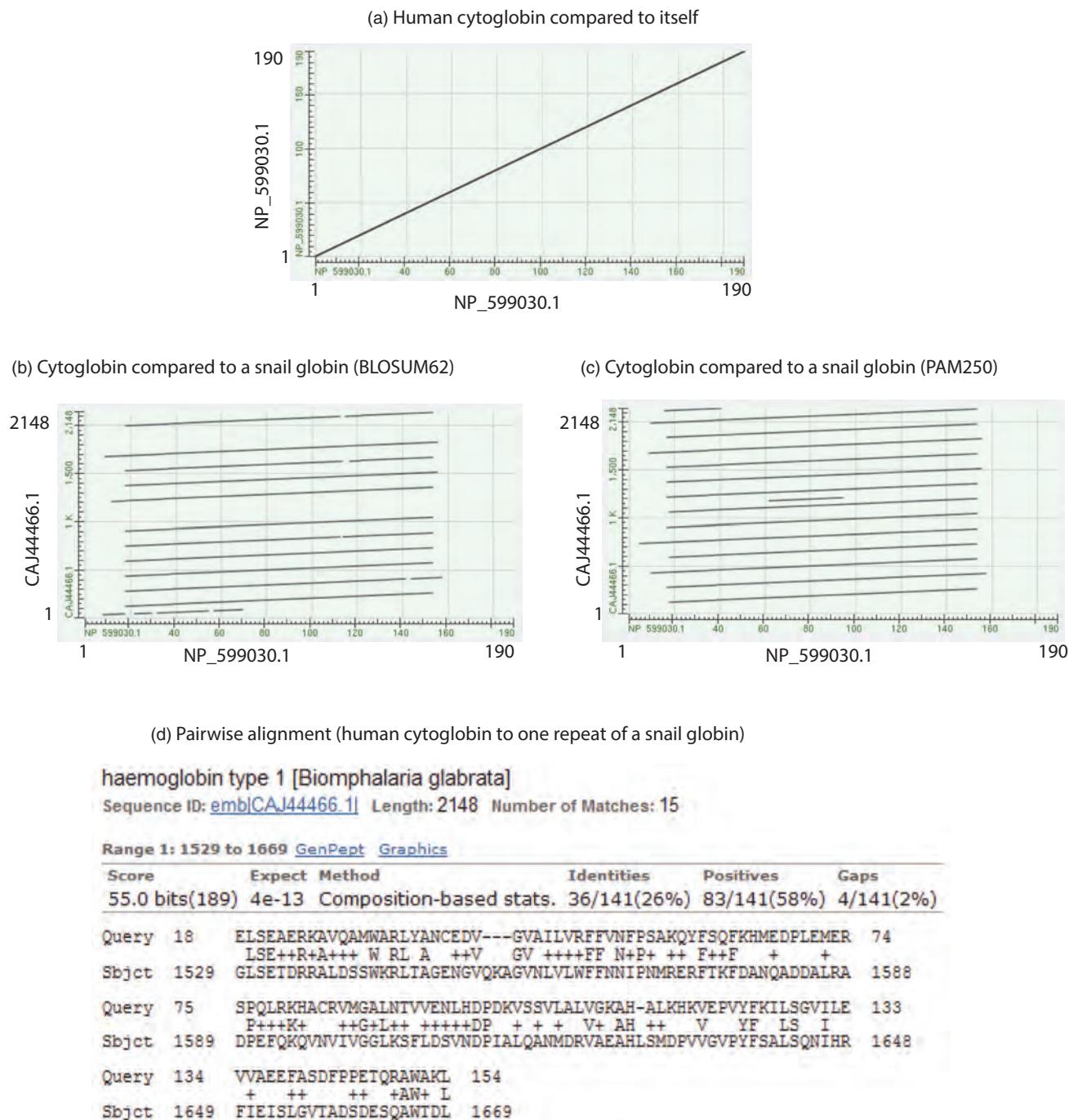
## Basic Local Alignment Search Tool (BLAST)

BLAST was introduced as a local alignment search tool that identifies alignments between a query sequence and a database without the introduction of gaps (Altschul *et al.*, 1990). The version of BLAST that is available today allows gaps in the alignment. We provided an example of an alignment of two proteins (Figs 3.4 and 3.5) and introduce BLAST in more detail in Chapter 4, where we describe its heuristic algorithm.

## Pairwise Alignment with Dotplots

In addition to displaying a pairwise alignment, the BLAST output includes a dotplot (or dot matrix), which is a graphical method for comparing two sequences. One protein or nucleic acid sequence is placed along the *x* axis and the other is placed along the *y* axis. Positions of identity are scored with a dot. A region of identity between two sequences results in the formation of a diagonal line. This is illustrated for an alignment of human cytoglobin with itself as part of the BLAST output (Fig. 3.25a). We also illustrate a dotplot using the web-based Dotlet program of Junier and Pagni (2000; Web Document 3.7). Dotlet features an adjustable sliding window size, a zoom feature, a variety of scoring matrices, and a histogram window to adjust the pixel intensities in order to manually optimize the signal-to-noise ratio.

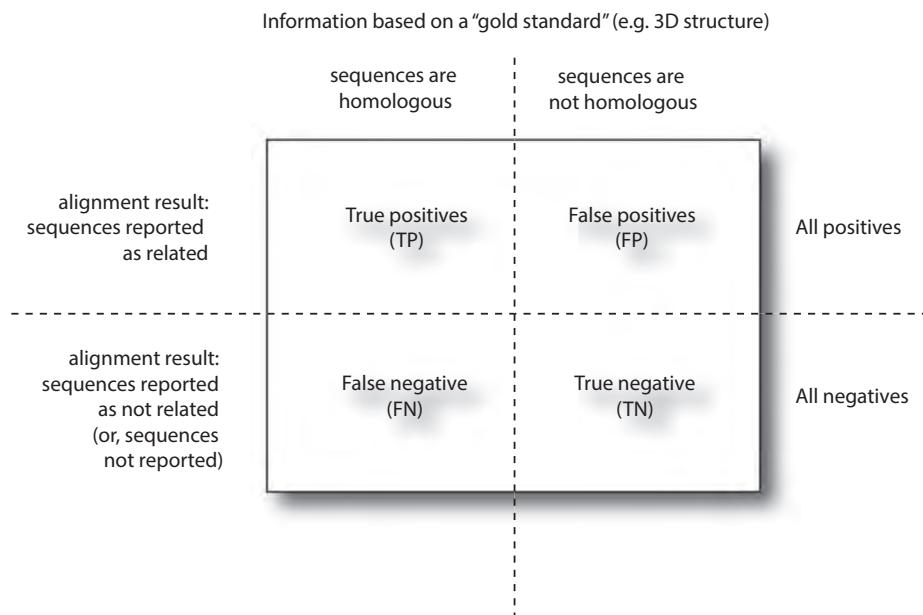
We can further illustrate the usefulness of dotplots by examining an unusual hemoglobin protein of 2148 amino acids from the snail *Biomphalaria glabrata*. It consists of 13 globin repeats (Lieb *et al.*, 2006). When we compare it to human cytoglobin (190 amino acids) with a default BLOSUM62 matrix, the BLAST output shows cytoglobin (*x* axis) matching the snail protein 12 times (*y* axis) (Fig. 3.25b); one repeat is missed. By changing the scoring matrix to BLOSUM45 we can now see all 13 snail hemoglobin repeats (Fig. 3.25c). The gap at the start of the dotplot (Fig. 3.25c, position 1 to the first red arrowhead on *x* axis) is evident in the pairwise alignment of that region (Fig. 3.25d): the first 128 amino acids of the snail protein are unrelated and therefore not aligned with cytoglobin. Using Dotlet, all 13 globin repeats are evident in a comparison of the snail protein with itself or with cytoglobin (Web Document 3.7).



**FIGURE 3.25** Dot matrix plots in the output of the NCBI BLASTP program permit visualization of matching domains in pairwise protein alignments. The program is used as described in **Figure 3.4**. (a) For a comparison of human cytoglobin (NP\_599030.1, length 190 amino acids) with itself, the output includes a dotplot shown with sequences 1 and 2 (both cytoglobin) on the x and y axes, and the data points showing amino acid identities appear as a diagonal line. (b) For a comparison of cytoglobin with a globin from the snail *Biomphalaria glabrata* (accession CAJ44466.1, length 2148 amino acids), the cytoglobin sequence (x axis) matches 12 times with internal globin repeats in the snail protein. This search uses the default BLOSUM62 scoring matrix. (c) Changing the scoring matrix to PAM250 enables all 13 globin repeats of the snail protein to be aligned with cytoglobin. (d) A pairwise alignment of the sequences shows that the snail globin repeats align with residues 18–154 of cytoglobin. This is reflected in the dotplots, where the portion on the x axis corresponding to cytoglobin residues 1–17 and 155–190 (see red arrowheads in (c)) do not align to the snail sequence. The BLASTP output produces a set of all the pairwise alignments of which the first is shown here.

Source: BLASTP, NCBI.

We encounter dotplots in Chapter 16 when we compare viral genome sequences to each other. We also see a dotplot in Chapter 18 (on fungi). Protein sequences from *Saccharomyces cerevisiae* chromosomes were systematically BLASTP searched against each other. The resulting dotplot (**Fig. 18.10**) showed many diagonal lines, indicating homologous regions. This provided evidence that, surprisingly, the entire yeast genome duplicated over 100 million years ago.



**FIGURE 3.26** Sequences alignments, whether pairwise (this chapter) or from a database search (Chapter 4), can be classified as true or false and positives or negatives. Statistical analyses of alignments provide the main method of evaluating whether an alignment represents a true positive, that is, an alignment of homologous sequences. Ideally, an alignment algorithm can maximize both sensitivity and specificity.

## THE STATISTICAL SIGNIFICANCE OF PAIRWISE ALIGNMENTS

How can we decide whether the alignment of two sequences is statistically significant? We address this question for local alignments and then for global alignments.

Consider two proteins that share limited amino acid identity (e.g., 20–25%). Alignment algorithms report the score of a pairwise alignment or the score of the best alignments of a query sequence against an entire database of sequences (Chapter 4). We need statistical tests to decide whether the matches are true positives (i.e., whether the two aligned proteins are genuinely homologous) or whether they are false positives (i.e., whether they have been aligned by the algorithm by chance; **Fig. 3.26**). For the alignments that are not reported by an algorithm, for instance because the score falls below some threshold, we would like to evaluate whether the sequences are true negatives (i.e., genuinely unrelated) or whether they are false negatives, that is, homologous sequences that receive a score suggesting that they are not homologous.

A main goal of alignment algorithms is therefore to maximize the sensitivity and specificity of sequence alignments (**Fig. 3.26**). Sensitivity is the number of true positives divided by the sum of true positive and false negative results. This is a measure of the ability of an algorithm to correctly identify genuinely related sequences. Specificity is the number of true negative results divided by the sum of true negative and false positive results. This describes the sequence alignments that are not homologous.

### Statistical Significance of Global Alignments

When we align two proteins, such as human beta globin and myoglobin, we obtain a score. We can use hypothesis testing to assess whether that score is likely to have occurred by chance. To do this, we first state a null hypothesis ( $H_0$ ) that the two sequences are not related. According to this hypothesis, the score  $S$  of beta globin and myoglobin represents a chance occurrence. We then state an alternative hypothesis ( $H_1$ ) that they are indeed

### BOX 3.10 STATISTICAL CONCEPTS: Z-SCORES

The familiar bell-shaped curve is a Gaussian distribution or normal distribution. The  $x$  axis corresponds to some measured values, such as the alignment score of beta globin versus 100 randomly shuffled versions of myoglobin. The  $y$  axis corresponds to the probability density (when considering measurements of an exhaustive set of shuffled myoglobins) or to the number of trials (when considering a number of shuffled myoglobins). The mean value is obtained simply by adding all the scores and dividing by the number of pairwise alignments; it is apparent at the center of a Gaussian distribution. For a set of data points  $x_1, x_2, x_3, \dots, x_n$  the mean  $\bar{x}$  is the sum divided by  $n$ , or:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

The sample variance  $s^2$  describes the spread of the data points from the mean. It is related to the squares of the distances of the data points from the mean, and it is given by:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The sample standard deviation  $s$  is the square root of the variance, so its units match those of the data points. It is defined:

$$s = \sqrt{\frac{\sum_{i=1}^N (Y_i - m)^2}{N-1}}.$$

Note that  $s$  is the sample standard deviation (rather than the population standard deviation,  $\sigma$ ) and  $s^2$  is the sample variance. Population variance refers to the average of the square of the deviations of each value from the mean, while the sample variance includes an adjustment from number of measurements  $N$ ;  $m$  is the sample mean (rather than the population mean,  $\mu$ ). Z-scores (also called standardized scores) describe the distance from the mean per standard deviation:

$$Z_i = \frac{x_i - \bar{x}}{s}.$$

If you compare beta globin to myoglobin, you can get a score (such as 43.9 as shown in Fig. 3.5a) based on some scoring system. Randomly scramble the sequence of myoglobin 1000 times (maintaining the length and composition of the myoglobin), and measure the 1000 scores of beta globin to these scrambled sequences. You can obtain a mean and standard deviation of the comparison to shuffled sequences. For more information on statistical concepts, see Motulsky (1995) and Cumming *et al.* (2007).

related. We choose a significance value  $\alpha$ , often set to 0.05, as a threshold for defining statistical significance. One approach to determining whether our score occurred by chance is to compare it to the scores of beta globin or myoglobin relative to a large number of other proteins (or DNA sequences) known to be not homologous. Another approach is to compare the query to a set of randomly generated sequences. A third approach is to randomly scramble the sequence of one of the two query proteins (e.g., myoglobin) and obtain a score relative to beta globin; by repeating this process 100 times, we can obtain the sample mean ( $\bar{x}$ ) and sample standard deviation ( $s$ ) of the scores for the randomly shuffled myoglobin relative to beta globin. We can express the authentic score in terms of how many standard deviations above the mean it is. A Z score (Box 3.10) is calculated as:

$$Z = \frac{x - \mu}{s} \quad (3.8)$$

where  $x$  is the score of two aligned sequences,  $\mu$  is the mean score of many sequence comparisons using a scrambled sequence, and  $s$  is the standard deviation of those measurements obtained with random sequences. We can do the shuffle test using an algorithm such as PRSS. This calculates the score of a global pairwise alignment, and also performs comparisons of one protein to a randomized (jumbled) version of the other.

If the scores are normally distributed, then the Z statistic can be converted to a probability value. If  $Z = 3$ , then we can refer to a table in a standard statistics resource to see

PRSS, written by William Pearson, is available online at [http://fasta.bioch.virginia.edu/fasta\\_www2/fasta\\_www.cgi?rm=shuffle](http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle) (WebLink 3.11). For an example of PRSS output for a comparison of human beta globin and myoglobin, see Web Document 3.8 at <http://www.bioinfbook.org/chapter3>.

For local pairwise alignments, the best approach to defining statistical significance is to estimate an expect value (*E* value) which is closely related to a probability value (*p* value). In contrast to the situation with global alignment, for local alignment there is a thorough understanding of the distribution of scores. An *E* value describes the number of matches having a particular score (or better) that are expected to occur by chance. For example, if a pairwise alignment of a beta globin and a myoglobin has some score with an associated *E* value of  $10^{-3}$ , that particular score (or better) can be expected one time in one thousand by chance. This is the approach taken by the BLAST family of programs; we discuss *E* values in detail in Chapter 4.

The accession numbers of rat and bovine odorant-binding proteins are NP\_620258.1 and P07435.2; the human protein closest to rat has accession EAW50553.1. The alignments of these proteins are shown in Web Document 3.9 at <http://www.bioinfbook.org/chapter3>.

that 99.73% of the population (i.e., of the scores) are within three standard deviations of the mean, and the fraction of scores that are greater than three standard deviations beyond the mean is only 0.13%. We can expect to see this particular score by chance about 1 time in 750 (i.e., 0.13% of the time). The problem in adopting this approach is that if the distribution of scores deviates from a Gaussian distribution the estimated significance level will be wrong. For global (but not local) pairwise alignments, the distribution is generally not Gaussian; there is therefore not a strong statistical basis for assigning significance values to pairwise alignments. What can we conclude from a *Z* score? If 100 alignments of shuffled proteins all have a score less than the authentic score of two aligned proteins, this indicates that the probability (*p*) that this occurred by chance is less than 0.01. (We can therefore reject the null hypothesis that the two protein sequences are not significantly related.) However, because of the concerns about the applicability of the *Z* score to sequence scores, conclusions about statistical significance should be made with caution.

Another consideration involves the problem of multiple comparisons. If we compare a query such as beta globin to one million proteins in a database, we have a million opportunities to find a high-scoring match between the query and some database entry. In such cases it is appropriate to adjust the significance level  $\alpha$ , that is, the probability at which the null hypothesis is rejected, to a more stringent level. One approach, called a Bonferroni correction, is to divide  $\alpha$  (nominally  $p < 0.05$ ) by the number of trials ( $10^6$ ) to set a new threshold for defining statistical significance of level of  $0.05/10^6$ , or  $5 \times 10^{-8}$ . The equivalent of a Bonferroni correction is applied to the probability value calculation of BLAST statistics (see Chapter 4), and we also encounter multiple comparison corrections in microarray data analysis (see Chapter 11).

## Statistical Significance of Local Alignments

Most database search programs such as BLAST (Chapter 4) depend on local alignments. Additionally, many pairwise alignment programs compare two sequences using local alignment.

## Percent Identity and Relative Entropy

One approach to deciding whether two sequences are significantly related from an evolutionary point of view is to consider their percent identity. It is very useful to consider the percent identity that two proteins share in order to obtain a sense of their degree of relatedness. As an example, a global pairwise alignment of odorant-binding protein from rat and cow reveals only 30% identity, although both are functionally able to bind odorants with similar affinities (Pevsner *et al.*, 1985). The rat protein shares just 26% identity to its closest human ortholog. From a statistical perspective the inspection of percent identities has limited usefulness in the “twilight zone;” it does not provide a rigorous set of rules for inferring homology and it is associated with false positive or false negative results. A high degree of identity over a short region might sometimes not be evolutionarily significant, and conversely a low percent identity could reflect homology. Percent amino acid identity alone is not sufficient to demonstrate (or rule out) homology.

Still, it may be useful to consider percent identity. Some researchers have suggested that if two proteins share 25% or more amino acid identity over a span of 150 or more amino acids, they are probably significantly related (Brenner *et al.*, 1998). If we consider an alignment of just 70 amino acids, it is popular to consider the two sequences “significantly related” if they share 25% amino acid identity. However, Brenner *et al.* (1998) have shown that this may be erroneous, partly because the enormous size of today’s molecular sequence databases increases the likelihood that such alignments occur by chance. For an alignment of 70 amino acid residues, 40% amino acid identity is a reasonable threshold to estimate that two proteins are homologous (Brenner *et al.*, 1998). If two proteins share

### BOX 3.11 RELATIVE ENTROPY

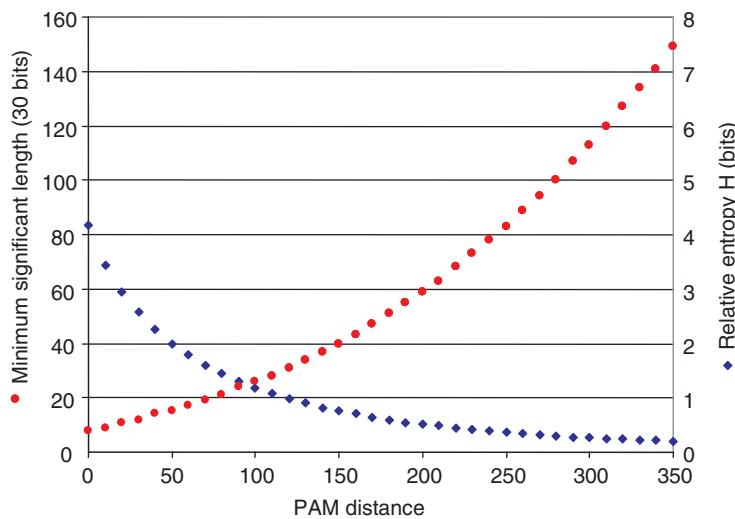
Altschul (1991) estimated that about 30 bits of information are required to distinguish an authentic alignment from a chance alignment of two proteins of average size (given that one protein is used against a database of a particular size). For each substitution matrix with its unique target frequencies  $q_{ij}$  and background distributions  $p_i p_j$ , it is possible to derive the relative entropy  $H$  as follows (Altschul, 1991):

$$H = \sum_{i,j} q_{i,j} s_{i,j} = \sum_{i,j} q_{i,j} \log_2 \frac{q_{i,j}}{p_i p_j}$$

where  $H$  corresponds to the information content of the target and background distributions associated with a particular scoring matrix (units nats). As shown in **Figure 3.27**, for higher  $H$  values it is easier to distinguish the target from background frequencies. This analysis is consistent with the analysis of the diagonals for the PAM1 and PAM250 mutation probability matrices (**Figs 3.9** and **3.13**) in which there is far less signal apparent in the PAM250 matrix.

about 20–25% identity over a reasonably long stretch (e.g., 70–100 amino acid residues), they are in the “twilight zone” (**Fig. 3.19**) and it is more difficult to be sure. Two proteins that are completely unrelated often share about 10–20% identity when aligned. This is especially true because the insertion of gaps can greatly improve the alignment of any two sequences.

Altschul (1991) evaluated alignment scores from an information theory perspective. Target frequencies vary as a function of evolutionary distance. Recall that an alignment of alanine with threonine is assigned a different score in a PAM10 matrix (-3; see **Fig. 3.15**) than in a PAM250 matrix (+1; see **Fig. 3.14**). The relative entropy ( $H$ ) of the target and background distributions measures the information that is available per aligned amino acid position that, on average, distinguishes a true alignment from a chance alignment (Box 3.11). For a PAM10 matrix, the value of  $H$  is 3.43 bits. Assuming that 30 bits of information are sufficient to distinguish a true rather than a chance alignment in a database search, an alignment of at least 9 residues is needed using a PAM10 matrix (**Fig. 3.27**).



**FIGURE 3.27** Relative entropy ( $H$ ) as a function of PAM distance. For PAM matrices with low value (e.g., PAM10), the relative entropy in bits is high and the minimum length required to detect a significantly aligned pair of sequences is short (e.g., about 10 amino acids). Using a PAM10 matrix, two very closely related proteins can therefore be detected as homologous even if only a relatively short region of amino acid residues is compared. For PAM250 and other PAM matrices with high values, the relative entropy (or information content in the sequence) is low, and it is necessary to have a longer region of amino acids (e.g., 80 residues) aligned in order to detect significant relationships between two proteins. Adapted from Altschul (1991).

**TABLE 3.4 Global pairwise alignment algorithms.**

Program	Site	URL
BLAST	NCBI	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>
Needle EMBOSS package (global pairwise alignment)	EBI	<a href="http://www.ebi.ac.uk/Tools/emboss/">http://www.ebi.ac.uk/Tools/emboss/</a>
Water EMBOSS package (local pairwise alignment)	EBI	<a href="http://www.ebi.ac.uk/emboss/align/">http://www.ebi.ac.uk/emboss/align/</a>
Pairwise	Two Sequence Alignment Tool (global and local options)	<a href="http://informagen.com/Applets/Pairwise/">http://informagen.com/Applets/Pairwise/</a>
Stretcher	Institut Pasteur; global alignment	<a href="http://bioweb2.pasteur.fr/docs/EMBOSS/stretcher.html">http://bioweb2.pasteur.fr/docs/EMBOSS/stretcher.html</a>

For a PAM250 matrix however, the relative entropy is 0.36 and an alignment of at least 83 residues are needed to distinguish authentic alignments.

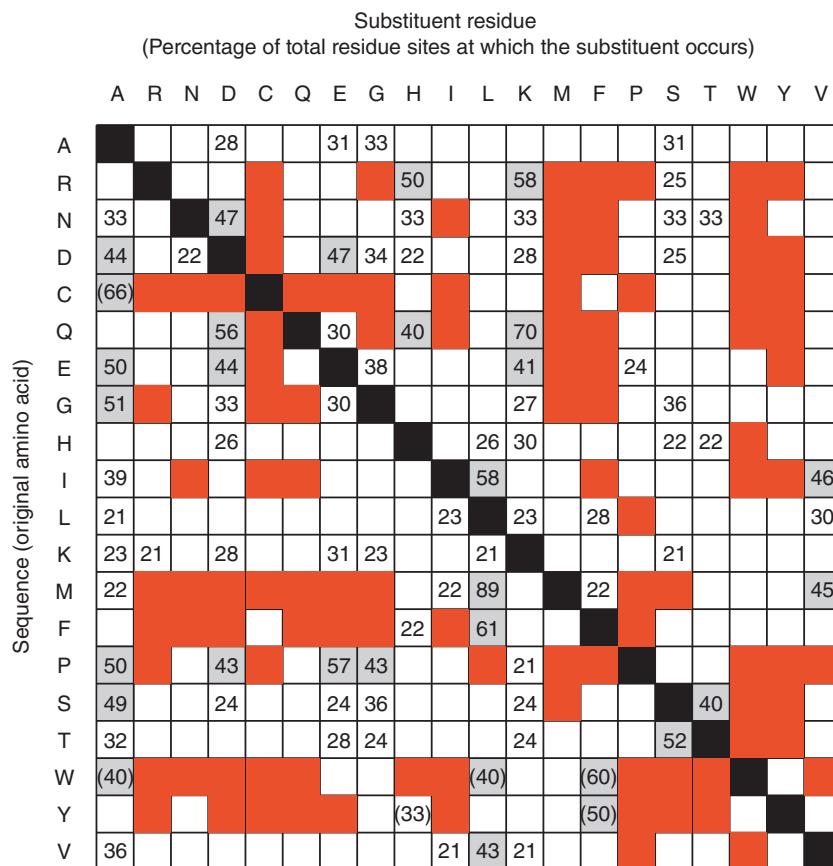
We see in Chapter 5 that scoring matrices (“profiles”) can be customized to a sequence alignment, greatly increasing the sensitivity of a search. We also see in Chapters 5 and 6 that multiple sequence alignments can offer far greater sensitivity than pairwise sequence alignment.

## PERSPECTIVE

The pairwise alignment of DNA or protein sequences is one of the most fundamental operations of bioinformatics. Pairwise alignment allows the relationship between any two sequences to be determined, and the degree of relatedness that is observed helps in the forming of a hypothesis about whether they are homologous (descended from a common evolutionary ancestor). Almost all of the topics in the rest of this book are heavily dependent upon sequence alignment. In Chapter 4, we introduce the searching of large DNA and/or protein databases with a query sequence. Database searching typically involves an extremely large series of local pairwise alignments, with results returned as a rank order beginning with most related sequences.

**TABLE 3.5 Local pairwise alignment algorithms.**

Resource	Description	URL
BLAST	At NCBI	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>
est2genome	EMBOSS program from the Institut Pasteur; aligns expressed sequence tags to genomic DNA	<a href="http://bioweb.pasteur.fr/docs/EMBOSS/est2genome.html">http://bioweb.pasteur.fr/docs/EMBOSS/est2genome.html</a>
LALIGN	Finds multiple matching subsegments in two sequences	<a href="http://www.ch.embnet.org/software/LALIGN_form.html">http://www.ch.embnet.org/software/LALIGN_form.html</a>
Pairwise	Two sequence alignment tool (global and local options)	<a href="http://informagen.com/Applets/Pairwise/">http://informagen.com/Applets/Pairwise/</a>
PRSS	From the University of Virginia (Bill Pearson)	<a href="http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle">http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle</a>
SIM	Alignment tool for protein sequences from ExPASy	<a href="http://web.expasy.org/sim/">http://web.expasy.org/sim/</a>
SSEARCH	At the Protein Information Resource	<a href="http://pir.georgetown.edu/pirwww/search/pairwise.shtml">http://pir.georgetown.edu/pirwww/search/pairwise.shtml</a>



**FIGURE 3.28** Substitution frequencies of globins (adapted from Zuckerkandl and Pauling, 1965, p. 118). Amino acids are presented alphabetically according to the three-letter abbreviations. The rows correspond to an original amino acid in an alignment of several dozen hemoglobin and myoglobin protein sequences from human, other primates, horse, cattle, pig, lamprey, and carp. Numbers represent the percentages of residue sites at which a given substitution occurs. For example, a glycine substitution was observed to occur in 33% of all the alanine sites. Substitutions that were never observed to occur are indicated by squares colored red. Rarely occurring substitutions (percentages <20%) are indicated by empty white squares (numerical values are not given). “Very conservative” substitutions (percentages  $\geq 40\%$ ) are in boxes shaded gray. For example, in 89% of the sites containing a methionine, leucine was also observed to be present. Identities are indicated by black solid squares. Values in parentheses indicate a very small available sample size, suggesting that conclusions about those data should be made cautiously.

Source: Zuckerkandl and Pauling (1965).

The algorithms used to perform pairwise alignment were developed in the 1970s, beginning with the global alignment procedure of Needleman and Wunsch (1970). Dayhoff (1978) introduced PAM scoring matrices that permit the comparison and evaluation of distantly related molecular sequences. Scoring matrices are an integral part of all pairwise (or multiple) sequence alignments, and the choice of a scoring matrix can strongly influence the outcome of a comparison. By the 1980s, local alignment algorithms were introduced (see the work of Sellers, 1974; Smith and Waterman, 1981; Smith *et al.*, 1981). Practically, pairwise alignment is performed today with a limited group of software packages, most of which are freely available.

The sensitivity and specificity of the available pairwise sequence alignment algorithms continue to be assessed. Recent areas in which pairwise alignment has been further

developed include methods of masking low-complexity sequences (to be discussed in Chapter 4) and theoretical models for penalizing gaps in alignments.

## PITFALLS

The optional parameters that accompany a pairwise alignment algorithm can greatly influence the results. A comparison of the homologs human RBP4 and bovine  $\beta$ -lactoglobulin using BLAST 2 Sequences results in no match detected if the default parameters are used.

Any two sequences can be aligned, even if they are unrelated. In some cases, two proteins that share even greater than 30% amino acid identity over a stretch of 100 amino acids are not homologous (evolutionarily related). It is always important to assess the biological significance of a sequence alignment. This may involve searching for evidence for a common cellular function, a common overall structure or, if possible, a similar three-dimensional structure.

Consider two aligned proteins, each of length 100 amino acids. When they share 50% amino acid identity, then on average 80 changes have occurred. I have found that this concept confuses many students. The explanation is that the observed number of changes (50 differences per 100 aligned residues) does not reflect the multiple substitutions that have occurred. For example, the proteins might be a mouse and human globin. About 90 million years ago a species of furry little creatures separated into two groups, eventually leading to speciation and the emergence of primate and rodent lineages. At one position the protein might have had an alanine in the common ancestor that mutated to a threonine and then to an asparagine in the rodent lineage. Two changes occurred at that particular position over a period of millions of years, although we observe only one. We further explore this concept in Chapter 7 on phylogeny and evolution.

When two proteins share 20% amino acid identity (and are in the “twilight zone”) they have 80 observed differences. However, Dayhoff (1978) estimated that 250 changes (on average) had occurred. The PAM250 matrix was therefore considered useful for detecting distantly related proteins.

## ADVICE FOR STUDENTS

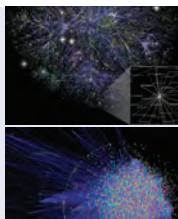
Begin using the BLAST website at NCBI to compare two sequences. Choose two closely related proteins and two that are very distantly related; what are the effects of changing scoring matrices or other parameters? For each topic we discussed, try to gain practical experience. For example, select members of a protein family that are locally aligned because they share a region of homology, and perform global alignment as well. Can you change the local alignment search parameters to include larger or smaller aligned regions? Also try different alignment tools, from various websites to R or Python. In Chapter 4 we introduce BLAST+ for performing any BLAST search on the command line, and you can also use BLAST+ for pairwise alignment.

## WEB RESOURCES

Pairwise sequence alignment can be performed using software packages that implement global or local alignment algorithms. In all cases, two protein or two nucleic acid sequences are directly compared.

Many websites offer web-based pairwise local alignment algorithms based upon global alignment (**Table 3.4**) or local alignment (**Table 3.5**). These sites include EBI and NCBI, the Baylor College of Medicine (BCM) launcher, the SIM program at ExPASy, and SSEARCH at the Protein Information Resource (PIR) at Georgetown University. Computer lab problem (3.4) introduces pairwise alignment in R.

Joshua Lederberg helped Zuckerkandl and Pauling (1965) make the matrix of **Figure 3.28**. They used an IBM 7090 computer, one of the first commercial computers based on transistor technology. The computer cost about US\$3 million. Its memory consisted of 32,768 binary words or about 131,000 bytes. To read about Lederberg's Nobel Prize from 1958, see [http://nobelprize.org/nobel\\_prizes/medicine/laureates/1958/](http://nobelprize.org/nobel_prizes/medicine/laureates/1958/) (WebLink 3.12).



## Discussion Questions

**[3-1]** If you want to compare any two proteins, is there any one “correct” scoring matrix to choose? Is there any way to know which scoring matrix is best to try?

**[3-2]** Many protein (or DNA) sequences have separate domains. (We discuss domains in Chapter 12.) Consider a protein that has one domain that evolves rapidly and a second domain that evolves slowly. In performing a pairwise alignment with another protein (or DNA) sequence, would you use two separate alignments with scoring matrices such as PAM40 and PAM250 or would you select one “intermediate” matrix? Why?

**[3-3]** Years before Margaret Dayhoff and colleagues published a protein atlas with scoring matrices, Emile Zuckerkandl and Linus Pauling (1965) produced a scoring matrix for several dozen available globin sequences (Fig. 3.28). The rows (y axis) of this figure show the original globin amino acid, and the columns show substitutions that were observed to occur. Numerical values are entered in cells for which the substitutions occur in at least 20% of the sites. Note that, for cells shaded red, these amino acid substitutions were never observed; for cells shaded gray the amino acid substitutions were defined as very conservative. How do the data in this matrix compare to those described by Dayhoff and colleagues? Which substitutions occur most rarely, and which most frequently? How would you go about filling in this table today?

**[3-4]** The first five computer lab problems (below) guide you to perform pairwise alignment using five different methods. If you want to align two protein or DNA sequences, how can you decide which tool(s) are most appropriate? In other words, what are some of the strengths and limitations of these various methods?

**[3-5]** The PAM1 matrix (Fig. 3.9) is nonreciprocal: the probability of changing an amino acid such as alanine to arginine is not equal to the probability of changing an arginine to an alanine. Why? Log-odds matrices such as PAM10 (Figure 3.15) are reciprocal.

### PROBLEMS/COMPUTER LAB

For problems (3.1)–(3.3) and (3.5) we perform pairwise alignments of globins using complementary approaches.

**[3-1]** Obtain the human HBA and HBB protein sequences. Perform pairwise alignment at the NCBI BLAST website. Then use a comparison tool from the EBI website. Vary the scoring matrix (e.g., try different

PAM and BLOSUM matrices) and record the effects on the score, the number of gaps, the percent identity, and the length of the aligned region. For the NCBI BLASTP program note that the output of a pairwise alignment includes a dot matrix view.

**[3-2]** Perform pairwise alignment at the UCSC website. (1) Go to <http://genome.ucsc.edu> (WebLink 3.13), follow the link to the genome browser, select the human genome hg19 build, and enter a query of hbb. This should direct you to chr11:5,246,696–5,248,301 (a region of 1606 base pairs encompassing the beta globin gene, *HBB*). (2) Click the box to set the view to default tracks. (3) Under “Comparative Genomics” select Placental Chain/Net and set the display to full. By clicking the Placental Chain/Net header you can view a series of options. Set Chains to full view and Nets to full view. Set the species to horse (deselect other species). Click submit. (4) The display now shows human/horse chained alignments and alignment nets.

**[3-3]** Perform pairwise alignment using EMBOSS tools via Galaxy and UCSC. In this exercise we perform global alignment with the EMBOSS package needle and local alignment with the EMBOSS package water. Both of these are available at the Galaxy public web server (along with over 100 other EMBOSS tools). Box 3.9 introduces EMBOSS and explains how to import beta globin (HBB) and alpha globin (HBA2) proteins from the UCSC Table Brower using Galaxy, and to then align them. This history is saved at <https://main.g2.bx.psu.edu/u/pevsner/h/pairwise-alignment-via-ucsc-and-emboss> (WebLink 3.14). Note that Galaxy is a web-based platform for using hundreds of bioinformatics tools, including next-generation sequence data analysis software. To use it visit <http://usegalaxy.org> then go to the public server. Be sure to create a username and log in. This will allow you to continue your work over time and at different work stations.

**[3-4]** View scoring matrices and perform pairwise alignment using R. In this exercise we begin by installing the `Biostrings` package. Instructions for installing R and RStudio are given in Chapter 2.

```
> getwd() # Get (show) the working directory
# Use setwd() to change it to any location
> source("http://bioconductor.org/biocLite.R")
> biocLite("Biostrings")
> library(Biostrings)
# Install the Biostrings library
> data(BLOSUM50)
# load the data for the BLOSUM50 matrix
> BLOSUM50[1:4,1:4]
# view the first four rows and
# columns of this matrix
> nw <- pairwiseAlignment(AAString("PAWHEAE"),
AAString("HEAGAWGHEE"), substitutionMatrix =
```

```
BLOSUM50, gapOpening = 0, gapExtension = -8)
# create object
# nw aligning two amino acid strings with the
# specified matrix and gap penalties
> nw # view the result.
# Try repeating this alignment with
# different gap penalties and scoring matrices.
# Biostrings includes 10 matrices (PAM30 PAM40,
# PAM70, PAM120, PAM250, BLOSUM45, BLOSUM50,
# BLOSUM62, BLOSUM80, and BLOSUM100).
> compareStrings(nwdemo) # view the alignment
```

**[3-5]** Perform pairwise alignment using Python, a freely available programming language. When implemented with Biopython it offers a broad range of computational tools (Cock *et al.*, 2009). You will need to install three programs: (1) Python; (2) Numpy (a package for scientific computing with Python); and (3) Biopython (this provides particular bioinformatics applications within the Python framework). The downloads can be obtained from <http://www.python.org> (WebLink 3.15), <http://www.numpy.org/> (WebLink 3.16), and <http://biopython.org> (WebLink 3.17). If you are working on a PC launch a user-friendly interface called IDLE (Python’s Integrated DeveLopment Environment). From a Mac, open a terminal window and type python to see the command prompt (>>>). For information on installing Biopython, and for a “cookbook” with many basic bioinformatics applications including pairwise alignment, visit <http://biopython.org/DIST/docs/tutorial/Tutorial.html> (WebLink 3.18). Try the following commands; my comments follow a hash (#) and are in green text.

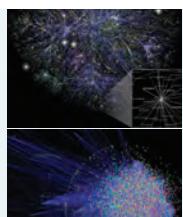
```
$ python # launch python from a terminal
Python 2.7.5 (default, Mar  9 2014, 22:15:05)
Type "copyright", "credits" or "license()" for
more information.
>>> from Bio import pairwise2
>>> from Bio.SubsMat import MatrixInfo as
matlist
>>> matrix = matlist.blosum62
# specify the scoring matrix
>>> help(matlist)
```

```
# This shows a list of available matrices
>>> gap_open = -10 # set the affine gap penal-
ties
>>> gap_extend = -1
>>> hbb = "VTALWGKVNVDEVGGGEALGRLL"
# This is part of beta globin from Fig. 3.5b
>>> mb = "VLNVWGKVEADIPGHGQEVLIRLF"
# This is part of myoglobin from Fig. 3.5b
>>> alns = pairwise2.align.globalds(hbb, mb, ma-
trix, gap_open, gap_extend)
>>> top_aln = alns[0]
>>> aln_hbb, aln_mb, score, begin, end = top_aln
>>> print aln_hbb+'\n'+aln_mb
# the '\n' command inserts a line break
VTALWGKVNVDEVGG--EALGRLL
VLNVWGKVEADIPGHGQEVLIRLF
```

We have used the pairwise2 module from Python. It is capable of both global and local pairwise alignments. Compare the result to **Fig. 3.5b** and note that the gap placement differs. Try raising the gap extend penalty from -0.5 to -2. What happens to the alignment? Documentation is available for the pairwise2 Python module (<http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html>).

**[3-6]** Using the amino acid explorer tool from NCBI. (1) Visit [http://www.ncbi.nlm.nih.gov/Class/Structure/aa\\_aa\\_explorer.cgi](http://www.ncbi.nlm.nih.gov/Class/Structure/aa_aa_explorer.cgi) (WebLink 3.19). (2) Select the Biochemical Properties table. Which amino acid is most abundant? (Is it leucine, at 9.94%?). Use this table to test yourself and make sure you know the one- and three-letter abbreviations for all 20 amino acids, as well as their structures. (3) Is tyrosine a hydrophobic amino acid? To decide, use the Common Substitutions table. Explore valine (a hydrophobic residue), sort the results by hydrophobicity, and see where tyrosine is located. You can also explore the Structure and Chemistry table.

**[3-7]** Many tools are available to manipulate sequences. Visit the Sequence Manipulation Suite (<http://www.bio-informatics.org/sms2/index.html>) (Weblink 3.20) to access a large number of tools. (Compare its tools to those in EMBOSS) What is the reverse complement of the sequence GGAATTCC?



## Self-Test Quiz

**[3-1]** Match the following amino acids with their single-letter codes:

Asparagine	Q
Glutamine	W
Tryptophan	Y
Tyrosine	N
Phenylalanine	F

**[3-2]** Orthologs are defined as:

- Homologous sequences in different species that share an ancestral gene.
- Homologous sequences that share little amino acid identity but share great structural similarity.
- Homologous sequences in the same species that arose through gene duplication.

- (d) Homologous sequences in the same species which have similar and often redundant functions.

**[3-3]** Which of the following amino acids is least mutable according to the PAM scoring matrix?

- (a) alanine;
- (b) glutamine;
- (c) methionine; or
- (d) cysteine.

**[3-4]** The PAM250 matrix is defined as having an evolutionary divergence in which what percentage of amino acids between two homologous sequences have changed over time?

- (a) 1%;
- (b) 20%;
- (c) 80%; or
- (d) 250%.

**[3-5]** Which of the following sentences best describes the difference between a global alignment and a local alignment between two sequences?

- (a) Global alignment is usually used for DNA sequences, while local alignment is usually used for protein sequences.
- (b) Global alignment has gaps, while local alignment does not have gaps.
- (c) Global alignment finds the global maximum, while local alignment finds the local maximum.
- (d) Global alignment aligns the whole sequence, while local alignment finds the best subsequence that aligns.

**[3-6]** You have two distantly related proteins. Which BLOSUM or PAM matrix is best suited to compare them?

- (a) BLOSUM45 or PAM250;
- (b) BLOSUM45 or PAM1;
- (c) BLOSUM80 or PAM250; or
- (d) BLOSUM80 or PAM1.

**[3-7]** How does the BLOSUM scoring matrix differ most notably from the PAM scoring matrix?

- (a) It is best used for aligning very closely related proteins.

- (b) It is based on global multiple alignments from closely related proteins.

- (c) It is based on local multiple alignments from distantly related proteins.

- (d) It combines local and global alignment information.

**[3-8]** True or false: Two proteins that share 30% amino acid identity are 30% homologous.

**[3-9]** A global alignment algorithm (such as the Needleman–Wunsch algorithm) is guaranteed to find an optimal alignment. Such an algorithm:

- (a) Puts the two proteins being compared into a matrix and finds the optimal score by exhaustively searching every possible combination of alignments.
- (b) Puts the two proteins being compared into a matrix and finds the optimal score by iterative recursions.
- (c) Puts the two proteins being compared into a matrix and finds the optimal alignment by finding optimal subpaths that define the best alignment(s).
- (d) Can be used for proteins but not for DNA sequences.

**[3-10]** In a database search or in a pairwise alignment, sensitivity is defined as:

- (a) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false positives (i.e., unrelated sequences having high similarity scores).
- (b) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false positives (i.e., homologous sequences that are not reported).
- (c) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false negatives (i.e., unrelated sequences having high similarity scores).
- (d) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false negatives (i.e., homologous sequences that are not reported).

## SUGGESTED READING

We introduced this chapter with the concept of homology, an often misused term. A one-page article by Reeck *et al.* (1987) provides authoritative, standard definitions of the terms homology and similarity. Other discussions of homology in relation to phylogeny are provided by Tautz (1998) and Pearson (2013). The William Pearson article provides an excellent introduction to sequence alignment (including *E* values, which we describe

in Chapter 4). His earlier article (Pearson, 1996) provides descriptions of the statistics of similarity scores, sensitivity and selectivity, and search programs such as Smith–Waterman and FASTA.

For studies of pairwise sequence alignment algorithms, an important historical starting point is the 1978 book by Margaret O. Dayhoff and colleagues (Dayhoff, 1978). Most of this book consists of an atlas of protein sequences with accompanying phylogenetic reconstructions. Chapter 22 of the *Atlas of Protein Sequence and Structure* introduces the concept of accepted point mutations, while chapter 23 describes various PAM matrices. Russell F. Doolittle (1981) also wrote a clear, thoughtful overview of sequence alignment. By the early 1990s, when far more protein sequence data were available, Steven and Jorja Henikoff (1992) described the BLOSUM matrices. This article provides an excellent technical introduction to the use of scoring matrices, usefully contrasting the performance of PAM and BLOSUM matrices. Later (in Chapters 4 and 5) we will use these matrices extensively in database searching.

The algorithms originally describing global alignment are presented technically by Needleman and Wunsch (1970) and later local alignment algorithms were introduced by Smith and Waterman (1981) and Smith *et al.* (1981). The problem of both sensitivity (the ability to identify distantly related sequences) and selectivity (the avoidance of unrelated sequences) of pairwise alignments was addressed by Pearson and Lipman in a 1988 paper introducing the FASTA program.

Marco Pagni and C. Victor Jongeneel (2001) of the Swiss Institute of Bioinformatics provide an excellent overview of sequence-scoring statistics. This includes a discussion of BLAST scoring statistics that is relevant to Chapters 4 and 5.

Finally, Steven Brenner, Cyrus Chothia, and Tim Hubbard (1998) have compared several pairwise sequence methods. This article is highly recommended as a way to learn how different algorithms can be assessed (we will see similar approaches for multiple sequence alignment in Chapter 6, for example). Reading this paper can help to show why statistical scores are more effective than other search parameters such as raw scores or percent identity in interpreting pairwise alignment results. For a more recent overview of sequence alignment, see Stormo (2009).

## REFERENCES

- Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* **219**(3), 555–565. PMID: 2051488.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.
- Altschul, S.F., Wootton, J.C., Gertz, E.M. *et al.* 2005. Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal* **272**(20), 5101–5109. PMID: 16218944.
- Anfinsen, C. 1959. *The Molecular Basis of Evolution*. John Wiley & Sons, Inc., New York.
- Brenner, S. E., Chothia, C., Hubbard, T. J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of National Academy of Sciences, USA* **95**, 6073–6078.
- Chothia, C., Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**, 823–826.
- Cock, P.J., Antao, T., Chang, J.T. *et al.* 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11), 1422–1423. PMID: 19304878.
- Cumming, G., Fidler, F., Vaux, D.L. 2007. Error bars in experimental biology. *Journal of Cell Biology* **177**, 7–11.
- Darwin, C. 1872. *The Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

- Dayhoff, M.O. (ed.) 1966. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
- Dayhoff, M. O. (ed.) 1978. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
- Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J., Jones, D.D. 1972. Gene duplications in evolution: the globins. In: *Atlas of Protein Sequence and Structure*, volume 5 (ed. Dayhoff, M.O.). National Biomedical Research Foundation, Washington, DC.
- Doolittle, R. F. 1981. Similar amino acid sequences: Chance or common ancestry? *Science* **214**, 149–159.
- Doolittle, R. F. 1987. *OF URFS AND ORFS: A Primer on How to Analyze Derived Amino Acid Sequences*. University of Science Books, Mill Valley, CA.
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G. 2000. *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
- Ewins, W.J., Grant, G.R. 2001. *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag, New York.
- Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology* **19**(2), 99–113. PMID: 5449325.
- Gonnet, G. H., Cohen, M. A., Benner, S. A. 1992. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.
- Gotoh, O. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology* **162**, 705–708.
- Hedges, S.B., Dudley, J., Kumar, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972.
- Henikoff, S., Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of National Academy of Sciences, USA* **89**, 10915–10919.
- Henikoff, J. G., Henikoff, S. 1996. Blocks database and its applications. *Methods in Enzymology* **266**, 88–105.
- Hossfeld, U., Olsson, L. 2005. The history of the homology concept and the “Phylogenetisches Symposium”. *Theory in Biosciences* **124**(2), 243–253. PMID: 1704635.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Jones, D.T., Taylor, W.R., Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275–282.
- Junier, T., Pagni, M. 2000. Dotlet: diagonal plots in a web browser. *Bioinformatics* **16**(2), 178–179. PMID: 10842741.
- Karlin, S., Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences, USA* **87**, 2264–2268.
- Lieb, B., Dimitrova, K., Kang, H.S. et al. 2006. Red blood with blue-blood ancestry: intriguing structure of a snail hemoglobin. *Proceedings of the National Academy of Sciences, USA* **103**(32), 12011–12016. PMID: 16877545.
- Motulsky, H. 1995. *Intuitive Biostatistics*. Oxford University Press, New York.
- Myers, E. W., Miller, W. 1988. Optimal alignments in linear space. *Computer Applications in the Biosciences* **4**, 11–17.
- Needleman, S. B., Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.
- Owen, R. 1843. *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals, Delivered at the Royal College of Surgeons in 1843*. Longman Brown Green and Longmans, London.
- Pagni, M., Jongeneel, C. V. 2001. Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics* **2**, 51–67.

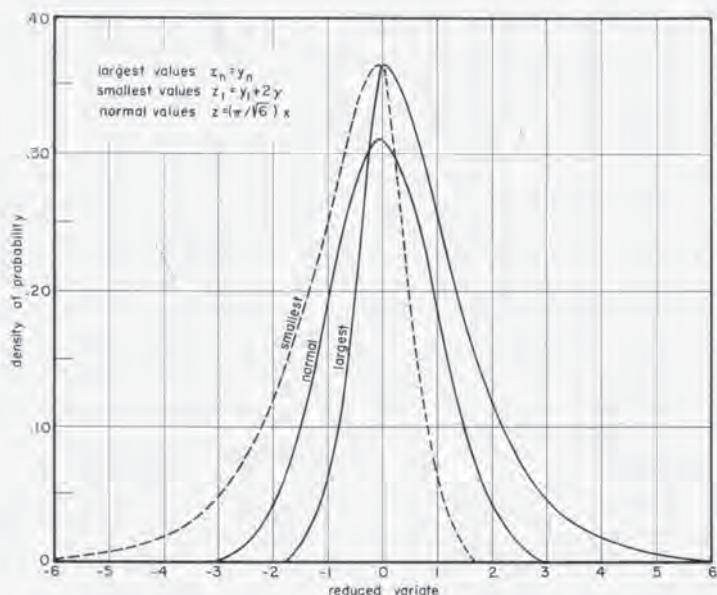
- Pearson, W. R. 1996. Effective protein sequence comparison. *Methods in Enzymology* **266**, 227–258.
- Pearson, W.R. 2013. An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics Chapter 3*, Unit 3.1. PMID:23749753.
- Pearson, W. R., Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, USA* **85**, 2444–2448.
- Pearson, W. R., Wood, T. C. 2001. Statistical significance in biological sequence comparison. In *Handbook of Statistical Genetics* (eds D. J.Balding, M.Bishop, C.Cannings). Wiley, London, pp. 39–65.
- Pevsner, J., Trifiletti, R.R., Strittmatter, S.M., Snyder, S.H. 1985. Isolation and characterization of an olfactory receptor protein for odorant pyrazines. *Proceedings of the National Academy of Sciences, USA* **82**, 3050–3054.
- Reeck, G. R., de Haen, C., Teller, D.C. et al. 1987. “Homology” in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* **50**, 667.
- Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: The European molecular biology open software suite. *Trends in Genetics* **16**, 276–277.
- Sedgewick, R. 1988. *Algorithms*. Addison-Wesley Longman, Reading, MA.
- Schopf, J.W. 2002. When did life begin? In: *Life’s Origin: The Beginnings of Biological Evolution* (ed. Schopf, J.W.). University of California Press, Berkeley.
- Sellers, P. H. 1974. On the theory and computation of evolutionary distances. *SIAM Journal of Applied Mathematics* **26**, 787–793.
- Smith, T. F., Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.
- Smith, T. F., Waterman, M. S., Fitch, W. M. 1981. Comparative biosequence metrics. *Journal of Molecular Evolution* **18**, 38–46.
- Stormo, G.D. 2009. An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics Chapter 3*, Unit 3.1 3.1.1-7. PMID: 19728288.
- Tatusova, T. A., Madden, T. L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters* **174**, 247–250.
- Tautz, D. 1998. Evolutionary biology. Debatable homologies. *Nature* **395**, 17, 19.
- Zuckerkandl, E., Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In: *Evolving Genes and Proteins* (eds Bryson, V., Vogel, H.J.). Academic Press, New York, pp. 97–166.



of the variate, the first double-exponential distribution has larger (smaller) densities than the normal one. The opposite is true for the second double exponential distribution.

Table 5.2.7. Selected Probabilities for Normal and Largest Values

Value	Reduced Variate		Probabilities		Return Periods	
	Largest	Normal	Largest	Normal	Largest	Normal
$\bar{x} - \sigma$	-7.0533	-1	.13206	.15866	7.57	6.30
$\bar{x} + \sigma$	1.85977	1	.85581	.84134	6.93	6.30
$\bar{x} \pm \sigma$	—	—	.72375	.68268	—	—
$\bar{x} - 2\sigma$	-1.98788	-2	.00068	.02275	1480.	43.96
$\bar{x} + 2\sigma$	3.14232	2	.95773	.97725	23.7	43.96
$\bar{x} \pm 2\sigma$	—	—	.95705	.95450	—	—
$\bar{x} - 3\sigma$	-3.27043	-3	$3.7 \cdot 10^{-12}$	.00135	$.27 \cdot 10^{11}$	741
$\bar{x} + 3\sigma$	4.42486	3	.98810	.99865	84.01	741
$\bar{x} \pm 3\sigma$	—	—	.98810	.99730	—	—



Graph 5.2.7(1). Extreme and Normal Distributions

In Graph 5.2.7(2) the probabilities of the largest and the smallest values and the normal probabilities for the same mean and standard deviation

BLAST search results allows you to assess whether a query sequence is significantly related to a match in the database.

Source: Gumbel (1958).

Chapter 4 describes the principal database search tool, BLAST. While BLAST was first described by Altschul *et al.* in 1990, the statistical interpretation of the scores obtained from a BLAST search are based on mathematical models developed in the 1950s. In many instances, the distribution of values in a population assumes a normal (Gaussian) distribution, as shown in this figure (see curve labeled "normal"). However, for a wide variety of natural phenomena the distribution of extreme values is not normal. Such is the case for database searches in which you search with a protein or DNA sequence of interest (the query) against a large database, as described in this chapter. The maximum scores fit an extreme value distribution (EVD) rather than a normal distribution.

In 1958 Emil Gumbel described the statistical basis of the EVD in his book *Statistics of Extremes*. This figure (Gumbel, 1958, p. 180) shows the EVD. Note that for the curve marked "largest" the tail is skewed to the right. Also, as shown in the table, for a normal distribution values that are up to three standard deviations above the mean occupy 99.865% of the area under the curve; for the EVD, values up to three standard deviations occupy only 98.810%. In other words, the EVD is characterized by a larger area under the curve at the extreme right portion of the plot. We see how this analysis applied to

# Basic Local Alignment Search Tool (BLAST)

# CHAPTER 4

*The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T.*

—Stephen Altschul et al. (1990)

*BLAST was the first program to assign rigorous statistics to useful scores of local sequence alignments. Before then people had derived many different scoring systems, and it wasn't clear why any should have a particular advantage. I had made a conjecture that every scoring system that people proposed using was implicitly a log-odds scoring system with particular 'target frequencies', and that the best scoring system would be one where the target frequencies were those you observed in accurate alignments of real proteins. It was the mathematician Sam Karlin who proved this conjecture and derived the formula for calculating the statistics of the scores [E-values] output by BLAST. This was the gravy to the algorithmic innovations of David Lipman, Gene Myers, Webb Miller and Warren Gish that yielded BLAST's unprecedented combination of sensitivity and speed.*

—Stephen Altschul, quoted in Altschul et al. (2013)

## LEARNING OBJECTIVES

Upon completing this chapter you should be able to:

- perform BLAST searches at the NCBI website;
- understand how to vary optional BLAST search parameters;
- explain the three phases of a BLAST search (compile, scan/extend, trace-back);
- define the mathematical relationship between expect values and scores; and
- outline strategies for BLAST searching.

## INTRODUCTION

Basic Local Alignment Search Tool (BLAST<sup>®</sup>) is the main NCBI tool for comparing a protein or DNA sequence to other sequences in various databases (Altschul *et al.*, 1990, 1997). BLAST searching is one of the fundamental ways of learning about a protein or gene: the search reveals what related sequences are present in the same organism and other organisms. The NCBI website includes several excellent resources for learning about BLAST.

In Chapter 3, we described how to perform a pairwise sequence alignment between two protein or nucleotide sequences. BLAST searching allows the user to select one sequence (termed the *query*) and perform pairwise sequence alignments between the

NCBI resources include a tutorial and a course that can be accessed through the main BLAST page (<http://blast.ncbi.nlm.nih.gov/>, WebLink 4.1 at <http://bioinfbook.org>).

query and an entire database (termed the *target*). Typically, this means that tens of millions of sequences are evaluated in a BLAST search (involving about 50 billion nucleotides in the case of a search against a default DNA database), and only the most closely related matches are returned. The Needleman–Wunsch (1970) global alignment algorithm is not used for database searches because we are usually more interested in identifying locally matching regions such as protein domains. The Smith–Waterman (1981) local alignment algorithm finds optimal pairwise alignments, but we cannot use it for database searches generally because it is too computationally intensive. BLAST offers a local alignment strategy having both speed and sensitivity, as described in this chapter. It also offers convenient accessibility on the World Wide Web or as a command-line tool.

BLAST is a family of programs that allows you to input a query sequence and compare it to DNA or protein sequences in a database. A DNA sequence can be converted into six potential proteins (see “Step 2: Selecting a BLAST Program”), and the BLAST algorithms include strategies to compare protein sequences to dynamically translated DNA databases or vice versa. The programs produce high-scoring segment pairs (HSPs) that represent local alignments between your query and database sequences. BLAST searching has a wide variety of uses:

- *Determining what orthologs and paralogs are known for a particular protein or nucleic acid sequence.* Besides alpha and beta globin and myoglobin, what other globins are known? When a new bacterial genome is sequenced and several thousand proteins are identified, how many of these proteins are paralogous? How many of the predicted genes have no significantly related matches in GenBank?
- *Determining what proteins or genes are present in a particular organism.* Are there any globins in plants? Are there any reverse transcriptase genes (such as HIV-1 Pol gene) in fish? In some cases searching for remote homologs requires the use of specialized BLAST-like approaches; we describe some of these in Chapter 5.
- *Determining the identity of a DNA or protein sequence.* For example, you may perform an RNAseq experiment (Chapter 11) and learn that a particular RNA sequence is dramatically regulated under the experimental conditions that you are using. This sequence may be searched against a protein database to learn what proteins are most related to the protein encoded by your nucleotide sequence.
- *Discovering new genes.* For example, a BLAST search of genomic DNA may reveal that the DNA encodes a protein that has not been described before. In this chapter, we show how BLAST searching can be used to find novel, previously uncharacterized genes.
- *Determining what variants have been described for a particular gene or protein.* For example, many viruses are extremely mutable; what HIV-1 Pol variants are known?
- *Investigating expressed sequence tags (ESTs) that may exhibit alternative splicing.* There is an EST database that can be explored by BLAST searching. Indeed, there are dozens of specialized databases that can be searched. For example, specialized databases consist of sequences from a specific organism, a tissue type, a chromosome, a type of DNA (such as untranslated regions), or a functional class of nucleic acids or proteins.
- *Exploring amino acid residues that are important in the function and/or structure of a protein.* The results of a BLAST search can be multiply aligned (Chapter 6) to reveal conserved residues such as cysteines that are likely to have important biological roles.

Visit the BLAST site at <http://blast.ncbi.nlm.nih.gov/> (WebLink 4.1), or go the main page of NCBI (<http://www.ncbi.nlm.nih.gov/>, WebLink 4.2) then select BLAST.

There are four components to performing any web-based BLAST search:

1. Selecting a sequence of interest and pasting, typing, or uploading it into the BLAST input box.
2. Selecting a BLAST program (most commonly BLASTP, BLASTN, BLASTX, TBLASTX, or TBLASTN).

3. Selecting a database to search. A common choice is the nonredundant (nr) database, but there are many other databases.
4. Selecting optional parameters, both for the search and for the format of the output. These options include choosing a substitution matrix, filtering of low-complexity sequences, and restricting the search to a particular set of organisms.

As we describe the steps of BLAST searching, we begin with a specific example. Select the link “Standard protein-protein BLAST [blastp].” You will see a box to enter the query sequence; enter the sequence of human beta globin (NP\_000509.1) then click the “BLAST” button (Fig. 4.1). The result lists the proteins that are most closely related to beta globin. We now describe the practical aspects of BLAST searching in detail.

As of February 2015, you can search a database of ~25 million protein sequences (and over 8 billion amino acid residues) within several seconds. For a DNA search, the default nonredundant database currently has ~30 million sequences and ~88 billion letters. Note that if you search with a query such as NP\_000509 without specifying a version number then, by default, the most recent version will be used.

The screenshot shows the NCBI Standard Protein BLAST search interface. The query sequence is entered as a FASTA-formatted string: >gi|4504349|ref|NP\_000509.1| hemoglobin subunit beta [Homo sapiens] MVRHTPEEKSAVITALWKGK[N]DEVGGEALGRLLVYYPWTQPFESFGDLSTPDAVMGNPKVKAH GKKVLIGAFSDGLAHLDNLKGTFAILSELHCDKLHVDPENFALLQNVLVCVLAHREGEKFIPVQ AAIQRVVAAGAAHALAHKYH. The database is set to "Reference proteins (refseq\_protein)". The search is restricted to the author Max Perutz (perutz mf[Author]). The algorithm selected is "blastp (protein-protein BLAST)". The search is set to search the database "Reference proteins (refseq\_protein)" using "Blastp (protein-protein BLAST)".

**FIGURE 4.1** Main page for a BLASTP search at NCBI. The sequence can be input as an accession number, GI identifier, or FASTA-formatted sequence as shown here (arrow 1). The database must be selected (arrow 2) if the default setting is not selected (as here, in which the database is set to RefSeq proteins); the choice is highlighted in yellow. The search can be restricted to a particular organism or taxonomic group, and Entrez queries can be used to further focus the search (arrow 3); here we limit the search to entries including the author Max Perutz. We discuss the BLASTP algorithm in this chapter (arrow 4), and PSI-BLAST, PHI-BLAST, and DELTA-BLAST in Chapter 5. Many of the search parameters can be modified (arrow 5).

Source: BLASTP, NCBI.

## BLAST SEARCH STEPS

The FASTA format is further described at <http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml> (WebLink 4.3). Do not confuse the FASTA format with the BLAST program, which we described briefly in Chapter 3. For BLAST searches, your query can be in uppercase or lowercase, with or without intervening spaces or numbers. If the query is DNA, BLAST algorithms will search both strands.

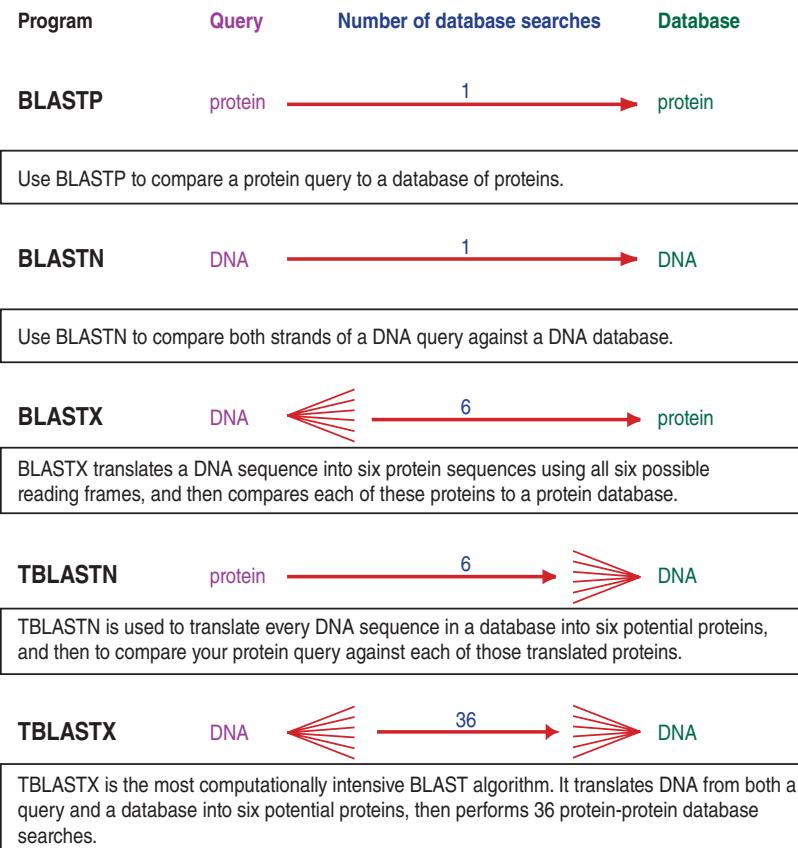
### Step 1: Specifying Sequence of Interest

A BLAST search begins with the selection of a DNA or protein sequence. There are two main forms of data input: (1) cutting and pasting DNA or protein sequence (e.g., in the FASTA format); and (2) using an accession number (e.g., a RefSeq or GenBank Identification (GI) number). A sequence in FASTA format begins with a single-line description followed by lines of sequence data. The description line is distinguished from the sequence data by a greater than (“>”) symbol in the first line. It is recommended that all lines of text be shorter than 80 characters in length. An example of a sequence in FASTA format was shown in **Figure 2.9**.

It is often convenient to input the accession number to a BLAST search. Note that the BLAST programs can recognize and ignore numbers that appear in the midst of the letters of your input sequence. The BLAST search also allows you to select a subset of an entire query sequence, such as a region or domain of interest.

### Step 2: Selecting BLAST Program

The NCBI BLAST family of programs includes five main programs, as summarized in **Figure 4.2**.



**FIGURE 4.2** Overview of the five main BLAST algorithms. Note that the suffix P refers to protein (as in BLASTP), N refers to nucleotide, and X refers to a DNA query that is dynamically translated into six protein sequences. The prefix T refers to “translating,” in which a DNA database is dynamically translated into six proteins.

- The BLASTP program compares an amino acid query sequence against a protein sequence database. Note that for this type of search there are optional parameters (see below) that are specifically relevant to protein searches, such as the choice of various PAM and BLOSUM scoring matrices.
- The BLASTN program is used to compare a nucleotide query sequence against a nucleotide sequence database.

Three additional BLAST algorithms rely on the fundamental relationship of DNA to protein. Any DNA sequence can be transcribed and translated into six potential reading frames (three on the top strand and three on the bottom strand; Fig. 4.3). For BLAST searching, the query DNA sequence may be translated into potential proteins, an entire DNA database may be translated, or both. In all three cases, these algorithms perform protein–protein alignments.

- The program BLASTX compares a nucleotide query sequence translated in all reading frames against a protein sequence database. If you have a DNA sequence and you want to know what protein (if any) it encodes, you can perform a BLASTX search. This automatically translates the DNA into six potential proteins (see Figs. 4.2 and 4.3). The BLASTX program then compares each of the six translated protein sequences to all the members of a protein database.
- The program TBLASTN compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. One might use this program to ask whether a DNA database encodes a protein that matches your protein query of interest. Does a query with beta globin yield any matches in a database of genomic DNA from the genome-sequencing project of a particular organism?

UniGene uses BLASTX to compare each nucleotide sequence in its database to all known proteins from organisms with sequenced genomes. The *E* value cutoff (discussed in “BLAST Algorithm: Local Alignment Search Statistics and *E* Value” below) is  $10^{-6}$ . See <http://www.ncbi.nlm.nih.gov/UniGene/help.cgi?item=protest> (WebLink 4.4).

### Homo sapiens hemoglobin, beta (HBB), mRNA

NCBI Reference Sequence: NM\_000518.4

[GenBank](#) [FASTA](#)

[Link To This Page](#) | [Feedback](#)



**FIGURE 4.3** DNA can potentially encode six different proteins. To demonstrate this, we view the NCBI Nucleotide entry for HBB and select the “graphics” view; The two strands of DNA sequence are shown (arrow 1). In this zoomed view, only a portion of the HBB sequence is displayed. From the top strand, three potential proteins are encoded (frames +1, +2, +3) with the corresponding amino acids indicated in gray using the single-letter amino acid abbreviations. In this case, frame +3 corresponds to the frame used for translation (arrow 2). Note that frames +1 and +2 as well as frame -3 include stop codons (asterisks shaded red). The lower portion of the display includes the amino acid sequence of the corresponding protein (arrow 3) as well as the corresponding nucleotides (matching frame +3); features indicated with black shading represent a site that may be acetylated or glycosylated and a globin domain.

Source: NCBI Nucleotide.

We discuss expressed sequence tags (ESTs) in Chapter 10. TBLASTX can help you identify frameshifts in ESTs, since all reading frames are compared.

- The program TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. The TBLASTX program is computationally intensive. Consider a situation in which you have a DNA sequence with no obvious database matches and you want to know if it encodes a protein with distant, statistically significant database matches in a database of expressed sequence tags. A BLASTX search would be more sensitive than BLASTN, and therefore useful to reveal genes that encode proteins homologous to your query.

### Step 3: Selecting a Database

The databases that are available for BLAST searching are listed on each BLAST page. For protein database searches (BLASTP and BLASTX), the default option is the nonredundant (nr) database. This consists of the combined protein records from GenBank, the Protein Data Bank (PDB), SwissProt, PIR, and PRF (see Chapter 2 for descriptions of these resources). Another option is to search only RefSeq proteins. **Table 4.1** summarizes the available protein databases for BLAST searching at NCBI, including the approximate number of sequences available in each database.

For DNA database searches (BLASTN, TBLASTN, TBLASTX) the default option is to search the nucleotide nr/nt database. This includes nucleotide sequences from GenBank, EMBL, DDBJ, PDB, and RefSeq. However, the nr database does not have records from the EST, sequence tagged site (STS), whole-genome sequence (WGS), genome survey sequence (GSS), transcriptome shotgun assembly (TSA), patents, or high-throughput genomic sequence (HTGS) databases. Other commonly used options include the human (or mouse) genomic plus transcript database or the EST database.

The nr databases are derived by merging several main protein or DNA databases. These databases often contain identical sequences. Generally only one of these sequences is retained by the nr database, along with multiple accession numbers. (Even if two sequences in the nr database appear to be identical, they should at least have some subtle difference.) The nr databases are often the preferred sites for searching the majority of available sequences.

A summary of all the nucleotide sequence databases that can be searched by standard BLAST searching at NCBI is provided in **Table 4.2**.

**TABLE 4.1 Protein sequence databases that can be searched by BLAST searching at NCBI. PDB, Protein Data Bank. # indicates approximate number of sequences in database. Adapted from BLAST, NCBI, <http://blast.ncbi.nlm.nih.gov/>.**

Database	Title	# sequences
nr	All nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects	65 million
Reference proteins	NCBI protein reference sequences	50 million
UniProtKB/SwissProt	Nonredundant UniProtKB/SwissProt sequences	450,000
Patented protein sequences	Protein sequences derived from the Patent division of GenBank	1.3 million
Protein Data Bank	PDB protein database	77,000
Metagenomic proteins	Proteins from WGS metagenomic projects (env_nr)	6.5 million
Transcriptome	Transcriptome Shotgun Assembly (TSA) sequences	770,000

**TABLE 4.2 Nucleotide sequence databases that can be searched using BLAST at NCBI. # indicates approximate number of sequences in database. Adapted from BLAST, NCBI, <http://blast.ncbi.nlm.nih.gov/>.**

Database	Title	# sequences
Human Genomic + Transcript	Homo sapiens NCBI Annotation Release 104 RNAs; Homo sapiens all assemblies	55,000
Mouse Genomic + Transcript	Mus musculus NCBI Annotation RNAs; Mus musculus all assemblies	N/A
nr/nt	All GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent sequences as well as phase 0, 1, and 2 HTGS sequences	25 million
refseq_rna	NCBI transcript reference sequences	3.5 million
refseq_genomic	NCBI genomic reference sequences	2.7 million
NCBI Genomes	NCBI chromosome sequences	28,000
Expressed sequence tags (EST)	Database of GenBank+EMBL+DDBJ sequences from EST Divisions	75 million
Genomic survey sequences (gss)	Genome survey sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences	36 million
High-throughput genomic sequences (HTGS)	Unfinished high-throughput genomic sequences; sequences: phases 0,1 and 2	153,000
Patent sequences	Nucleotide sequences derived from the Patent division of GenBank	21 million
Protein Data Bank	PDB nucleotide database	8000
alu	Human Alu repeat elements	325
Sequence tagged sites (STS)	Database of GenBank+EMBL+DDBJ sequences from STS Divisions	1.3 million
Whole-genome shotgun (wgs)	Whole-genome-shotgun contigs	116 million
Transcriptome Shotgun Assembly (TSA)	Transcriptome shotgun assembly (TSA) sequences	15 million
16S ribosomal RNA sequences (Bacteria and Archaea)	16S ribosomal RNA sequences (bacteria and archaea)	7500

### Step 4a: Selecting Optional Search Parameters

We initially focus our attention on a standard protein–protein BLAST search. In addition to deciding on which sequence to input and which database to search, there are many optional parameters that you can adjust (see Figs. 4.1 and 4.4).

1. *Query.* In addition to a choice of formats (accession number, GI identifier, or FASTA format), you can select a range of amino acid or nucleotide residues to search.
2. *Limit by Entrez Query.* Any NCBI BLAST search can be limited using any terms that are used in an Entrez search. Enter the term “perutz mf[Author]” and perform a BLASTP search using beta globin as a query (Fig. 4.1, arrow 3). Instead of obtaining hundreds of hits, the matches are to entries that refer to Nobel laureate Max Perutz. BLAST searches can also be restricted by organism. Some popular groups are Archaea, Metazoa (multicellular animals), Bacteria, Vertebrata, Eukaryota, Mammalia, Embryophyta (higher plants), Rodentia, Fungi, and Primates. BLAST searches can be restricted to any genus and species or other taxonomic grouping.

If you want to restrict your BLAST search to a particular organism (or group of organisms), use the box labeled “organism” and type at least part of the name to access a dynamic pull-down menu. You can also access a specific taxonomy identifier. To do this, try beginning at the home page of NCBI and selecting Taxonomy Browser from the top bar (or visit <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>, WebLink 4.5). Select from the list of commonly studied organisms or perform a query in the taxonomy page. You can find the appropriate taxonomy identifier (txid) for any organism in this way. Examples include txid10090 for mouse, txid9606 for human, and txid33090 for Viridiplantae (the plant kingdom).

The expect value is sometimes also referred to as the expectation value. We discuss practical examples of interpreting  $E$  values later in this chapter. Note that  $E$  values much higher than 0.05 may represent biologically relevant, homologous matches, as discussed below.

We illustrate some effects of applying optional features of BLASTP by using human insulin (NP\_000198.1) as a query; this is a preprotein (a peptide that is processed, as depicted in Fig. 7.3) of 110 amino acid residues. We restrict the output to RefSeq proteins from *Drosophila melanogaster* (type this into the Organism search box, or enter txid:7227). The BLAST web form listing various options is shown in Figure 4.4.

3. *Max target sequences*. You can select fewer or more than the default value of 100.
4. *Short queries*. If you select this option, the expect value and word size are automatically adjusted.
5. *Expect threshold*. The expect value  $E$  is the number of different alignments with scores equal to or greater than some score  $S$  that are expected to occur in a database search by chance. Look at the best match in Figure 4.5a (a match between human insulin and insulin-like peptide 3 from *Drosophila*). The score is 31.6 bits, and the  $E$  value is 0.05. This indicates that, based on the particular search parameters used (including the size of the database and the choice of the scoring matrix), a score of 31.6 bits or better is expected to occur by chance 1 in 20 times. A reasonable general guideline is that database matches having  $E$  values of  $\leq 0.05$  are statistically significant.

The default setting for the expect value is 10 for BLASTN, BLASTP, BLASTX, and TBLASTN. At this  $E$  value, 10 hits with scores equal to or better than the alignment score  $S$  are expected to occur by chance. (This assumes that you search the database using a random query with similar length to your actual query.) By changing the expect option to a lower number (such as 0.01), fewer database hits are returned; fewer chance matches are reported. Increasing  $E$  returns more hits. Consider a very short protein or nucleotide query (e.g., 10 amino acids). There is no opportunity for that query to accumulate a large score

The screenshot shows the 'Algorithm parameters' section of the BLASTP web interface. The 'General Parameters' group contains the following settings:

- Max target sequences: 100 (arrow 1)
- Short queries:  Automatically adjust parameters for short input sequences (arrow 2)
- Expect threshold: 10 (arrow 3)
- Word size: 3 (arrow 4)
- Max matches in a query range: 0 (arrow 5)

The 'Scoring Parameters' group contains:

- Matrix: BLOSUM62 (arrow 6)
- Gap Costs: Existence: 11 Extension: 1 (arrow 7)
- Compositional adjustments: Conditional compositional score matrix adjustment (arrow 8)

The 'Filters and Masking' group contains:

- Filter:  Low complexity regions (arrow 9)
- Mask:  Mask for lookup table only  
 Mask lower case letters (arrow 10)

At the bottom, there is a 'BLAST' button and a search field: 'Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)'.

**FIGURE 4.4** Optional BLASTP parameters. Numbered arrows refer to discussion in the text.

Source: BLASTP, NCBI.

## (a) Default: conditional compositional score matrix adjustment

Insulin-like peptide 3 [Drosophila melanogaster]  
Sequence ID: ref|NP\_648360.2| Length: 120 Number of Matches: 1

Range 1: 32 to 114 GenPept Graphics				
Score	Expect	Method	Identities	Positives
31.6 bits(70)	0.050	Compositional matrix adjust.	21/88(24%)	40/88(45%)
			12/88(13%)	
Query 29	LCGSHLVEALYLVCGERGFFYTPKTRREADELQVGQVELGGGPGAGSLQPLALEGSLO--	87		
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q			
Sbjct 32	KLCGRKLPETL SKLCV-- YGFNANTKRTLD PVNFNQID-- GFEDRSLLERLLSDSSVQM	86		
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 87	KTRRL RDGVFDEC CLKSCTMDEVL RYC 114			

## (b) No adjustment (by default, filter low complexity regions)

Insulin-like peptide 3 [Drosophila melanogaster]  
Sequence ID: ref|NP\_648360.2| Length: 120 Number of Matches: 1

Range 1: 33 to 114 GenPept Graphics				
Score	Expect	Identities	Positives	Gaps
33.5 bits(75)	0.009	21/87(24%)	40/87(45%)	12/87(13%)
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREADELQVGQVELGGGPGAGSLQPLALEGSLO--	87		
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q			
Sbjct 33	KLCGRKLPETL SKLCV-- YGFNANTKRTLD PVNFNQID-- GFEDRSLLERLLSDSSVQM	87		
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 88	KTRRL RDGVFDEC CLKSCTMDEVL RYC 114			

## (c) Composition-based statistics

Insulin-like peptide 3 [Drosophila melanogaster]  
Sequence ID: ref|NP\_648360.2| Length: 120 Number of Matches: 1

Range 1: 33 to 114 GenPept Graphics				
Score	Expect	Method	Identities	Positives
30.4 bits(67)	1e-04	Composition-based stats.	21/87(24%)	40/87(45%)
			12/87(13%)	
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREADELQVGQVELGGGPGAGSLQPLALEGSLO--	87		
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q			
Sbjct 33	KLCGRKLPETL SKLCV-- YGFNANTKRTLD PVNFNQID-- GFEDRSLLERLLSDSSVQM	87		
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 88	KTRRL RDGVFDEC CLKSCTMDEVL RYC 114			

**FIGURE 4.5** Pairwise alignments from BLASTP searches illustrating the effects of changing compositional matrices and filtering options. Human insulin (NP\_000198.1) was used as a query in a BLASTP search restricted to RefSeq proteins in *Drosophila*. (a) Default settings show a match to a *Drosophila* insulin protein with a score of 31.6 bits and an *E* value of 0.05. Results are shown using (b) no compositional adjustments and (c) composition based statistics. The expect values for these three searches are indicated (red boxes).

and, since the score is inversely related to the expect value (see Equation (4.5) below), the *E* value cannot be very small. Indeed, an *E* value of 50 or 100 might occur for a database match of considerable biological interest. When you select the optional parameter “short queries” in BLASTP, the *E* value is therefore set to 200,000 or *E* = 1000 in BLASTN. We describe the *E* value in more detail in a discussion of BLAST search statistics (see section “BLAST Algorithm: Local Alignment Search Statistics and *E* Value” below), including a comparison of searches with varying *E* values.

6. *Word size.* For protein searches, a window size of 3 (default) or 2 may be set. When a query is used to search a database, the BLAST algorithm first divides the query into a series of smaller sequences (words) of a particular length (word size). For BLASTP, a larger word size yields a more accurate search. For any word size, matches made to each word are then extended to produce the BLAST output. In practice, the word size can remain at 3 and should be reduced to 2 only when your query is a very short

peptide (i.e., a short string of amino acids). Changing the size from 3 to 2 has no effect on the alignment (or the scores) of human insulin with its nematode homolog.

For nucleotide searches, the default word size is 11 and can be raised (word size 15) or reduced (word size 7). Lowering the word size yields a more accurate but slower search. Raising the word size is applied in MegaBLAST and discontiguous MegaBLAST (see Chapter 5), two alternate programs at NCBI that perform nucleotide searches. For MegaBLAST the default word size is 28, and can be set as high as 256. Very long word sizes match relatively infrequently, encouraging a much faster search. This is useful for speed when searching with long queries (e.g., many thousands of nucleotides) for nearly exact matches in a database.

7. *Max matches in a query range.* Matches to one region of interest can be obscured by frequent matches to a different region of a protein. This feature offers a solution in which redundant database hits are discarded (Berman *et al.*, 2000).
8. *Matrix.* Eight amino acid substitution matrices are available for BLASTP protein–protein searches: PAM30, PAM70, and PAM250; BLOSUM45, BLOSUM50, BLOSUM62 (default), BLOSUM80, and BLOSUM90. Some alternative BLAST servers (discussed in Chapter 5 on advanced BLAST searching) offer many more choices for substitution matrices. It is sometimes advisable to try a BLAST search using several different scoring matrices. For example, as described in Chapter 3 PAM40 and PAM250 matrices (Fig. 3.16) have entirely distinct properties as scoring matrices for sequences sharing varying degrees of similarity. For very short queries (e.g., 15 or fewer amino acid residues), a PAM30 matrix is recommended (and is automatically invoked at the NCBI BLASTP site).

For BLASTN, the default scoring system is +2 for a match and –3 for a mismatch. A variety of other scoring schemes are available, including the default +1, –1 for Megablast (Chapter 5). For each scoring system, the BLAST family offers appropriate gap opening and extension penalties.

9. *Gap costs.* A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another (Chapter 3). Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is frequently ascribed more significance than the length of the gap. The gap introduction is therefore penalized heavily, whereas a lesser penalty is ascribed to each subsequent residue in the gap. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

Gap scores are typically calculated as the sum of  $G$ , the gap-opening penalty, and  $L$ , the gap extension penalty. For a gap of length  $n$ , the gap cost would be  $G + Ln$ . The choice of gap costs is typically 10–15 for  $G$  and 1–2 for  $L$ . These are called affine gap penalties, in which the penalty for introducing a gap is far greater than the penalty for extending one.

MegaBLAST (Chapter 5) uses non-affine gap penalties, that is, there is no cost for opening a gap. We further discuss the problem of gaps in multiple sequence alignments in Chapter 6.

10. *Compositional adjustments.* The “conditional compositional score matrix adjustment,” which is selected as default, generally improves the calculation of the  $E$  value statistic (see section “BLAST Algorithm: Local Alignment Search Statistics and  $E$  Value” below). Some proteins (whether queries or database matches) have nonstandard compositions such as having hydrophobic or cysteine-rich regions. For some organisms, the entire genome has a very high guanine plus cytosine (GC) or adenine plus thymine (AT) content. For example, the entire genome of the malaria parasite *Plasmodium falciparum* is 80.6% AT, biasing its proteins towards having amino acids encoded by AT-rich codons. A standard matrix such

as BLOSUM62 is not appropriate for the comparison of two proteins with non-standard composition, and the target frequencies  $q_{ij}$  (see Equation (3.7)) need to be adjusted in the context of new background frequencies  $p_i p_j$  (Yu *et al.*, 2003; Yu and Altschul, 2005). In performing a BLASTP search, a default option is to use composition-based statistics. This implements a slightly different scoring system for each database sequence in which all scores are scaled by an analytically determined constant (Schäffer *et al.*, 2001). It is applicable to any BLAST protein search including the position-specific scoring matrix of PSI-BLAST or DELTA-BLAST (Chapter 5).

Compositional adjustments generally increase the accuracy of BLAST searches considerably (Schäffer *et al.*, 2001; Altschul *et al.*, 2005). The improvement can be quantified using receiver operating characteristic (ROC) curves that plot the number of true positives (based on an independent criterion such as expert manual curation) versus false positives (Gribskov and Robinson, 1996). In addition to using composition-based statistics, a conditional compositional score matrix adjustment can be applied to BLASTP searches. This can reduce false positive search results in specialized circumstances such as subjects matching queries of very different lengths (Altschul *et al.*, 2005). In that case the longer sequence may have a substantially different composition than the shorter.

In the example of our insulin search versus *Drosophila*, removing compositional adjustments lowers the  $E$  value from 0.05 to 0.009 (Fig. 4.5b). Invoking a composition-based statistics option improves the  $E$  value by 500-fold to  $1 \times 10^{-4}$  (Fig. 4.5c). The magnitude of these effects depends on the composition of the particular query you choose and, for some searches, it is helpful to try a series of compositional adjustments. Note that both of the altered settings (Fig. 4.5b, c) had minimal effects on the lengths of the aligned regions or the gaps.

11. *Filters.* Filtering masks portions of the query sequence that have low complexity (or highly biased compositions; Wootton and Federhen, 1996). Low-complexity sequences are defined as having commonly found stretches of amino acids (or nucleotides) with limited information content. Examples are dinucleotide repeats (e.g., the repeating nucleotides CACACACA...), *Alu* sequences, or regions of a protein that are extremely rich in one or two amino acids. Stretches of hydrophobic amino acid residues that form a transmembrane domain are very common, and a database search with such sequences results in many database matches that are statistically significant but biologically irrelevant. Other motifs that are masked by filtering include acidic-, basic-, and proline-rich regions.

The BLASTP and BLASTN programs offer several main options. Note that filtering is applied to the query sequence, and not to the entire database. One approach is to filter low-complexity regions. For protein sequence queries, the SEG program is used; for nucleic acid sequences, the DUST program is employed. Another approach is to filter repeats (for BLASTN only). This is useful to avoid matching a query with *Alu* repeats or other repetitive DNA to spurious database entries.

12. *Masking.* The “mask for lookup table only” option masks the matching of words above threshold to database hits. This avoids matches to low-complexity sequences or repeats. BLAST extensions then occur without masking (so hits can be extended even if they contain low-complexity sequence). The “mask lower case letters” option allows you to enter a query in the FASTA format using upper case characters for the search but filtering those residues you choose to filter by entering them in lower case. These particular options have little effects on our insulin search of *Drosophila* proteins. For some queries (including those having transmembrane spans that can potentially match thousands of database entries) the results can be dramatic.

For examples of proteins that are highly hydrophobic, very cysteine-rich, or adenine and thymine (AT)-rich sequences from *P. falciparum*, see Web Documents 4.1, 4.2, and 4.3 at <http://www.bioinfbook.org/chapter4>. We discuss *P. falciparum* in Chapter 19; it is responsible for up to 1 million deaths a year.

We explore repetitive DNA sequences in Chapter 8. Web Document 4.4 (at <http://www.bioinfbook.org/chapter4>) offers over a dozen spectacular examples of repetitive DNA and protein sequences.

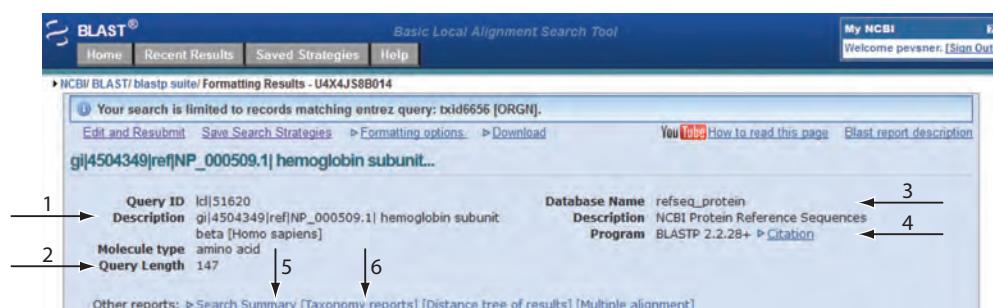
## Step 4b: Selecting Formatting Parameters

There are many options for formatting the output of a BLAST search (Boratyn *et al.*, 2013). These are illustrated by performing a protein–protein BLASTP search with human beta globin (NP\_000509.1) as a query and restricting the search to RefSeq proteins from the mouse (*Mus musculus*). The results of the search occur in several main parts. At the top (Fig. 4.6), details of the search are provided including the type of BLAST search, a description of the query and the database, and a taxonomy link to the results organized by species. By clicking “Search Summary,” details of the search such as the word size, expect value threshold, scoring matrix, and choice of composition-based statistics can be viewed (Fig. 4.7).

The middle portion of a typical BLAST output provides a graphic summary of the results (Fig. 4.8). This includes conserved domains followed by a color-coded summary, with the length of the query sequence represented across the *x* axis. Each bar drawn below the map represents a database protein (or nucleic acid) sequence that matches the query sequence. The position of each bar relative to the linear map of the query allows the user to see the extent to which the database matches align with a single or multiple regions of the query. The most similar hits are shown at the top in red. Hatched areas (when present) correspond to the nonsimilar sequence between two or more distinct regions of similarity found within the same database entry.

The alignments are next described in a table (Fig. 4.9). The description lines are sorted by increasing *E* value; the most significant alignments (lowest *E* values) are therefore at the top. The table includes columns listing the description (name and species), score, *E* value, percent identity, and accession number. If the user checks boxes at the left of each row, those entries are selected for further analyses such as a distance tree or multiple alignment. An example is shown in Figure 4.10. Following a search of human beta globin against arthropod (insect) RefSeq proteins, the top eight hits were checked and sent to a multiple alignment. Here various links are shown (e.g., to Map Viewer, Gene, and UniGene) and the multiple alignment is displayed. A tree can also be produced (not shown).

The lower portion of a BLAST search output consists of a series of pairwise sequence alignments, such as those in Figure 4.5. Here, the pairwise match between the query (input sequence) and the subject (i.e., the particular database match that is aligned to the query) can be inspected. Four scoring measures are provided: the bit score, the expect score, the percent identity, and the positives (percent similarity).



**FIGURE 4.6** Top portion of a BLAST output describes the search that was performed including the query (arrow 1), the query length (arrow 2), the database that was searched (arrow 3), and the program that was employed (BLASTP 2.2.28 in this case; arrow 4). At the bottom, additional links include a search summary showing details of the search statistics (arrow 5) and taxonomy reports of the results (arrow 6).

Source: BLAST, NCBI.

Search Parameters	
Program	blastp
Word size	3
Expect value	10 ← 1
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62 ← 2
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11 ← 3
Composition-based stats	2

Database	
Posted date	Jun 12, 2013 10:46 AM
Number of letters	6,910,040,539 ← 4
Number of sequences	19,996,853
Entrez query	txid10090 [ORGN]

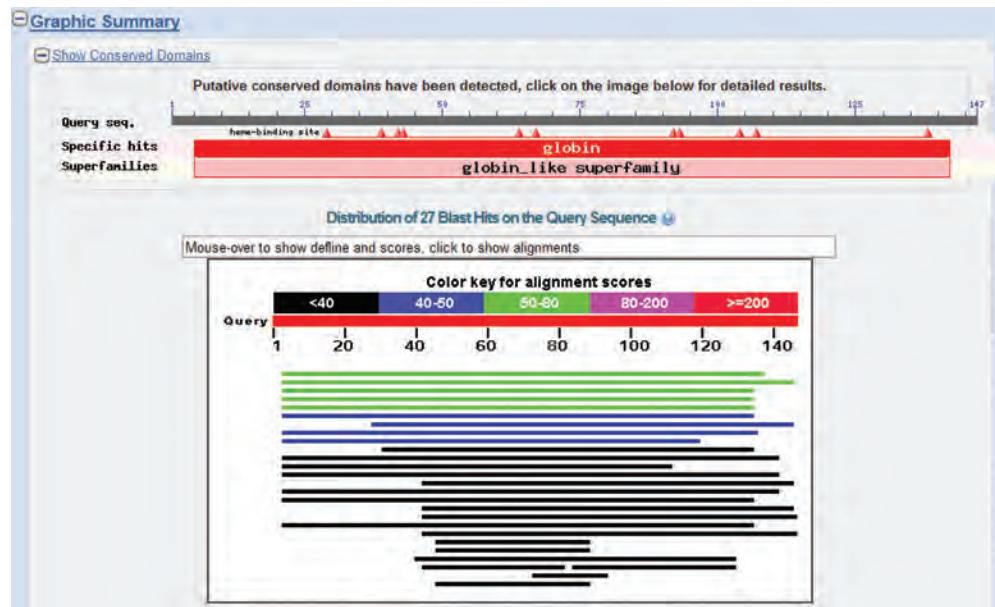
  

Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

**FIGURE 4.7** BLAST search summary. The upper portion shows the search parameters (e.g., the program that was used, the expect value (arrow 1), the scoring matrix (arrow 2), any filters that were applied, the threshold (arrow 3)). The middle portion describes the database; in this example it includes about 6.9 billion amino acid residues (arrow 4), and the output has been restricted to txid10090 (i.e., mouse). The bottom portion shows Karlin–Altschul statistics including lambda, K, and H.

Source: BLAST, NCBI.

Without reperforming an entire BLAST search, the output can be reformatted to provide a range of different output options. The number of descriptions and of alignments can be modified. There are several options for visualizing the aligned sequences (including a multiple sequence alignment). This is an especially useful way to identify specific amino acid residues that are conserved (or divergent) within a protein or DNA family. For nucleotide searches (e.g., BLASTN), by selecting the CDS (coding sequence)



**FIGURE 4.8** The graphic summary of BLAST results includes a display of conserved domains (here showing a match to the globin protein family), then a color-coded distribution of hits. Here the *x* axis corresponds to the length of the query (147 amino acid residues for beta globin), with each database match characterized by a color-coded score (e.g., five matches shaded green have scores of 50–80) and lengths (one of the five green database hits includes an aligned region that extends fully to the carboxy-terminus of the HBB query, while the other four do not). This graphic can be useful to summarize the regions in which database matches align to the query.

Source: BLAST, NCBI.

feature the pairwise alignments also show the positions of the corresponding protein (when that information is available). For example, a search of human beta globin DNA (NM\_000518.4) against human RefSeq nucleotide sequences includes a match to epsilon 1 globin (NM\_005330.3). That alignment includes information about the corresponding proteins (Fig. 4.11).

Sequences producing significant alignments:							
		Description	Max score	Total score	Query cover	E value	Max ident
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1  PREDI	59.7	59.7	91%	1e-10	29%	XP_003396832.1
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1  PREDI	58.5	58.5	97%	3e-10	28%	XP_003494219.1
<input type="checkbox"/>	PREDICTED: globin-like [Megachile rotundata]	57.8	57.8	89%	6e-10	29%	XP_003707185.1
<input type="checkbox"/>	PREDICTED: globin-like [Apis florea]	53.9	53.9	89%	1e-08	30%	XP_003690810.1
<input type="checkbox"/>	globin 1 [Apis mellifera]	52.8	52.8	89%	4e-08	30%	NP_001071291.1
<input type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1  PREDI	45.1	45.1	89%	2e-05	26%	XP_003396830.1
<input type="checkbox"/>	PREDICTED: neuroglobin-like partial [Acrythosiphon pisum]	42.4	42.4	80%	2e-04	23%	XP_001946608.2
<input type="checkbox"/>	globin putative [Ixodes scapularis]	42.7	42.7	90%	2e-04	25%	XP_002414906.1

**FIGURE 4.9** A typical BLASTP output includes a list of database sequences that match the query. Links are provided to that database entry (e.g., an NCBI Protein entry) and to the pairwise alignment to the query. The bit score and *E* value for each alignment are also provided. Note that the best matches at the top of the list have large bit scores and small *E* values.

Source: BLASTP, NCBI.

The screenshot shows the COBALT interface with the following details:

- Top Bar:** COBALT, Constraint-based Multiple Alignment Tool, My NCBI, Welcome pevner, [Sign Out].
- Header:** Phylogenetic Tree, Edit and Resubmit, Back to Blast Results, Download.
- Title:** Multiple Alignment Results - gi|4504349|ref|NP\_000509.1| hemoglobin subunit... - Cobalt RID U57PC4Y5211 (8 seqs).
- Section:** Descriptions, Select All, Re-align, Alignment parameters.
- Legend:** Legend for links to other resources: UniGene (U), GEO (E), Gene (G), Structure (S), Map Viewer (M).
- Table (Descriptions):**

Accession	Description	Links
XP_003396832	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1  PREDICTED: cytoglobin	G M
XP_003494219	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1  PREDICTED: cytoglobin	G M
XP_003707185	PREDICTED: globin-like [Megachile rotundata]	G
XP_003690810	PREDICTED: globin-like [Apis florea]	G
NP_001071291	globin 1 [Apis mellifera] >emb CAJ43389.1  globin 1 [Apis mellifera] >emb CAJ43388.1  globin 1 [Apis mellifera]	U G M
XP_003396830	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1  PREDICTED: cytoglobin	G M
XP_001946608	PREDICTED: neuroglobin-like, partial [Acyrthosiphon pisum]	G M
XP_002414906	globin, putative [Ixodes scapularis] >gb EEC18571.1  globin, putative [Ixodes scapularis]	G
- Section:** Alignments, Select All, Re-align, Mouse over the sequence identifier for sequence title.
- View Format:** Compact, Conservation Setting: 2 Bits.
- Sequence Alignments:**

Accession	Length
XP_003396832	80
XP_003494219	80
XP_003707185	79
XP_003690810	80
NP_001071291	80
XP_003396830	79
XP_001946608	42
XP_002414906	75

**FIGURE 4.10** The lower part of a BLASTP search (or other BLAST family search) consists of a series of pairwise sequence alignments such as those shown in **Figure 4.5**. Using the reformat option, the results can be displayed as a multiple sequence alignment as shown here for a group of globins. Other output format options are available, allowing the user to inspect regions of similarity as well as divergent regions within protein families.

Source: BLASTP, NCBI.

## Stand-Alone BLAST

The web-based BLAST programs at the NCBI website are extremely popular. As an alternative, you can also download BLAST+, a set of command-line executables for local use (Camacho *et al.*, 2009). You might do this in order to search custom databases, to perform bulk searches (i.e., using very large numbers of queries), implement complex search strategies using custom scripts, or to use your own computer cluster for improved performance.

Instructions for downloading BLAST+ are available at the NCBI BLAST website and in an NCBI help manual. You can install various files and directories including a binary (bin) directory filled with various executables. Copy these executable programs (such as those called `blastp` and `blastn`) to your home directory bin folder (e.g., from the BLAST directory type `$ cp * ~/bin` where \$ indicates a UNIX or Mac terminal command-line prompt, cp is the copy utility, and \* indicates all files in that directory will be copied). This will allow you to execute the copied programs from any location.

From the BLAST homepage (<http://blast.ncbi.nlm.nih.gov/>, WebLink 4.1), click the “Help” tab then follow the link to “Download BLAST Software and Databases.” For an NCBI online book on BLAST+ visit <http://www.ncbi.nlm.nih.gov/books/NBK1762/> (WebLink 4.6).

Download ▾ GenBank Graphics ▾ Next ▾ Previous ▾ Descriptions

Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA

Sequence ID: ref|NM\_00530| Length: 816 Number of Matches: 1

Range 1: 203 to 705 GenBank Graphics					Next Match	Previous Match
Score	Expect	Identities	Gaps	Strand		
410 bits(454)	5e-113	393/503(78%)	3/503(0%)	Plus/Plus		
CDS:hemoglobin subun	1				M V H	
Query	3	ATTIGCTTCTGACACAACITGTGTCAGCAACCTCAAA---CAGACACCATGGTCAT				
Sbjct	203	AICTGCTTCCGACACAGCTGCAATCACTAGCAAGCTCTCAGGCCCTGGCAATCATGGTCAT			M V H	
CDS:hemoglobin subun	1					
CDS:hemoglobin subun	4	L T P E E K S A V T A L W G K V N V D E				
Query	60	CTGACTCTTGAGGGAGAAGCTGCGCTACTGCCCTGIGGGCAAGGTGAACGTGGATGAA				
Sbjct	263	TTTACTGCTGAGGGAGAAGGCTGCGCTACTAGCTGIGGGAGCAAGAAGAATGTTGAAAGAG				
CDS:hemoglobin subun	4	P T A E E K A A V T S L W S K M N V E E				
CDS:hemoglobin subun	24	V G G E A L G R L L V V Y P W T Q R F F				
Query	120	GTGCGGTGAGGCCCTGGGCAGGCTGCTGGCTCTACCCCTGGACCCAGAGGTCTT				
Sbjct	323	GCTGGGGTGAAGCCTGGGCAGACTCTCTGTTACCCCTGGACCCAGAGAATT				
CDS:hemoglobin subun	24	A G G E A L G R L L V V Y P W T Q R F F				
CDS:hemoglobin subun	44	E S F G D L S T P D A V M G N P K V K A				
Query	180	GAGTCCTTGGGATCTGCTTCACTCTGATGCTGTTATGGCAACCTTAAGGTGAAGGCT				
Sbjct	383	GACAGCTTGGAAACCTGTCGTCCTCCCTGCTGGCAATCTGGCAACCCCAAGGCTAACGCC				
CDS:hemoglobin subun	44	D S F G N L S S P S A I L G N P K V K A				
CDS:hemoglobin subun	64	H G K K V L G A F S D G L A H L D N L K				
Query	240	CATGGCAAGAAAGTGCTCGGTGCTTAGTGAIGGCTGGCTCACCTGGACAACCTCAAG				
Sbjct	443	CAIGGCAGAAGGCTGACTCTCTGGAGATGCTATTAAAACATGGACAACCTCAAG				
CDS:hemoglobin subun	64	H G K K V L T S F G D A I K N M D N L K				
CDS:hemoglobin subun	84	G T F A T L S E L H C D K L H V D P E N				
Query	300	GGCACCTTGGCACACTGAGTGAGCTGCACTGTGACARGCTGACGTGGATCTGAGAAC				
Sbjct	503	CCCGCCTTCTGCTAAGCTGAGTGAGCTGCACTGTGACAAGCTGATGTTGGATCTGAGAAC				
CDS:hemoglobin subun	84	P A F A K L S E L H C D K L H V D P E N				
CDS:hemoglobin subun	104	F R L L G N V L V C V L A H H F G K E F				
Query	360	TTCAAGCTCTGGCAACGTGCTGGCTGTGCTGGCCCATACITGGCAAAGAACATT				
Sbjct	563	TICAAGCTCTGGTAACGTGAAGGTGATIAITCTGGCTACTCACTTGGCAAGGAGTC				
CDS:hemoglobin subun	104	F K L L G N V M V I I L A T H F G K E F				
CDS:hemoglobin subun	124	T P P V Q A A Y Q K V V A G V A N A L A				
Query	420	ACCCACCACTGCAAGCTGCTGGCTATCAGAAAGCTGGCTGGCTTAATGCCCTGGCC				
Sbjct	623	ACCCCTGAAGTGCAGGCTGCTGGCAGAAAGCTGGCTGCTGCGCAATGCCCTGGCC				
CDS:hemoglobin subun	124	T P E V Q A A W Q K L V S A V A I A L A				
CDS:hemoglobin subun	144	H K Y H				
Query	480	CACAAAGTATCAGTCAAGCTCGCTT	502			
Sbjct	683	CATAAGTACCACTGAGTCTCTT	705			
CDS:hemoglobin subun	144	H K Y H				

**FIGURE 4.11** For BLASTN searches, the coding sequence (CDS) option in the reformat page allows the amino acid sequence of the coding regions of the query and the subject (i.e., the database match) to be displayed. Here, human beta globin DNA (NM\_000518) was used as a query, and a match to the closely related epsilon 1 globin is shown. The corresponding protein sequences are provided, including mismatches in purple.

Source: BLASTN, NCBI.

To demonstrate BLAST+ we search human beta globin protein against a protein database, working on the Linux operating system (which is closely similar to the Mac terminal). Our three tasks are: (1) to obtain a protein database, (2) obtain a query protein, and (3) perform the search.

1. Let's choose the RefSeq protein database. One approach is to navigate to the Download section of NCBI main page (or BLAST page) to find the link. You can then use a utility such as `wget` to download the database. As a different, preferred approach we will use a Perl script (called `update_blastdb.pl` and included with your BLAST+ installation) to download a database. We make a new directory (I've called

### Related Information

[Gene](#) - associated gene details  
[UniGene](#) - clustered expressed sequence tags  
[Map Viewer](#) - aligned genomic context  
[GEO Profiles](#) - microarray expression data

119

322

179

382

239

442

299

502

359

562

419

622

479

682

it database), then navigate into it to perform our download. Note that the # sign refers to comments I add to the commands.

```
$ mkdir database # this creates a new directory
$ cd database/ # we navigate into that directory
# Enter the following, without arguments, to see a help document.
$ update_blastdb.pl
# Next get a list of all available databases
$ update_blastdb.pl --showall
$ update_blastdb.pl --showall | less
```

There were too many rows of information to view conveniently on a typical computer monitor, so we use the pipe (|) to send the output of `--showall` to the `less` utility. This shows us one page of the output at a time. (For information about `less` or any other utility, enter `$ man less` and view the manual.) The choices include the `refseq_protein` database that we will use, as well as dozens of other databases.

```
$ update_blastdb.pl refseq_protein
```

The result is that the requested database is downloaded in the form of a series of ~600 GB files having extensions `.tar.gz`, indicating they are compressed. We unpack (i.e., decompress) them using the `tar` utility (`-x` is to extract from the archive to the disk, `-v` is verbose output, `-z` filters the archive through `gzip`, and `-f` uses an archive file). This extracts the files to the current directory.

```
$ tar -zxvf refseq_protein.00.tar.gz
```

This unpacked database has a variety of file extensions. However, we simply use the `-db refseq_protein` argument.

2. The query we will use is the human beta globin RefSeq protein (NP\_000509). As one approach you could find the protein sequence on the NCBI (or other) website, copy it, and paste it into a text editor (such as `vim` or `nano` in Linux). We instead use EDirect (introduced in Chapter 2). We output the protein sequence in the FASTA format in a file we call `hbb.txt`.

```
$ esearch -db protein -query "NP_000509" | efetch -format fasta > hbb.txt
$ cat hbb.txt # cat is the concatenate utility that we use to print the
# file
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

3. Now we can perform a search. We first invoke the help document to see the usage of the program.

```
$ blastp --h # Get help
$ blastp -query hbb.txt -db ./database/refseq_protein -out mysearch1
# Note that we use ./ to specify the directory location of the
# executable which is within the executable directory
```

When the search completes, we can view the results:

```
$ less mysearch1
```

The output resembles the text portion of a web-based BLASTP search (there is a list of sequences producing significant alignments with identifiers, bit scores and E values as well as a set of pairwise alignments).

You can reach an FTP site for BLAST databases from

• <http://www.ncbi.nlm.nih.gov/guide/data-software/> (WebLink 4.7), or go directly to • <ftp://ftp.ncbi.nlm.nih.gov/blast/db/> (WebLink 4.8). These databases are formatted

for BLAST searches.

The downloaded database files include md5 checksums. These are useful to confirm that the downloads are complete.

At this point re-visit `blastp --h` to explore the many command-line options that are available. For example, this database currently has a size of 27,715,879 sequences and 9,753,871,274 total letters. The best match is the alignment of the query to itself with score 301 bits and an  $E$  value of  $2 \times 10^{-102}$ . You can change dozens of parameters such as the  $E$  value threshold, word size, gap open and extend penalties, and scoring matrix. You can also change the formatting (including the option to produce an HTML output). Set the database size to be about 1000 times smaller than the default database size.

```
$ blastp -query hbb.txt -db ./database/refseq_protein -dbsize 9750000 -out mysearch2
```

What result would you expect for a search of a database that is 1000 times smaller? Any findings would likely be more significant because they are found among a much smaller pool of possible matches. Indeed, the best result has an expect value of  $2 \times 10^{-105}$ , which is 1000 times smaller than in our first search. This is explained mathematically by Equation (4.5) where we made the right-hand side of the equation 1000 times smaller, explaining why the  $E$  value on the left side was similarly reduced.

## BLAST ALGORITHM USES LOCAL ALIGNMENT SEARCH STRATEGY

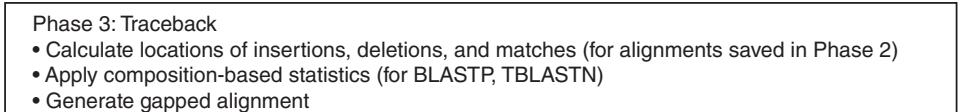
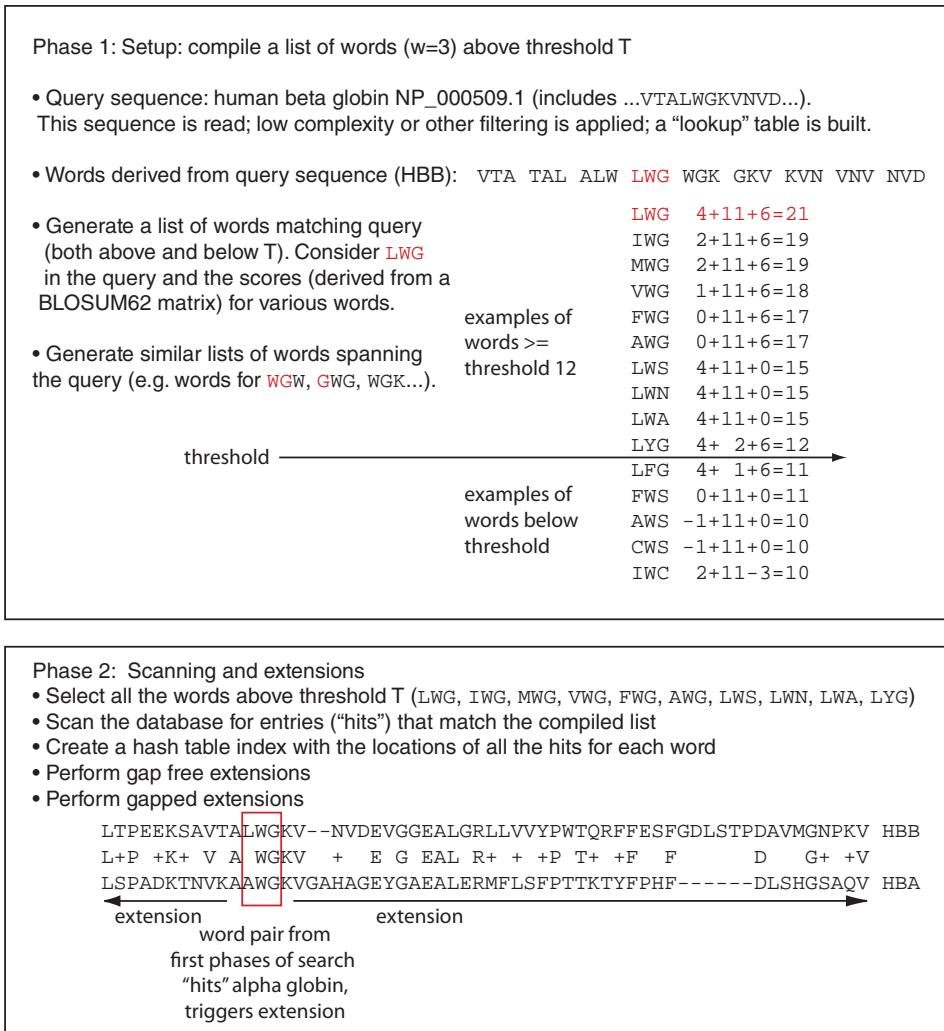
The BLAST search identifies the matches in a database to an input query sequence. Global similarity algorithms optimize the overall alignment of two sequences. These algorithms, including the GAP program (Chapter 3), are best suited for finding matches consisting of long stretches of low similarity. In contrast, local similarity algorithms such as BLAST identify relatively short alignments. Local alignment is a useful approach to database searching because many query sequences have domains, active sites, or other motifs that have local but not global regions of similarity to other proteins. Databases typically also have fragments of DNA and protein sequences that can be locally aligned to a query.

### BLAST Algorithm Parts: List, Scan, Extend

The BLAST search algorithm finds a match between a query and a database sequence and then extends the match in either direction (Altschul *et al.*, 1990, 1997). The search results consist of both highly related sequences from the database as well as marginally related sequences, along with a scoring scheme to describe the degree of relatedness between the query and each database hit. The BLASTP algorithm can be described in three phases (Camacho *et al.*, 2009; **Fig. 4.12**):

1. For protein searches, BLAST compiles a preliminary list of pairwise alignments called word pairs.
2. The algorithm scans a database for word pairs that meet some threshold score  $T$ . When this occurs, such hits are extended using ungapped and gapped alignments. BLAST extends the word pairs to find those that surpass a cutoff score  $S$ , at which point those hits will be reported to the user. Scores are calculated from scoring matrices (such as BLOSUM62) along with gap penalties.
3. A trace-back procedure is performed in which the locations of insertions, deletions and mismatches are assigned.

In the first phase, the BLASTP algorithm compiles a list of “words” of a fixed length  $w$  that are derived from the query sequence. A threshold value  $T$  is established for the score of aligned words. Those words either at or above the threshold are collected and used to identify database matches; those words below threshold are not further pursued. For protein searches the word size typically has a default value of 3. Since there are



**FIGURE 4.12** Schematic of the original BLAST algorithm. In the setup phase a query sequence (such as human beta globin) is analyzed with a given word size (e.g.,  $w = 3$ ), and a list of words is compiled having a threshold score (e.g.,  $T = 11$ ). Several possible words derived from the query sequence are listed in the figure (from LWG to IWC); in a BLAST search there are 8000 words compiled for  $w = 3$ . For a given word, such as the portion of the query sequence consisting of LWG, a list of words is compiled with scores greater than or equal to some threshold  $T$  (e.g., 12). In this example, 15 words are shown along with their scores from a BLOSUM62 matrix; 10 of these are above the threshold, and 5 are below. In phase 2, a database is scanned to find entries that match the compiled word list. Ungapped and gapped extensions are performed, although (to increase efficiency) positions are not saved. The database hits are extended in both directions to obtain high-scoring segment pairs (HSPs). If a HSP score exceeds a particular cutoff score  $S$ , it is reported in the BLAST output. In phase 3, a trace-back is performed and locations of insertions and deletions are recorded. Note that in this particular example the word pair that triggers the extension step is not an exact match (see boxed residues LWG aligned to AWG). The main idea of the threshold  $T$  for protein searches is to also allow both exact and related but nonexact word hits to trigger an extension. For nucleotide BLASTN searches, exact matches are required rather than words above a threshold.

In the BLAST papers by Steven Altschul, David Lipman, and colleagues, the threshold parameter is denoted  $T$  (Altschul *et al.*, 1990, 1997). In the command-line BLAST+ program (described below), the threshold parameter is controlled by the `-threshold` option. You can see the threshold in the output of a web-based BLAST search (Fig. 4.7, arrow 3).

For the parameter  $A$ , the default value is 0 (for BLASTN and megablast) and 40 (for other programs such as BLASTP). This is shown in the Window Size entry of Figure 4.7.

20 amino acids, there are  $20^3 = 8000$  possible words. The word size parameter can be modified by the BLAST user, as described above (see option 3). The threshold score  $T$  can be lowered to identify more initial pairwise alignments. This will increase the time required to perform the search and may increase the sensitivity.

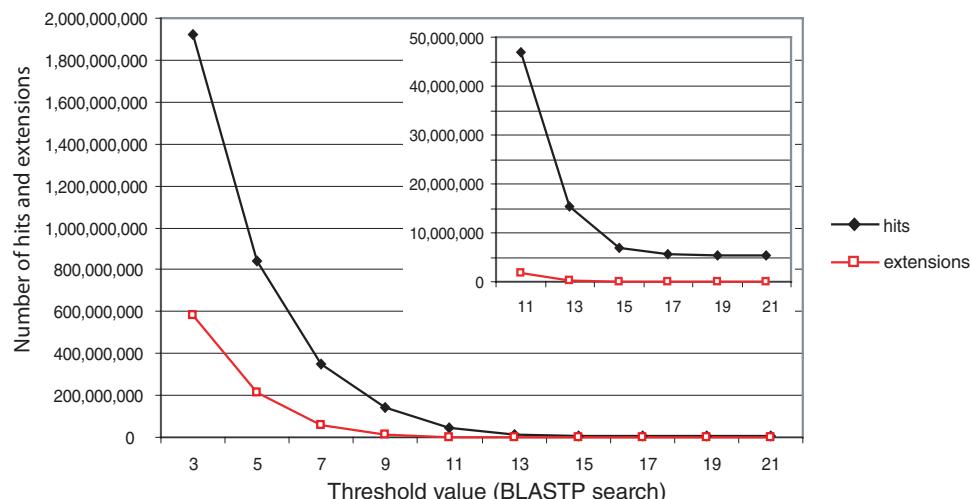
For BLASTN, the first phase is slightly different. Threshold scores are not used in association with words. Instead, the algorithm demands exact word matches. The default word size is 11 (and can be adjusted by the user to values of 7 or 15). Lowering the word length effectively achieves the same aim as lowering the threshold score. Specifying a smaller word size induces a slower, more accurate search.

In the second phase, after compiling a list of word pairs at or above threshold  $T$ , the BLAST algorithm scans a database for hits. This requires BLAST to search an index of the database to find entries that correspond to words on the compiled list. In the original implementation of BLAST, one hit was sufficient. In the current versions of BLAST, the algorithm seeks two separate word pairs (i.e., two nonoverlapping hits) within a certain distance  $A$  from each other. It then generates an ungapped extension of these hits (Altschul *et al.*, 1997). The two-hit approach greatly speeds up the time required to do a BLAST search. Compared to the one-hit approach, the two-hit method generates on average about three times as many hits, but the algorithm then needs to perform only one-seventh as many extensions (Altschul *et al.*, 1997). BLAST extends hits to find alignments called high-scoring segment pairs (HSPs). For sufficiently high-scoring alignments, a gapped extension is triggered. The extension process is terminated when a score falls below a cutoff.

In the third phase, a trace-back is performed in which the locations of insertions, deletions, and matches are assigned. Composition-based statistics are applied.

In summary, the main strategy of the BLAST algorithm is to compare a protein or DNA query sequence to each database entry and to form pairwise alignments (HSPs). As a heuristic algorithm, BLAST is designed to offer both speed and sensitivity. When the threshold parameter is raised, the speed of the search is increased but fewer hits are registered; distantly related database matches may be missed. When the threshold parameter is lowered, the search proceeds far more slowly but many more word hits are evaluated as sensitivity is increased.

We can demonstrate the effect of different threshold levels on a BLASTP search by changing the  $f$  parameter from its default value (11) to a range of other values. The results are dramatic (Fig. 4.13). With the default threshold value of 11, there are about 47 million hits to the database and 1.8 million extensions. When the threshold is lowered to just 3,



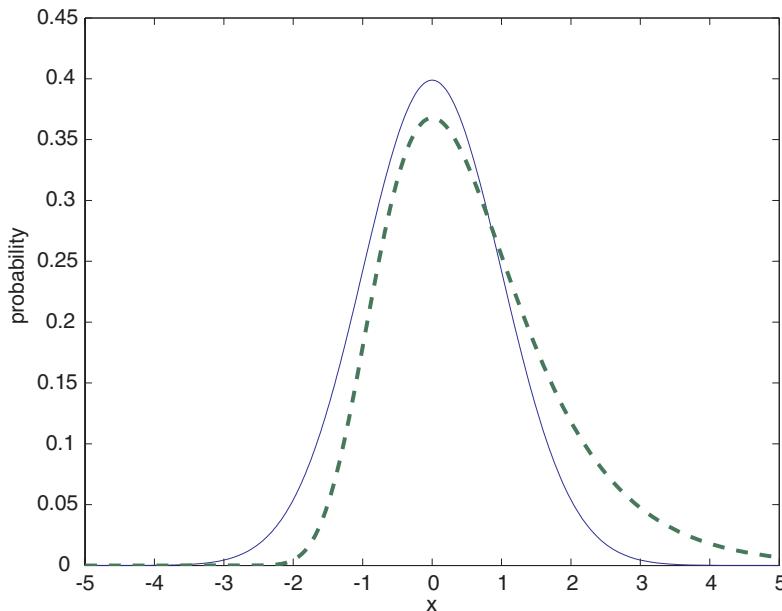
**FIGURE 4.13** The effect of varying the threshold ( $x$  axis) on the number of database hits (black line) and extensions (red line). BLASTP searches were performed using human beta globin as a query.

there are about 1.9 billion hits to the database and 582 million extensions. This occurs because many additional words have scores above  $T$ . With the threshold raised to 15 or higher, there are only about 6 million hits and 50,000 extensions. The final results of the search are not dramatically different with the default value compared to the lowered or raised threshold values, as the number of gapped HSPs is comparable. With the high threshold some matches were missed, although the reported matches are more likely to be true positives; with the lower threshold values there were somewhat more successful extensions. This supports the conclusion that a lower threshold parameter yields a more accurate search, though a slower one. This trade-off between sensitivity and speed is central to the BLAST algorithm. Practically, for most users of BLAST the default threshold parameters are always appropriate.

### BLAST Algorithm: Local Alignment Search Statistics and $E$ Value

We care about the statistical significance of a BLAST search because we want some quantitative measure of whether the alignments represent significant matches or whether they would be expected to occur by chance. For local alignments (including BLAST searches), rigorous statistical tests have been developed (Altschul *et al.*, 1990, 1994, 1997; Altschul and Gish, 1996; Pagni and Jongeneel, 2001).

We have described how local, ungapped alignments between two protein sequences are analyzed as HSPs. Using a substitution matrix, specific probabilities are assigned for each aligned pair of residues and a score is obtained for the overall alignment. For the comparison of a query sequence to a database of random sequences of uniform length, the scores can be plotted and shown to have the shape of an extreme value distribution (see Fig. 4.14, where it is compared to the normal distribution). The normal or Gaussian distribution forms the familiar, symmetric bell-shaped curve. The extreme value distribution is skewed to the right, with a



**FIGURE 4.14** Normal distribution (solid line) is compared to the extreme value distribution (dotted line). Comparing a query sequence to a set of uniform-length random sequences usually generates scores that fit an extreme-value distribution (rather than a normal distribution). The area under each curve is 1. For the normal distribution, the mean ( $\mu$ ) is centered at zero, and the probability  $Z$  of obtaining some score  $x$  is given in terms of units of standard deviation ( $\sigma$ ) from  $x$  to the mean:  $Z = (x - \mu)/\sigma$ . In contrast to the normal distribution, the extreme value distribution is asymmetric with a skew to the right. It is fit to the equation  $f(x) = (e^{-x})(e^{-e^{-x}})$ . The shape of the extreme value distribution is determined by the characteristic value  $u$  and the decay constant  $\lambda$  ( $u = 0$ ;  $\lambda = 1$ ).

## BOX 4.1. THE EXTREME VALUE DISTRIBUTION

The shape of the extreme value distribution shown in **Figure 4.14** is described by two parameters: the characteristic value  $\mu$  and the decay constant  $\lambda$ . The extreme value distribution is sometimes called the Gumbel distribution, after the person who described it in 1958. The application of the extreme value distribution to BLAST searching has been reviewed by Altschul *et al.* (1994), Altschul and Gish (1996), and Pagni and Jongeneel (2001). For two random sequences  $m$  and  $n$ , the cumulative distribution function of scores  $S$  is defined:

$$P(S < x) = \exp(-e^{-\lambda(x-u)}). \quad (4.1)$$

(Note that the characteristic value  $u$  relates to the maximum of the distribution, although it is not the mean  $\mu$ .) To use this equation, we need to know (or estimate) the values of the parameters  $u$  and  $\lambda$ . For ungapped alignments, the parameter  $u$  is dependent on the lengths of the sequences being compared and is defined:

$$u = \frac{\ln Kmn}{\lambda} \quad (4.2)$$

where  $m$  and  $n$  refer to the lengths of the sequences being compared and  $K$  is a constant. Combining Equations (4.1) and (4.2), the probability of observing a score equal to or greater than  $x$  by chance is given by:

$$P(S \geq x) = 1 - \exp(-kmne^{-\lambda x}). \quad (4.3)$$

Our goal is to understand the likelihood that a BLAST search of an entire database produces a result by chance alone. The number of ungapped alignments with a score of at least  $x$  is described by the parameter  $Kmn e^{-\lambda x}$ . In the context of a database search,  $m$  and  $n$  refer to the length (in residues) of the query sequence and the length of the entire database, respectively. The product  $m n$  defines the size of the search space. The search space represents all the sites at which a query sequence can be aligned to any sequence in the database. Because the ends of a sequence are not as likely to participate in an average-sized alignment, the BLAST algorithm calculates the effective search space in which the average length of an alignment  $L$  is subtracted from  $m$  and  $n$  (Altschul and Gish, 1996):

$$\text{Effective Search Space} = (m - L)(n - L). \quad (4.4)$$

As described above (“Stand-Alone BLAST”), you can adjust the effective search space using BLAST+.

Because of the rapid tailing of the normal distribution in  $x^2$ , if we tried to use the normal distribution to describe the significance of a BLAST search result (e.g., by estimating how many standard deviations above the mean a search result occurs) we would tend to overestimate the significance of the alignment.

Equation (4.5) is described online in the document “The Statistics of Sequence Similarity Scores,” available in the help section of the NCBI BLAST site (<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>, WebLink 4.9).

tail that decays in  $x$  (rather than  $x^2$ , which describes the decay of the normal distribution). The properties of this distribution are central to our understanding of BLAST statistics because they allow us to evaluate the likelihood that the highest scores from a search (i.e., the values at the right-hand tail of the distribution) occurred by chance.

From the equations defining the extreme value distribution (Box 4.1, Equations (4.1)–(4.4)) we can derive a formula that describes the likelihood that a particular BLAST score occurs by chance. The expected number of HSPs having some score  $S$  (or better) by chance alone is defined:

$$E = Kmne^{-\lambda S} \quad (4.5)$$

where  $E$  refers to the expect value, which is the number of different alignments with scores equivalent to or better than  $S$  that are expected to occur by chance in a database search. This provides an estimate of the number of false positive results from a BLAST search. From Equation (4.5) we see that the  $E$  value depends on the score and on  $\lambda$ , which is a parameter that scales the scoring system.  $E$  also depends on the length of the query sequence and the length of the database. The parameter  $K$  is a scaling factor for the search space. The parameters  $K$  and  $\lambda$  are described by Karlin and Altschul (1990), and are often referred to as Karlin–Altschul statistics.

Note several important properties of Equation (4.5):

- The value of  $E$  decreases exponentially with increasing  $S$ . The score  $S$  reflects the similarity of each pairwise comparison and is partly based upon the scoring matrix selected. Higher  $S$  values correspond to better alignments and lower  $E$  values.

As  $E$  approaches zero, the probability that the alignment occurred by chance approaches zero. We relate the  $E$  value to probability ( $P$ ) values below.

- The expected score for aligning a random pair of amino acids must be negative. Otherwise, very long alignments of two sequences could accumulate large positive scores and appear to be significantly related when they are not.
- The size of the database that is searched – as well as the size of the query – influences the likelihood that particular alignments will occur by chance. Consider a BLAST result with an  $E$  value of 1. This value indicates that, in a database of this particular size, one match with a given score (or better) is expected to occur by chance. If the database were twice as big, there would be twice the likelihood of finding a score equal to or greater than  $S$  by chance.
- The theory underlying Equation (4.5) was developed for ungapped alignments. For these, BLAST calculates values for  $\lambda$ ,  $K$ , and  $H$  (entropy; see Fig. 3.27). Equation (4.5) can be successfully applied to gapped local alignments as well (such as the results of a BLAST search). For gapped alignments however,  $\lambda$ ,  $K$  and  $H$  cannot be calculated analytically, but instead are estimated by simulation and looked up in a table of precomputed values.

### Making Sense of Raw Scores with Bit Scores

A typical BLAST output reports  $E$  values, raw scores, and bit scores. Raw scores are calculated from the substitution matrix and gap penalty parameters that are chosen. The bit score  $S'$  is calculated from the raw score by normalizing with the statistical variables that define a given scoring system. Bit scores from different alignments, even those employing different scoring matrices in separate BLAST searches, can therefore be compared. A raw score from a BLAST search must be normalized to parameters such as the size of the database being queried. The raw score is related to the bit score by:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \quad (4.6)$$

where  $S'$  is the bit score, which has a standard set of units. The  $E$  value corresponding to a given bit score is defined:

$$E = mn \times 2^{-S'} \quad (4.7)$$

Why are bit scores useful? First, raw scores are unitless and have little meaning alone. Bit scores account for the scoring system that was used and describe the information content inherent in a pairwise alignment. They therefore allow scores to be compared between different database searches, even if different scoring matrices are employed. Second, bit scores can tell you the  $E$  value if you know the size of the search space,  $m \times n$ . (The BLAST algorithms use the effective search space size, described above.)

### BLAST Algorithm: Relation Between $E$ and $p$ Values

The  $p$  value is the probability of a chance alignment occurring with the score in question or better. It is calculated by relating the observed alignment score  $S$  to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant  $p$  values are those close to zero. The  $p$  and  $E$  values are different ways of representing the significance of the alignment. The probability of finding an HSP with a given  $E$  value is

$$p = 1 - e^{-E} \quad (4.8)$$

**TABLE 4.3 Relationship of  $E$  to  $p$  values in BLAST using Equation (4.8). Small  $E$  values (0.05 or less) correspond closely to the  $p$  values.**

$E$	$p$
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001000

**Table 4.3** lists several  $p$  values corresponding to  $E$  values. While BLAST reports  $E$  values rather than  $p$  values, the two measures are nearly identical, especially for very small values associated with strong database matches. An advantage of using  $E$  values is that it is easier to think about  $E$  values of 5 versus 10 rather than 0.99326205 versus 0.99995460.

A  $p$  value below 0.05 is traditionally used to define statistical significance (i.e., to reject the null hypothesis that your query sequence is not related to any database sequence). If the null hypothesis is true, then 5% of all random alignments will result in an apparently significant score. An  $E$  value of 0.05 or less may therefore be considered significant.

It is also possible to approach  $E$  values with conservative corrections. We discussed probability ( $p$ ) values in Chapter 3, and we return to the topic in Chapter 11 when we discuss microarray data analysis. The significance level  $\alpha$  is typically set to 0.05, such that a  $p$  value of 0.05 suggests that some observation (e.g., the score of a protein query to a match in a database) is likely to have occurred by chance 1 time in 20. The null hypothesis is that your query is not homologous to the database match, and the alternate hypothesis is that they are homologous. If the  $p$  value is sufficiently small (e.g.,  $p < 0.05$ ), we can reject the null hypothesis. When you search a database that has one million proteins, there are many opportunities for your query to find matches. Five percent of 1 million proteins is 50,000 proteins, and we might expect to obtain that many matches (with  $p = 0.05$ ) by chance. A related issue arises in microarray data analysis when we compare two conditions (e.g., normal versus diseased sample) and measure the RNA transcript levels of 20,000 genes: 1000 transcripts (i.e., 5%) may be differentially expressed by chance.

This situation involves multiple comparisons: you are not hypothesizing that your query will match *one* particular database entry, you are interested in knowing if it matches *any* entries. A solution is to correct for multiple comparisons by adjusting the  $\alpha$  level. A very conservative way to do this (called the Bonferroni correction) is to divide  $\alpha$  by the number of measurements (e.g., divide  $\alpha$  by the size of the database). In the case of BLAST searches this is automatically done as shown in Equation (4.5), because the  $E$  value is divided by the effective search space.

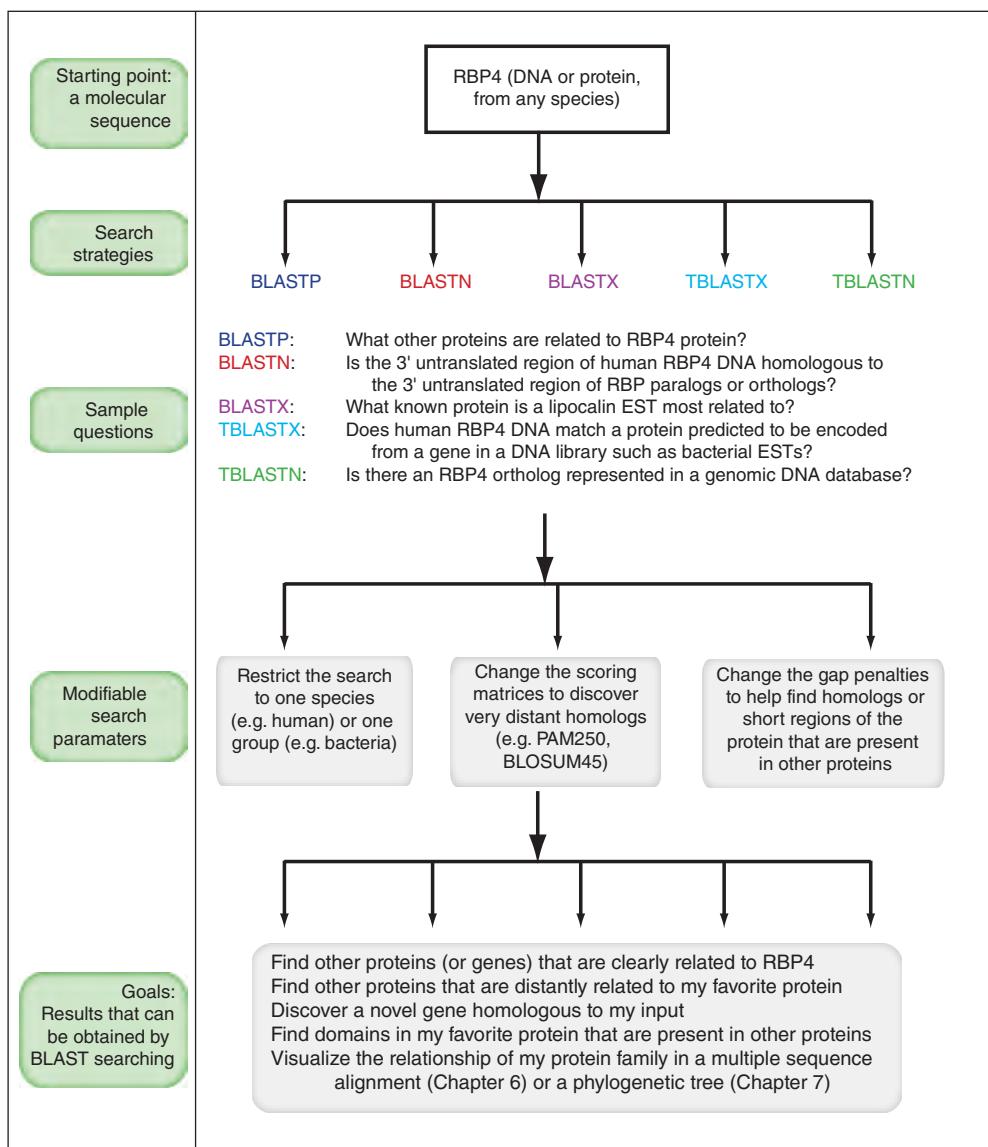
Beyond this multiple comparison correction inherent in BLAST, some researchers consider it appropriate to adjust the significance level  $\alpha$  for search results from 0.05 to some even lower value. In analyses of completed microbial genomes, BLAST or FASTA search  $E$  values were reported as significant if they were below  $10^{-4}$  (Ferretti *et al.*, 2001) or below  $10^{-5}$  (Ermolaeva *et al.*, 2001; Colbourne *et al.*, 2011; Huang *et al.*, 2013). In the public consortium analysis of the human genome, Smith–Waterman alignments were reported with an  $E$  value threshold of  $10^{-3}$  and TBLASTN searches used a threshold of  $10^{-6}$  (International Human Genome Sequencing Consortium, 2001). You can choose how conservatively to interpret BLAST results.

Some BLAST servers use  $p$  values in the output.

## BLAST SEARCH STRATEGIES

### General Concepts

BLAST searching is a tool to explore databases of protein and DNA sequence. We have introduced the procedure: it is essential that you define the question you want answered, the DNA or protein sequence you want to input, the database you want to search, and the algorithm you want to use. We now address some basic principles regarding strategies for BLAST searching (Altschul *et al.*, 1994). We illustrate these issues with globin, lipocalin and HIV-1 Pol searches. Key issues include how to evaluate the statistical significance of BLAST search results, and how to modify the optional parameters of the BLAST programs when your search yields too little or too much information. An overview of the types of searches that can be performed with RBP4 DNA (NM\_006744.3) or protein (NP\_006735.2) sequence is depicted in **Figure 4.15**.



**FIGURE 4.15** Overview of BLAST searching strategies. There are many hundreds of questions that can be addressed with BLAST searching, from characterizing the genome of an organism to evaluating the sequence variation in a single gene.

## Principles of BLAST Searching

### *How to Evaluate the Significance of Results*

When you perform a BLAST search, which database matches are authentic? To answer this question, we first define a true positive as a database match that is homologous to the query sequence (descended from a common ancestor). Homology is inferred based on sequence similarity, with support from statistical evaluation of the search results. The error rate of database search algorithms is reduced by using statistical scores such as expect values rather than relying on percentage identity (or percent similarity) (Gotoh, 1996; Brenner, 1998; Park *et al.*, 1998). We therefore focus on inspection of *E* values.

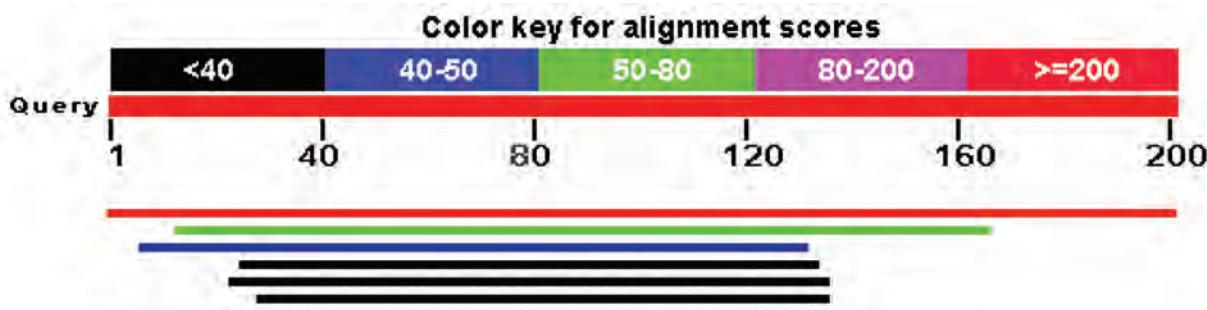
The problem of assigning homology between genes or proteins is not solved by sequence analysis alone: it is also necessary to apply biological criteria to support the inference of homology. BLAST results can be implemented with evaluations of protein structure and function. The sequences of genuinely related proteins can diverge greatly, even while these proteins retain a related three-dimensional structure. We therefore expect that database searches (and pairwise protein alignments) result in a number of false negative matches. Many members of the lipocalin family, such as RBP4 and odorant-binding protein (OBP), share very limited sequence identity, although their three-dimensional structures are closely related and their functions as carriers of hydrophobic ligands are thought to be the same.

The accession of the human RBP4 protein is NP\_006735.2. The DELTA-BLAST program (Chapter 5) generates a score of over 52 bits, an aligned region of 182 amino acid residues, and an *E* value of  $2 \times 10^{-8}$  for the same match of RBP4 to complement component 8 gamma.

Consider a BLASTP search of the RefSeq database restricted to human entries using human RBP4 protein as a query. There are 6 entries in this case (Fig. 4.16a,b). The best *E* value score ( $1 \times 10^{-150}$ ) is RBP4 itself. Subsequent matches have significant *E* values ( $1 \times 10^{-9}$ ,  $5 \times 10^{-4}$ , 0.034) but then two entries have nonsignificant values. The alignment to complement component 8 gamma (NP\_000597) has a nonsignificant *E* value of 0.18, and that protein shares only 25% amino acid identity with RBP over a span of 114 amino acid residues (including three gap regions in the alignment; Fig. 4.16c). It might be concluded that these two proteins are not homologous; in this case they are, however. In deciding whether two proteins (or DNA sequences) are homologous, several questions can be asked:

- Is the expect value significant? In this particular case it is not, because the proteins are distantly related. Search techniques such as DELTA-BLAST or HMMER (based on hidden Markov models), introduced in Chapter 5, can typically assign higher scores and lower *E* values to distant matches.
- Are the two proteins approximately the same size? It is not at all required that homologous proteins have similar sizes, and it is possible for two proteins to share only a limited domain in common. Indeed, local alignments search tools such as BLAST are specialized to find limited regions of overlap. However, it is also important to develop a biological intuition about the likelihood that two proteins are homologous. A 1000-amino-acid protein with transmembrane domains is relatively unlikely to be homologous to RBP, and the vast majority of lipocalins are approximately 200 amino acids in length (20–25 kilodaltons).
- Do the proteins share a common motif or signature? In this case, both RBP4 and complement component 8 gamma have a glycine-X-tryptophan (GXW; X signifies any residue, as indicated in the boxed region of Fig. 4.16c) signature that is characteristic of the lipocalin superfamily.
- Are the proteins part of a reasonable multiple sequence alignment? Note that you can create a multiple sequence alignment of selected BLAST results, as indicated in the results table (Fig. 4.16b).
- Do the proteins share a similar biological function? Like all lipocalins, both proteins are small, hydrophilic, abundant, secreted molecules.

(a) Graphical overview



(b) List of alignments

**Sequences producing significant alignments:**

Select: All None Selected: 6

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input checked="" type="checkbox"/>	<a href="#">retinol-binding protein 4 precursor [Homo sapiens]</a>	420	420	100%	1e-150	100%	NP_006735.2
<input checked="" type="checkbox"/>	<a href="#">apolipoprotein D precursor [Homo sapiens]</a>	55.5	55.5	76%	1e-09	28%	NP_001638.1
<input checked="" type="checkbox"/>	<a href="#">glycodeulin precursor [Homo sapiens] &gt;ref NP_002562.2  glycodeulin precursor [Homo s</a>	40.0	40.0	62%	5e-04	26%	NP_001018059.1
<input checked="" type="checkbox"/>	<a href="#">protein AMBP preproprotein [Homo sapiens]</a>	35.0	35.0	54%	0.034	23%	NP_001624.1
<input checked="" type="checkbox"/>	<a href="#">complement component C8 gamma chain precursor [Homo sapiens]</a>	32.3	32.3	56%	0.18	25%	NP_000597.2
<input checked="" type="checkbox"/>	<a href="#">lipocalin-15 precursor [Homo sapiens]</a>	28.5	28.5	53%	3.4	23%	NP_976222.1

(c) Pairwise alignment of RBP4 and C8G

**complement component C8 gamma chain precursor [Homo sapiens]**Sequence ID: [ref|NP\\_000597.2|](#) Length: 202 Number of Matches: 1Range 1: 33 to 139 [GenPept](#) [Graphics](#)[Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
32.3 bits(72)	0.18	Compositional matrix adjust.	28/114(25%)	49/114(42%)	8/114(7%)
Query 24	VSSFRVKENFDKARFSGTWTYAMAKKDPEGLFLQDNIVAEFSVDETG-QMSATAKGRVRLL				82
	+S+ + K NFD +F+GTW +A + AE + Q +A A R L				
Sbjct 33	ISTIQPKANFDAQQFAGTWLLAVGSACRFLQEQGHRAEAITLHVAPQGTAMAVSTFRKL				92
Query 83	NNWDVCADMVGTFITDTEPDKMVKYWGVASFLQKGNDHWIVDTDYDTYAVQY				136
	+ +C + + DT +F ++ +G + +TDY ++AV Y				
Sbjct 93	DG--ICWQVRQLYGDTGVLRFLLQARD-----RGAVHVVVAETDYQSFAVLY				139

**FIGURE 4.16** Results of a BLASTP nr search using human RBP as a query, restricting the output to human RefSeq proteins. (a) The graphical overview shows that there are 6 hits, only one of which (RBP4 itself) has a high score (bar shaded red) extending across the length of the query. (b) The BLASTP output includes a list of alignments. Inspection of the E values suggests that, in addition to RBP itself, several authentic paralogs may have been identified by this search. Is complement component 8 gamma (C8G), having an alignment E value of 0.18, likely to be homologous to RBP? (c) Pairwise alignment of RBP4 and C8G, provided as part of the BLASTP output, includes 25% amino acid identity and alignment of a GXW motif (red rectangle) that is consistently conserved among lipocalin carrier proteins such as RBP4.

Source: BLASTP, NCBI.

An accession number for the three-dimensional structure of human complement protein C8 $\gamma$  is 1IW2, while for RBP4 an accession is 1RBP. We discuss Protein Data Bank accession numbers (such as these) in Chapter 13.

We define motifs and signatures in Chapter 12 and trees in Chapter 7.

- Do the proteins share a similar three-dimensional structure? Although there is great diversity in lipocalin sequences, they share a remarkably well-conserved structure. This structure, a cup-like calyx, allows them to transport hydrophobic ligands across an aqueous compartment (see Chapter 13).
- Is the genomic context informative? The human complement component gamma gene has a similar number and length of exons as other lipocalins (Kaufman and Sodetz, 1994). It is mapped to chromosome 9q34.3, immediately adjacent to another lipocalin gene (*LCN12*) in the vicinity of 10 other other lipocalin genes on 9q34. This information suggests that the BLASTP match is biologically significant, even if the *E* value is not statistically significant.
- If a BLAST search results in a marginal match to another protein, perform a new BLAST search using that distantly related protein as a query. A BLASTP nr search using complement component 8 gamma (C8G) as a query results in the identification of several proteins (complex-forming glycoprotein HC and  $\alpha$ -1-microglobulin/bikunin) that are also detected by RBP4 (Fig. 4.17a,b). This finding increases our confidence that RBP4 and complement component 8 gamma are in fact homologous members of a protein superfamily. If the BLASTP search using complement component 8 gamma had shown that protein to be part of another characterized family, this would have suggested that it is not related to RBP4.

We can see examples of matches to nonhomologous proteins in our search using complement component C8 gamma as a query. One match is to a protein (tenascin-X isoform 1) that is 4242 amino acids in length, and does not include the GXW motif (Fig. 4.17c). Another match is to a neuroblastoma-amplified sequence that has 44% amino acid identity to the query but only over a span of 41 residues. We can perform a BLASTP query with this neuroblastoma-amplified sequence to see that it is a member of a Sec39 superfamily with no annotated relationship to lipocalins.

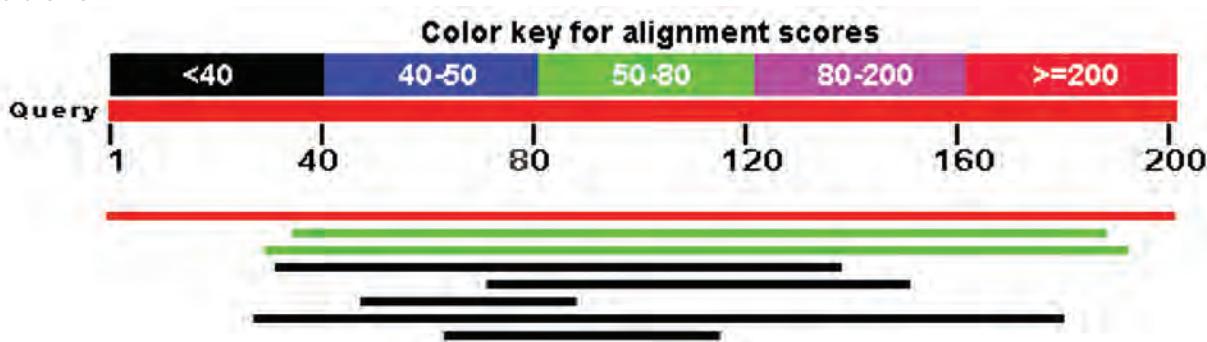
Historically, early database searches yielded results that were entirely unexpected. In 1984, the  $\beta$ -adrenergic receptor was found to be homologous to rhodopsin (Dixon *et al.*, 1986). This was surprising because of the apparent differences between these receptors in terms of function and localization: rhodopsin is a retina-specific receptor for light, and



**FIGURE 4.17** Results of a BLASTP search against human proteins via the nonredundant database, using human complement component 8 gamma (C8G) as a query. (a) The graphical overview shows 8 matches, including the query to itself (red bar) and several alignments with low scores (black bars) spanning just short stretches of amino acids. (b) The list of alignments includes RBP4 and other members of the lipocalin family. This “reciprocal” search supports the hypothesis that C8G, identified in a previous RBP4 search, is an authentic homolog. Here three database matches are not homologous (arrows). The *E* values are unconvincingly high, and the proteins are members of protein families other than lipocalins (as can be confirmed by separate BLAST searches). (c) Inspection of pairwise alignments between C8G and two putative nonhomologous proteins shows that these proteins are far larger than typical lipocalins (4242 and 2371 amino acid residues). The tenascin X isoform 1 does not overlap the highly conserved GXW motif. The neuroblastoma-amplified sequence does match the GXW motif, but the region of overlap extends to only 41 residues. These results highlight the need to inspect each pairwise alignment from a BLAST search. The *E* value provides a statistical argument for evaluating possible homology, but it should be complemented by knowledge of the biological properties of the sequences. Here RBP4, C8G, and other lipocalins are soluble, hydrophilic, abundant proteins that probably share similar functions as carrier proteins; they also share similar three-dimensional structures (see Chapter 13).

Source: BLASTP, NCBI.

## (a) Graphical overview



## (b) List of alignments

## Sequences producing significant alignments:

Select: All None Selected: 0

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
<input type="checkbox"/>	<a href="#">complement component C8 gamma chain precursor [Homo sapiens]</a>	412	412	100%	3e-147	100%	<a href="#">NP_000597.2</a>
<input type="checkbox"/>	<a href="#">lipocalin-15 precursor [Homo sapiens]</a>	69.7	69.7	76%	1e-14	34%	<a href="#">NP_976222.1</a>
<input type="checkbox"/>	<a href="#">protein AMBP preproprotein [Homo sapiens]</a>	68.9	68.9	80%	1e-13	25%	<a href="#">NP_001624.1</a>
<input type="checkbox"/>	<a href="#">retinol-binding protein 4 precursor [Homo sapiens]</a>	33.1	33.1	52%	0.12	25%	<a href="#">NP_006735.2</a>
<input type="checkbox"/>	<a href="#">tenascin-X isoform 1 precursor [Homo sapiens]</a> ← Not homologous	30.0	30.0	39%	1.5	31%	<a href="#">NP_061978.6</a>
<input type="checkbox"/>	<a href="#">neuroblastoma-amplified sequence [Homo sapiens]</a> ← Not homologous	29.6	29.6	20%	2.1	44%	<a href="#">NP_056993.2</a>
<input type="checkbox"/>	<a href="#">neutrophil gelatinase-associated lipocalin precursor [Homo sapiens]</a>	28.9	28.9	75%	2.9	21%	<a href="#">NP_005555.2</a>
<input type="checkbox"/>	<a href="#">HBS1-like protein isoform 1 [Homo sapiens]</a> ← Not homologous	28.5	28.5	25%	5.4	33%	<a href="#">NP_006611.1</a>

## (c) Pairwise alignments with nonhomologous proteins

[Download](#) [GenPept](#) [Graphics](#)

tenascin-X isoform 1 precursor [Homo sapiens]

Sequence ID: [ref|NP\\_061978.6|](#) Length: 4242 Number of Matches: 1

Range 1: 3255 to 3330 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
30.0 bits(66)	1.5	Compositional matrix adjust.	25/81(31%)	36/81(44%)	6/81(7%)

Query 73    TTLHVAPQGTAMAVSTFRKLD-GICWQVRQLYGDTGVLGRFLQARDARGAVHVVVAETD    131  
           I L V P+ +AV+                    G+ W V Q                    G                    FL+Q RDA+G            V            D  
 Sbjct 3255    TPLPVEPRLGEVAAVTSDSVGLSWIVAQ----GPFDSFLVQYRDAQQQPQAVPVSGD    3309

Query 132    YQSFAVLYLERAGQLSVKLYA    152  
           ++ AV L+ A +                    L+  
 Sbjct 3310    LRAVAVSGLDPARKYKFLLFG    3330

[Download](#) [GenPept](#) [Graphics](#)

neuroblastoma-amplified sequence [Homo sapiens]

Sequence ID: [ref|NP\\_056993.2|](#) Length: 2371 Number of Matches: 1

Range 1: 2323 to 2360 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
29.6 bits(65)	2.1	Compositional matrix adjust.	18/41(44%)	23/41(56%)	3/41(7%)

Query 49    GTWLLVAVG SACRFI LQE QGHRAE ATT LHVAPQGTAMAVSTF    89  
           G W            +G            R L+E GH AEA +L +A +GT A TF  
 Sbjct 2323    GRWDAAEELG---RHLREAGHEAEAGSLLLAVRGTHQAFRIF    2360

Go to NCBI Gene and enter “rhodopsin,” restricting the organism to human. There are >700 entries, mostly consisting of members of this family of receptors thought to have seven transmembrane spans. We learn how to explore protein families in Chapter 12.

the adrenergic receptors were known to bind epinephrine (adrenalin) and norepinephrine, stimulating a signal transduction cascade that results in cyclic adenosine monophosphate (cAMP) production. Alignment of the protein sequences revealed that they share similar structural features (seven predicted transmembrane domains). It is now appreciated that rhodopsin and the  $\beta$ -adrenergic receptor are prototypic members of a superfamily of receptors that bind ligands, initiating a second messenger cascade. Another surprising finding was that some viral genes that are involved in transforming mammalian cells are actually derived from the host species. The human epidermal growth factor receptor was sequenced and found to be homologous to an avian retroviral oncogene, *v-erb-B* (Downward *et al.*, 1984). There are many more examples of database searches that revealed unexpected relationships. In many other cases, the reported relationships represented false positive results. The false positive error rate will yield occasional matches that are not authentic, and comparison of the three-dimensional structures of the potential homologs can be used as a criterion for deciding whether two proteins are in fact homologous.

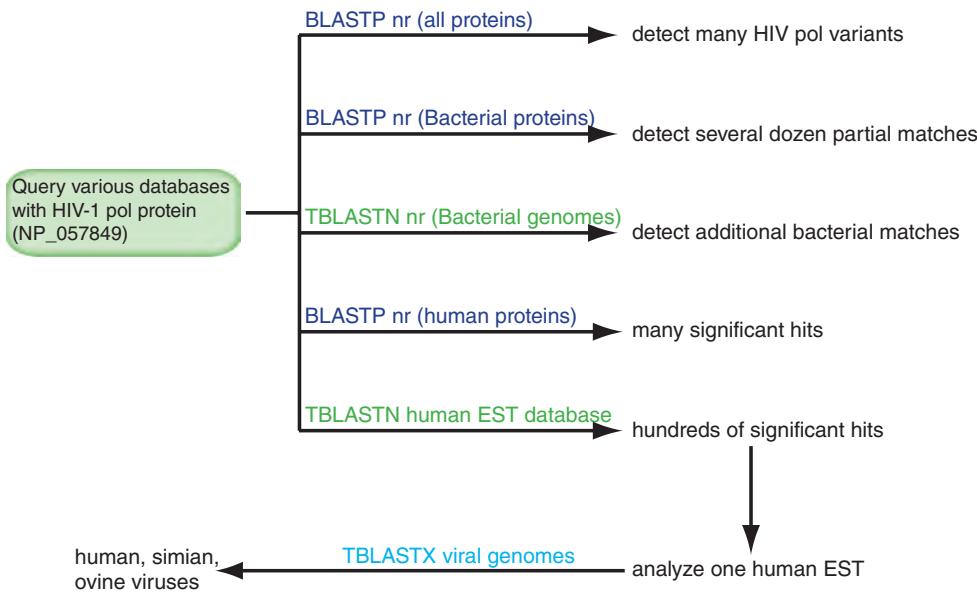
### How to Handle Too Many Results

A common situation that is encountered in BLAST searching is that too many results are returned. There are many strategies available to limit the number of results but, to make the appropriate choices, you must focus on the question you are trying to answer.

- Select a “refseq” database and all the hits that are returned will have RefSeq accession numbers. This will often eliminate redundant database matches.
- Limit the database returns by organism, when applicable. One convenient approach is to select the taxonomy identifier (txid) of interest. This may eliminate extraneous information. If you use the options feature of the BLAST server to limit a search by organism, the same size search is performed; in contrast, if you choose an organism-specific database, this may increase the speed of the search. (We present some organism-specific BLAST servers in Chapter 5.) You can use the “exclude” feature (Fig. 4.1) to ignore matches from an organism or group of interest.
- Use just a portion of the query sequence, when appropriate. A search of a multidomain protein can be performed with just the isolated domain sequence. If you are studying HIV-1 Pol, you may be interested in the entire protein or in a specific portion such as the reverse transcriptase domain.
- Adjust the scoring matrix to make it more appropriate to the degree of similarity between your query and the database matches.
- Adjust the expect value; lowering *E* reduces the number of database matches that are returned.

### How to Handle Too Few Results

Many genes and proteins have no significant database matches or have very few. As new microbial and viral genomes are sequenced, half the predicted proteins may have no matches to any other proteins (Chapters 16 and 17). Some strategies to increase the number of database matches from BLAST searching are obvious: remove Entrez limits, raise the expect values, and try scoring matrices with higher PAM or lower BLOSUM values. A large variety of additional databases can also be searched. Within the NCBI website, all available databases (e.g., HTGS and GSS) can be searched. Many genome-sequencing centers for a variety of organisms maintain separate databases that can be searched by BLAST. These are described in Chapter 5 (advanced BLAST searching). Additionally, there are many database-searching algorithms that are more sensitive than BLAST.



**FIGURE 4.18** Overview of BLAST searches beginning with HIV-1 Pol protein. A series of BLAST searches can often be performed to pursue questions about a particular gene, protein, or organism. The number of database matches returned by a BLAST search can vary from none to thousands and depends on the nature of the query, the database, and the search parameters.

These include position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs), and are also described in Chapter 5.

### BLAST Searching with Multidomain Protein: HIV-1 Pol

The Gag-Pol protein of HIV-1 (NP\_057849.4) has 1435 amino acid residues and includes separate protease, reverse transcriptase, and integrase domains; it is therefore an example of a multidomain protein. **Figure 4.18** previews the kinds of searches we can perform with a viral protein such as this one.

What happens upon BLASTP searching the nonredundant protein database with this protein? We input the RefSeq accession number (NP\_057849) and click submit. The program reports that putative conserved domains have been detected, and a schematic of the protein indicates the location of each domain (**Fig. 4.19a**). Clicking on any of these domains links to the NCBI Conserved Domain Database as well as to the Pfam and SMART databases (see Chapters 6 and 12). Continuing with the BLAST search, we see that there are many hits, all with extremely low expect values, and all correspond to HIV matches from various isolates. Reformatting the output to “query-anchored with dots for identities” is one way to view the dramatic conservation of these viral proteins (**Fig. 4.19b**). This view also highlights particular amino acid substitutions that are empirically observed, such as five arginine residue positions that are perfectly conserved, one arginine that is substituted with lysine in a half dozen instances, and one arginine that is rarely substituted by glutamine (**Fig. 4.19b**, arrows). Such position-specific differences in substitution frequencies reflect selective evolutionary pressures and are the basis of PSI-BLAST and DELTA-BLAST approaches (Chapter 5).

These highly conserved HIV-1 variants of Gag-Pol obscure our ability to evaluate non-HIV-1 matches. We can repeat the BLASTP search, setting the database to RefSeq proteins. Now Gag-Pol orthologs are evident across a variety of virus species. Clicking Taxonomy Reports from the main BLASTP search result page shows that, surprisingly,

there are even some homologs in the wild boar *Sus scrofa*, the rust-red flour beetle *Tribolium castaneum*, and a group of fungi (**Fig. 4.20**).

To learn more about the distribution of Pol proteins throughout the tree of life, we may further ask what bacterial proteins are related to the viral HIV-1 Pol polyprotein. Repeat the BLASTP search with NP\_057849 as the query, but limit the search to “Bacteria” (txid2[Organism]). Here, the graphical overview of BLAST search results is extremely helpful to show that two domains of viral Pol have the majority of matches to known bacterial sequences, corresponding to amino acids 500–800, 1000–1150, and 1200–1300 of Pol (**Fig. 4.21**). Comparison of this output to the domain architecture of HIV-1 Pol (**Fig. 4.19a**) suggests that in particular the ribonuclease H and integrase core domains of HIV-1 match many dozens of bacterial proteins. You can inspect pairwise alignments to confirm that the viral and bacterial proteins are homologous, often sharing about 30% amino acid identity over spans of over 150 amino acids.

Let us now turn our attention to human proteins that may be homologous to HIV-1 Pol. The BLASTP search is identical to our search of bacteria, except that we restrict the organism to *Homo sapiens* nonredundant proteins. Interestingly, there are many human matches (**Fig. 4.22a**), and a number of these span the majority of the viral protein. These human proteins have been annotated as gag-ropvirus-pol-env proteins, polymerases, endogenous retrovirus proteins, reverse transcriptases, and cellular nucleic acid-binding proteins.

Are these human genes expressed? If so, they should produce RNA transcripts that may be characterized as ESTs from cDNA libraries. Perform a search of human ESTs with the viral Pol protein; it is necessary to use the translating BLAST website with the TBLASTN algorithm, and the database must be set to EST while the organism is restricted to human. There are hundreds of human transcripts, actively transcribed, that are predicted to encode proteins homologous to viral Pol (**Fig. 4.22b**). These correspond to three regions of HIV-1 Gag-Pol. In Chapter 10, we see how to evaluate these human ESTs to determine where in the body they are expressed and when during development they are expressed.

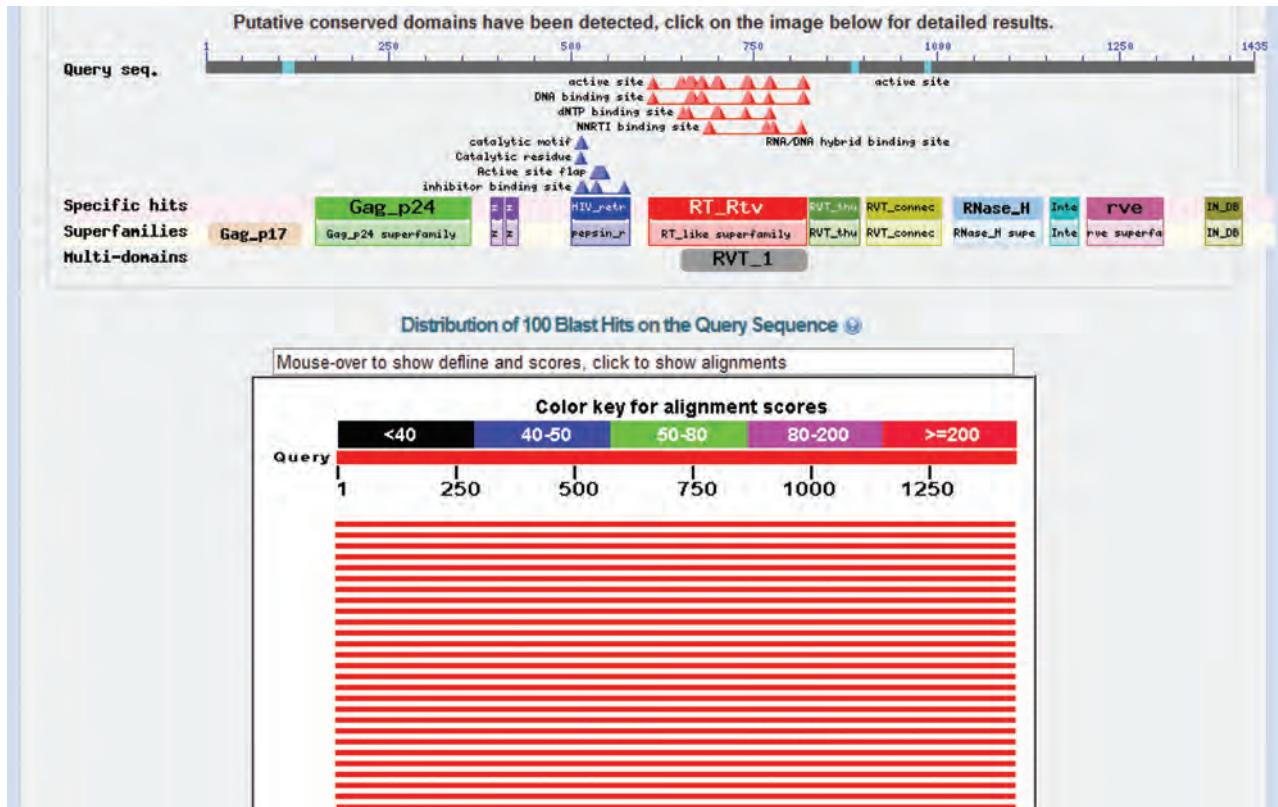
Could the human ESTs that are homologous to HIV-1 Pol be even more closely related to other viral Pol genes? To answer this question, select a human EST that we found to be related to HIV-1 Pol (from **Fig. 4.22b**; we will choose accession BX509809.1 because it has the lowest *E* value,  $5 \times 10^{-29}$ ). Perform a TBLASTX search using this EST’s accession as an input and restrict the database to refseq\_genomic and

---

**FIGURE 4.19** A BLASTP search with HIV-1 viral Pol (NP\_057849). (a) Graphical overview shows conserved domains in the protein. These blocks are clickable and link to the Conserved Domain Database at NCBI (Chapters 5 and 6). The links are to protein domains (Gag\_p17, Gag\_p24) and abbreviations include rvp, retroviral aspartyl protease; rvt, reverse transcriptase (RNA-dependent DNA polymerase); rnaseH, ribonuclease H; rve, integrase core domain. The red horizontal bars indicate many close matches to viral proteins. (b) The BLAST alignment options include formats such as query-anchored, in which dots correspond to residues in database entries that match the query. This view highlights the occasional sequence differences in viral proteins. Arrows indicate arginine (R) positions in the query that are perfectly conserved, or that are sometimes substituted with lysine (K) or glutamine (Q).

Source: BLASTP, NCBI.

(a) Graphical overview



(b) List of alignments (query-anchored with dots for identities)

Query	1	MGARASVLSGGELDRWEKIRLRPGGKKYKLKHIVWASRELERFAVNPGLETSEGCRQI	60
NP_057849	1	.....	60
P0C6F2	1	.....K.....	60
P03366	1	.....	60
P03367	1	.....	60
P04587	1	.....	60
AAD03191	1	.....Q.R.....	60
P35963	1	....A....K.....Q.R.....D.....	60
P12497	1	.....K.....Q.....	60
P20875	1	.....R.....R.....Q.....S.....	60
AAD03200	1	.....R.....R.....Q.....S.....	60
P20892	1	.....K.....Q.....I.....	60
Q73368	1	.....S.....	60
BAB85751	1	.....Q.....M.....	60
AFB39387	1	.....Q.....R.....A.....	60
P03369	1	.....K.....	60
P05959	1	.....K.K.....R.R.....S.A.....	60
AAG30116	1	....I.....K.....R.L.....Q.I.....A.....	60
AAD03217	1	....I.....Q.....	60

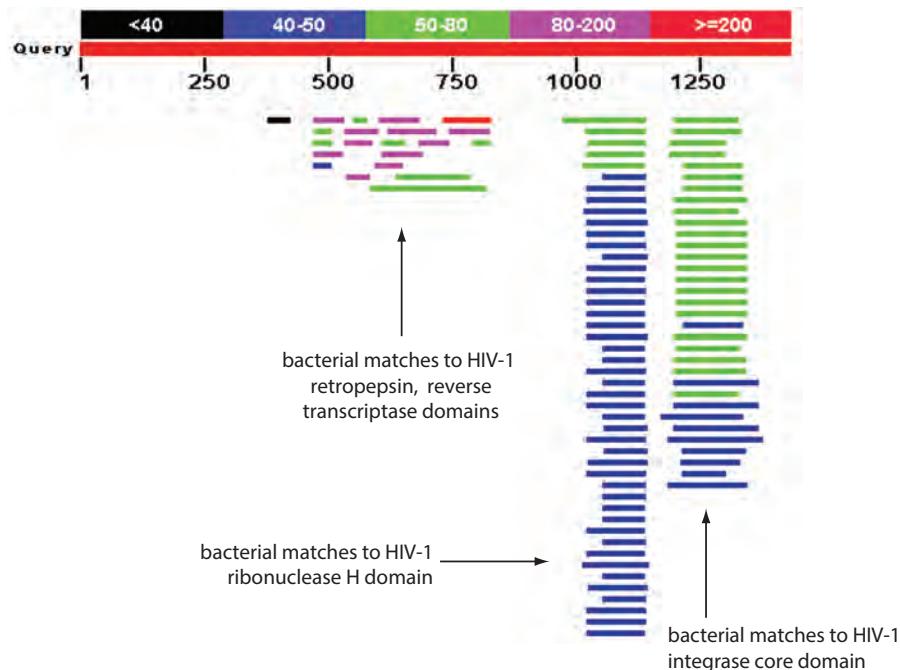
Annotations below the alignment show anchor points: R (red arrow), R,K (blue arrow), R,R (red double-headed arrow), R (red arrow), R,Q (green arrow), and R (red arrow).

<b>Human immunodeficiency virus 1 [viruses]</b> taxid 11676			
ref NP_057849.4  Gag-Pol [Human immunodeficiency virus 1]	2971	0.0	
ref NP_789740.1  Pol [Human immunodeficiency virus 1]	2052	0.0	
ref NP_705927.1  reverse transcriptase [Human immunodefici...]	1149	0.0	
ref YP_001856242.1  reverse transcriptase [Human immunodef...	1149	0.0	
ref NP_789739.1  reverse transcriptase p51 subunit [Human ...]	912	0.0	
ref NP_057850.1  Pr55(Gag) [Human immunodeficiency virus 1]	908	0.0	
ref NP_705928.1  integrase [Human immunodeficiency virus 1]	602	0.0	
ref YP_001856243.1  integrase [Human immunodeficiency viru...]	602	0.0	
ref NP_579880.1  capsid [Human immunodeficiency virus 1]	481	4e-156	
ref NP_579876.2  matrix [Human immunodeficiency virus 1]	271	7e-81	
ref NP_705926.1  retropepsin [Human immunodeficiency virus 1]	204	2e-57	
ref YP_001856241.1  retropepsin [Human immunodeficiency vi...]	204	2e-57	
ref NP_579881.1  nucleocapsid [Human immunodeficiency viru...]	130	5e-32	
ref NP_787043.1  Gag-Pol Transframe peptide [Human immunod...	119	4e-28	
<b>Simian immunodeficiency virus [viruses]</b> taxid 11723			
ref NP_687035.1  Gag-Pol [Simian immunodeficiency virus]	1687	0.0	
ref NP_054369.1  gag protein [Simian immunodeficiency virus]	502	1e-159	
<b>Human immunodeficiency virus 2 [viruses]</b> taxid 11709			
ref NP_663784.1  gag-pol fusion polyprotein [Human immunod...	1675	0.0	
ref NP_056837.1  gag polyprotein [Human immunodeficiency v...]	523	3e-167	
<b>Simian immunodeficiency virus SIV-mnd 2 [viruses]</b> taxid 159122			
ref NP_758887.1  pol protein [Simian immunodeficiency viru...]	1377	0.0	
ref NP_758886.1  gag protein [Simian immunodeficiency viru...]	486	2e-153	
<b>Feline immunodeficiency virus [viruses]</b> taxid 11673			
ref NP_040973.1  pol polyprotein [Feline immunodeficiency ...]	489	2e-148	
ref NP_040972.1  gag protein [Feline immunodeficiency virus]	158	8e-38	
<b>Equine infectious anemia virus [viruses]</b> taxid 11665			
ref NP_056902.1  pol polyprotein [Equine infectious anemia...]	424	1e-123	
ref NP_056901.1  gag protein [Equine infectious anemia virus]	154	2e-36	
<i>///</i>			
<b>Candida albicans SC5314 [ascomycetes]</b> taxid 237561			
ref XP_888860.1  hypothetical protein Ca019_6468 [Candida ...]	90	2e-15	
ref XP_721310.1  hypothetical protein Ca019.6468 [Candida ...]	86	1e-14	
<b>Sus scrofa</b> (wild boar, ...) [even-toed ungulates] taxid 9823			
ref XP_003482346.1  PREDICTED: hypothetical protein LOC100...	90	2e-15	
<b>Tribolium castaneum</b> (rust-red flour beetle) [beetles] taxid 7070			
ref XP_001815322.1  PREDICTED: similar to orf [Tribolium c...]	89	5e-15	
ref XP_001808495.1  PREDICTED: similar to orf [Tribolium c...]	88	8e-15	
<b>Candida dubliniensis CD36</b> [ascomycetes] taxid 573826			
ref XP_002421195.1  retrovirus-related Pol polyprotein fro...	88	6e-15	
<b>Moniliophthora perniciosa FA553</b> [basidiomycetes] taxid 554373			
ref XP_002387985.1  hypothetical protein MPER_13056 [Monil...]	88	7e-15	

**FIGURE 4.20** The taxonomy report for a BLASTP search shows an overview of which species have proteins matching the HIV-1 query. Most matches are viral, but others include rabbit, fungal, pig, and insect sequences. The *///* symbols indicate a series of other matches (not shown).

Source: BLASTP, NCBI.

the organism of the search to viruses. At the present time, this search results in the identification of viruses having significant but limited relatedness to our query (e.g., koala retrovirus, banana streak virus). We initially performed a BLAST search with an HIV query and have used a further series of BLAST searches to gain insight into the biology of HIV-1 Pol.



**FIGURE 4.21** Result of a BLASTP search with HIV-1 Pol as a query, restricting the output to bacteria. The graphical output of the BLAST search allows identification of the domains within HIV-1 that have bacterial matches. The length of overlap and the number of bacterial sequences are also evident.

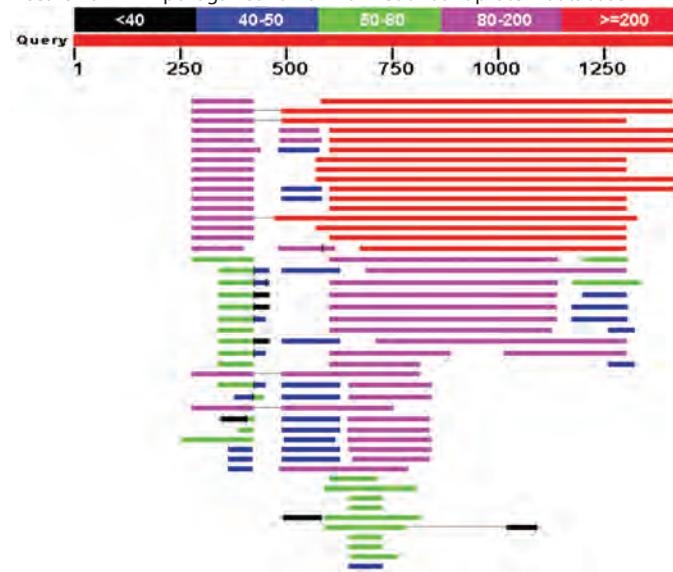
Source: BLASTP, NCBI.

## USING BLAST FOR GENE DISCOVERY: FIND-A-GENE

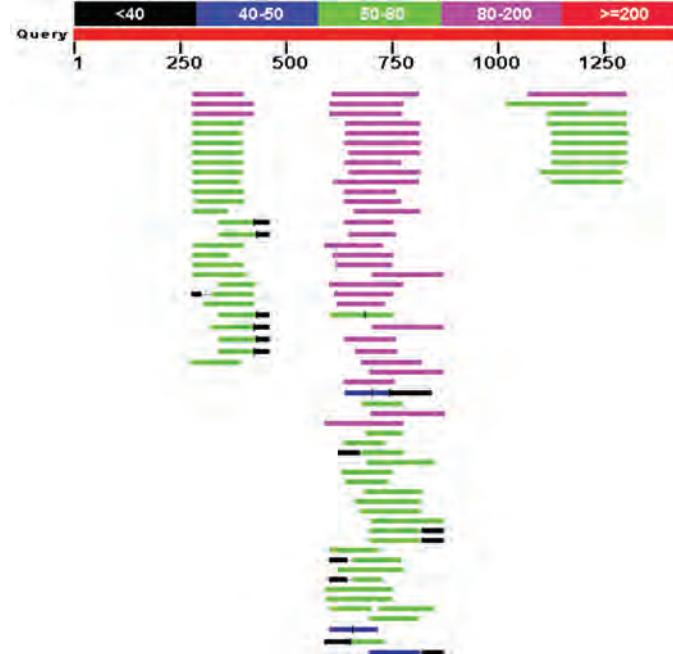
A common problem in biology is finding a new gene. Traditionally, genes and proteins were identified using the techniques of molecular biology and biochemistry. Complementary DNAs were cloned from libraries, or proteins were purified then sequenced based upon some biochemical criteria such as enzymatic activity. Such experimental biology approaches will always remain essential. Bioinformatics approaches can also be useful to provide evidence for the existence of new genes. For our purposes a “new” gene refers to the discovery of some DNA sequence in a database that is not annotated (described). You may want to find new genes for many reasons:

- You want to study a globin or lipocalin that no one has characterized before, perhaps in a specific organism of interest such as a plant or archaeon.
- You are interested in the lipocalins, and you see that one has been described in the tears of hamsters. Could there be a new, undiscovered gene that encodes a lipocalin protein expressed in human tears? (At present, there is one!)
- You want to know if viruses have globins or lipocalins. If so, this might give you insight into the evolution of these families of carrier proteins.
- You study diseases in which sugars are not processed properly and, as part of this research, you study sugar transport in cell lines from some organism. You know that glucose transporters have been characterized by biochemical assays (e.g., sugar uptake). You also know that there is a family of glucose transporter genes (and proteins) that have been deposited in GenBank. You cloned all the known transporters, expressed them in cells, and found that none of the recombinant proteins transports your sugar. You hypothesize that there must be at least one more transporter that has not yet been described. Is there a way to search the database to find genes encoding novel transporters?

(a) BLASTP search of HIV-1 pol against human non-redundant protein database



(b) TBLASTN search of HIV-1 pol against human expressed sequence tags



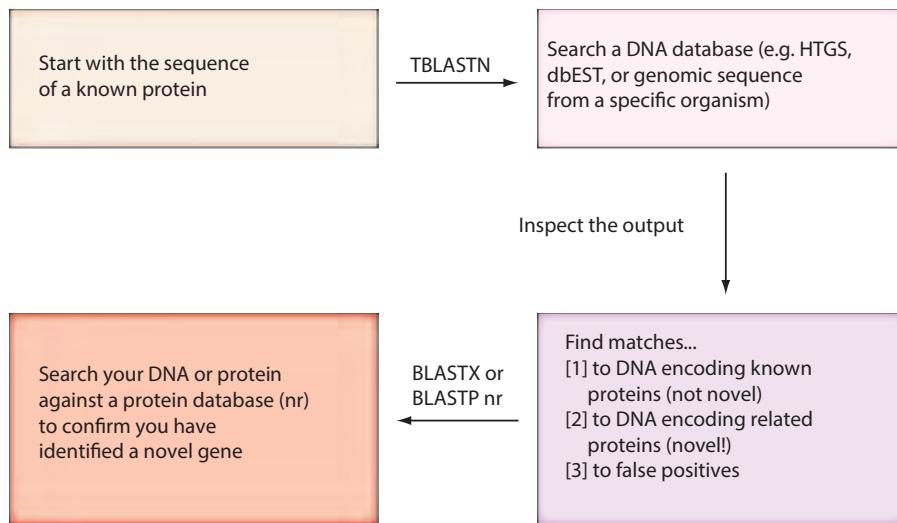
**FIGURE 4.22** (a) Graphical output of a BLASTP search using HIV-1 Pol protein to search for matches against human proteins. Note that some human hits have very high scores. (b) Are human transcripts expressed that encode proteins homologous to HIV-1 Pol protein? The results of a TBLASTN search with viral Pol protein against a human EST database are shown. Many human genes are actively transcribed to generate transcripts predicted to make proteins homologous to HIV-1 Pol.

Source: BLASTP, NCBI.

The “find-a-gene project” is summarized at Web Document 4.5. The beta globin “find-a-gene project” described here is available as Web Document 4.6 (<http://www.bioinfbook.org/chapter4/>).

A general strategy to solve any of these problems is presented in **Figure 4.23**. I have called this the “find-a-gene” project and have used it as a teaching exercise since the year 2000. All of the hundreds of students who attempted it completed it successfully. Each student summarizes the results in a word document. The steps are as follows.

1. Choose the name of a favorite protein you are interested in. Include the species and the accession number. As an example (below), we will select human beta globin and search for a novel globin gene.



**FIGURE 4.23** How to discover a novel gene by BLAST searching. Begin with the sequence of a known protein such as human beta globin. Perform a TBLASTN search of a DNA database. It is unlikely that there are many “novel” genes in the well-characterized genomes of organisms such as human, yeast, or *E. coli*. It may therefore be helpful to search databases of organisms that are poorly characterized or not fully annotated. The TBLASTN search may result in two types of significant matches: (1) matches of your query to known proteins that are already annotated; and (2) homologous proteins that have not yet been annotated (“novel” genes and corresponding novel proteins). (3) The DNA sequence corresponding to the putative novel gene may be searched using the BLASTX algorithm against the nonredundant (nr) database. This may confirm that the DNA does indeed encode a protein that has no perfect match to any described protein.

2. Perform a TBLASTN search against a DNA database consisting of genomic DNA or expressed sequence tags (ESTs). The BLAST server can be at NCBI or elsewhere. Include the output of that BLAST search in your document.

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. It is important to be able to inspect the pairwise alignment you have selected, including the *E* value and score. In general, this step is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e., a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of step (4) below), and a nonhomologous result.

As an example of finding a “novel” globin, use beta globin (NP\_000509) as a query and perform a TBLASTN search of ESTs restricted to nematodes. We introduce ESTs in Chapter 10; they are short fragments of DNA (typically up to 800 base pairs) corresponding to genes that have been expressed in a particular organism in some region and at some time of development. For example, libraries of ESTs are available from human fetal liver or adult mouse brain. By restricting the output to nematode ESTs, we find a match with a significant *E* value (Fig. 4.24a; accession JK511422.1, a 559 base pair clone from *Anguillicola crassus*). This nematode EST encodes a protein that shares 47% amino acid identity with human beta globin, with a convincing *E* value of  $6 \times 10^{-44}$ . Is this nematode protein “novel” in the sense that it has never been annotated as a globin? Follow the link to the nematode accession, choose BLAST, and perform a BLASTX search against the nonredundant (NR) database. The best match is not to any nematode protein, confirming

If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output. It is not necessary to print out all of the BLAST results if there are many pages.

(a) Result of TBLASTN against nematode ESTs using human beta globin as a query

Ac_EH1r_01A07_M13 Adult <i>Anguillilcola crassus</i> <i>Anguillilcola crassus</i> cDNA clone Ac_EH1r_01A07					
Sequence ID: <a href="#">gb JK511422.1</a> Length: 559 Number of Matches: 1					
Range 1: 40 to 483 GenBank Graphics					
Score	Expect	Method	Identities	Positives	Gaps
149 bits(375)	6e-44	Compositional matrix adjust.	69/148(47%)	97/148(65%)	1/148(0%) +1
Query 1	MVHLTPEEKSAVTALNGKVNDEVGGEALGRLLVVYPWTQRFESFGDLSPTDAVMGNPK	60			
MV	T	+E+	+LW	K+IV+E+G	+A+RLL+V
Sbjct 40	MVEWTDAEHTAILSLWKKINVVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAIMNNEK	219			
Query 61	VKAHGKKVLGAFSDGLAHLNDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAAHHFG	120			
V	HG	V+G	++D+K	+LS	+H +KLHVDP+NFRL
Sbjct 220	VAKHGTTVMGGGLDRAIQNMDDIKNAYRELSEKLVHDPDNFRLLSEHITLCMAAKFG	399			
Query 121	-KEFTPVQAYQKVVAAGVANALAHKYH	147			
EFT	VQ	A+QK	+V	+AI	+YH
Sbjct 400	TEFTADVQEAWQKFLMAVTISALGRQYH	483			

(b) BLASTX result with a nematode EST showing its closest known protein match is in a vertebrate

RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain  
Sequence ID: [sp|P80946\\_1|HBBA\\_ANGAN](#) Length: 147 Number of Matches: 1

Range 1: 1 to 147 GenPept Graphics					
Score	Expect	Method	Identities	Positives	Gaps
290 bits(742)	2e-97	Compositional matrix adjust.	136/147(93%)	141/147(95%)	0/147(0%) +1
Query 43	VEWIDAETAILSWKINVVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAIMNNEK	222			
VEW+I	E	TAI	S	W	KIN+EEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAIMNNEK+KV
Sbjct 1	VENTEDERTAIKSWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAIMNNDKV	60			
Query 223	AKHGTTVMGGGLDRAIQNMDDIKNAYRELSEKLVHDPDNFRLLSEHITLCMAAKFGP	402			
AKHG	TTVMGGGLDRAI	QNMDDIKNA	YR+LSV	MSEKLHVDPDNFRLL+ERITLCMAAKFGP	
Sbjct 61	AKHGTTVMGGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFGP	120			
Query 403	TEFTADVQEAWQKFLMAVTISALGRQYH	483			
TEFTADVQEAWQKFLMAVTISALGRQYH					
Sbjct 121	TEFTADVQEAWQKFLMAVTISALARQYH	147			

**FIGURE 4.24** The find-a-gene project was demonstrated using human beta globin (NP\_000509) as a query and searching a database of expressed sequence tags (ESTs) restricted to nematodes. (a) The matches included one to an EST from *Anguillilcola crassus* (GenBank accession JK511422.1). (b) Using this accession as a query, a BLASTX nr search revealed matches to known beta globins. The best match, shown here, was to a vertebrate globin. However, since there was not a match to an *A. crassus* globin, this suggests that the find-a-gene project resulted in the identification of a DNA sequence that encodes a previously undescribed nematode globin. This novel globin can then be characterized in terms of its full-length sequence, homologs, evolution, structure, and function.

Source: NCBI.

that this is indeed novel, but instead it is to a globin from the European eel (Fig. 4.24b), confirming that this is a globin homolog.

3. Gather information about this “novel” protein. At a minimum, identify the protein sequence of the “novel” protein as displayed in the BLAST results from step (2). In some cases, you will be able to perform further BLAST searches to obtain even more sequence of your novel gene.

Propose a name for the novel protein (e.g., “*Anguillilcola* globin”), and report the species from which it derives. It is very unlikely (but still possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human, or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or mosses or protozoa.

4. Demonstrate that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Use the DNA sequence of the EST and perform a BLASTX query against the nonredundant (nr) database

as described above. As an alternative strategy, take the encoded protein sequence (step (3)), and use it as a query in a BLASTP search of the nonredundant (nr) database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the best match is to a protein with less than 100% identity to your query, then it is likely that your protein is novel and you have succeeded.
- If there is a match with 100% identity but to a different species than the one you started with, then you have succeeded in finding a novel gene.
- If there are no database matches to the original query from step (1), this indicates that you have found a DNA/protein that is not homologous to the original query. You should start over.

There are several further steps for this project, involving themes we will cover in later chapters.

5. Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a reasonable maximum is 30. We describe multiple sequence alignment in Chapter 6.
6. Create a phylogenetic tree using a method such as neighbor-joining, maximum parsimony, maximum likelihood, or Bayesian inference (see Chapter 7). Bootstrapping and tree rooting are optional. Use any program such as MEGA, Phylip, or MrBayes.
7. Predict the secondary and tertiary structure of your novel protein (see Chapter 13), and compare it to that of a known structure.
8. Determine whether this gene is under positive or negative evolutionary selection (see Chapter 7).
9. Discuss the significance of your novel gene. What have you learned about this gene/protein family?

The main benefits of the find-a-gene project as a teaching tool are: (1) it requires that you know when and how to use the main family of BLAST programs (e.g., TBLASTN, BLASTX); (2) it allows you to become familiar with a variety of searchable databases (e.g., EST, genomic DNA, and nonredundant); and (3) it requires you to interpret different kinds of BLAST output. For many initial TBLASTN searches with a protein query of interest, it is easy to find “novel” genes; for some cases it is not easy to find new genes, perhaps because relevant homologs do not exist or because the appropriate database is not searched. Begin again with a different protein query.

## PERSPECTIVE

BLAST searching has emerged as an indispensable tool to analyze the relation of a DNA or protein sequence to millions or even trillions of sequences in public databases. All database search tools confront the issues of sensitivity (i.e., the ability to minimize false negative results), selectivity (i.e., the ability to minimize false positive results), and time. As the size of the public databases has grown exponentially in recent years, the BLAST tools have evolved to provide a rapid, reliable way to screen the databases. For protein searches we have focused on BLASTP. However, for most biologists performing even routine searches with a protein query, the DELTA-BLAST or HMMER programs described in Chapter 5 are strongly preferred. This is because of their more optimally constructed scoring matrices.

## PITFALLS

There are several common pitfalls to avoid in BLAST searching. The most common error among novice BLAST users is to search protein or DNA sequences against the wrong database. It is also important to understand the basic BLAST algorithms. These concepts are summarized in **Figure 4.2**.

An important issue in BLAST searching is deciding whether an alignment is significant. Each potential BLAST match should be compared to the query sequence to evaluate whether it is reasonable from both a statistical and a biological point of view. It is more likely that two proteins are homologous if they share similar domain architecture (i.e., motifs or domains; Chapter 12) or other common features.

## ADVICE FOR STUDENTS

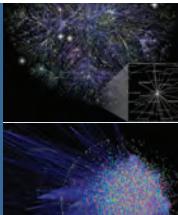
BLAST searches are quick and easy. Practice doing many, many searches. Explore all the optional parameters and read the NCBI or other documentation to learn what they do.

If you have not used Linux before, find a Linux machine and install BLAST+. (Mac O/S and Windows machines also work.) Its installation is very well documented. Become comfortable downloading a database (it's easy with the Perl script described above, although it is possible to use the `wget` command in Linux or to perform downloads on a PC or Mac). Use EDirect (as described above) to obtain your query or set of queries in the FASTA format (you can also use an editor such as `nano` or `vim`). Then perform a series of searches on the command line. Once you can use one program in Linux, it becomes much easier to access many other programs. If you have questions, try searching Biostars (<http://www.biostars.org>) for related questions and answers; if needed, post your own question.

## WEB RESOURCES

The main website for BLAST searching is that of the National Center for Biotechnology Information (<http://blast.ncbi.nlm.nih.gov/>). Within this site are links to the main programs (BLASTN, BLASTP, BLASTX, TBLASTN, and TBLASTX). There are other specialized BLAST programs at NCBI that are discussed in Chapter 5.

An important web resource is the set of BLAST tutorials, courses, and references available at the NCBI BLAST site. A practical BLAST tutorial is offered at <http://www.ncbi.nlm.nih.gov/books/NBK1734/> (WebLink 4.10).



## Discussion Questions

**[4-1]** Why doesn't anyone offer "Basic Global Alignment Search Tool" (BGAST) to complement BLAST? Would BGAST be a useful tool? What computational difficulties might there be in setting it up? (Note that some tools do combine global and local search strategies. We describe HMMER software in Chapter 5, and another example is USEARCH from Robert Edgar at <http://www.drive5.com>.)

**[4-2]** Should you consider a significant expect value to be 1, 0.05, or  $10^{-5}$ ? Does this depend on the particular search you are doing?

**[4-3]** Why is it that database programs such as BLAST must make a trade-off between sensitivity and selectivity? How does the BLASTP algorithm address this issue?

### PROBLEMS/COMPUTER LAB

**[4-1]** In this problem we explore the effect of a short protein query on the BLASTP parameters. Perform a BLASTP search at NCBI using the following query of just 12 amino acids: PNLHGLFGRKTG. By default, the parameters are adjusted for short queries. Inspect the output. What is the *E* value cutoff? What is the word size? What is the scoring matrix? How do these settings compare to the default parameters?

**[4-2]** Protein searches are usually more informative than DNA searches. Perform a BLASTP search using RBP4 (NP\_006735), restricting the output to Arthropoda (insects). Next, carry out a BLASTN search using the RBP4 nucleotide sequence (NM\_006744). For this query, select only the nucleotides corresponding to the coding region of the DNA. (To do this visit the NCBI Nucleotide page, follow the link to the coding sequence (CDS), then choose the FASTA format.) Which search is more informative? How many database matches have an *E* value less than 1.0 in each search?

**[4.3]** This problem introduces batch queries. It is possible to search many queries simultaneously, either using the web-based BLAST (as in this problem) or via locally installed BLAST+. Mosses are plants of the phylum Bryophyta, including the non-seed plant *Physcomitrella patens* that had its genome sequenced (Rensing *et al.*, 2008). Do mosses have any globin proteins and, if so, which human globin(s) are they most closely related to? (1) First obtain the accession numbers of all human globins. There are several approaches to doing this, including BLASTP using beta globin and neuroglobin as queries. Other approaches involve DELTA-BLAST (Chapter 5) or Pfam (Chapter 6). These accession numbers are provided in Web Document 4.7. (2) Perform a BLASTP search using all accession numbers as queries, entering them into the query box. Restrict the output to RefSeq proteins of the mosses. (3) Results for each query are shown (one at a time) via a pull-down menu. Currently there are significant, although distant, matches of all human globins to moss proteins except for hemoglobin subunit mu. (See for example the match between human epsilon globin and predicted moss protein XP\_001786089.1 with an *E* value of 0.01. A BLASTP search with that moss protein confirms it is related to many annotated plant globins.) Notably, only one

human protein (neuroglobin, NP\_001030585.1) has very strong matches to moss proteins such as *P. patens* predicted protein XP\_001764902.1 (*E* value  $2 \times 10^{-10}$ , 27% identity across a span of 138 amino acid residues).

**[4.4]** Use the BLAST+ suite to run BLASTP on the command line. Start with default settings, then change the effective database size for your search making it 1000 times smaller than 1000 times larger. What are the *E* values? Explore different output formats by using the help function; for example, use `-outfmt 2` for a multiple alignment format.

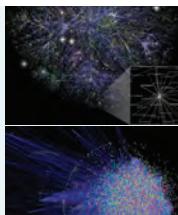
**[4.5]** BLAST+ is useful for performing batch queries. Create a text file named 3proteins.txt having three protein sequences: human beta globin, bovine odorant-binding protein, and cytochrome b from the malaria parasite *Plasmodium falciparum*. (These are available at Web Document 4.8.) Search them with BLASTP against the RefSeq protein database. The output file includes the results of three separate BLASTP searches.

**[4.6]** For the search you just performed in problem (4.5), what happens if you use a scoring matrix that is more suited to finding distantly related proteins?

**[4.7]** Is the Pol protein of HIV-1 more closely related to the Pol protein of HIV-2 or to the Pol protein of simian immunodeficiency virus (SIV)? Use the BLASTP program to decide. Hint: try the Entrez command “NOT hiv-1 [organism]” to focus the search away from HIV-1 matches.

**[4.8]** “The Iceman” is a man who lived 5300 years ago and whose body was recovered from the Italian Alps in 1991. Some fungal material was recovered from his clothing and sequenced. To what modern species is the fungal DNA most related?

**[4.9]** You perform a BLAST search and a result has an *E* value of about  $1 \times 10^{-4}$ . What does this *E* value mean? Name some parameters on which an *E* value depends.



## Self-Test Quiz

**[4-1]** You have a reasonably short, typical, double-stranded DNA sequence. Basically, how many proteins can it *potentially* encode?

- (a) 1;
- (b) 2;
- (c) 3; or
- (d) 6.

**[4-2]** You have a DNA sequence. You want to know which protein in the main protein database (“nr,” the nonredundant database) is most similar to some protein encoded by your DNA. Which program should you use?

- (a) BLASTN;
- (b) BLASTP;
- (c) BLASTX;
- (d) TBLASTN; or
- (e) TBLASTX.

**[4-3]** Which output from a BLAST search provides an estimate of the number of false positives from a BLAST search?

- (a) *E* value;
- (b) bit score;
- (c) percent identity; or
- (d) percent positives.

**[4-4]** Match up the following BLAST search programs with their correct descriptions:

- |         |  |
|---------|--|
| BLASTP  | (a) Nucleotide query against a nucleotide sequence database              |
| BLASTN  | (b) Protein query against a translated nucleotide sequence database      |
| BLASTX  | (c) Translated nucleotide query against a protein database               |
| TBLASTN | (d) Protein query against a protein database                             |
| TBLASTX | (e) Translated nucleotide query against a translated nucleotide database |

**[4-5]** Changing which of the following BLAST parameters would tend to yield fewer search results?

- (a) turning off the low-complexity filter;
- (b) changing the expect value from 1 to 10;
- (c) raising the threshold value; or
- (d) changing the scoring matrix from PAM30 to PAM70.

**[4-6]** You can limit a BLAST search using any Entrez term. For example, you can limit the results to those containing a researcher's name.

- (a) true; or
- (b) false.

**[4-7]** An extreme value distribution:

- (a) describes the distribution of scores from a query against a database;
- (b) has a larger total area than a normal distribution;
- (c) is symmetric; or
- (d) has a shape that is described by two constants:  $\mu$  (the mean) and  $\lambda$  (a decay constant).

**[4-8]** As the *E* value of a BLAST search becomes smaller:

- (a) the value  $K$  also becomes smaller;
- (b) the score tends to be larger;
- (c) the probability  $p$  tends to be larger; or
- (d) the extreme value distribution becomes less skewed.

**[4-9]** The BLAST algorithm compiles a list of “words” typically of three amino acids (for a protein search). Words at or above a threshold value  $T$  are defined as:

- (a) “hits,” and are used to scan a database for exact matches that may then be extended;
- (b) “hits,” and are used to scan a database for exact or partial matches that may then be extended;
- (c) “hits,” and are aligned to each other; or
- (d) “hits,” and are reported as raw scores.

**[4-10]** Normalized BLAST scores (also called bit scores):

- (a) are unitless;
- (b) are not related to the scoring matrix that is used;
- (c) can be compared between different BLAST searches, even if different scoring matrices are used; or
- (d) can be compared between different BLAST searches, but only if the same scoring matrices are used.

## SUGGESTED READING

BLAST searching was introduced in a classic paper by Stephen Altschul and colleagues (1990). This paper describes the theoretical basis for BLAST searching and describes basic issues of BLAST performance, including sensitivity (accuracy) and speed. Fundamental modifications to the original BLAST algorithm were later introduced, including the introduction of gapped BLAST (Altschul *et al.*, 1997). This paper includes a discussion of specialized position-specific scoring matrices that we consider in Chapter 5.

Ian Korf, Mark Yandell, and Joseph Bedell (2003) have written an excellent book called *BLAST*. A useful older description of database searching is the article entitled “Effective protein sequence comparison” by William Pearson (1996). Altschul *et al.* (1994) provide a highly recommended article, “Issues in searching molecular sequence

databases.” Marco Pagni and C. Victor Jongeneel (2001) of the Swiss Institute of Bioinformatics provide a technical overview of sequence alignment statistics. This article includes sections on the extreme value distribution, the use of random sequences, local alignment with and without gaps, and BLAST statistics. See also a review of alignment statistics was written by Stephen Altschul and Warren Gish (1996).

NCBI offers online books including “BLAST® Help” at <http://www.ncbi.nlm.nih.gov/books/NBK1762/> (WebLink 4.6). “The Statistics of Sequence Similarity Scores” from the help section of the NCBI BLAST site provides an excellent resource (<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>, WebLink 4.9).

## REFERENCES

- Altschul, S. F., Gish, W. 1996. Local alignment statistics. *Methods in Enzymology* **266**, 460–480.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.
- Altschul, S. F., Boguski, M. S., Gish, W., Wootton, J. C. 1994. Issues in searching molecular sequence databases. *Nature Genetics* **6**, 119–129.
- Altschul, S. F., Madden, T.L., Schäffer, A.A. *et al.* 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.
- Altschul, S.F., Wootton, J.C., Gertz, E.M. *et al.* 2005. Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal* **272**, 5101–5109.
- Altschul, S., Demchak, B., Durbin, R. *et al.* 2013. The anatomy of successful computational biology software. *Nature Biotechnology* **31**(10), 894–789. PMID: 24104757.
- Berman, P., Zhang, Z., Wolf, Y.I., Koonin, E.V., Miller, W. 2000. Winnowing sequences from a database search. *Journal of Computational Biology* **7**(1–2), 293–302. PMID: 10890403.
- Boratyn, G.M., Camacho, C., Cooper P.S. *et al.* 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Research* **41**(Web Server issue), W29–33. PMID: 23609542.
- Brenner, S. E. 1998. Practical database searching. *Bioinformatics: A Trends Guide* **1998**, 9–12.
- Camacho, C., Coulouris, G., Avagyan, V. *et al.* 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). PMID: 20003500.
- Colbourne, J.K., Pfrender, M.E., Gilbert, D. *et al.* 2011. The ecoresponsive genome of Daphnia pulex. *Science* **331**(6017), 555–561. PMID: 21292972.
- Dixon, R. A., Kobilka, B.K., Strader, D.J. *et al.* 1986. Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin. *Nature* **321**, 75–79.
- Downward, J., Yarden, Y., Mayes, E. *et al.* 1984. Close similarity of epidermal growth factor receptor and v-erb-B oncogene protein sequences. *Nature* **307**, 521–527.
- Ermolaeva, M. D., White, O., Salzberg, S. L. 2001. Prediction of operons in microbial genomes. *Nucleic Acids Research* **29**, 1216–1221.
- Ferretti, J. J., McShan, W.M., Ajdic, D. *et al.* 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Science, USA* **98**, 4658–4663.
- Gotoh, O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology* **264**, 823–838.
- Gribskov, M., Robinson, N.L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computational Chemistry* **20**, 25–33.
- Gumbel, E. J. 1958. *Statistics of Extremes*. Columbia University Press, New York.
- Huang, Y., Li, Y., Burt, D.W. *et al.* 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genetics* **45**(7), 776–783. PMID: 23749191.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

- Karlin, S., Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Science, USA* **87**, 2264–2268.
- Kaufman, K.M., Sodetz, J.M. 1994. Genomic structure of the human complement protein C8 gamma: homology to the lipocalin gene family. *Biochemistry* **33**(17), 5162–5166. PMID: 8172891.
- Korf I., Yandell M., Bedell, J. 2003. *BLAST*. O'Reilly Media, Sebastopol, CA.
- Needleman, S. B., Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.
- Pagni, M., Jongeneel, C. V. 2001. Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics* **2**, 51–67.
- Park, J., Karplus, K., Barrett, C. *et al.* 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology* **284**, 1201–1210.
- Pearson, W. R. 1996. Effective protein sequence comparison. *Methods in Enzymology* **266**, 227–258.
- Rensing S.A., Lang, D., Zimmer, A.D. *et al.* 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**(5859), 64–69. PMID: 18079367.
- Schäffer, A.A., Aravind, L., Madden, T.L. *et al.* 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**, 2994–3005.
- Smith, T. F., Waterman, M. S. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197.
- Wootton, J. C., Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology* **266**, 554–571.
- Yu, Y.-K., Altschul, S.F. 2005. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics* **21**, 902–911.
- Yu, Y.-K., Wootton, J.C., Altschul, S.F. 2003. The compositional adjustment of amino acid substitution matrices. *Proceedings of the National Academy of Science, USA* **100**, 15688–15693.



NOTE 8. *Ultimate composition of fibrin from ox-blood.* (Mulder.)

		Atoms.	Calculated.
Carbon	.	54·56	400
Hydrogen	.	6·90	310
Nitrogen	.	15·72	50
Oxygen	.	22·13	120
Phosphorus	.	0·33	1
Sulphur	.	0·36	1

Hence, in its composition, it is identical with the albumen of eggs.

NOTE 9. *Ultimate composition of casein from cows' milk.* (Mulder.)

		Atoms.	Calculated.
Carbon	.	54·96	400
Hydrogen	.	7·15	310
Nitrogen	.	15·80	50
Oxygen	.	21·73	120
Sulphur	.	0·36	1

NOTE 10. *Ultimate composition of crystallin from the eye.* (Mulder.)

Carbon	.	55·39	
Hydrogen	.	6·94	
Nitrogen	.	16·51	hence it closely resembles casein.
Oxygen	.	20·91	
Sulphur	.	0·25	

NOTE 11. *Ultimate composition of globulin.*

The analysis referred to in the text was published by Mulder in the 'Bulletin' for 1839. In his recent work on the 'Chemistry of Animal and Vegetable Physiology,' he states that, although a protein-compound, its real composition is not yet known.

NOTE 12. *Ultimate composition of pepsin.* (Vogel.)

Carbon	.	57·718	
Hydrogen	.	5·666	
Nitrogen	.	21·088	
Oxygen	.	16·064	

Gerardus Johannes Mulder and other biochemists in the 1830s and 1840s hypothesized that all proteins had the same composition of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. Simon (1846, vol. 2, p. 505) summarized the composition of known proteins including fibrin, casein, crystalline, and pepsin. He noted that the composition of globulin (i.e., hemoglobin) was not yet known.

Source: Simon, transl. Day (1846).

# Advanced Database Searching

# CHAPTER 5

*For many queries, the PSI-BLAST extension can greatly increase sensitivity to weak but biologically relevant sequence relationships. PSI-BLAST retains the ability to report accurate statistics, per iteration runs in times not much greater than gapped BLAST, and can be used both iteratively and fully automatically. These developments should enhance significantly the utility of database search methods to the molecular biologist.*

—Altschul et al. (1997)

*Instead of finding seeds by searching a data structure derived from the query [as in BLAST], one could instead find seeds by searching a data structure derived from the database.*

—Morgulis et al. (2008) on BLAT, SSAHA, and MegaBLAST

*We describe a block-sorting, lossless data compression algorithm... The algorithm works by applying a reversible transformation to a block of input text. The transformation does not itself compress the data, but reorders it to make it easy to compress with simple algorithms such as move-to-front coding.*

—Burrows and Wheeler (1994)

## LEARNING OBJECTIVES

After completing this chapter you should be able to:

- define a position-specific scoring matrix (PSSM);
- explain how position-specific iterated BLAST (PSI-BLAST) and DELTA-BLAST greatly improve the sensitivity of BLAST protein searches;
- describe profile hidden Markov models (HMMs) and explain their advantages over BLAST for database searching;
- explain how spaced seed strategies improve the sensitivity of DNA searches; and
- describe how millions of next-generation sequencing reads are aligned to a reference genome.

## INTRODUCTION

In Chapters 3 and 4 we introduced pairwise alignments and BLAST. BLAST searching allows a database to be searched for proteins or genes. BLAST searches can be very versatile, and in this chapter we cover several advanced database-searching techniques.

Let us introduce three problems for which the five main NCBI BLAST programs are not sufficient.

Using human myoglobin (NP\_005359) as a query in a BLASTP result against human RefSeq proteins, beta globin does not appear.

1. We know that myoglobin is homologous to alpha globin and beta globin; all are vertebrate members of a globin superfamily. We have seen in **Figure 3.1** that myoglobin shares a very similar three-dimensional structure with alpha and beta globin. However, if you use beta globin (NP\_000509.1) as a query and perform a BLASTP search (restricting the output to human and setting the database to nr (nonredundant) or RefSeq), myoglobin does not appear in the results. Fortunately there are programs such as DELTA-BLAST and HMMER that can easily find such homologous but distantly related proteins.
2. Suppose we want to compare long query sequences (e.g., 20,000 base pairs or more) against a database. We might also want to perform a pairwise alignment between two long sequences, such as human chromosome 20 (62 million base pairs long) versus mouse chromosome 2. We need an algorithm that is faster than BLASTN, and we need to explore both global and local strategies. For this problem we can expect some regions of the alignment to have regions of high conservation, but other regions will have diverged substantially. Finding solutions to such searching and alignment problems becomes more critical with the recent availability of thousands of completed genome sequences.
3. A typical next-generation sequencing experiment generates hundreds of millions of short reads (100–400 base pairs) that must be aligned to a single reference genome (e.g., the ~3 billion base pairs of a human genome reference). It would take BLASTN literally weeks to accomplish this alignment problem, but fast aligners can accomplish this in minutes or hours.

This chapter begins with a brief overview of the kinds of specialized BLAST resources that are available to help solve many kinds of research questions. We then introduce PSI-BLAST, DELTA-BLAST and hidden Markov models as tools to find distantly related proteins. We then consider BLAST-like tools for the alignment of genomic DNA.

## SPECIALIZED BLAST SITES

So far, we have used BLAST resources from the NCBI website (Chapters 3 and 4). Other related programs are available, including organism-specific BLAST sites, BLAST sites that allow searches of specific molecules, and specialized database search algorithms.

### Organism-Specific BLAST Sites

We have seen that for standard BLAST searches at the NCBI website the output can be restricted to a particular organism. BLAST searches focused on dozens of prominent organisms can also be performed through the NCBI Map Viewer site.

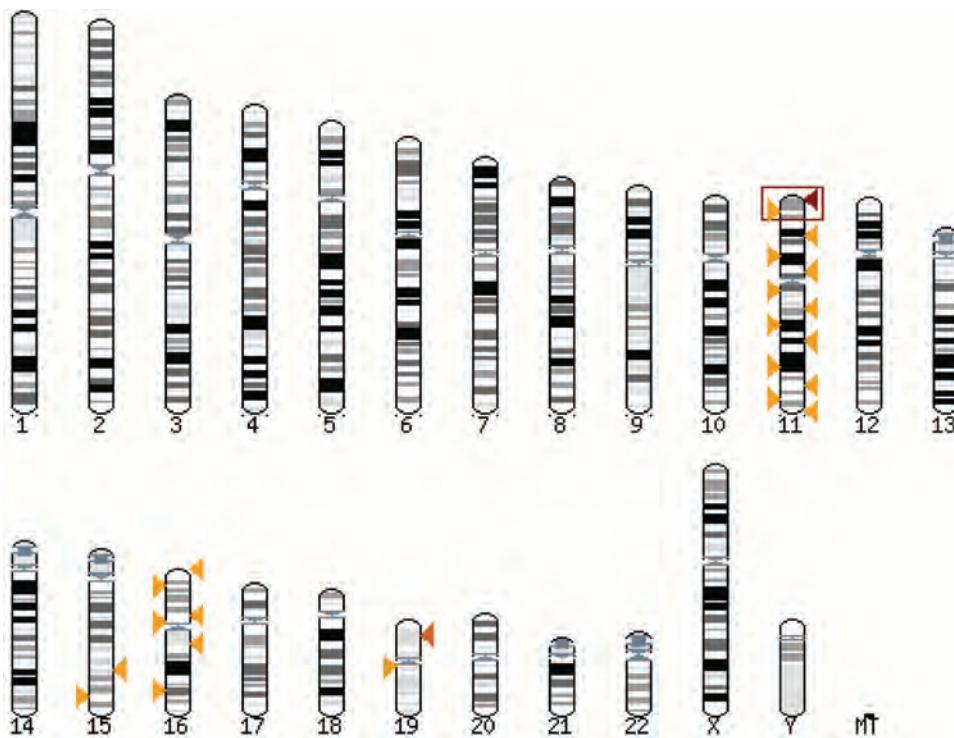
Many databases include molecular sequence data from a specific organism, and these often offer organism-specific BLAST servers. In some cases the data include unfinished sequences that have not yet been deposited in GenBank. If you have a protein or DNA sequence with no apparent matches in standard NCBI BLAST searches, then searching these specialized databases can provide a more exhaustive search. Also, as described in the section on “Specialized BLAST-Related Algorithms”, some of these databases also present unique output formats and/or search algorithms.

Access the Map Viewer from the home page of NCBI or <http://www.ncbi.nlm.nih.gov/mapview/> (WebLink 5.1). It currently offers about 150 separate organism-specific BLAST sites.

Web Document 5.1 at <http://www.bioinfbook.org/chapter5> lists organism-specific BLAST servers.

### *Ensembl BLAST*

Project Ensembl is a joint effort of the Wellcome Trust Sanger Institute (WTSI) and the European Bioinformatics Institute (EBI). The Ensembl website provides a comprehensive resource for studying the human genome and other genomes (see Chapters 15 and 19–20).



**FIGURE 5.1** Output of a TBLASTN search of the Ensembl human database using human beta globin as a query. The results are presented in a graphical format by chromosome, showing the best match to the short arm of chromosome 11 (red box and arrowhead). Weaker matches to paralogs on other chromosomes are also evident (orange arrowheads).

Source: Ensembl Release 76; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

The Ensembl BLAST server allows the user to search the Ensembl database. As an example, paste in the FASTA-formatted amino acid sequence of human beta globin (accession NP\_000509) and perform a TBLASTN search. The output also consists of a graphical output showing the location of the database matches by chromosome (Fig. 5.1). This conveniently shows the chromosomal location of the best hits, including chromosome 11 for beta globin. An alignment summary is provided (Fig. 5.2) with an emphasis on genomic loci. Reasonably high-scoring matches to chromosome 16 can be seen, corresponding to

Links	Query Start	End	Ori	Chromosome Name	Start	End	Ori	Stats			
								Score	E-val	%ID	Length
[A] [S] [G] [C]	31	106	+	Chr:11	5247804	5248031	-	652	4.0e-94	98.68	76
[A] [S] [G] [C]	31	124	+	Chr:11	5255155	5255445	-	646	2.5e-65	81.63	98
[A] [S] [G] [C]	31	110	+	Chr:11	5275504	5275746	-	532	2.4e-82	75.31	81
[A] [S] [G] [C]	13	121	+	Chr:11	5290606	5290980	-	529	9.2e-41	56.25	128
[A] [S] [G] [C]	31	110	+	Chr:11	5270580	5270822	-	527	7.3e-82	75.31	81
[A] [S] [G] [C]	32	104	+	Chr:11	5264339	5264557	-	436	5.9e-73	72.97	74
[A] [S] [G] [C]	101	147	+	Chr:11	5246831	5246962	-	360	4.0e-94	91.49	47
[A] [S] [G] [C]	65	147	+	Chr:11	5254197	5254418	-	323	7.2e-35	55.95	84
[A] [S] [G] [C]	1	45	+	Chr:11	5248123	5248251	-	272	9.1e-42	80.00	45
[A] [S] [G] [C]	105	147	+	Chr:11	5289702	5289830	-	266	1.1e-25	74.42	43
[A] [S] [G] [C]	65	147	+	Chr:11	5274510	5274728	-	263	2.3e-25	50.59	85
[A] [S] [G] [C]	31	143	+	Chr:16	226926	227237	+	260	1.7e-15	35.54	121
[A] [S] [G] [C]	31	143	+	Chr:16	223122	223433	+	256	4.4e-15	35.59	118

The Wellcome Trust Sanger Institute website is <http://www.sanger.ac.uk/> (WebLink 5.2). The EBI is at <http://www.ebi.ac.uk/> (WebLink 5.3). Ensembl's human BLAST server is at [http://www.ensembl.org/Homo\\_sapiens/blastview](http://www.ensembl.org/Homo_sapiens/blastview) (WebLink 5.4), and Ensembl BLAST servers for mouse and other organisms can also be found through <http://www.ensembl.org/> (WebLink 5.5).

**FIGURE 5.2** The Ensembl BLAST server provides an output summary with scores, E values, and links to pairwise alignments (A), the query sequence (S), the genome (matching) sequence (G), and an Ensembl ContigView (C).

Source: Ensembl Release 76; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

alpha globin. The output links include pairwise alignments between the query and each match, and a link to the Contig View. That is the genome browser consisting of assorted graphics and dozens of fields of information (e.g., an ideogram of the chromosome band, a view of neighboring genes, protein and DNA database links, polymorphisms, mouse homologies, and expression data).

#### *Wellcome Trust Sanger Institute*

The WTSI BLAST resources are available at <http://www.sanger.ac.uk/resources/software/blast/> (WebLink 5.6). The VEGA homepage is <http://vega.sanger.ac.uk/> (WebLink 5.7).

#### **Specialized BLAST-Related Algorithms**

We have focused on the standard BLAST algorithms at NCBI, but many other algorithms are available.

#### *WU BLAST 2.0*

WU BLAST 2.0, called AB-BLAST, is licensed by Advanced Biocomputing, LLC at <http://www.advbiocomp.com/> (WebLink 5.8).

Developed by Warren Gish at Washington University, WU BLAST 2.0 is related to the traditional NCBI BLAST algorithms; both were developed from the original NCBI BLAST algorithms that did not permit gapped alignments. WU BLAST 2.0 may provide faster speed and increased sensitivity, and includes a variety of options such as a full Smith–Waterman alignment on some pairwise alignments of database matches. The command-line version of WU BLAST 2.0 offers dozens of options, comparable in scope to the client BLAST+ at NCBI (Chapter 4).

#### *European Bioinformatics Institute (EBI)*

The EBI website provides access to BLAST and other related database search tools (**Table 5.1**):

- BLAST tools include WU BLAST 2.0 as well as NCBI BLAST and PSI-BLAST.
- FASTA (FAST-All), like BLAST, is a heuristic algorithm for searching DNA or protein databases. A set of global and local alignment tools are available, including an implementation of the Smith–Waterman algorithm with SSEARCH. While the run time is relatively slow, this provides a more sensitive algorithm than BLAST or FASTA.
- A search of the European Nucleotide Archive is available, allowing you to find matches to a sequence of interest from next-generation sequence data. Currently, a query with beta globin quickly searches a database of ~1 terabase pairs.

EBI tools are available at <http://www.ebi.ac.uk/Tools/ss/> (WebLink 5.9).

#### *Specialized NCBI BLAST Sites*

The main BLAST site at NCBI offers access to specialized searches of immunoglobulins, vectors, single-nucleotide polymorphisms (SNPs; see Chapter 8), or the trace archives of raw genomic sequence (see Chapter 15). For example, IgBLAST reports the three germline V genes, two D, and two J genes that show the closest match to the query sequence (Ye *et al.*, 2013).

#### *BLAST of Next-Generation Sequence Data*

In Chapter 9 we introduce next-generation sequencing (NGS) and the Sequence Read Archive (SRA) that stores NGS data. You can perform web-based BLAST searching of NGS reads. From the home page of NCBI, enter the search term NA12878; this is the identifier for the well-studied genome (using many sequencing technologies)

**TABLE 5.1 Sequence similarity searching tools at EBI. P, protein; N, nucleotide; G, genomes; WGS, whole-genome shotgun.**

Category	Tool	Query	Description
FASTA	FASTA	P, N, G, WGS	Fast, heuristic, local alignment searching
	SSEARCH	P, N, G, WGS	Optimal (not heuristic-based) local alignment search tool (uses Smith–Waterman)
	PSI-SEARCH	P	Combines SSEARCH with PSI-BLAST profile construction to detect distant relationships
	GGSEARCH	P, N	Optimal global alignment using Needleman–Wunsch algorithm
	GLSEARCH	P, N	Optimal alignment using (global in the query, local in the database sequence).
	FASTM/S/F	P, N, Proteomes	Analyzes short peptide queries
BLAST	NCBI BLAST	P, N, Vectors	Fast, heuristic, local alignment
	WU-BLAST	P, N	Higher-sensitivity alternative to NCBI BLAST
	PSI-BLAST	P	Position-specific iterated BLAST to detect distant relationships
ENA Sequence Search		N	Fast search of European Nucleotide Archive

Source: <http://www.ebi.ac.uk/Tools/ssss/>. Accessed April 2015.

of an individual. Follow the link to SRA where there are currently >400 entries. In the results list, check one or more boxes (e.g., select the result “High-coverage whole-exome sequencing of CEPH/UTAH female individual (HapMap: NA12878)”), then select the link “Send to” and “BLAST.” A standard BLAST interface page appears, and you can search that set of NGS reads using a query of interest such as NM\_000518.4 for beta globin.

## FINDING DISTANTLY RELATED PROTEINS: POSITION-SPECIFIC ITERATED BLAST (PSI-BLAST) AND DELTA-BLAST

Many homologous proteins share only limited sequence identity. Such proteins may adopt the same three-dimensional structures (based on methods such as X-ray crystallography), but in pairwise alignments they may have no apparent similarity. We have seen that scoring matrices are sensitive to protein matches at various evolutionary distances. For example, we compared the PAM250 to the PAM10 log-odds matrices (Figs 3.14 and 3.15) and saw that the PAM250 matrix provides a superior scoring system for the detection of distantly related proteins. In performing a database search with BLAST, we can adjust the scoring matrix to try to detect distantly related proteins. Even so, many proteins in a database are too distantly related to a query to be detected using a standard BLASTP search. In many other cases, protein matches are detected but are so distant that the inference of homology is unclear. We saw that a BLASTP search using RBP4 as a query returned a statistically questionable match ( $E = 0.18$ ) to an authentic homolog, complement component 8 gamma (Fig. 4.16). We would like an algorithm that can assign statistical significance to distantly related proteins that are true positives, while minimizing the numbers of both false positive results (e.g., reporting two proteins as related when they are not) and false negative results (e.g., failing to report that two proteins are significantly related).

You can access PSI-BLAST via the protein BLAST page at  
 ⓘ <http://www.ncbi.nlm.nih.gov/BLAST> (WebLink 5.10) and at other servers such as EBI  
 ⓘ <http://www.ebi.ac.uk/Tools/ssb/psiblast/>, WebLink 5.11) and the Pasteur Institute ⓘ <http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::psiblast>, WebLink 5.12).

We have seen a multiple sequence alignment from a BLAST output in **Figure 4.10**, and examine this topic further in Chapter 6.

PSSM is sometimes pronounced “possum.”

Position-specific iterated BLAST (abbreviated PSI-BLAST or  $\psi$ -BLAST) is a specialized kind of BLAST search that is often more sensitive than a regular BLAST search (Altschul *et al.*, 1997; Zhang *et al.*, 1998; Schäffer *et al.*, 2001). The purpose of using PSI-BLAST is to look deeper into the database to find distantly related proteins that match your protein of interest. In many cases, when a complete genome is sequenced and the predicted proteins are analyzed to search for homologs, PSI-BLAST has been the algorithm of choice.

PSI-BLAST is performed in five steps:

1. A normal BLASTP search uses a scoring matrix (such as BLOSUM62, the default scoring matrix) to perform pairwise alignments of your query sequence (such as RBP) against the database. PSI-BLAST also begins with a protein query that is searched against a database at the NCBI website.
2. PSI-BLAST constructs a multiple sequence alignment from an initial BLASTP-like search using composition-based statistics (Schäffer *et al.*, 2001). It then creates a specialized, individualized search matrix (also called a profile) based on that multiple alignment.
3. This position-specific scoring matrix (PSSM) is then used as a query to search the database again. (Your original query is not used.)
4. PSI-BLAST estimates the statistical significance of the database matches, essentially using the parameters we described for gapped alignments.
5. The search process is continued iteratively, typically about five times. At each step a new profile is used as the query. You must decide how many iterations to perform; simply click on “Run PSI-BLAST Iteration.” You can stop the search process at any point, for example, whenever few new results are returned or when the program reports convergence because no new results are found.

When we view a multiple sequence alignment of proteins, we can generally see column positions in which a given residue has its own specific patterns of substitution. We highlighted an example in **Figure 4.19** showing how arginine residues in an alignment are in some positions perfectly conserved, and in other positions they may be substituted with other residues. This kind of information is captured by a PSSM.

For a query of length  $L$ , PSI-BLAST generates a PSSM of dimension  $L \times 20$ . The rows of each matrix have a length  $L$  equal to the query sequence. Redundant sequences (having at least 94% amino acid identity in a pairwise alignment of any two sequences in the matrix) are eliminated. This ensures that a group of very closely related sequences will not overly bias the construction of the PSSM. The same gap scores are applied as in BLASTP, rather than implementing position-specific gap scores. A unique scoring matrix (profile) is derived from the multiple sequence alignment (Box 5.1). For each iteration of PSI-BLAST, a separate scoring matrix is created.

What does a PSSM look like? The NCBI Education site offers a PSSM viewing tool. The result for an alignment of human beta globin to a family of globin proteins is shown in **Figure 5.3**. The consensus sequence is arranged in a column (see arrow; the first 20 residue positions of the PSSM are numbered in the column labeled P). The query sequence (beta globin protein) is next given in a column (see arrow), followed by 20 columns for the amino acids. The numeric values are scores assigned to the residues at each position of the PSSM. Consider alanine (column indicated by arrow 1). In our query, alanine occurs two times and is assigned scores of +2 and +3. In the consensus alanine occurs three times (at positions 3, 7, and 16) where it is assigned scores of +3, +3, and +1. The fact that the score assigned to an alanine match varies according to position exemplifies the main feature of the PSSM. Where the score is +3 alanine tends to occur frequently for many proteins in this family; where the score is only +1 alanine is less well conserved (and our beta globin query protein has a valine here).

## BOX 5.1 PSI-BLAST TARGET FREQUENCIES

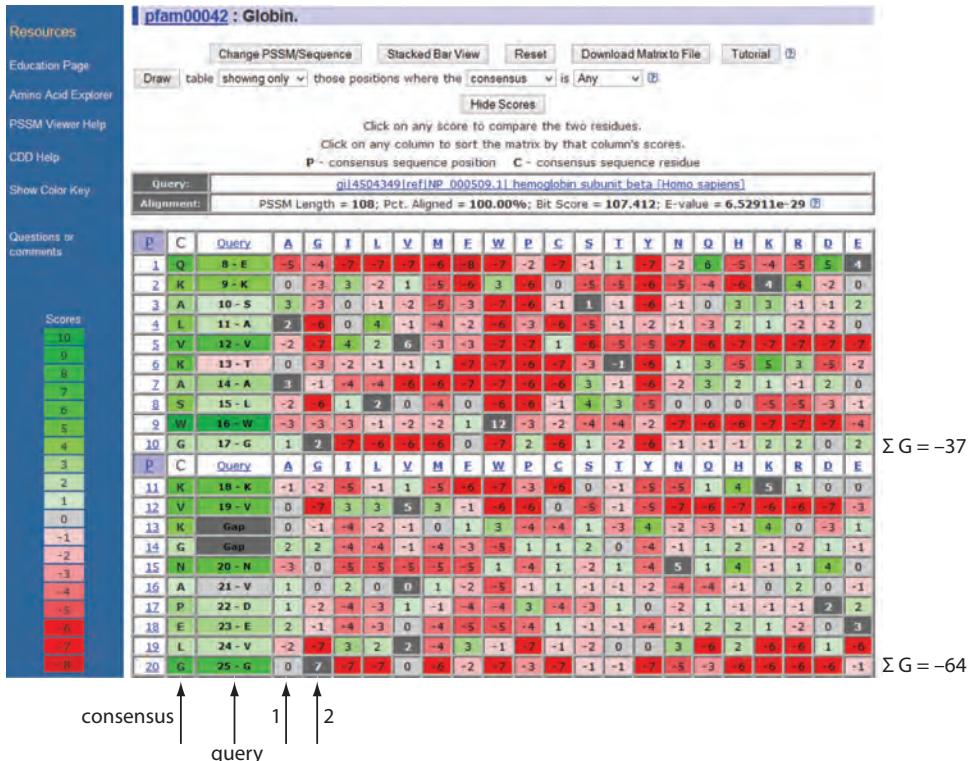
Scores are derived for each specific column position in the form  $\log(q_i / p_i)$ , where  $q_i$  is the estimated probability of residue  $i$  being found in that column position and  $p_i$  is the background probability for that residue (Altschul *et al.*, 1997). The key problem is to estimate the target frequencies  $q_i$ . This is accomplished in two steps using a method of pseudocounts (Tatusov *et al.*, 1994). First, pseudocount frequencies  $g_i$  are obtained for each column position as follows:

$$g_i = \sum_j \frac{f_j}{p_j} q_{ij} \quad (5.1)$$

where  $f_j$  are the observed frequencies,  $p_j$  are the background frequencies, and  $q_{ij}$  are the target frequencies implicit in the substitution matrix (as described in Equations (3.4), (3.7)). Next, the target frequencies  $q_i$  (corresponding to the likelihood of observing residue  $i$  in the position of a column) are defined:

$$q_i = \frac{\alpha f_i + \beta g_i}{\alpha + \beta} \quad (5.2)$$

where  $\alpha$  and  $\beta$  are relative weights assigned to the observed frequencies  $f_i$  and the pseudocount residue frequencies  $g_i$ . Having estimated the target frequencies, it is now possible to assign a score for a given aligned column as  $\ln(q_i / p_i) / \lambda$ . Altschul *et al.* (2009) further showed that more highly conserved column positions require fewer pseudocounts, a correction that improves retrieval accuracy.



**FIGURE 5.3** Matrix view of a position-specific scoring matrix (PSSM). The NCBI PSSM visualization tool was queried for pfam00042 (corresponding to the Conserved Domain Database globin protein family), and the accession number for human beta globin protein (NP\_000509) was entered as a specific protein to align to the PSSM. The matrix view option was selected. The rows are numbered by position (column header P, showing up to the 20th position) for the consensus sequence (column header C) and this particular query. The columns include the 20 amino acids (labeled from A to E). Arrows 1–2 indicate amino acids that are assigned dramatically different scores depending on their positions in the protein (see text for details). The right side of the figure shows that for two glycine residues the sum of the scores is far more negative for one (at position 20) than the other (at position 10), reflecting the greater conservation of glycine at position 20. At position 10 the score for matching a glycine is only +2, while at position 20 the score is +7.

Source: PSSM Viewer, NCBI.

For the NCBI PSSM viewer visit  
[http://www.ncbi.nlm.nih.gov/  
 Class/Structure/pssm/pssm\\_viewer.cgi](http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi) (WebLink 5.13).

You can adjust the inclusion threshold. Try  $E$  values of 0.5 or 0.00005 to see what happens to your search results. If you set the  $E$  value too low, you will only see very closely related homologs. If you set  $E$  too high, you will often find false positive matches.

For glycine (arrow 2) the assigned score is +2 (see consensus position 10) or +7 (consensus position 20). This difference is enormous. Looking across the row at position 10, the sum of scores assigned to all 20 possible amino acids is -37 (negative scores indicate that alignment of a glycine with any other residue occur less frequently than expected by chance). Next looking across the row at position 20, the sum of scores is -64, showing that the occurrence of glycine is heavily favored and the penalties for not selecting a glycine are severe. Similar patterns occur for other amino acids. These examples illustrate one of the main advantages of PSI-BLAST: the PSSM reflects a more customized estimate of the probabilities with which amino acid substitutions occur at all positions.

We can illustrate the dramatic results of the PSI-BLAST process as follows. Go to the protein BLAST page at NCBI, enter the protein accession number of human beta globin (NP\_000509), and select the PSI-BLAST option and the RefSeq database restricted to fungi. Using the default parameters, there are over 60 hits (Table 5.2). Nine of these have significant  $E$  values lower than the inclusion threshold (set as a default at  $E = 0.005$ ). By inspection these are called hypothetical proteins (from various fungal species), so it is not clear from their names alone whether they are globins. There are also dozens of database matches that are worse than the inclusion threshold: these do not have significant  $E$  values. Some of these distantly related matches (such as flavohemoproteins and an *Aspergillus* protein called “bacterial hemoglobin”) are authentic globins, based on criteria such as having similar three-dimensional structures and related biological functions as carrier proteins. Other proteins on this list appear to be true negatives.

Through this initial step, the PSI-BLAST search is performed in a manner nearly identical to a standard BLASTP search, using some amino acid substitution matrix such as BLOSUM62. However, upon selecting all of the hits better than threshold, the program creates a multiple sequence alignment from the initial database matches. By analyzing this alignment, the PSI-BLAST program then creates a PSSM. The original query sequence serves as a template for this profile.

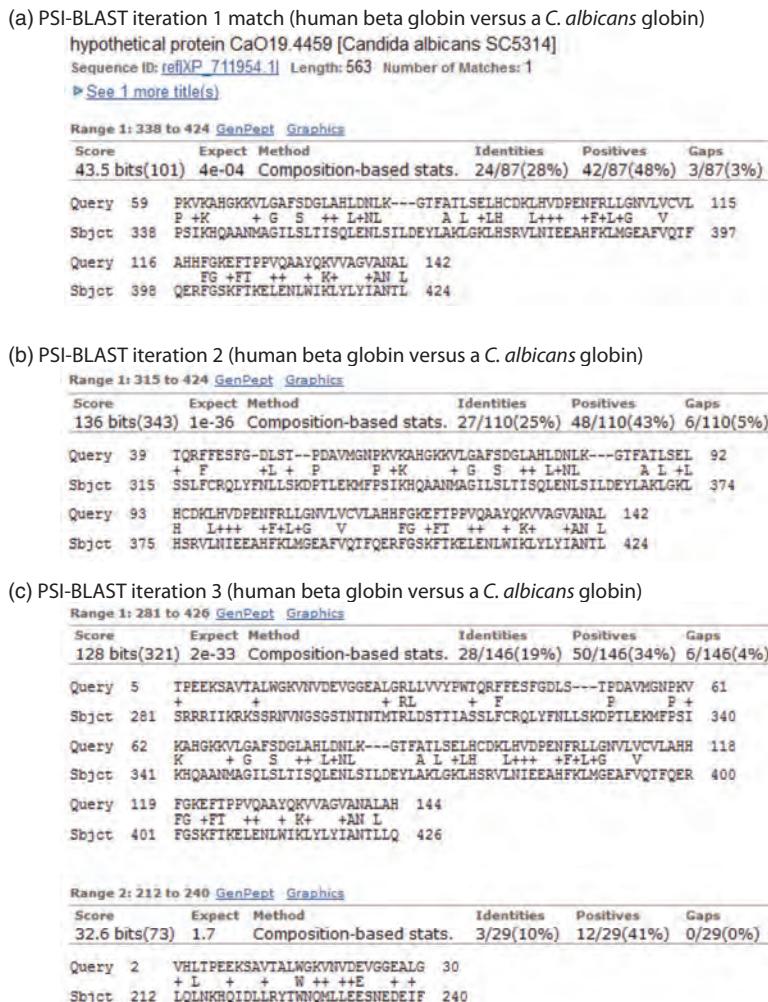
The unique profile that PSI-BLAST identifies is next used to perform an iterative search. Click the box “run PSI-BLAST iteration 2.” The search is repeated using the customized profile, and new proteins are often added to the alignment. This is seen in the second iteration as the number of database hits better than the threshold rises from 9 to 182 (Table 5.2). In subsequent iterations, the number of database hits better than the threshold rises slightly to 206. By inspection, all of these are authentic members of the globin family. The search can be halted once such a plateau is reached, or the iteration process continued until the program reports that convergence has been reached. This indicates that no more database matches are found, and the PSI-BLAST search is ended.

**TABLE 5.2** PSI-BLAST produces dramatically more hits with significant  $E$  values than BLASTP. Human beta globin (NP\_000509) was used as a query in a PSI-BLAST search of the RefSeq database restricted to fungi (txid:4751; February 2015). Results of iterations 5–10 (not shown) resembled those of iterations 3–4.

Iteration	Hits with $E \leq 0.005$	Hits with $E > 0.005$
1	9 (hbb fungi)	54
2	182	22
3	206	41
4	207	24

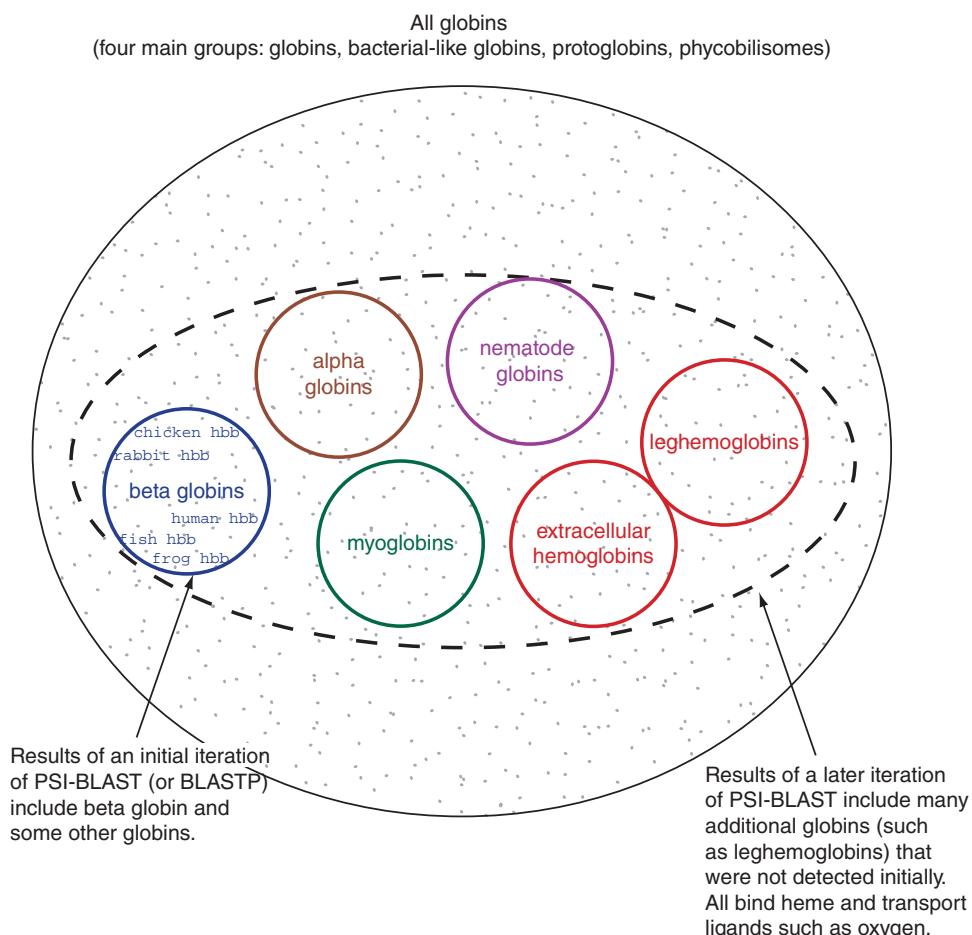
What did this search achieve? After a series of position-specific iterations, over 200 additional database matches were identified. Many distantly related proteins are now shown in the alignment. We can understand how the sensitivity of the search increased by examining the pairwise alignment of the query (HBB) with the best match, a hypothetical protein from *Candida albicans*. In the first PSI-BLAST iteration, the bit score was 43.5, the expect value was 4e-04 (i.e.,  $4 \times 10^{-4}$ ), and there were 24 identities and 3 gaps across an alignment of 87 residues (**Fig. 5.4a**). After the second iteration the score rose to 136 bits, the *E* value dropped (to  $10^{-30}$ ), and the length of the alignment increased (to 110 residues), although the number of gaps increased with this longer alignment (**Fig. 5.4b**). In the third PSI-BLAST iteration the *E* value for this pairwise alignment was  $2 \times 10^{-33}$  (**Fig. 5.4c**). (The *E* value did not reduce further, probably because of the particular nature of the adjusted PSSM which likely favored the inclusion of other fungal globins.) Relative to the first iteration the *E* value was dramatically lower as a result of using a scoring matrix specially constructed for this family of proteins. Note

The *Candida albicans* protein has accession XP\_711954.1.



**FIGURE 5.4** PSI-BLAST search detects distantly related proteins using progressive iterations with a PSSM. (a) A search with human beta globin as a query (NP\_000509.1) detects a fungal globin that is annotated as a hypothetical protein (XP\_711954.1) in the first iteration. (b) As the search progresses to the second iteration, the length of the alignment increases, the bit score becomes higher, and the expect value decreases. (c) By the third iteration, the match to human protein spans two regions: 146 and 29 residues. In the fourth iteration (not shown) these are pieced together into a single alignment spanning 146 residues (with 19% amino acid identity).

Source: BLAST, NCBI.



**FIGURE 5.5** PSI-BLAST algorithm increases the sensitivity of a database search by detecting homologous matches with relatively low sequence identity. In this figure, each dot represents a single globin protein. There are four related families of globins (globins, bacterial-like globins, protoglobins, and phycobilisomes; see Chapter 12). The ellipse represents the globins (such as alpha and beta globins, myoglobins, and leghemoglobins). All these proteins are homologous by virtue of their membership in the same family. A standard BLASTP search with human beta globin returns matches that are relatively close to the query in sequence identity, and the result (represented by the circle on the left) may include additional matches to proteins such as alpha globins. However, many other homologous proteins such as leghemoglobins are not detected. The fundamental limitation in standard BLAST search sensitivity is the reliance on standard PAM and BLOSUM scoring matrices. In a PSI-BLAST search, a PSSM generates a scoring system that is specific to the group of matches detected using the initial query sequence. While the initial iteration of a PSI-BLAST search results in the same number of database matches as a standard BLAST search, subsequent PSI-BLAST iterations (represented by the dashed oval) using a customized matrix extend the results to allow the detection of more distantly related homologs.

that the PSSM allowed the amino terminal residues of the human query to align with the *Candida* globin only in the third iteration (Fig. 5.4c, bottom). By the fourth iteration these two regions are stitched together, resulting in aligned span of 146 residues with only 6 gaps.

We can visualize the PSI-BLAST process by imagining each globin protein in the database as a point in space (Fig. 5.5). An initial BLASTP search with beta globin, not restricted to any taxonomic group, detects other globin orthologs (e.g., chicken, fish) and paralogs (e.g., alpha globins). The PSSM of PSI-BLAST facilitates the detection of many

other globins. Fewer than a dozen fungal globins are detected by a BLASTP search using HBB as a query, but hundreds are found by PSI-BLAST.

The number of iterations that a PSI-BLAST search performs relates to the number of hits (sequences) in the database that running the program reports. After each PSI-BLAST iteration, the results that are returned describe which sequences match the input PSSM.

### PSI-BLAST Errors: Problem of Corruption

The main source of error in PSI-BLAST searches is the spurious amplification of sequences that are unrelated to the query. This problem most often arises when the query (or the profile generated after PSI-BLAST iterations) contains regions with highly biased amino acid composition. Once the program finds even one new protein hit having an *E* value even slightly above the inclusion threshold, that new hit will be incorporated into the next profile and will reappear in the next PSI-BLAST iteration. If the hit is to a protein that is not homologous to the original query sequence, then the PSSM has been corrupted. We can define corruption as occurring when, after five iterations of PSI-BLAST, the PSSM produces at least one false positive alignment of  $E < 10^{-4}$ .

This definition of corruption is adapted from Schäffer *et al.* (2001).

There are three main approaches to stopping corruption of PSI-BLAST queries.

1. Apply a filtering algorithm that removes biased amino acid regions. These “low-entropy” regions include stretches of amino acids that are highly basic, acidic, or rich in a residue such as proline. The NCBI PSI-BLAST site employs the SEG program for this purpose, applying the filtering algorithm to database sequences that are detected by the query.
2. Adjust the expect level from its default value (e.g.,  $E = 0.005$ ) to a lower value (e.g.,  $E = 0.0001$ ). This may suppress the appearance of false positives, although it could also interfere with the detection of true positives.
3. Visually inspect each PSI-BLAST iteration. Each protein listed in the PSI-BLAST output has a checkbox; select and remove suspicious ones. As an example, your query protein may have a generic coiled-coil domain, and this may cause other proteins sharing this motif (such as myosin) to score better than the inclusion threshold even though they are not homologous.

SEG was described by Wootton and Federhen (1996).

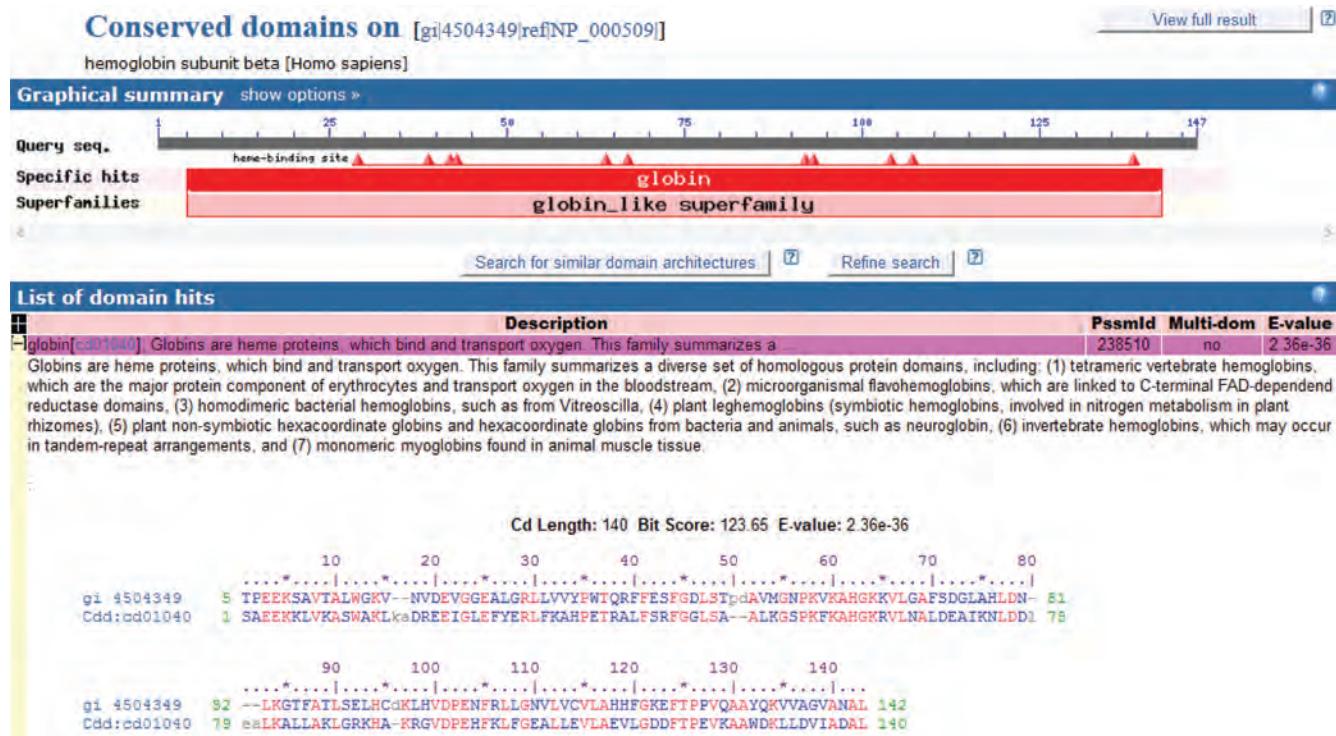
### Reverse Position-Specific BLAST

Reverse position-specific BLAST (RPS-BLAST) is used to search a single protein query against a large database of predefined PSSMs. The purpose is to identify conserved protein domains in the query. RPS-BLAST searches are implemented in the Conserved Domain Database (CDD) at NCBI (Marchler-Bauer *et al.*, 2013). A typical result, based on using human beta globin as a query, shows the globin family (Fig. 5.6). Annotations are by CDD and by the protein family database PFAM which we will describe in Chapter 6. CDD includes manually curated PSSMs that include information from three-dimensional structures (Chapter 13).

CDD is available at <http://www.ncbi.nlm.nih.gov/cdd> (WebLink 5.14) or through the main BLAST page (<http://blast.ncbi.nlm.nih.gov/>, WebLink 5.15). Currently there are over 50,000 alignment models (PSSMs) in the CDD database (v3.10). We discuss protein domains in Chapter 12.

### Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)

Now that we have introduced PSI-BLAST and RPS-BLAST, we can describe the most sensitive and accurate protein search tool at NCBI: DELTA-BLAST (Boratyn *et al.*, 2012). Each tool uses a single protein as a query. PSI-BLAST automatically creates multiple sequence alignments and then generates PSSMs in an iterative manner. DELTA-BLAST begins with an RPS-BLAST search against a library of pre-computed PSSMs,



**FIGURE 5.6** Reverse position-specific BLAST is used to search a query (here human beta globin) against a collection of predefined position-specific scoring matrices. The result includes an *E* value, a pairwise alignment of the query to the consensus PSSM, and a description of the family of proteins in the PSSM. This BLAST tool is available at NCBI as part of the Conserved Domain Database.

Source: PSSM Viewer, NCBI.

then uses a resulting PSSM to search a protein database. There are important advantages to DELTA-BLAST:

- It yields larger, more complete PSSMs than PSI-BLAST. This is because it relies on the strength of the well-curated CDD database.
- It is more sensitive than BLASTP and PSI-BLAST, including searches of distantly related proteins.
- It is fast. While DELTA-BLAST was designed to allow multiple iterations, its performance surprisingly declines with more than one iteration. (Boratyn *et al.* suggest that this may sometimes occur because a PSSM loses sensitivity when applied to overly divergent members of a protein family.)
- It produces better-quality alignments than BLASTP.

As an example of DELTA-BLAST performance, first use beta globin (NP\_000509.1) as a query in a BLASTP search of human RefSeq proteins. The result includes 10 hemoglobin proteins having low *E* values. Repeat the search using DELTA-BLAST and the same proteins are identified as well as two more divergent globins (neuroglobin and myoglobin). The search is therefore complete.

As another example, search plant RefSeq proteins using beta globin as a query. Currently there is only one significant match (leghemoglobin from the barrel medic *Medicago truncatula*, a eudicot; *E* value 0.024). Repeat the search with DELTA-BLAST and there are 58 significant plant matches (*E* values ranging from  $1 \times 10^{-25}$  to  $5 \times 10^{-6}$ ).

In sum, DELTA-BLAST vastly outperforms BLASTP. In the minority of cases in which a protein query matches no PSSMs from CDD, the DELTA-BLAST result simply resembles that of a typical BLASTP search. In addition to the web-based version, the stand-alone BLAST+ package includes the program *deltablast*.

## Assessing Performance of PSI-BLAST and DELTA-BLAST

To assess the performance of PSI-BLAST and DELTA-BLAST, it is useful to search databases containing structural information such as structural classification of proteins (SCOP) or ASTRAL. These databases included distantly related proteins that are known to be homologous (and that adopt similar three-dimensional structures even if particular protein pairs share very limited percent amino acid identity). The homologous relationships in these databases constitute a “gold standard” of true positive results and also allow the number of false positive results to be assessed. This information can be plotted to generate receiver operating characteristic (ROC) scores (Gribskov and Robinson, 1996). Such ROC curves were reported for PSI-BLAST by Schäffer *et al.* (2001) and Park *et al.* (1998) and for DELTA-BLAST by Boratyn *et al.* (2012). The sensitivity of DELTA-BLAST was superior to PSI-BLAST, BLASTP, and several other programs: at a given number of false positives (e.g., 1 per query), DELTA-BLAST identified about three times as many homologs as BLASTP.

## Pattern-Hit Initiated BLAST (PHI-BLAST)

Sometimes a protein you are interested in contains a pattern or “signature” of amino acid residues that help define that protein as part of a family. For example, a signature might be an active site of an enzyme, a string of amino acid residues that define a structural or functional domain of a protein family, or even a characteristic signature of unknown function (such as the three amino acids GXW that are almost always present in the lipocalin family, where X refers to any residue). Pattern-hit initiated BLAST (PHI-BLAST) is a specialized BLAST program that allows you to search with a query and to find database matches that both match a pattern and are significantly related to the query (Zhang *et al.*, 1998). PHI-BLAST may be preferable to simply searching a database with a short query corresponding to a pattern, because such a search could result in the detection of many random matches or proteins that are unrelated to your query protein. While DELTA-BLAST is highly sensitive, it will not report information about user-selected patterns.

Consider a BLASTP search of bacterial sequences using human RBP4 as a query (NP\_006735), restricted to the refseq database. The result (at the time of writing in February 2015) is that there are seven database matches having small *E* values (<0.05). We know that there are many more bacterial lipocalins distantly related to human RBP4 (this could be confirmed using DELTA-BLAST). Select the three best-scoring bacterial lipocalin protein sequences and align them with human RBP4 (Fig. 5.7a). This alignment shows us which amino acid residues are actually shared between RBP4 and the bacterial proteins. Focusing on the three conserved residues NFD, as well as the GXW motif that is shared between almost all lipocalins, we can try to define a pattern (or signature) of amino acids that is shared by RBP4, the three bacterial lipocalins, and possibly many other bacterial lipocalins. The purpose of defining a signature is to customize a PSI-BLAST algorithm to search for proteins containing that signature.

How is the signature or pattern defined? We do not expect the signature to be exactly identical between all bacterial lipocalins, and so we want to include freedom for ambiguity. We can define any pattern we want; as an example we examine the multiple alignment in Figure 5.7a and create the pattern NDFX(5)GXW[YF]. X(5) indicates that five positions can assume any amino acid residue; [YF] indicates that either a tyrosine or phenylalanine is accepted at the last position of the string. Note that the pattern you choose must not occur too commonly; the algorithm only allows patterns that are expected to occur less frequently than once every 5000 database residues. In general, it is acceptable to choose any pattern with four completely specified residues or three residues with average background frequencies of ≤5.8% (Zhang *et al.*, 1998).

SCOP (Chapter 13) is available at

<http://scop.mrc-lmb.cam.ac.uk/scop/> (WebLink 5.16). It was developed by Cyrus Chothia and colleagues. Park *et al.* (1998) used the PDBD40-J database, which contains proteins of known structure with ≤40% amino acid identity. DELTA-BLAST was tested against several databases including ASTRAL. The current version of SCOP+ASTRAL (1.75B) includes over 4000 protein families, ~2000 superfamilies and ~1200 folds. See <http://scop.berkeley.edu/> (WebLink 5.17) and Chapter 13.

PHI-BLAST is launched from the NCBI BLASTP web page.

Look at the alignment in Figure 5.7a and then try to create and test your own pattern. The syntax for a PHI-BLAST pattern is derived from the Prosite dictionary (Chapter 12) and is described at <http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html> (WebLink 5.18). A detailed example of using PHI-BLAST is provided in Web Document 5.2 at <http://www.bioinfbook.org/chapter5>.

(a) Multiple alignment of human RBP4 and three bacterial homologs

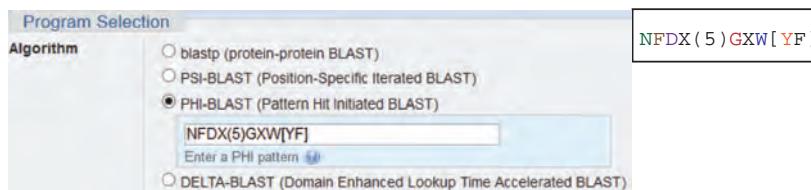
MUSCLE (3.8) multiple sequence alignment

```

NP_006735.2      -MKWVWALLLAALGSGRAERDCRVSSFRVK--ENFDKARFSGTWYAMAKK
WP_010388720.1   --MKLAFKTALFITAMFLSACTSAPEGITPVKNFDLEKYQGKWYEIARL
WP_008992866.1   MKA  
K  
N  
K  
I  
L  
I  
A  
A  
C  
A  
I  
G  
L  
G  
A  
L  
L  
N  
S  
C  
A  
I  
P  
K  
N  
A  
K  
A  
V  
KNFDIDRYLGTWYEIARF
YP_003021245.1   -MKKLSLLSLIFTG-----CVGIPENVPVDNFDVHRYLGKWYEIARL
: * . . . : * . ** : * .

```

(b) PHI pattern



(c) Example of a PHI-BLAST result (asterisks match PHI pattern)

outer membrane lipoprotein (lipocalin) [Pseudoalteromonas sp. SM9913]

Sequence ID: ref|YP\_004064995.1| Length: 177 Number of Matches: 1

[► See 1 more title\(s\)](#)

Range 1: 31 to 109 GenPept Graphics				
Score	Expect	Identities	Positives	Gaps
21.4 bits(63)	8e-05	21/80(26%)	40/80(50%)	1/80(1%)
Pattern *****				
Query 31	ENFDKARFSGT <b>WYAMAKK</b> DPEGLFLQDNIVAEFSVDET <b>GQMSATAKGRVRLNNWDVCAD</b> 90			
	+NFD ++ G WY +A+ D + + A +S+++ G + KG + WD A+			
Sbjct 31	KNF <b>D</b> LEYQ <b>GK</b> WYEIARLDHSFEQGM <b>E</b> QTATYSINDDGTVKVLNKGFISKE <b>Q</b> KNDE-AE 89			
Query 91	MVGTFIDIEDPAKEFKM <b>KYWG</b> 110			
	+ E + D EK+ ++G			
Sbjct 90	GLAKFVENADTG <b>H</b> FKVSFFG 109			

**FIGURE 5.7** Choosing a pattern and performing a PHI-BLAST search. (a) Human RBP4 (accession NP\_006735) was used as a query in a BLASTP search against bacterial sequences, then multiply aligned with three bacterial lipocalins (accessions WP\_010388720.1 from gammaproteobacterium *Pseudoalteromonas undina*, WP\_008992866 from the flavobacterium *Galibacter*, and delta-proteobacterium *Geobacter* YP\_003021234). The purpose of evaluating these four protein sequences together is to try to identify a short, sequential pattern of amino acid residues that consistently occurs in a protein family. This pattern is then included in a new PHI-BLAST search to increase its sensitivity and specificity. The alignment was performed using MUSCLE (Chapter 6), and a portion of the alignment is shown. The invariant GXW motif that is typical of lipocalins is evident (within the boxed region). A PHI pattern can be selected that includes these residues and several more. As an example, we select the pattern NFDX(5)GXW[YF] in which, following NFD, any five residues are allowed, then the next three positions are GXW, and the final position contains either Y or its closely related residue F. The user can select any pattern by trial and error. (b) A PHI-BLAST search is selected from the NCBI protein blast page, and the PHI pattern is entered. (c) The database will then be searched. All database matches include the selected pattern, indicated by asterisks. In some cases the use of a PHI pattern returns results not found with other search tools.

Source: BLAST, NCBI.

We use PHI-BLAST and enter the “PHI pattern” NFDX(5)GXW[YF] (Fig. 5.7b). The BLAST search output is restricted to a subset of the database consisting of proteins that contain that specified pattern. The result of this search is 28 database matches consisting of bacterial lipocalins having *E* values less than 0.05. The pairwise alignment output of the PHI-BLAST search has the identical format to the PSI-BLAST output, except that information about where both the query and each database sequence match the PHI pattern is shown by a series of asterisks (Fig. 5.7c). The ensuing PSI-BLAST iteration, which no longer uses the PHI pattern but instead uses a search-specific PSSM, will successfully identify a large family of bacterial lipocalins.

The PHI-BLAST algorithm employs a statistical analysis based on identifying alignment  $A_0$  spanned by the input pattern and regions  $A_1$  and  $A_2$  to either side of the pattern, which are scored by gapped extensions. Scores  $S_0$ ,  $S_1$ , and  $S_2$  corresponding to these regions are calculated, and PHI-BLAST scores are ranked by the score  $S' = S_1 + S_2$  (ignoring  $S_0$ ). The alignment statistics are closely related to those used for BLASTP searches (Zhang *et al.*, 1998).

## PROFILE SEARCHES: HIDDEN MARKOV MODELS

DELTA-BLAST employs scoring matrices that are customized because of their position-specific nature in a manner that is dependent on the particular input sequence(s). DELTA-BLAST is therefore more sensitive at detecting significantly related aligned residues than PAM or BLOSUM matrices. PSSMs are examples of profiles, a concept introduced by Gribskov and others (Gribskov *et al.*, 1987, 1990). Profile hidden Markov models (HMMs) are even more versatile than PSSMs to generate a position-specific scoring system useful for the detection of remote sequence similarities (Baldi *et al.*, 1994; Eddy, 1998, 2004; Krogh, 1998; Birney, 2001; Schuster-Böckler and Bateman, 2007; Yoon, 2009). HMMs have been widely used in a variety of signal detection problems ranging from speech detection to sonar. Within the field of bioinformatics they have been used for applications as diverse as sequence alignment, prediction of protein structure, prediction of transmembrane regions in proteins, analysis of chromosomal copy number changes, and gene-finding algorithms.

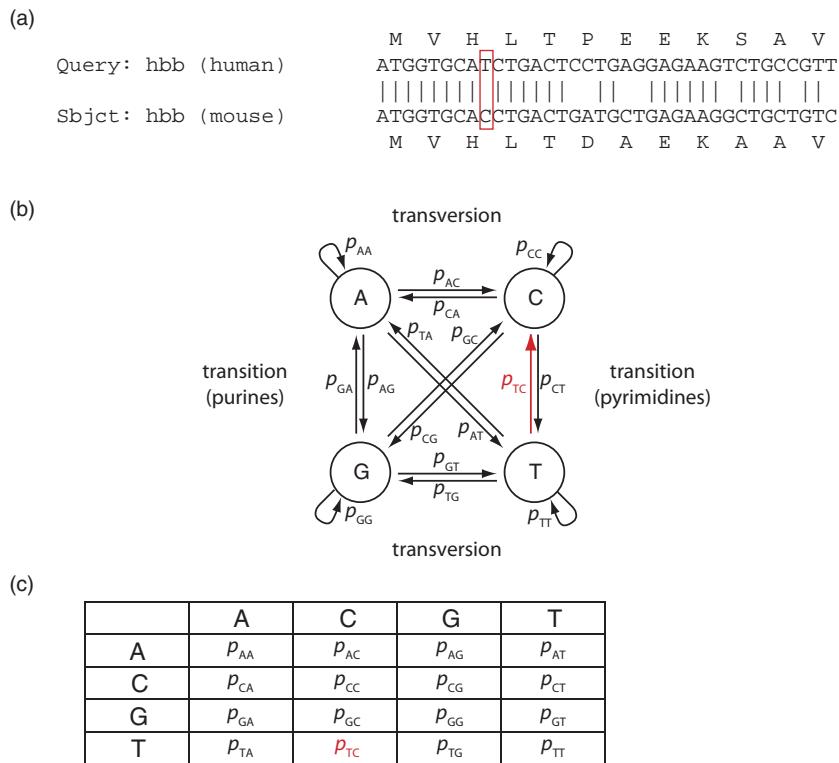
The main strength of a profile HMM is that it is a probabilistic model. This means that it assesses the likelihood of matches, mismatches, insertions, and deletions (i.e., gaps) at a given position of an alignment. By developing a statistical model that is based on known sequences, we can use a profile HMM to describe the likelihood that a particular sequence (even one that was previously unknown) matches the model. In contrast, DELTA-BLAST does not specify a full probabilistic model.

A profile HMM can convert a multiple sequence alignment into a position-specific scoring system. A common application of profile HMMs is the query of a single protein sequence of interest against a database of profile HMMs. Another application is to use a profile HMM as the query in a database search. PFAM and SMART (Chapters 6 and 12) are examples of prominent databases that are based on HMMs.

A Markov chain is a data structure that consists of a computational model with a start state, a finite, discrete set of possible states, and transition functions that describe how to move from one state to the next. This type of computational model is also called a finite-state machine. A basic feature of Markov chains is that the process occupies one state at any given unit of time, and remains in that state or moves to another allowable state. Consider an nucleotide alignment of human and mouse beta globin, beginning with the start codons, focusing on a position at which a T (in human) is aligned with a C (in mouse) (Fig. 5.8a). A Markov model displays the transition probabilities for any nucleotide (A, C, G, T) changing to any other (Fig. 5.8b). Each circle containing a nucleotide represents a state. There are 16 arrows indicating the probabilities of making a transition to another state. These 16 probabilities can be summarized in a tabular transition matrix (Fig. 5.8c). This resembles the mutation probability matrix for amino acids that were developed by Dayhoff and colleagues (Chapter 3).

In the case of a hidden Markov model (HMM), we cannot observe the states directly. However, we do have observations from which we can infer the hidden states. In the case of molecular sequences, the observed states are the positions of amino acids (or nucleotides) in a multiple sequence alignment. The hidden states are the match states, insert states, and delete states. Together, such states define a model for the sequence of that protein or nucleotide family.

Markov chains were introduced by Andrei Andreyevich Markov (1856–1922), a Russian mathematician. HMMs were introduced into the field of bioinformatics by Gary Churchill (1989) and Anders Krogh, David Haussler and colleagues (Krogh *et al.*, 1994).

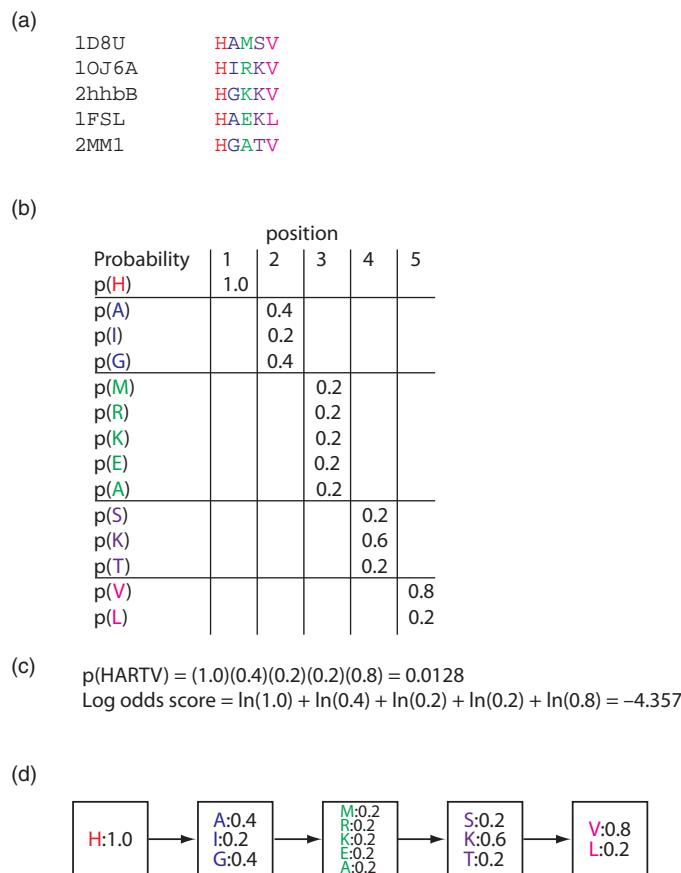


**FIGURE 5.8** A hidden Markov model describes the transition probabilities for the alignment of nucleotides (shown here) or amino acids. For proteins this is conceptually related to the position-specific scoring matrix used by PSI-BLAST. (a) Human beta globin protein (NP\_000509.1) and corresponding DNA (NM\_000518.1) are aligned to mouse hemoglobin, beta adult major chain (Hbb-b1) (NP\_001265090.1 and NM\_001278161.1). A transition is indicated in the red rectangle. (b) The four nucleotides GATC (in circles) are represented as states with 16 arrows showing potential state changes (nucleotide substitutions). The observed alignment includes a substitution of T->C (a transition represented by a red arrow). (c) The probabilities of each of the 16 changes is displayed as a transition matrix. Adapted from Schuster-Böckler and Bateman (2007) with permission from John Wiley & Sons.

An HMM therefore consists of a series of defined states. Consider the five amino acid residues taken from an alignment of five globin proteins (Fig. 5.9a). An HMM can be calculated by estimating the probability of occurrence of each amino acid in the five positions (Fig. 5.9b). In this sense, the HMM approach resembles the position-specific scoring matrix calculation of PSI-BLAST. From the HMM probabilities, a score can be derived for the occurrence of any specific pattern of a related query, such as HARTV (Fig. 5.9c). The HMM is a model that can be described in terms of “states” at each position of a sequence (Fig. 5.9d).

A profile HMM is constructed from an initial multiple sequence alignment to define a set of probabilities. The structure of a profile HMM is shown in Figure 5.10a (Krogh *et al.*, 1994; Krogh, 1998). Along the bottom row is a series of main states (from “begin” to  $m_1-m_5$ , then “end”). These states might correspond to residues of an amino acid sequence such as HAEKL. The second row consists of insert states (Fig. 5.10, diamond-shaped objects labeled  $i_1-i_5$ ). These states model variable regions in the alignment, allowing sequences to be inserted as necessary. The third row (at the top) consists of circles called delete states, which correspond to gaps. They provide a path to skip a column (or columns) in the multiple sequence alignment. The emissions lead to the observed sequences in the alignment.

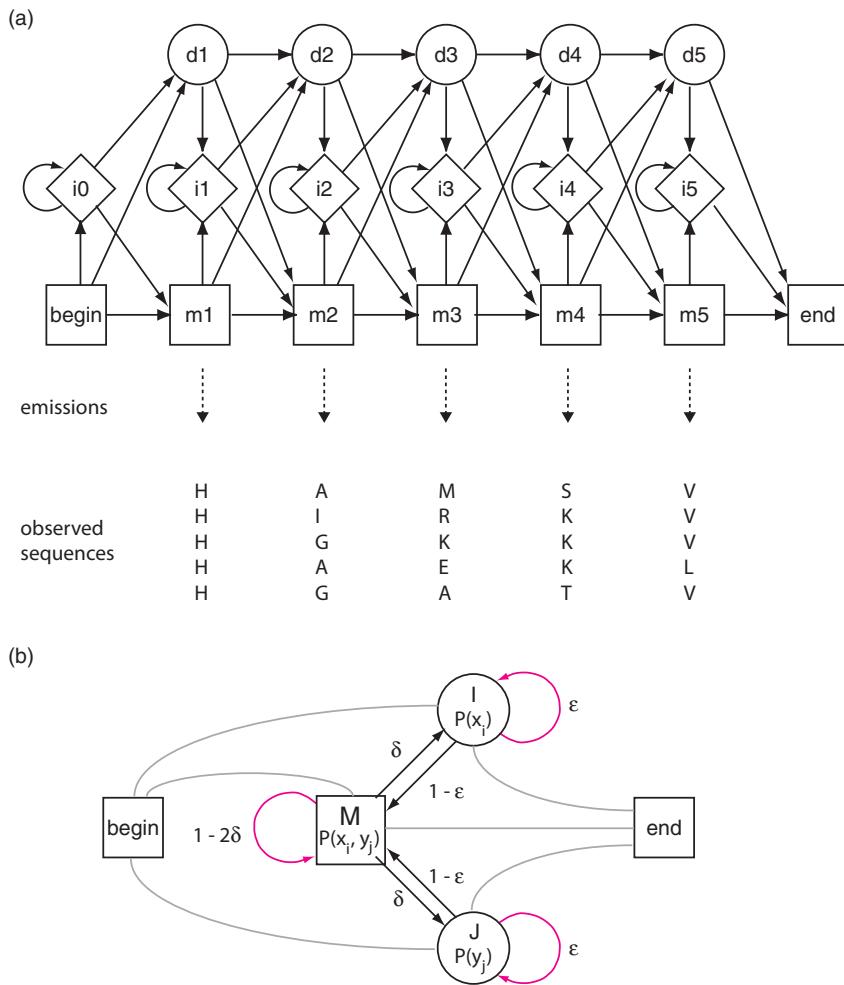
The sequence of an HMM is defined by a series of states that are influenced by two main parameters: the transition probability and the emission probability. The transition



**FIGURE 5.9** An alignment of five globins is shown. The five proteins are a nonsymbiotic plant hemoglobin from rice (*Oryza sativa*) (Protein Data Bank accession 1D8U), human neuroglobin (1OJ6A), human beta globin (2hhbB), leghemoglobin from the soybean *Glycine max* (1FSL), and human myoglobin (2MM1). (b) The probability of each residue occurring in each aligned column of residues is calculated. (c) From these probabilities, a score is derived for any query such as HARTV. Note that the actual score will also account for gaps and other parameters. Also note that this is a position-specific scoring scheme; for example, there is a different probability of the amino acid residue lysine occurring in position 3 versus 4. (d) The probabilities associated with each position of the alignment can be displayed in boxes representing states.

probability describes the path followed along the hidden state sequence of the Markov chain (Fig. 5.10a, solid arrows). Each state also has a “symbol emission” probability distribution for matching a particular amino acid residue. The symbol sequence of an HMM is an observed sequence that resembles a consensus for the multiple sequence alignment. Note that profile HMMs, unlike PSI-BLAST, include probabilities associated with insertions and deletions. The HMM is called a “hidden” model because it consists of both observed symbols (such as the amino acid residues in a sequence modeled by the HMM) and a hidden state sequence which is inferred probabilistically from the observed sequence.

HMMs can also be applied to pairwise alignments (Fig. 5.10b). In addition to beginning and end states, there is a match state  $M$  with emission probability  $x_i, y_j$  for emitting an aligned pair of residues  $i, j$ . State  $I$  has the emission probability  $p_i$  for emitting a symbol  $i$  aligned to a gap; state  $J$  corresponds to residue  $j$  aligned to a gap. The gaps may be extended with probability  $\epsilon$ . The alignment is modeled through a process of choosing sequential states from beginning to end according to the highest transition probabilities, with aligned residues added according to the emission probabilities.



**FIGURE 5.10** The structure of a hidden Markov model. (a) The HMM consists of a series of states associated with probabilities. The “main states” are shown in boxes along the bottom (from begin to end, with  $m_1-m_5$  in between). These main states model the columns of a multiple sequence alignment, and the probability distribution is the frequency of amino acids (see Fig. 5.9d). The “insert states” are in diamond-shaped objects and represent insertions. For example, in a multiple sequence alignment some of the proteins might have an inserted region of amino acids, and these would be modeled by insert states. The “deletion states” ( $d_1-d_5$ ) represent gaps in the alignment. Adapted from Krogh *et al.* (1994) with permission from Elsevier. (b) Pair hidden Markov model (Pair-HMM) for the alignment of sequences  $X$  and  $Y$  having residues  $x_i$  and  $y_j$ . State  $M$  corresponds to the alignment between two amino acids; this state emits two letters. State  $I$  corresponds to a position in which a residue  $x_i$  is aligned to a gap, while state  $J$  corresponds to an alignment of  $y_j$  to a gap. The logarithm of the emission probability function  $P(x_i, y_j)$  at state  $M$  corresponds to a substitution scoring matrix. The transition penalties  $\delta$  and  $\epsilon$  define the transition probabilities.

HMMER software comes with detailed tutorials and documentation to help you run these commands yourself.

### HMMER Software: Command-Line and Web-Based

Profile HMMs are important because they provide a powerful way to search databases for distantly related homologs; HMM methods therefore complement the BLAST search tools. Profile HMMs can define a protein or gene family, and databases of profile HMMs are searchable. Practically, HMMs can be created using the HMMER program (available for Windows, Mac OS/X or Linux platforms; for the example below we use Linux and fol-

low an example in the User’s Guide). You can build a profile HMM with the `hmmbuild` program, which reads a multiple sequence alignment as input.

```
$ ./hmmbuild -h # provides brief help documentation
$ ./hmmbuild globins4.hmm ../tutorial/globins4.sto
```

The `hmmbuild` function creates an HMM (called `globins4.hmm`) using as input a set of four multiply aligned globin proteins. We can search this HMM against a database. We first find a suitable database, in this case human RefSeq proteins in the FASTA format.

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.faa.gz
$ gunzip human.protein.faa.gz
$ wc -l human.protein.faa
302761 human.protein.faa
```

We therefore find a database online at a resource such as NCBI, Ensembl, or UCSC. We download it with `wget`, and decompress it with `gunzip`. We can check how large it is with `ls -lh`, inspect its contents with programs such as `less` or `head`, and count how many rows it has with `wc -l` (this file has 302,761 lines).

Next we search the HMM against the database.

```
$ ./hmmssearch globins4.hmm human.protein.faa > globins4.out
```

We use the `hmmssearch` function, specifying the HMM and the database, and send the results to a file called `globins4.out`. A portion of the result is shown in **Figure 5.11**. This search strategy successfully finds all human globins (and no other proteins). As an exercise at the end of this chapter, we can experiment with building different HMMs and searching different databases. For example, building an HMM from a group of very diverse globin proteins will be more effective when we search for bacterial globins.

By default, for each model that is built the resulting profile HMM is global with respect to the HMM and local with respect to the sequences it matches in a database search. The HMM model does not invoke Needleman–Wunsch (global) and Smith–Waterman (local) algorithms separately, but rather uses a model that has the properties of both (and has sometimes been called “glocal”). You can adjust the sensitivity of a HMMER search by building an HMM that is, for example, local with respect to both the sequence and the HMM, focusing on local alignments rather than on complete domains.

The `hmcalibrate` program matches a set of 5000 random sequences to the profile HMM, fits the scores to an extreme value distribution (Chapter 4), and calculates the parameters that are necessary to estimate the statistical significance of database matches. The profile HMM can then be used to search a database using the `hmmssearch` program.

Sean Eddy and colleagues have introduced two major improvements to HMMER3 software. First, its speed is now comparable to BLAST due to a series of innovations including a “multiple segment Viterbi” heuristic (Eddy, 2011). This is an ungapped version of the algorithm using a profile that produces ungapped alignments. This profile is comparable to that shown in **Figure 5.10a** without insertion and deletion states (all match–match transition probabilities are set to 1.0). This heuristic resembles that of BLAST. Second, a HMMER web server has been introduced, offering search speeds comparable to BLAST (Finn *et al.*, 2011). **Figure 5.12** shows an example of a web-based HMMER search for beta globin matches.

Visit [http://ftp.ncbi.nlm.nih.gov/refseq/H\\_sapiens/mRNA\\_Prot/](http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/) (WebLink 5.19) to find a database of human proteins in the FASTA format (suffix `.faa`). The suffix `.gz` indicates that the file is compressed; the `gunzip` command decompresses it. You can reach this and other databases through the home page of NCBI (search for downloads). When you find the database of interest, copy the link location (right-click on a Windows machine) then use the `wget` program to download the database to your Linux machine. Access it from Windows through software such as PuTTY.

HMMER is available at <http://hmmer.janelia.org/> (WebLink 5.20). It was written by Sean Eddy. The program is designed to run on UNIX, Windows, or MacOS platforms. We discuss how to create multiple sequence alignments (used as an input to HMMER) in Chapter 6.

A HMMER web server is available at <http://hmmer.janelia.org/> (WebLink 5.21). You can also use `hmmalign` software at the Pasteur Institute (<http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::hmmalign>, WebLink 5.22).

```

# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query HMM file:          globins4.hmm
# target sequence database: /mnt/reference/human.protein.faa
# -----


Query:      globins4 [M=149]
Scores for complete sequences (score includes all domains):
--- full sequence ---
  E-value   score   bias     Sequence           Description
  -----  -----
  3.3e-64  216.6   0.0    ref|NP_000509.1| hemoglobin subunit beta [Homo sa
  7e-61   205.8   0.0    ref|NP_000510.1| hemoglobin subunit delta [Homo s
  2.3e-60  204.2   1.3    ref|NP_000508.1| hemoglobin subunit alpha [Homo s
  2.3e-60  204.2   1.3    ref|NP_000549.1| hemoglobin subunit alpha [Homo s
  6.2e-60  202.8   0.3    ref|NP_976311.1| myoglobin [Homo sapiens]
  6.2e-60  202.8   0.3    ref|NP_976312.1| myoglobin [Homo sapiens]
  6.2e-60  202.8   0.3    ref|NP_005359.1| myoglobin [Homo sapiens]
  4.8e-55  186.9   0.0    ref|NP_000175.1| hemoglobin subunit gamma-2 [Homo
  1.4e-54  185.4   0.4    ref|NP_005321.1| hemoglobin subunit epsilon [Homo
  2.1e-54  184.8   0.1    ref|NP_000550.2| hemoglobin subunit gamma-1 [Homo
  4.9e-48  164.2   0.2    ref|NP_005323.1| hemoglobin subunit zeta [Homo sa
  1.7e-40  139.7   0.1    ref|NP_005322.1| hemoglobin subunit theta-1 [Homo
  1.8e-39  136.4   0.2    ref|NP_599030.1| cytoglobin [Homo sapiens]
  5e-35   121.9   0.3    ref|NP_001003938.1| hemoglobin subunit mu [Homo sapi
  3e-08   35.0    0.0    ref|NP_067080.1| neuroglobin [Homo sapiens]
----- inclusion threshold -----
  0.14   13.4   0.0    ref|NP_001371.1| dedicator of cytokinesis protein
  0.25   12.6   0.8    ref|NP_006737.2| sex comb on midleg-like protein
  0.28   12.4   0.8    ref|NP_001032629.1| sex comb on midleg-like protein

```

**FIGURE 5.11** The HMMER program can be used to create a profile HMM using a multiple sequence alignment as input. The program was obtained from <http://hmmer.janelia.org/> and downloaded on a Linux server. Four vertebrate globin proteins were multiply aligned, and the hmmbuild program was used to build a profile HMM. The output of a HMMER search against all human RefSeq proteins includes all known globins. Results may vary when the same database is searched with different HMMs or when other databases are searched.

Source: Howard Hughes Medical Institute. Reproduced with permission from HHMI.

## BLAST-LIKE ALIGNMENT TOOLS TO SEARCH GENOMIC DNA RAPIDLY

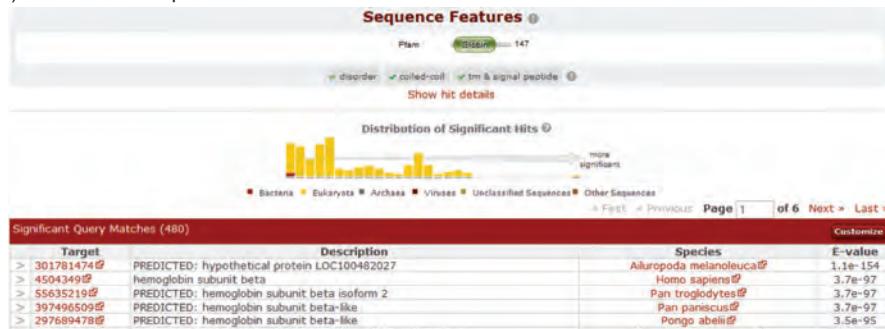
As genomic DNA databases grow in size, it becomes increasingly common to search them using protein queries or long DNA sequences from the same or other species. This is a specialized problem.

We discuss exons and introns in Chapters 8 (on the eukaryotic chromosome) and 10 (on gene expression).

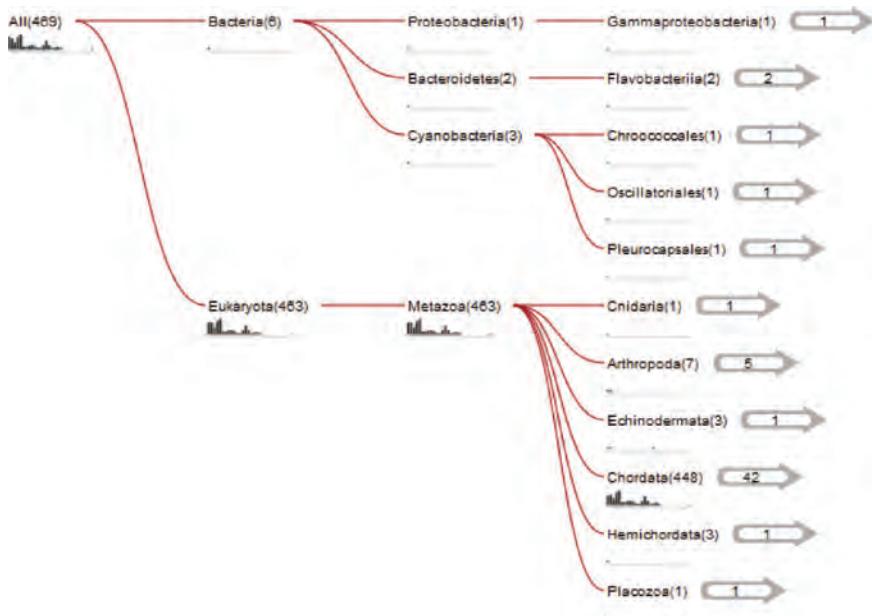
1. The genomic DNA includes both exons (regions corresponding to the coding sequence) and introns (intervening, noncoding regions of genes). Ideally, an alignment tool should find the exons in genomic DNA.
2. Genomic DNA often has sequencing errors that should be taken into account.
3. We may want to compare genomic DNA between closely related organisms such as mouse and rat or distantly related organisms, for example fish and tomato. In any comparison, genomic changes may have occurred such as deletions, duplications, inversions, or translocations. Algorithms should solve problems such as the alignment of 10 million base pairs (Mb) containing a 1 Mb inversion.
4. Algorithms are needed to find small differences between DNA sequences, such as single-nucleotide polymorphisms (SNPs; Chapter 8).

Several BLAST-like algorithms have been written to address these needs. The algorithms are available in programs that are useful for pairwise alignments and/or searches

(a) HMMER web output



(b) HMMER phylogenetic output



**FIGURE 5.12** The HMMER program is available online. Human beta globin protein sequence (NP\_000509) was submitted to the server. The output includes descriptions of significant matches and several tools showing the phylogenetic distribution of hits.

Source: Howard Hughes Medical Institute, 2013. Reproduced with permission from HHMI.

of entire databases with a query. We illustrate several of these programs using a query of 50,000 base pairs from human chromosome 11p (the short arm of chromosome 11). This region contains five globin genes (*HBE1*, *HBD*, *HBB*, *HBG2*, and *HBG1* corresponding to  $\epsilon$ ,  $\delta$ ,  $\beta$ ,  $\gamma 2$ , and  $\gamma 1$  globins) and a beta globin pseudogene (*HBBP1*). A convenient way to view this region of 50,000 base pairs is to visit the UCSC Genome Browser.

### Benchmarking to Assess Genomic Alignment Performance

Throughout this book we describe benchmark datasets that allow the specificity and sensitivity of a method to be assessed. We discussed this for PSI-BLAST and DELTA-BLAST in the earlier sections of this chapter. For the multiple sequence alignment of proteins (Chapter 6), several databases contain information on the trusted members of homologous protein families based on their three-dimensional structures as rigorously determined by X-ray crystallography. In finding genes in genomic DNA (Chapter 8), we describe the EGASP project that provides a “gold standard” for assessing gene

The UCSC Genome Browser is available at <http://genome.ucsc.edu> (WebLink 5.23). Set the genome to human (GRCh37 build), and enter chr11:5,245,001–5,295,000 to specify the genomic position spanning these 50 kb.

For some software such as BLAT (see below), the query cannot be longer than 25,000 base pairs; files with both 50 kilobase pairs (kb) and 25 kb queries are available in Web Documents 5.3 and 5.4.

BLASTZ is now called LASTZ by its authors, although various publications and websites continue to use either name.

The Pollard *et al.* data are available at <http://rana.lbl.gov/AlignmentBenchmarking/data.html> (WebLink 5.24). ROSE is available at <http://bibiserv.techfak.uni-bielefeld.de/rose/> (WebLink 5.25).

Other implementations of PatternHunter use slightly different models such as 111010010100110111. PatternHunter software is commercially available at <http://www.bioinfor.com> (WebLink 5.26).

prediction software. In the case of tools for the alignment of genomic DNA, often including large regions of noncoding DNA, there are no experimentally derived databases of correct alignments of large genomic regions. Nonetheless, in each case sensitivity and specificity are evaluated. For example, Schwartz *et al.* (2003) compared human to mouse genomic DNA using BLASTZ, finding that about 39% of the human and mouse genomes could be aligned. Then they reversed the mouse sequence (without complementing it), obtaining a mouse test set with the same size and compositional complexity as the real mouse sequences; only 0.164% of the human sequence now aligned to this reversed set.

A benchmark dataset for noncoding genomic DNA can be created using a strategy of computational simulations rather than using experimentally obtained standards. Pollard *et al.* (2004a) examined noncoding DNA in the fruit fly *Drosophila melanogaster* (a well-characterized genome that lacks many of the ancestral repeats and lineage-specific transposition events found in vertebrates), assembling a group of 10 kilobase fragments. They used the random model of sequence evolution (ROSE) software package (Stoye *et al.*, 1998) to create a set of simulated sequences having a variety of insertions, deletions, point substitutions, and interspersed blocks of constrained sequences, that is, variations in evolutionary rate estimated across a range of species divergence times. They then tested the ability of eight pairwise genomic alignment tools including BLASTZ (Pollard *et al.*, 2004b). They concluded that global alignment tools (such as LAGAN) have the highest overall sensitivity, while local alignment tools (such as BLASTZ) more accurately align variable regions.

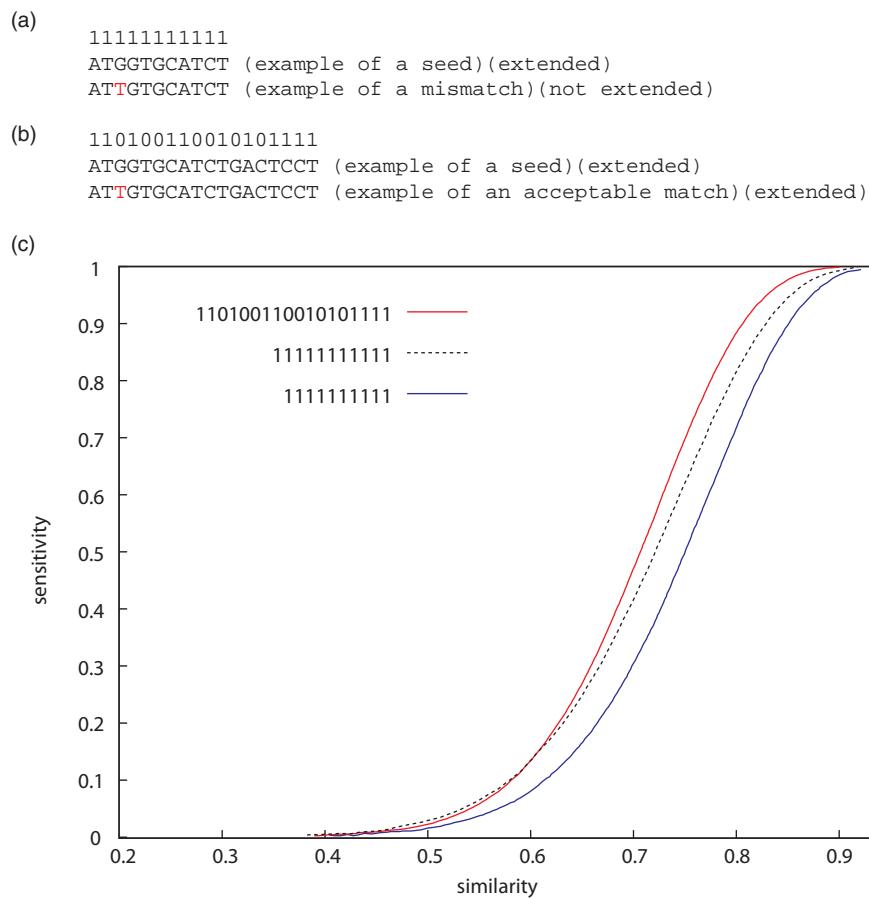
### PatternHunter: Nonconsecutive Seeds Boost Sensitivity

BLASTN uses a short seed, typically consisting of a word size of 11 consecutive nucleotides. Exact matches are identified in a DNA database and then extended into longer alignments (Chapter 4). PatternHunter (Ma *et al.*, 2002) achieves improvements in both speed and sensitivity by creatively using nonconsecutive letters as seeds. If we denote 1 for a match and 0 for a mismatch, the BLASTN word ( $w = 11$ ) has a form 11111111111 (Fig. 5.13a). No mismatches are tolerated. For PatternHunter, the pattern of its seed is 11010011001010111 (Fig. 5.13b). There are still 11 matches, but they are distributed over a range of 18 nucleotide positions. A query may align with a database entry having in its sequence a mismatch corresponding to a 0 position (or any of the seven 0 positions). In this case, the mismatch is ignored and an extension can still occur. The reason for the improved sensitivity of a nonconsecutive mismatch becomes clear if we consider a particular region of length 64 nucleotides having 70% identity, as described by Ma *et al.* (2002). For BLASTN the probability of having at least one hit is 0.30, while for the nonconsecutive seed model the probability is 0.466. This is illustrated in Figure 5.13c which shows greater sensitivity for a given amount of similarity. Within some region of 64 nucleotides, the consecutive seed model is disrupted for a mismatch across a group of adjacent seeds which all share a group of 1s. For the nonconsecutive seed model, the seed matches occur at different positions, helping to increase sensitivity. This occurs because fewer bases are shared between neighboring seed matches, making the matches more independent than for a consecutive seed strategy.

The innovative approach to seed models introduced by Ma *et al.* (2002) has been adopted by other homology search algorithms including BLASTZ and MegaBLAST, discussed in the following sections.

### BLASTZ

BLASTZ was developed to align human and mouse genomic DNA sequences based on modifications of the gapped BLAST program (Schwartz *et al.*, 2003). It is useful for



**FIGURE 5.13** Nonconsecutive seeds in PatternHunter improve its sensitivity in database searches. (a) In a typical BLASTN search with a word size of 11, the matching nucleotides occur consecutively and may be represented with a series of 1s. An example of a seed from a database query is shown; if the database target has a single-nucleotide substitution, there is no perfect match and an extension does not occur. (b) The Ma *et al.* (2002) approach uses nonconsecutive letters as seeds. The values 1 correspond to matches, while the 0 positions are ignored. For some nucleotide mismatches, as shown here, the seed nonetheless matches successfully and extension occurs. (c) A plot of similarity versus sensitivity for the consecutive model with 10 letters (blue line), 11 letters (black dotted line), or the spaced model having 11 matches (red line). The sensitivity is higher over a range of similarities for the nonconsecutive seed approach. Adapted from Ma *et al.* (2002), with permission from Oxford University Press on behalf of the International Society for Computational Biology.

comparing long genomic sequences from a variety of organisms. As for gapped BLAST, it searches for short near-exact matches, extends them without allowing gaps, and then performs further extensions using dynamic programming. BLASTZ functions as follows (Schwartz *et al.*, 2003):

1. Lineage-specific interspersed repeats (further described in Chapter 8) are removed from both sequences. To improve its execution speed, when one region of the human genome aligns to multiple regions of the mouse genome, that human segment is dynamically masked. This was helpful in processing regions of the mouse genome that have a large number of highly related genes (e.g., zinc finger genes or olfactory receptor genes).
2. Matches are identified using a word size of 12 (either identically matching or allowing one transition), and are extended without allowing gaps. When the score exceeds

Transitions are substitutions between the purines (the nucleotides A↔G) and between the pyrimidines (C↔T). Transversions are substitutions between purines and pyrimidines (A↔C, A↔T, G↔C, or G↔T). Transitions occur more commonly than transversions (see Chapter 7).

We return to this aligned region in Chapter 6 to view its multiple sequence alignment.

some threshold, extensions are repeated with gaps allowed. Following the innovation introduced in PatternHunter, BLASTZ uses a seed of 12 matches in 19 consecutive positions having the string 1110100110010101111.

- Step (2) is repeated for regions adjacent to successful alignments using a lower (more sensitive) word size, such as 7.

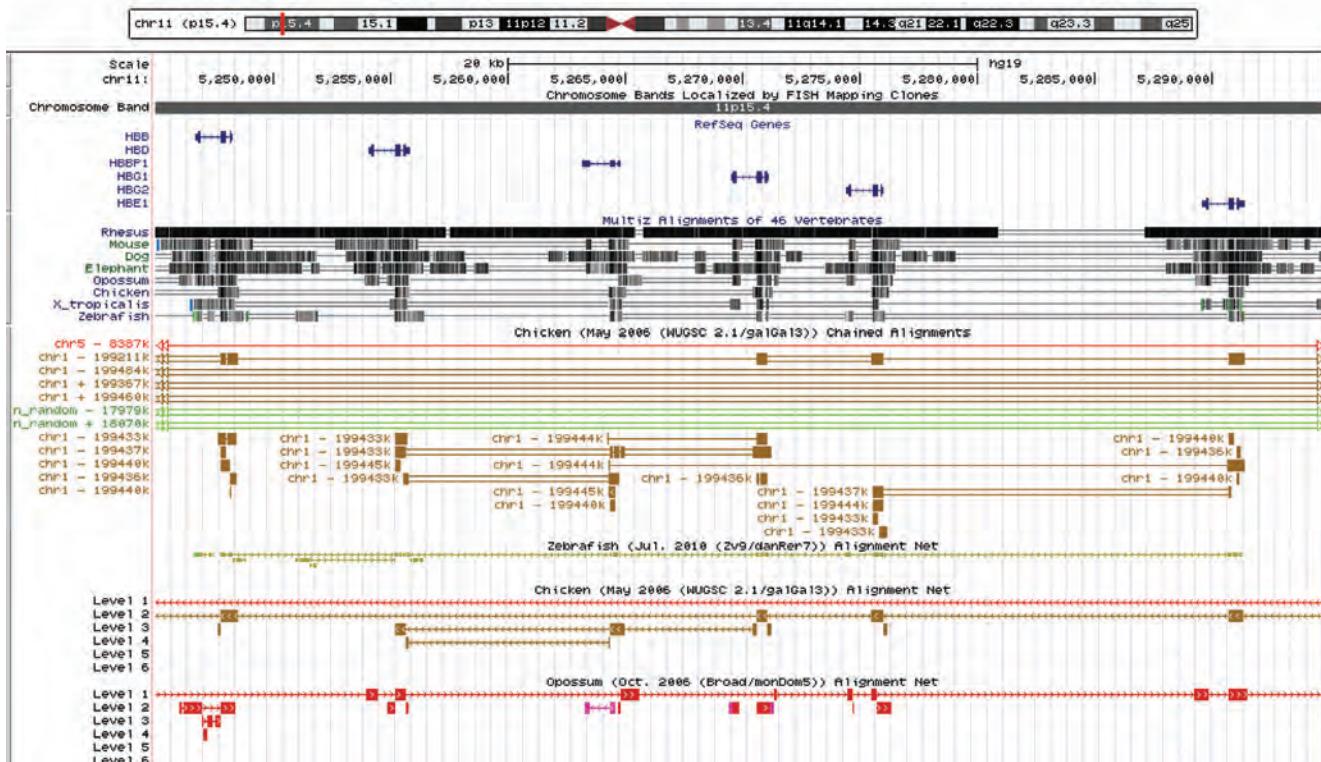
BLASTZ was used to align the mouse genome (2.5 billion base pairs or gigabases (Gb)) with the human genome (2.8 Gb of sequence; Schwartz *et al.*, 2003). To accomplish this, the human genome was divided into ~3000 segments of about 1 megabase (1 Mb) each, while the mouse genome was divided into ~100 segments of 30 Mb. BLASTZ alignments between a variety of species are represented as tracks on the UCSC Genome Browser. For the 50,000 base pairs containing the human HBB region, examples of the features which can be viewed are (Fig. 5.14): (1) the chromosome band (11p15.4); (2) the genes in the region (*HBB*, *HBD*, *HBG1*, *HBE1*); (3) a

### UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr11:5,245,001-5,295,000 50,000 bp. enter position, gene symbol or search terms

go



**FIGURE 5.14** Precomputed alignments of genomic sequence, aligned by BLASTZ, can be visualized using the UCSC Genome Browser. The genome browser is set to the GRCh37/hg19 assembly of the human genome, and 50,000 base pairs on chromosome 11p are displayed. The tracks include the following: (1) base pair positions; (2) the chromosome band (11p15.4); (3) RefSeq genes in this region (there are six); (4) vertebrate Multiz alignment and conservation (precomputed BLASTZ results showing an overall conservation score as well as alignments from human to a subset of 46 available genomes including rhesus monkey, mouse, dog, elephant, opossum, chicken, frog, and zebrafish); (5) chicken chained alignments, showing BLASTZ alignment results to chicken; and (6) alignment nets showing a summary of the highest-scoring alignments between genomic DNA from humans and other species (zebrafish, chicken, opossum) using BLASTZ. Note that the UCSC Genome Browser annotation tracks can be interactively added or removed, and information can be displayed in a more or less compressed form. Here it is evident that the most highly conserved segments in this 50 kilobase pair region correspond to the globin genes, while intergenic regions tend to be less well conserved.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

vertebrate conservation track, showing in particular high-scoring regions of conservation across multiple species at the location of the globin genes (and in some non-coding regions, having regulatory functions); (4) chicken chains, showing the regions aligned well by BLASTZ; and (5) nets for zebrafish (clade Teleostei), chicken (clade Dinosauria), and opossum (clade Mammalia), showing a summary of the best-scoring chains. By clicking on the chains or nets you can access the pairwise sequence alignments. In this example, there are dozens of distinct blocks of aligned genomic DNA sequence for each comparison between species, separated by regions that could not be reliably aligned.

BLASTZ has been employed for various projects including an analysis of 13 million base pairs of DNA from the extinct woolly mammoth (*Mammuthus primigenius*) to the modern African elephant (*Loxodonta africana*) (Poinar *et al.*, 2006) and an analysis of transcription units on human chromosome 22 (Lipovich and King, 2006). The BLASTZ program is available for local use.

You can obtain LASTZ and BLASTZ at Webb Miller's web site at Pennsylvania State University, [http://www.bx.psu.edu/miller\\_lab/](http://www.bx.psu.edu/miller_lab/) (WebLink 5.27).

## Enredo and Pecan

In Chapter 6 we discuss software for the multiple sequence alignment of genomic DNA that is used at Ensembl: Enredo (to create a large-scale homology map for pairwise whole-genome alignment); Pecan (to make multiple sequence alignments using the principle of consistency); and Ortheus (to reconstruct ancestral sequences; Paten *et al.*, 2008). The resulting alignments are more accurate than those produced by other software by several criteria.

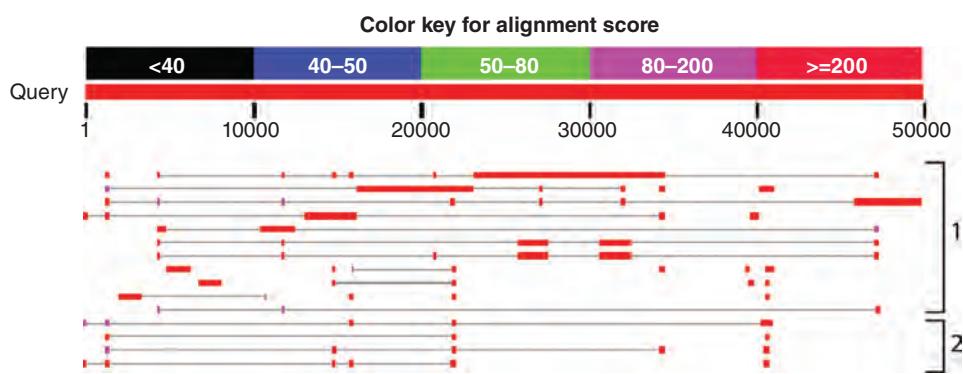
## MegaBLAST and Discontinuous MegaBLAST

MegaBLAST is an NCBI program optimized for the rapid alignment of very large DNA queries (Zhang *et al.*, 2000). The program offers a default word size of 28 (and can accommodate a word size as large as 256), in contrast to the default word size of 11 for BLASTN. This greatly increases the speed of MegaBLAST, since the word size corresponds to the minimal length of an exact match required to initiate an extension. With its smaller word size, BLASTN is more sensitive but slower. For MegaBLAST you can also specify the percent identity threshold to be reported (e.g., only alignments sharing values such as 99%, 90%, or 80% identity) as well as the corresponding match and mismatch scores. For example, for sequences sharing 95–99% identity, a match score of +1 and mismatch of −3 is applied; for alignments sharing 85–90% identity the mismatch score is instead set to −2. Non-affine gapping parameters are used: the gap opening penalty is 0 (causing MegaBLAST to have alignments with more gaps but with the benefit of enhanced speed), and the gap extension penalty is based on the selected match and mismatch scores.

Discontiguous MegaBLAST is a related algorithm at NCBI that is designed to align more distantly related genomic sequences. It employs a “discontiguous word” strategy of Ma *et al.* (2002) described for PatternHunter. It is useful for comparing relatively divergent sequences (e.g., from different organisms).

We can demonstrate the use of MegaBLAST selecting 50,000 base pairs of DNA from the short arm of human chromosome 11 as a query, and selecting an orang-utan (*Pongo pygmaeus*) nonredundant nucleotide collection (abbreviated nr/nt). This query region contains five globin genes (*HBE1*, *HBD*, *HBB*, *HBG2*, and *HBG1*) and a beta globin pseudogene (*HBBP1*). Using the default settings of MegaBLAST (word size 28, match score +1, mismatch score −2, and gap opening and extension penalties zero), we find matches ranging from about 80% to 97% nucleotide identity to the human genomic DNA query (Fig. 5.15).

Try this using *Macaca mulatta* (taxid:9544), the *Anubis* baboon (taxid:9555), or the bonobo *Pan paniscus* (taxid:9597). You can also try varying the word size up to as large as 256.



**FIGURE 5.15** Megablast is an NCBI tool specialized for rapidly searching long DNA queries against genomic DNA databases. Here 50 kilobases of DNA spanning the human beta globin genes were used as a query restricted to nonredundant sequences of *Pongo pygmaeus*. Matches are to orang-utan globin genes and pseudogenes (area 1) as well as to repetitive sequences (area 2).

Source: Megablast, NCBI.

### BLAST-Like Tool (BLAT)

BLAT is accessible on the web at <http://genome.ucsc.edu> (WebLink 5.28).

BLAT is designed to perform extremely rapid genomic DNA searches (Kent, 2002). Like SSAHA2 (see section of this title below), the BLAT algorithm is in some ways a mirror image of BLAST. BLAST parses a query sequence into words and then searches a database with words above a threshold score. Two proximal hits are extended. BLAT parses an entire genomic DNA database into an index of words. These words consist of all non-overlapping 11-mers in the genome (excluding repetitive DNA sequences). BLAT then searches a query using words from the database. The BLAT strategy of database indexing has been adopted by SSAH2 (see below) and subsequently by Megablast (Morgulis *et al.*, 2008).

BLAT offers a variety of additional features (Kent, 2002):

- While BLAST triggers an extension with two hits, BLAT triggers extensions on multiple strong hits.
- BLAT is designed to find matches between queries that share 95% nucleotide identity or more. While it is in some ways similar to the Megablast, Sim4, and SSAHA programs, it is orders of magnitude faster.
- BLAT searches for intron–exon boundaries, essentially building a model of a gene structure. It uses each nucleotide derived from an mRNA query once (as is appropriate from a biological perspective), rather than searching for highest scoring segment pairs.

A BLAT search using human beta globin protein as a query is shown in Figure 5.16. Human genomic DNA is translated in six frames, and the best match is to the *HBB* gene on chromosome 11 that encodes the HBB protein. By adjusting the coordinates on the genome browser to display 50,000 base pairs in the beta globin locus region, we can see that the BLAT search also resulted in matches to genes encoding other closely related globin proteins.

### LAGAN

LAGAN (Limited Area Global Alignment of Nucleotides) is a pairwise alignment tool for genomic DNA (Brudno *et al.*, 2003). We discuss its companion Multi-LAGAN in Chapter 6 (multiple sequence alignment). LAGAN proceeds in three steps to create a global pairwise alignment (Fig. 5.17a). First, it generates local alignment between two sequences, therefore identifying a set of anchors (Fig. 5.17b). This strategy permits the matching of

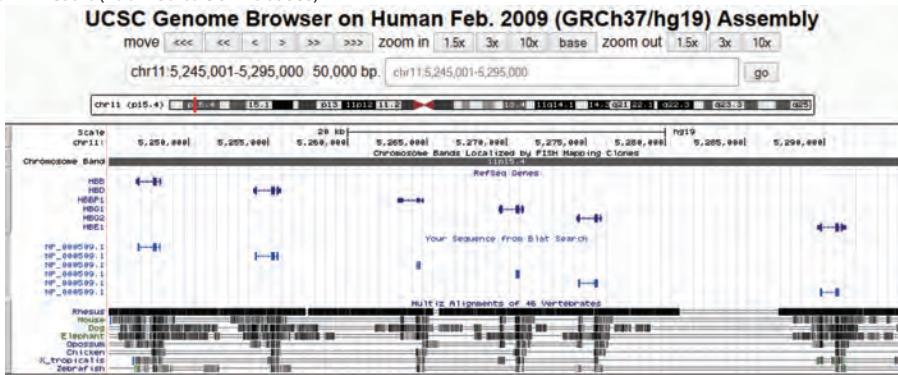
(a) BLAT query (protein or DNA)

### BLAT Search Genome

Genome: Human      Assembly: Feb. 2009 (GRCh37/hg19)      Query type: BLAT's guess      Sort output: query.score      Output type: hyperlink

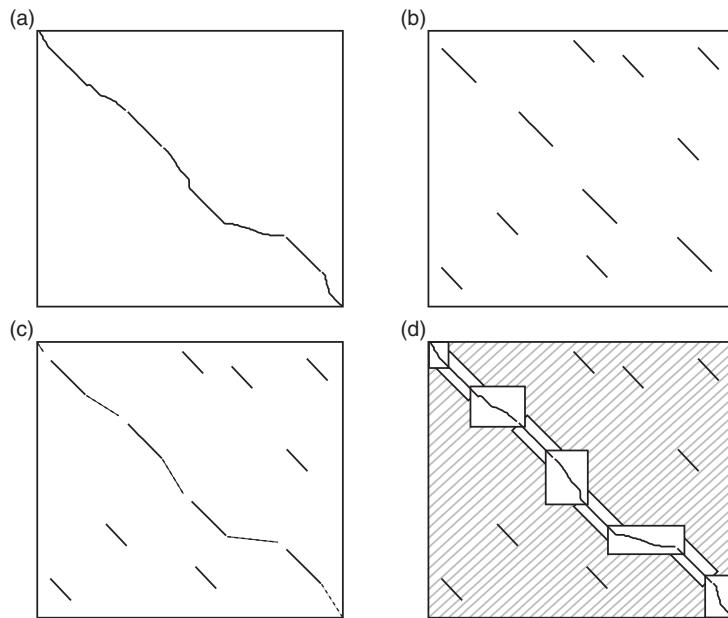
```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTEEEKSAVIALWNGKVNVDEVGSEALGRLLVVYPNITQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLNDNLKGTFATLSELHCDKLHVDPENFRLLGIVLVCVLAHHFGKEFTPVQAAAYQKVVAGVAN
ALAHKYH
```

(b) BLAT result (zoomed to 50 kilobases)



**FIGURE 5.16** The BLAST-Like Tool (BLAT) at the UCSC Genome Bioinformatics website. (a) DNA or protein sequence can be pasted or uploaded from a text file. The output settings include a hyperlink option to access the Genome Browser view. (b) The Browser view includes a custom track (“Your Sequence from BLAT Search”) which shows a series of matches to five globin genes (*HBB*, *HBD*, *HBG1*, *HBG2*, and *HBE1*) in a 50,000 base pair segment of human chromosome 11.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.



**FIGURE 5.17** The LAGAN algorithm for pairwise alignment of genomic DNA sequences. (a) LAGAN uses a combined local/global strategy to produce a global alignment of two sequences. The *x* and *y* axes correspond to the physical position (e.g., chromosomal coordinates) of two DNA queries. (b) A local alignment search strategy identifies conserved regions (solid downward-slanting lines). Note that an inversion in one of the sequences would be represented by a line having a positive slope. (c) Locally aligned segments are joined in chains. Anchors, or maximally scoring ordered subsets of locally aligned regions, are identified and joined to create a rough global map. (d) LAGAN computes an optimal alignment within the boxed areas, ignoring the hatched regions. Adapted from Brudno *et al.* (2003).

SSAHA2 is available at the Ensembl web server <http://www.ensembl.org> (WebLink 5.29). The SSAHA2 home page is available at <http://www.sanger.ac.uk/resources/software/> (WebLink 5.30). A hash table contains data (e.g., a list of words having a length of 14 nucleotides in a DNA database) and associated information (e.g., the positions in genomic DNA of each of those words).

multiple short inexact words rather than long exact words. Second, LAGAN creates a rough global map consisting of a maximally scoring ordered subset of the alignments (anchors). Third, it computes a final global alignment, restricting the operation to the limited area defined by the rough map. This focused search strategy avoids the inefficiency of performing a global alignment with the Needleman–Wunsch algorithm on the two input sequences.

## SSAHA2

Sequence Search and Alignment by Hashing Algorithm, abbreviated SSAHA, is designed to search large DNA databases very rapidly (Ning *et al.*, 2001). It is commonly used to map next-generation sequence reads to a reference genome. An input file includes the genomic reference sequence (such as the human genome) in the FASTA format. SSAHA2 converts such a DNA database into a hash table with a fixed word length (user-selected  $k$ -mers). This hash table can then be searched quickly for matches by pairwise alignment. Sequencing reads in the FASTQ format (introduced in Chapter 9) are mapped against the genomic reference. Exactly matching seeds in the sequence reads are identified and aligned using a modified Smith–Waterman algorithm.

## ALIGNING NEXT-GENERATION SEQUENCE (NGS) READS TO A REFERENCE GENOME

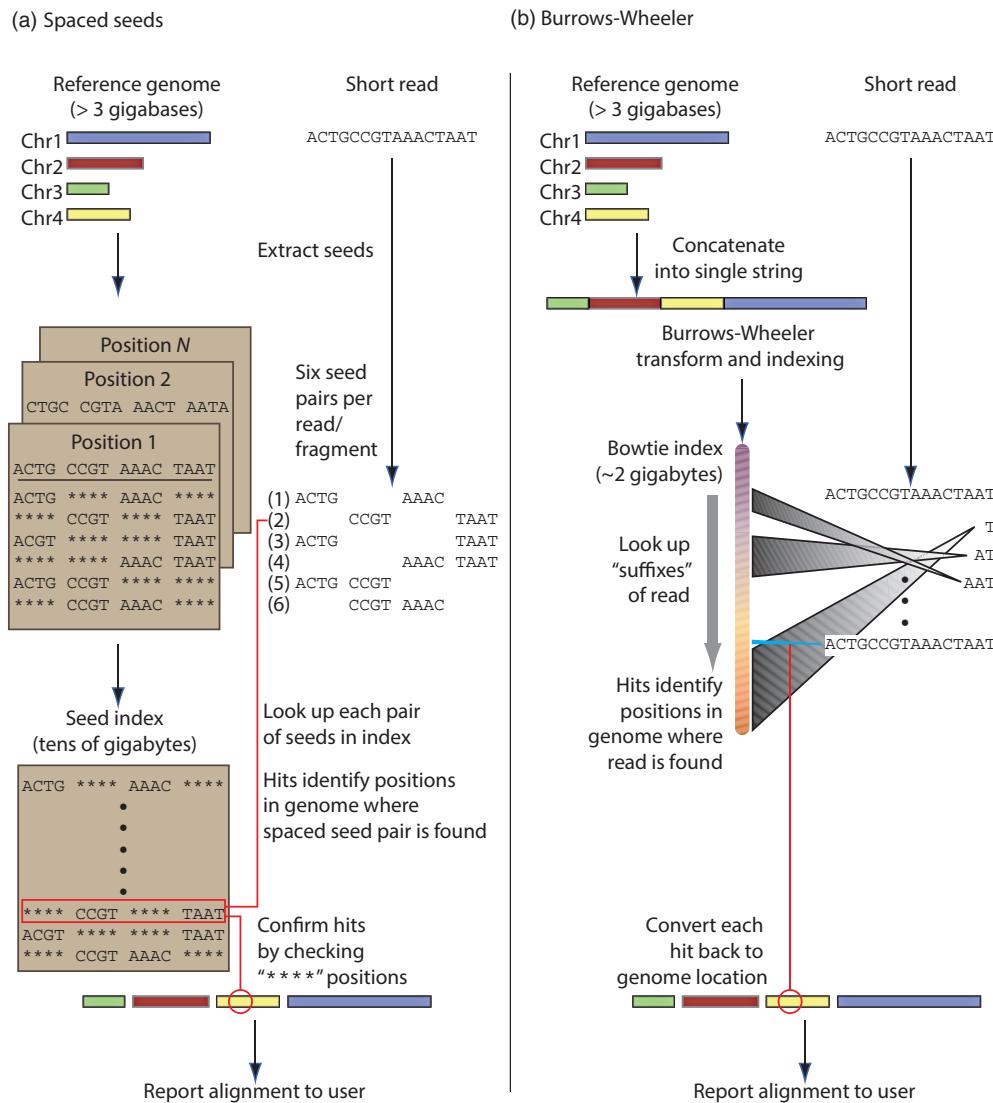
Beginning in the 1970s and continuing to the present, dideoxynucleotide sequencing (also called Sanger sequencing) has been an important way to determine the nucleotide sequence of DNA. Next-generation sequencing (NGS) was introduced in 2005, beginning a technological revolution that has massively increased available DNA sequence (see Fig. 2.3; Chapter 9). When we sequence an individual haploid human genome, which has ~3 billion bases of DNA, it is necessary to obtain adequate depth of coverage such as an average of 30-fold redundancy for each base. (This is necessary to obtain reliable base calls and because reads are not distributed evenly across the genome.) For 90 billion base pairs of sequence, each read may be 100 base pairs in length. In such a case there would be 900 million (i.e.,  $9 \times 10^{11}$ ) reads. We address the technology by which these reads are generated and perform alignment in Chapter 9. Here we ask: how are these reads aligned to a reference human genome? The reference genome is typically available in the FASTA format. The output we seek is a set of genomic coordinates for each read.

In performing alignment we need to consider matches and mismatches. Not all reads will map to the genome at unique positions. In some cases this is because of duplications of genomic regions (e.g., ~5% of the human genome has segmental duplications and about half the genome has other kinds of repetitive DNA). It is also expected that each genome is likely to have single-nucleotide variants, as well as variants that may be due to technical error.

We must also consider speed. Both genome sizes and cumulative read sizes are so large that dynamic programming (e.g., with the Smith–Waterman algorithm) is too slow to be feasible. Some form of indexing is therefore required. Two of the main approaches to alignment are those based on hash tables and suffix trees (Fig. 5.18) (Trapnell and Salzberg, 2009; Li and Homer, 2010).

### Alignment Based on Hash Tables

Hash table indexing adopts the seed-and-extend strategy that we described for BLAST (Fig. 4.12). The approach is outlined in Figure 5.18a. There are two inputs: a reference genome and a large set of short reads. Hash table indexing begins with the approach of BLAST: the positions of  $k$ -mers (e.g., 11-mers using the default BLASTN word size) are stored in a hash table and scanned for  $k$ -mer exact matches (seeds) which are then



**FIGURE 5.18** Two strategies by which short read aligners take a large number of short reads (e.g., 500 million reads, each of length 150 base pairs) and align them to a reference genome (e.g., the haploid human genome reference sequence of ~3 billion base pairs). (a) Spaced seed indexing algorithms use hash tables. The reference genome and the short reads are cut into equal-sized segments called seeds. Seeds from the short reads are paired and stored in a lookup table (hash table) and used to scan the reference. Matches to the seed index have an assigned genomic location. (b) The Burrows–Wheeler transform is used to efficiently represent the reference genome in software such as BWA2 and Bowtie. The reference genome is concatenated into a string, transformed using the Burrows–Wheeler transform (see Fig. 5.19), and indexed. Reads are aligned beginning with the base at the 3' end and continuing towards the 5' end. This approach is very fast relative to spaced seed approaches.

Source: Redrawn from Trapnell and Salzberg (2009). Used with permission from Macmillan Publishers.

extended using dynamic programming. Spaced seeds are commonly used because they offer increased sensitivity.

One of the earliest programs to use this approach was MAQ (Li *et al.*, 2008). It builds multiple hash tables to index the reads, and scans the reference database against the hash tables to identify hits. Using multiple hash tables ensures that all reads having zero, one, or two mismatches will be identified. (Consider a 16 base pair read split into four smaller seeds. If there are no mismatches, all four seeds will align. If there is one mismatch, three

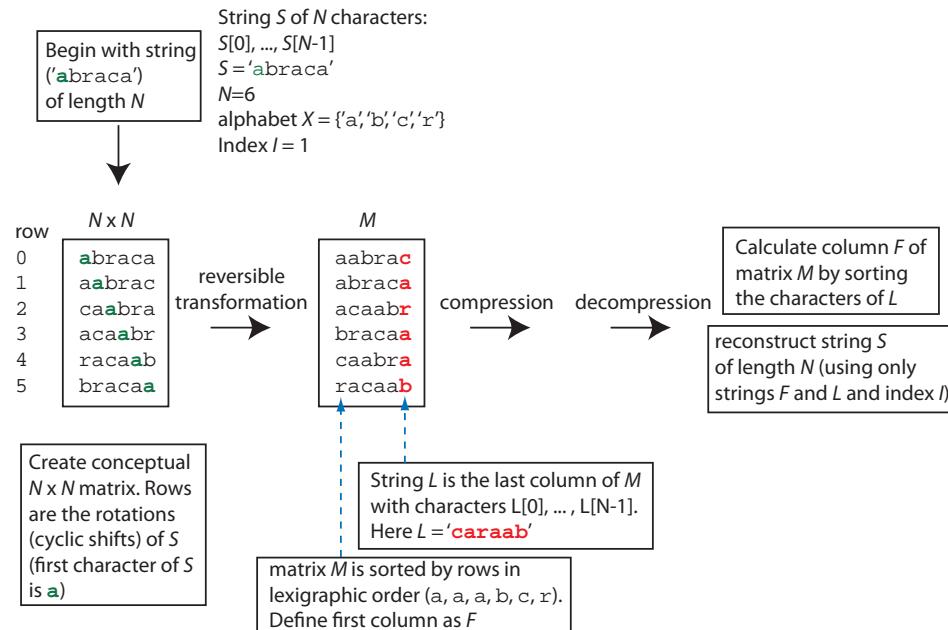
seeds will align perfectly; if there are two mismatches, either two or three seeds will align perfectly. By aligning all six possible seed pairs to the reference genome, the location of that read can be mapped allowing as many as two mismatches. Given three mismatches, 57% of the hits can still be identified.)

Other software tools adopting a hash table approach include ELANDv2 (alignment software from Illumina) and SSAHA2. A limitation of this approach is that it can require tens of gigabytes of memory to store the indexed reads.

### Alignment Based on the Burrows–Wheeler Transform

A faster approach to alignment is offered by suffix trees and suffix arrays. Two of the most popular aligners are BWA and Bowtie2, each of which matches sequences against a reference genome. BWA uses BWA-backtrack for short sequence reads (up to 100 base pairs) (Li and Durbin, 2009) while BWA-MEM is more accurate for longer reads (Li and Durbin, 2010). Bowtie2 is an ultrafast, memory-efficient aligner (Langmead *et al.*, 2009; Langmead and Salzberg, 2012). Both BWA and Bowtie take into consideration read lengths, sequencing error rates, gap penalties, and local versus global alignment of reads.

The key feature of this class of aligners is a method used to index a reference genome as large as the human genome into <2 Gb of memory. The reference genome sequence is first modified and compressed using the Burrows–Wheeler transform (BWT). This is a lossless method, that is, one that allows data to be compressed then fully retrieved in the original decompressed form. We can explain the transform using the example provided in the original article from Burrows and Wheeler (1994; Fig. 5.19). Given a string of characters we first build an  $N \times N$  matrix in which each row corresponds to a cyclic shift or rotation of the sequence. We reorder this into a new matrix  $M$  having rows sorted by the first character of each line. The first column is defined as  $F$  and the last column as  $L$ .



**FIGURE 5.19** The Burrows–Wheeler Transform (BWT). A string, such as the genomic DNA sequence of a reference genome, undergoes compression transformation then decompression to recover the original string. We begin with an input of a string  $S$  of  $N$  characters ( $N = 6$  here). We create an  $N \times N$  matrix consisting of cyclic rotations of  $S$ , then sort lexicographically. Following compression and later decompression, the original string can be recovered. Adapted from Burrows and Wheeler (1994).

This matrix can be effectively compressed and, surprisingly, it can then be reconstructed using only the information in strings  $F$  and  $L$  as well as an index. The BWT method does not itself compress data, but it stores data in a way that makes compression very fast and efficient.

Why is the BWT effective? Burrows and Wheeler (1994) give the example of a text document sorted using the BWT, having many instances of the word “the.” When the list of cyclic shifts is sorted, all the rotations starting with “he” will sort together, and many will end in “t.” Any local region of string  $L$  is likely to have a disproportionately high number of a limited set of characters. This facilitates compression and decompression.

When short reads are mapped to a BWT-transformed genome, the 3' base is first searched against the genome (the wide swath corresponding to the single T in **Figure 5.18b** corresponds to a T matching many locations of a genome). Next genomic positions matching AT are identified, then AAT, until finally the entire short read is assigned one or more genomic positions. If perfect matches are not found, BWA and Bowtie2 can backtrack to tolerate mismatches (tolerating two mismatches in a read by default). Bowtie (the predecessor to Bowtie 2) is 30-fold faster than MAQ.

## PERSPECTIVE

While BLAST searching has emerged as a fundamental tool for studying proteins and genes (Chapter 4), many specialized BLAST applications have also been developed. These applications include variant algorithms (such as the PSSM of DELTA-BLAST and the hidden Markov models of HMMER) and specialized databases (such as a variety of organism-specific databases). Currently, a BLASTP search using human beta globin as a query fails to identify human myoglobin as a significant match. In contrast, using DELTA-BLAST or HMMER, myoglobin is easily detected. This highlights the need for position-specific scoring matrices as well as databases built upon HMMs. We highlight one such database, Pfam, in Chapters 6 and 12.

The exponential rise in DNA sequence data (**Fig. 2.3**) presents us with massive amounts of information about genes and proteins. BLAST is not able to search large amounts of genomic DNA, and alternative strategies include the use of longer word sizes (as in Mega-BLAST), spaced seeds, and indexing of databases and/or queries. Short-read aligners are specialized for aligning tens or hundreds of millions of short reads to a reference genome. Typical uses are to identify single-nucleotide variants or structural variants in a genome. We use a short-read aligner in Chapter 9. These tools will continue to be fundamentally important to biology for many years to come, especially as the pace of genomic sequencing continues to accelerate.

## PITFALLS

As with any bioinformatics problem, it is essential to define the purpose of a database search. What are you trying to accomplish? Once you have decided this, you can select the appropriate database and search algorithm.

For PSI-BLAST (and DELTA-BLAST), the biggest problem is obtaining false positives. Once a spurious sequence has been detected that is better than some expect value cutoff, it will be included in the PSSM for the next iteration. This iteration will almost certainly find the spurious sequence again, and will probably expand the number of database matches. To avoid this problem:

- Inspect the results for apparently spurious database matches. If you see them, remove such spurious matches by deselecting them.
- Adjust the expect value as appropriate.

- Perform “reverse” searches in which you evaluate a potentially spurious PSI-BLAST result by using that sequence as a query in a BLAST search.
- Further evaluate a marginal database match by performing pairwise sequence alignment as described in Chapter 3.

For PHI-BLAST, the most common problem encountered is that new users do not have a feel for the rules involved in creating a PHI-BLAST pattern. The best approach is to practice using a variety of signatures.

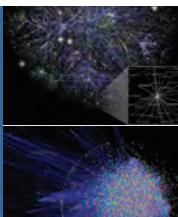
For studies of genomic DNA (as for any bioinformatics problem), it is important to select the appropriate tool for the task at hand. The various tools we described offer trade-offs in speed and sensitivity. Some are designed to analyze regions of a particular length or composition.

## ADVICE FOR STUDENTS

We discussed a series of programs (e.g., PSI-BLAST, DELTA-BLAST, HMMER) for protein (or in some cases, DNA) database searches. Take your favorite protein (or take beta globin) and get a feel for the performance of each tool. How did the developers of these tools measure the sensitivity and specificity? Consider a small family of well-characterized proteins (such as the globins) in which you have a clear definition of true positive and true negative (as well as false positive and false negative) findings. Read the key papers on each of these search tools. Approximately how many authentic database matches are likely to be missed with each tool?

## WEB RESOURCES

We have described different kinds of BLAST and related search tools, including organism-specific databases for BLAST searching, BLAST sites that focus on specialized molecules, and alternative algorithms for database searching including DELTA-BLAST, HMMER, MegaBLAST, and BLAT. Links to these resources are provided at <http://www.bioinfbook.org/chapter5>.



## Discussion Questions

**[5-1]** BLAT is an extremely fast, accurate program. Why will it not replace BLAST or at least become as commonly used as BLAST? Is it applicable to protein searches?

**[5-2]** In the original implementation of PSI-BLAST, the algorithm performed a multiple sequence alignment and deleted all but one copy of aligned sequence segments having  $\geq 98\%$  identity (Altschul *et al.*, 1997). In a recent modification, the program now purges segments having  $\geq 94\%$  identity. What do you think would happen if this percentage were adjusted to  $\geq 75\%$ ? How could you test this idea in practice?

### PROBLEMS/COMPUTER LAB

**[5-1]** Create an artificial protein sequence consisting of human RBP4 followed by the C2 domain of human protein

kinase C $\alpha$ . An example of this is shown in Web Document 5.5. Enter this combined sequence into a PSI-BLAST or DELTA-BLAST search. In general, are multiple domains always detected by these programs? Do any naturally occurring proteins have both lipocalin and C2 domains?

**[5-2]** The purpose of this problem is to compare BLASTP to DELTA-BLAST. The malarium parasite *Plasmodium vivax* has a multigene family called *vir* that is specific to that organism (del Portillo *et al.*, 2001). There are 600–1000 copies of these genes, and they may have a role in causing chronic infection through antigenic variation. Select *vir1* and perform a BLASTP search of the nonredundant protein database (restricting the species to *Plasmodium vivax*). Then perform a DELTA-BLAST search with the same entry. For each search, approximately how many proteins have an *E* value less than  $1 \times 10^{-10}$ ?

**[5-3]** Are there globins in fungi? Perform a PSI-BLAST search using human beta globin (NP\_000509) as a query, restricting the output to sequences from fungi (taxid:4751) in the nr database. What is the approximate range of lengths of fungal proteins having globin domains? What nonglobin domains are often present in fungal globins? Does the presence of these unrelated domains lead to corruption? Why or why not? In the first iteration there are several hits (with *E* values below the 0.005 threshold). After several more iterations there are many dozens of hits, including flavohemoproteins that include a globin domain. These fungal proteins have globin domains that are more related to bacterial than vertebrate orthologs. Most of the fungal flavohemoproteins are quite long (over 400 amino acids and sometimes about 1000 amino acids long), having multiple domains. However, only the globin domain is used for the continued PSI-BLAST iterations.

**[5-4]** Perform HMMER searches. First make two different HMMs. You can obtain sets of vertebrate globin and bacterial/fungal/vertebrate globin sequences from Web Documents 5.6 and 5.7 at <http://www.bioinfbook.org/chapter5>. The multiple sequence alignments that we use as input to HMMER are in these documents.

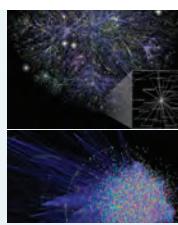
When the profile HMM is built from a multiple sequence alignment of vertebrate alpha and beta globins and used to search the human RefSeq database there are many database matches, including myoglobin (that we could not detect with BLASTP). In contrast, when an alignment of bacterial and fungal globins is used to generate a profile HMM, the output consists of one result with a nonsignificant expect value. Combining several human globins with the bacterial and fungal globins in a multiple sequence alignment results in the creation of a HMM that readily detects human

globins. The profile HMM is therefore a model that is sensitive to the choice of sequences that are used as input for the multiple sequence alignment.

The full results of the HMMER searches for (1) vertebrate, (2) bacterial plus fungal, and (3) bacterial plus fungal plus vertebrate globins are shown in Web Documents 5.8, 5.9, and 5.10 at <http://www.bioinfbook.org/chapter5>. The HMM match to human myoglobin had a higher score and lower *E* value in search (3) than in (1). HMMER searches are run locally. This search was run against all human RefSeq proteins. You can download NCBI databases such as RefSeq by visiting the file transfer protocol (FTP) site from the home page of NCBI or going directly to <http://www.ncbi.nlm.nih.gov/Ftp/> (WebLink 5.31). Place the downloaded database into the same directory as your input sequences for HMMER.

**[5-5]** We previously performed a series of BLAST searches using HIV-1 Pol as a query (NP\_057849). Perform a BLASTP search using this query. Look at the taxonomy report to see which viruses match this query. Next, repeat the search using DELTA-BLAST. Compare this taxonomy report to that of the BLASTP search. What do you observe? Are there any nonviral sequences detected in the DELTA-BLAST search? Did you expect to find any?

**[5-6]** Explore PHI-BLAST using human RBP4 (NP\_006735) as a query, restricting the output to bacteria and the RefSeq database. Use the PHI pattern GXW[YF] X[VILMAFY]A[RKH]. Perform this search and save the results. Then repeat the search using the PHI pattern GXW[YF][EA][IVLM]. How do the results differ? Select one protein that appears as a bacterial protein in a pairwise alignment with the human RBP4 query; what are the *E* values, and why do they differ?



## Self-Test Quiz

**[5-1]** A DELTA-BLAST search is most useful when you want to:

- find the rat ortholog of a human protein;
- extend a database search to find additional proteins;
- extend a database search to find additional DNA sequences; or
- use a pattern or signature to extend a protein search.

**[5-2]** Which of the following BLAST programs uses a signature of amino acids to find proteins within a family?

- PSI-BLAST;
- PHI-BLAST;
- MS BLAST; or
- WormBLAST.

**[5-3]** Which of the following BLAST programs is best used for the analysis of immunoglobulins?

- (a) RPS-BLAST;
- (b) PHI-BLAST;
- (c) IgBLAST; or
- (d) ProDom.

**[5-4]** In a position-specific scoring matrix, the column headings can have the 20 amino acids and the rows can represent the residues of a query sequence. Within the matrix, the score for any given amino acid residue is assigned based on:

- (a) a PAM or BLOSUM matrix;
- (b) its frequency of occurrence in a multiple sequence alignment;
- (c) its background frequency of occurrence; or
- (d) the score of its neighboring amino acids.

**[5-5]** As part of a PSI-BLAST search, a score is assigned to an alignment between a query sequence and a database match over some length (such as 50 amino acid residues). It is possible for this pairwise alignment to receive a higher or lower score over successive PSI-BLAST iterations, even though there is no change in which amino acid residues are aligned:

- (a) true; or
- (b) false.

**[5-6]** A position-specific scoring matrix is said to be “corrupted” when it incorporates a spurious sequence (i.e., a false positive result). Which of the following choices is the best way to reduce corruption?

- (a) lower the  $E$  value;
- (b) remove filtering;
- (c) use a shorter query; or
- (d) run fewer iterations.

**[5-7]** What is a major advantage of using HMMER for protein searches?

- (a) it uses full probabilistic models, including models of amino acid substitutions, insertions, and deletions, to accurately find even distant databases matches;
- (b) it converts amino acid substitution matrices to log-odds forms, increasing sensitivity and specificity;
- (c) it employs position-specific scoring matrices; or
- (d) it uses successive search iterations.

**[5-8]** If you want to find proteins that are distantly related to your query, which of these strategies is most likely to be successful?

- (a) using DELTA-BLAST, because you can specify a signature that is selective for the proteins related to your query;
- (b) using PSI-BLAST, because its strategy of using a position-specific scoring matrix is likely to be most sensitive;
- (c) using BLASTP, because you can adjust the scoring matrices to maximize the sensitivity of your search; or
- (d) using organism-specific databases, because they are most likely to include distantly related sequences.

**[5-9]** Which of the following steps is crucial to validating a sequence you believe to be that of a novel gene?

- (a) performing a PSI-BLAST search;
- (b) checking the EST database to see where this gene might be expressed;
- (c) checking NCBI Gene to see if other family members of this gene have been annotated; or
- (d) BLAST searching your novel sequence into the appropriate database to determine whether anyone else has described your protein.

## SUGGESTED READING

In this chapter we introduced a variety of BLAST servers and BLAST-related software. In most cases the websites contain documentation online. PSI-BLAST was introduced in an excellent paper by Altschul *et al.* (1997) (see also Suggested Reading for Chapter 4). Further modifications of PSI-BLAST are introduced by Schäffer *et al.* (2001). For DELTA-BLAST see Boratyn *et al.* (2012). Benjamin Schuster-Böckler and Alex Bateman (2007) contributed an excellent introduction to hidden Markov models. Trapnell and Salzberg (2009) introduce short-read alignment.

## REFERENCES

- Altschul, S. F., Madden, T.L., Schäffer, A.A. *et al.* 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402. PMID: 9254694.
- Altschul, S.F., Gertz, E.M., Agarwala, R., Schäffer, A.A., Yu, Y.K. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Research* **37**(3), 815–824. PMID: 19088134
- Baldi, P., Chauvin, Y., Hunkapiller, T., McClure, M.A. 1994. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences, USA* **91**(3), 1059–1063. PMID: 8302831.
- Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., Madden, T.L. 2012. Domain enhanced lookup time accelerated BLAST. *Biology Direct* **7**(1), 12. PubMed PMID: 22510480.
- Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., NISC Comparative Sequencing Program, Green, E.D., Sidow, A., Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**, 721–731.
- Burrows, M., Wheeler, D.J. 1994. A block-sorting lossless data compression algorithm. *SRC Research Report* **124**, 1–18.
- Churchill, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**(1), 79–94. PMID: 2706403.
- del Portillo, H. A., Fernandez-Becerra, C., Bowman, S. *et al.* 2001. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410**, 839–842. PMID: 11298455.
- Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14**(9), 755–763. PMID: 9918945.
- Eddy, S.R. 2004. What is a hidden Markov model? *Nature Biotechnology* **22**(10), 1315–1316. PMID: 15470472.
- Eddy, S.R. 2011. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**(10), e1002195. PMID: 22039361.
- Finn, R.D., Clements, J., Eddy, S.R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011). PMID: 21593126.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Gribskov, M., Robinson, N.L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computational Chemistry* **20**, 25–33.
- Gribskov, M., McLachlan, A.D., Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proceedings of National Academy of Sciences, USA* **84**(13), 4355–4358. PMID: 3474607.
- Gribskov, M., Lüthy, R., Eisenberg, D. 1990. Profile analysis. *Methods in Enzymology* **183**, 146–159. PMID: 2314273.
- Kent, W. J. 2002. BLAT—the BLAST-like alignment tool. *Genome Research* **12**, 656–664.
- Krogh, A. 1998. An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology* (eds S. L.Salzberg, D. B.Searls, S.Kasif), pp. 45–63. Elsevier, Amsterdam.
- Krogh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology* **235**, 1501–1531.
- Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359. PMID: 22388286.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3), R25. PMID: 19261174.
- Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760. PMID: 19451168.

- Li, H., Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5), 589–595. PMID: 20080505.
- Li, H., Homer, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* **11**(5), 473–483. PMID: 20460430.
- Li, H., Ruan, J., Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**(11), 1851–1858. PMID: 18714091.
- Lipovich, L., King, M.C. 2006. Abundant novel transcriptional units and unconventional gene pairs on human chromosome 22. *Genome Research* **16**, 45–54.
- Ma, B., Tromp, J., Li, M. 2002. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440–445. PMID: 11934743.
- Marchler-Bauer, A., Zheng, C., Chitsaz, F. et al. 2013. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Research* **41**(D1), D348–D352. PMID: 23197659.
- Morgulis, A., Coulouris, G., Raytselis, Y. et al. 2008. Database indexing for production MegaBLAST searches. *Bioinformatics* **24**(16), 1757–1764. PMID: 18567917.
- Ning, Z., Cox, A. J., Mullikin, J. C. 2001. SSAHA: A fast search method for large DNA databases. *Genome Research* **11**, 1725–1729.
- Park, J., Karplus, K., Barrett, C. et al. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology* **284**, 1201–1210. PMID: 9837738.
- Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E. 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research* **18**(11), 1814–1828. PMID: 18849524.
- Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Rampp, M., Miller, W., Schuster, S.C. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**, 392–394.
- Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., Eisen, M.B. 2004a. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 1–17.
- Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E., Eisen, M.B. 2004b. Correction: Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 73.
- Schäffer, A.A., Aravind, L., Madden, T.L. et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**, 2994–3005. PMID: 11452024.
- Schuster-Böckler, B., Bateman, A. 2007. An introduction to hidden Markov models. *Current Protocols in Bioinformatics Appendix 3*, Appendix 3A. PMID: 18428778.
- Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Research* **13**, 103–107.
- Simon, J.F. 1846. *Animal Chemistry with Reference to the Physiology and Pathology of Man*. G.E. Day, transl. Sydenham Society, London.
- Stoye, J., Evers, D., Meyer, F. 1998. Rose: generating sequence families. *Bioinformatics* **14**, 157–163.
- Tatusov, R.L., Altschul, S.F., Koonin, E.V. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences, USA* **91**, 12091–12095.
- Trapnell, C., Salzberg, S.L. 2009. How to map billions of short reads onto genomes. *Nature Biotechnology* **27**(5), 455–457. PMID: 19430453.
- Wilming, L.G., Gilbert, J.G., Howe, K. et al. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Research* **36**(Database issue), D753–760. PMID: 18003653.
- Wootton, J. C., Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods in Enzymology* **266**, 554–571.
- Ye, J., Ma, N., Madden, T.L., Ostell, J.M. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* **41**(Web Server issue), W34–40. PMID: 23671333.

- Yoon, B.J. 2009. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics* **10**(6), 402–415. PMID: 20190955.
- Zhang, Z., Schäffer, A. A., Miller, W. *et al.* 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research* **26**, 3986–3990.
- Zhang, Z., Schwartz, S., Wagner, L., Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**, 203–214.

Table 1

	Hæmoglobin	Myoglobin	Hæmoglobin	$\beta$ -	Myoglobin	
1	Val	Val	Leu	Leu	Asp	19
2	Leu	His	Leu	Ala	Arg	20
3	Lys	Leu	Leu	His	-	
4	Ser	Thr	Leu	Leu		
5	Pro	Ala	Asp	Asp		
6	Ala	Glu (e) (f)	Gly	Asp		
7	Asp	Glu (g)	Glu	Asp		
8	Lys	Lys	Tyr	Leu	Lys	EF1
9	Thr	Ser (l)	5	5	Lys	2
10	Asp	Asp	6	84	Gly	3
11	Val	Val	7	85	Thr	4
12	Lys	Thr (m)	8	86	Leu	5
13	Ala	Ala	9	87	His	
14	Ala	Leu	Met		Glu	6
15	Arg	Tyr	10	88	Leu	7
16	Gly	Gly	11	89	Asp	8
17	Lys (a)	Lys	12	90	Leu	9
18	Val	Leu	13	91	Asp	10
19	Gly	Asp	14	92	Leu	11
20	Ala		15	93	Asp	12
21	His		16	94	Pro	13
22	Ala	Leu	17	95	Pro	14
23	Val	Val	18	96	Thr	15
24	Gly	Asp	19	97	Ala	16
25	Glu (n)	Gly	20	98	Val	17
26	Tyr	Val	21	99	Val	18
27	Gly	Asp	22	100	Asp	19
28	Ala	Glu (h)	23	101	His	20
29	Ala	Ala	24	102	Lys	21
30	Leu	Thr	25	103	Leu	22
31	Gly	Leu	26	104	Leu	23
32	Arg	Ser	27	105	Asp	24
33	Met	Leu	28	106	Asp	25
34	Phe	Leu	29	107	Pro	26
35	Leu	Phe	30	108	Pro	27
36	Ser	Val	31	109	Val	28
37	Phe	Ser	32	110	Asp	29
38	Pro	His	33	111	Arg	30
39	Thr	Gly	34	112	Leu	31
40	Thr	Thr	CD1	117	Asp	32
41	Lys	Tyr	2	118	Val	33
42	Thr	Arg	3	119	Leu	34
43	Tyr	Phe	4	120	His	35
44	Phe	Phe	5	121	CysH	36
45	Glu	Glu	6	122	Ala	37
46	Pro	Asp	7	123	Ala	38
47	His	Ser	8	124	Thr	39
48	Phe	Phe	9	125	Gly	40
49	Gly	Gly	10	126	CysH	41
50	His	Asp	11	127	His	42
51	Leu	His	12	128	GH1	43
52	Ser	Leu	13	129	Asp	2
53	Ser	Ser	14	130	Asp	3
54	Thr (o)	Thr	15	131	Gly	4
55	Pro	Pro	16	132	Thr	5
56	Asp	Glu	17	133	Gly	6
57	Ala	Glu	18	134	Pro	7
58	Met	Met	19	135	Pro	8
59	Val	Val	20	136	Ala	9
60	Asp	Asp	21	137	Val	10
61	Glu	Gly	22	138	Leu	11
62	Val	Val	23	139	Asp	12
63	Lys	Lys	24	140	Glu	13
64	Gly (b)	Asp	25	141	Phe	14
65	His (c)	Ala	26	142	Pro	15
66	Gly	Val	27	143	Pro	16
67	Lys	Gly	28	144	Ala	17
68	Lys	Lys	29	145	Val	18
69	Leu	Ileu	30	146	Leu	19
70	Leu	Gly	31	147	Asp	20
71	Asp	Asp	32	148	Leu	21
72	Ala	Ala	33	149	Leu	22
73	Leu	Phe	34	150	Asp	23
74	Thr	Ser	35	151	Leu	24
75	Asp (d)	Gly	36	152	Leu	25
76	Ala	Ileu	37	153	Gly	26
			38	154	Arg	27
			39	155	Tyr	28
			40	156	Tyr	29
			41	157	Arg	30
			42	158	His	31
			43	159		32

Source: Watson and Kendrew (1961). Reproduced with permission from Macmillan Publishers.

The determination of the primary amino acid sequences of myoglobin, alpha globin, and beta globin were milestones in the history of biochemistry. (See the quote on horse versus human hemoglobin at the start of Chapter 21.) Each protein was purified, cleaved with endopeptidases, and peptide fragments were overlaid to generate full length sequences. There were many ambiguities; for example, the sequencing technology did not allow aspartate and glutamate to be discriminated. Once the sequences were identified Watson and Kendrew (1961) performed one of the earliest multiple sequence alignments. This clearly showed the relatedness of the three globin proteins. In this figure the correct sequences of the three proteins are also indicated. In some cases the correct (or nearly correct) short peptide sequence was positioned incorrectly in the protein.

# Multiple Sequence Alignment

# CHAPTER 6

*A progressive alignment method is described that utilizes the Needleman and Wunsch pairwise alignment algorithm iteratively to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The sequences are assumed a priori to share a common ancestor, and the trees are constructed from different matrices derived directly from the multiple alignment. The thrust of the method involves putting more trust in the comparison of recently diverged sequences than in those evolved in the distant past.*

—Da-Fei Feng and Russell F. Doolittle (1987, p. 351)

## LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- explain the three main stages by which ClustalW performs multiple sequence alignment (MSA);
- describe several alternative programs for MSA (such as Muscle, ProbCons, and TCoffee), explain how they work, and contrast them with ClustalW;
- explain the significance of performing benchmarking studies and describe several of their basic conclusions for MSA; and
- explain the issues surrounding MSA of genomic regions.

## INTRODUCTION

When we consider a protein (or gene), one of the most fundamental questions is what other proteins are related. Biological sequences often occur in families. These families may consist of related genes within an organism (paralogs), sequences within a population (e.g., polymorphic variants), or genes in other species (orthologs). Sequences diverge from each other for reasons such as duplication within a genome or speciation leading to the existence of orthologs. We have studied pairwise comparisons of two protein (or DNA) sequences (Chapter 3), and we have also seen multiple related sequences in the form of profiles or as the output of a BLAST or other database search (Chapters 4 and 5). We will also explore multiple sequence alignments in the context of molecular phylogeny (Chapter 7), protein domains (Chapter 12), and protein structure (Chapter 13).

In this chapter, we consider the general problem of multiple sequence alignment from three perspectives. First, we describe five approaches to making multiple sequence alignments from a group of homologous proteins of interest. Second, we explore databases of multiply aligned sequences such as Pfam, the protein family database. Third, we discuss

multiple alignment of genomic DNA. This is typically a comparative genomics problem of aligning large chromosomal regions from different species or from distinct, repeated regions of a single genome.

Multiple sequence alignments are of great interest because homologous sequences often retain similar structures and functions (Edgar and Batzoglou, 2006; Do and Katoh, 2008; Pirovano and Heringa, 2008; Kemena and Notredame, 2009). Compared to pairwise alignments, multiple sequence alignments are very powerful because two sequences that may not align well to each other can be aligned via their relationship to a third sequence, thereby integrating information in a way not possible using only pairwise alignments. We can therefore define members of a gene or protein family, and identify conserved regions. If we know a feature of one of the proteins (e.g., hemoglobin transports oxygen), then when we identify homologous proteins, we can predict that they may have a similar function. The overwhelming majority of proteins have been identified through the sequencing of genomic DNA or complementary DNA (cDNA; Chapter 8). The function of most proteins is therefore assigned on the basis of homology to other known proteins rather than on the basis of results from biochemical or cell biological (functional) assays.

### Definition of Multiple Sequence Alignment

Domains or motifs that characterize a protein family are defined by the existence of a multiple sequence alignment of a group of homologous sequences. A multiple sequence alignment is a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned. Homologous residues are aligned in columns across the length of the sequences. These aligned residues are homologous in an evolutionary sense: they are presumably derived from a common ancestor. The residues in each column are also presumed to be homologous in a structural sense: aligned residues tend to occupy corresponding positions in the three-dimensional structure of each aligned protein.

Multiple sequence alignments are easy to generate, even by eye, for a group of very closely related protein (or DNA) sequences. We have seen an alignment of closely related sequences (**Fig. 3.10**, GAPDH). As soon as the sequences exhibit some divergence, the problem of multiple alignment becomes extraordinarily difficult to solve. In particular, the number and location of gaps is difficult to assess. We saw an example of this with kappa caseins (**Fig. 3.11**), and in this chapter we examine a challenging region of five distantly related globins. Practically, you must: (1) choose homologous sequences to align; (2) choose software that implements an appropriate objective scoring function (i.e., a metric such as maximizing the total score of a series of pairwise alignments); (3) choose appropriate parameters such as gap opening and gap extension penalties; and (4) interpret the output and re-run the analyses as needed.

There is not necessarily one “correct” alignment of a protein family (Löytynoja, 2012). This is because while protein structures tend to evolve over time, protein sequences generally evolve even more rapidly than structures. Looking at the sequences of human beta globin and myoglobin, we saw that they share only 25% amino acid identity (**Fig. 3.5**) but the three-dimensional structures are nearly identical (**Fig. 3.1**). In creating a multiple sequence alignment, it may be impossible to identify the amino acid residues that should be aligned with each other as defined by the three-dimensional structures of the proteins in the family. We do not often have high-resolution structural data available, and we rely on sequence data to generate the alignment. Similarly, we do not often have functional data to identify domains (such as the specific amino acids that form the catalytic site of an enzyme), so again we rely on sequence data. It is possible to compare the results of multiple sequence alignments that are generated solely from sequence data and then to examine known structures for those proteins. For a given pair of divergent but significantly related protein sequences (e.g., for two proteins sharing 30% amino acid identity),

Chothia and Lesk (1986) found that about 50% of the individual amino acid residues are superposable in the two structures.

Aligned columns of amino acid residues characterize a multiple sequence alignment. This alignment may be determined due to features of the amino acids, such as:

- There are highly conserved residues such as cysteines that are involved in forming disulfide bridges.
- There are conserved motifs such as a transmembrane span or an immunoglobulin domain. We encounter examples of protein domains and motifs (such as the PROSITE dictionary) in Chapter 12.
- There are conserved features of the secondary structure of the proteins, such as residues that contribute to  $\alpha$  helices,  $\beta$  sheets, or transitional domains.
- There are regions that show consistent patterns of insertions or deletions.

## Typical Uses and Practical Strategies of Multiple Sequence Alignment

When and why are multiple sequence alignments used?

- If a protein (or gene) you are studying is related to a larger group of proteins, this group membership can often provide insight into the likely function, structure, and evolution of that protein.
- Most protein families have distantly related members. Multiple sequence alignment is a far more sensitive method than pairwise alignment to detect homologs (Park *et al.*, 1998). Profiles (such as those described for DELTA-BLAST and hidden Markov models in Chapter 5) depend on accurate multiple sequence alignments.
- When the output of any database search (such as a BLAST search) is examined, a multiple sequence alignment format can be extremely useful to reveal conserved residues or motifs in the output.
- Each human genome harbors ~11,000 nonsynonymous single-nucleotide variants (causing an amino acid substitution) of which ~300 are predicted to be deleterious (see Chapters 9 and 21). Algorithms that predict whether variants are harmful often rely on DNA and/or protein multiple sequence alignments to assess cross-species conservation. Deleterious variants tend to occur at more conserved positions.
- Analysis of population data can provide insight into many biological questions involving evolution, structure, and function.
- When the complete genome of any organism is sequenced, a major portion of the analysis consists of defining the protein families to which all the gene products belong. Database searches effectively perform multiple sequence alignments, allowing comparisons of each novel protein (or gene) to the families of all other known genes.
- We see in Chapter 7 how phylogeny algorithms begin with multiple sequence alignments as the raw data with which to generate trees. The most critical part of making a tree is to produce an optimal multiple sequence alignment.
- The regulatory regions of many genes contain consensus sequences for transcription factor-binding sites and other conserved elements. Many such regions are identified based on conserved noncoding sequences that are detected using multiple sequence alignment.

## Benchmarking: Assessment of Multiple Sequence Alignment Algorithms

We describe five different approaches to creating multiple sequence alignments. How can we assess the accuracy and performance properties of the various algorithms? The performance depends on factors including the number of sequences being aligned, their similarity, and the number and position of insertions or deletions. Benchmarking provides an important answer. There are databases with information about protein secondary or tertiary structure (introduced in Chapter 13), including distantly related proteins that are

known to be homologous and for which structural data can be used to decide the extent to which multiple sequence alignment programs are accurate. We first describe a series of prominent alignment methods, then describe the results of benchmarking.

## FIVE MAIN APPROACHES TO MULTIPLE SEQUENCE ALIGNMENT

There are many approaches to multiple sequence alignment; in the past decade many dozens of programs have been introduced (reviewed in Batzoglou, 2005; Do and Katoch, 2008). We consider five algorithmic approaches: (1) exact methods; (2) progressive alignment (e.g., ClustalW); (3) iterative approaches (e.g., PRALINE, IterAlign, MUSCLE); (4) consistency-based methods (e.g., MAFFT, ProbCons); and (5) structure-based methods that include information about one or more known three-dimensional protein structures to facilitate creation of a multiple sequence alignment (e.g., Expresso). The programs we describe in categories (3) to (5) are often overlapping; for example, all rely on progressive alignment and some combine iterative and structure-based approaches. All the programs offer tradeoffs in speed and accuracy. MUSCLE and MAFFT are fastest, and are therefore most useful for aligning large numbers of sequences. ProbCons and T-COFFEE, although slower, are more accurate in many applications.

We explore how the same set of globin sequences can be aligned differently using various programs, and we try to assess which alignments are most accurate. A related question is the consequence of a misalignment. Potentially, the conservation of critical residues (such as active site amino acids of an enzyme, the heme-binding residues of a globin, or conserved residues that cause disease when mutated) may be missed. Phylogenetic inference (Chapter 7) may be compromised because all molecular phylogeny algorithms depend on a multiple sequence alignment as input. Protein structure prediction (Chapter 13) is severely compromised by faulty multiple sequence alignment, which is often a first step in homology-based modeling.

The programs we explore can be used by web interfaces, although local installation of the programs typically allows you access to a more complete package of options. All the web interfaces allow you to paste in a set of DNA, RNA, or protein sequences in the FASTA format, or to upload a text file containing these sequences.

We explore sets of distantly and closely related globin sequences in the FASTA format. These are available as Web Documents 6.1 and 6.2 at <http://www.bioinfbook.org/chapter6>.

There are many ways that you can easily obtain a group of sequences in the FASTA format. Examples include HomoloGene at NCBI (for eukaryotic proteins), or you can select any subset of the results of a BLAST search and view the sequences in NCBI Protein (or NCBI Nucleotide) in the FASTA format.

### Exact Approaches to Multiple Sequence Alignment

Dynamic programming as described by Needleman and Wunsch (1970) for pairwise alignment is guaranteed to identify the optimal global alignment(s). Exact methods for multiple sequence alignment employ dynamic programming, although the matrix is multidimensional rather than two-dimensional. The goal is to maximize the summed alignment score of each pair of sequences. Exact methods generate optimal alignments but are not feasible in time or space for more than a few sequences. For  $N$  sequences, the computational time that is required is  $O(2^N L^N)$  where  $N$  is the number of sequences and  $L$  is the average sequence length. An exact multiple sequence alignment of more than four or five average-sized proteins would consume prohibitively too much time. Non-exact methods, which we discuss next, are computationally feasible. For example, ClustalW has time complexity  $O(N^4 + L^2)$  and MUSCLE has time complexity  $O(N^4 + NL^2)$ . Although they are faster, these heuristic approaches are not guaranteed to produce optimal alignments.

### Progressive Sequence Alignment

The most commonly used algorithms that produce multiple alignments are derived from the progressive alignment method. This was proposed by Fitch and Yasunobu (1975) and described by Hogeweg and Hesper (1984) who applied it to the alignment of 5S ribosomal RNA sequences. The method was popularized by Da-Fei Feng and Russell Doolittle

(1987, 1990). It is called “progressive” because the strategy entails calculating pairwise sequence alignment scores between all the proteins (or nucleic acid sequences) being aligned, then beginning the alignment with the two closest sequences and progressively adding more sequences to the alignment. A benefit of this approach is that it permits the rapid alignment of hundreds or even thousands of sequences. A major limitation is that the final alignment depends on the order in which sequences are joined. It is therefore not guaranteed to provide the most accurate alignments.

From the 1990s until recently the most popular web-based program for performing progressive multiple sequence alignment has been ClustalW (Thompson *et al.*, 1994; Larkin *et al.*, 2007). While most experts recommend newer programs (such as MAFFT, ProbCons, MUSCLE, and T-COFFEE) which offer improved performance, we introduce the ClustalW algorithm to explain progressive alignment. It proceeds in three stages. We illustrate the procedure by aligning five distantly related globins, selected from NCBI protein and pasted into a text document in the FASTA format (Fig. 6.1). The results are

Note that while most database searches such as BLAST rely on local alignment strategies, many multiple sequence alignments focus on global alignments or a combination of global and local strategies.

**STEP 1 - Enter your input sequences**

Enter or paste a set of **Protein** sequences in any supported format:

```
>beta_globin 2hhbB NP_000509.1 [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCD
KLHVDPENFRLLGNVLCVLAHIFGFKEFTPPVQAAYQKVVAGVANALAHKYH
>myoglobin 2MM1 NP_005359.1 [Homo sapiens]
MGLSDGEWQQLVNVW/GKVEADIPGHGQEVLIRLFKGHPETLEKFDFKHKLKSEDEMKASEDLKKHGATVLTALGGILKKKGHIHEAEIKPLAQSHAT
KHKIPVKYLEFISECIIQVLQSKHPGDFGADAQQGAMNKAELFRKDMASNYKELGFQG
>neuroglobin 1OJ6A NP_067080.1 [Homo sapiens]
MERPEPELIRQSWRAVSRSPLLEHGTIVLARLFALEPDPLPLFQYNCQFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHR
```

Or, upload a file:

**STEP 2 - Set your Pairwise Alignment Options**

Alignment Type:  Slow  Fast

**Slow Pairwise Alignment Options**

Protein Weight Matrix	GAP OPEN	GAP EXTENSION
Gonnet	10	0.1

**STEP 3 - Set your Multiple Sequence Alignment Options**

Protein Weight Matrix	GAP OPEN	GAP EXTENSION	GAP DISTANCES	NO END GAPS
BLOSUM	10	0.20	5	no

ITERATION	NUMITER	CLUSTERING
none	1	NJ

**Output Options**

FORMAT	ORDER
Clustal w/ numbers	input

**STEP 4 - Submit your job**

Be notified by email (Tick this box if you want to be notified by email when the results are available)

**FIGURE 6.1** Multiple sequence alignment of five distantly related globins using ClustalW. Five distantly related globin proteins were pasted in using the FASTA format from Entrez (NCBI).

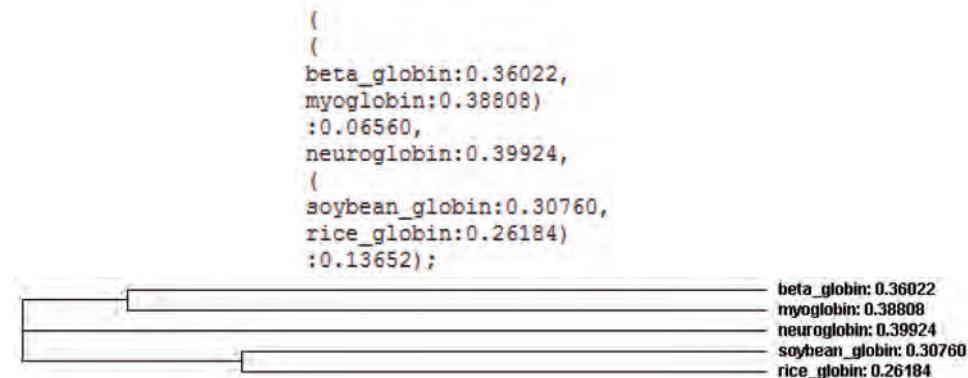
Source: ClustalW, European Bioinformatics Institute.

(a) Stage 1: series of pairwise alignments

SeqA	Name	Length	SeqB	Name	Length	Score
1	beta_globin	147	2	myoglobin	154	25.17
1	beta_globin	147	3	neuroglobin	151	15.65
1	beta_globin	147	4	soybean_globin	144	13.19
1	beta_globin	147	5	rice_globin	166	21.09
2	myoglobin	154	3	neuroglobin	151	16.56
2	myoglobin	154	4	soybean_globin	144	8.33
2	myoglobin	154	5	rice_globin	166	12.99
3	neuroglobin	151	4	soybean_globin	144	17.36
3	neuroglobin	151	5	rice_globin	166	18.54
4	soybean_globin	144	5	rice_globin	166	43.06

1

(b) Stage 2: create a guide tree (calculated from a distance matrix)



The 1994 ClustalW paper has been cited over 48,000 times (as of February 2015). It is present but no longer actively maintained at the European Bioinformatics Institute (EBI) website (<http://www.ebi.ac.uk/Tools/msa/>, WebLink 6.1). In its place, Clustal Omega has been introduced (Sievers et al., 2011) with an emphasis on aligning thousands of sequences. ClustalW continues to be maintained at dozens of websites, such as the EMBOSS program emma at servers such as <http://embossgui.sourceforge.net/demo/emma.html> (WebLink 6.2) and Galaxy (<http://usegalaxy.org>, WebLink 6.3), and through MEGA software (see Computer Lab problem 6.1) at <http://www.megasoftware.net/> (WebLink 6.4).

**FIGURE 6.2** Progressive alignment method of Feng and Doolittle (1987) used by many multiple alignment programs such as ClustalW. In stage 1, a series of pairwise alignments is generated for five distantly related globins (see Fig. 6.1). Note that the best score is for an alignment of two plant globins (score = 43; arrow 1). In stage 2, a guide tree is calculated describing the relationships of the five sequences based upon their pairwise alignment scores. A graphical representation of the guide tree is shown using the JalView tool at the ClustalW web server. Branch lengths (rounded off) reflect distances between sequences and are indicated on the tree; compare to Figure 6.4.

Source: Kyoto University Bioinformatics Center, Courtesy of Kanehisa Laboratories.

shown in Figures 6.2 and 6.3. Later we also align five closely related globins (Figs 6.4 and 6.5). In this particular example we select proteins for which the corresponding three-dimensional structure has been solved by X-ray crystallography. This will help us to interpret the accuracy of the alignment from a structural perspective as well as an evolutionary perspective.

- In stage 1, the global alignment approach of Needleman and Wunsch (1970; Chapter 3) is used to create pairwise alignments of every protein that is to be included in a multiple sequence alignment (Fig. 6.2, stage 1). As shown in the figure, for an alignment of 5 sequences, 10 pairwise alignment scores are generated.

Algorithms that perform pairwise alignments generate raw similarity scores. Note that for the default setting of ClustalW, the scores are simply the percent identities. Many progressive sequence alignment algorithms including ClustalW use a distance

## CLUSTAL 2.1 multiple sequence alignment

```

beta_globin      -----MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVY PWTQRFFESFG- 47
myoglobin       -----MGLSDGEWQLVLNVGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin     -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean_globin  -----MVAFTEKQDALVSSSEAFKANIPQYSVVFYTSILEKAPAACKDLFSFLA- 49
rice_globin     MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
: : : : . : : * * .
beta_globin     DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAT-----LSELHCDKLHVDP 101
myoglobin       HLKSEDEMKASEDLKKHGATVLTALGGTLKKKGHEABIKP-----LAQSHATKHKIPV 102
neuroglobin     QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSIEY---LASLGRKHRAVGVKLS 104
soybean_globin --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVAD---AALGSVHAQKAVTDP 101
rice_globin     --NSDVPLEKNPKLKTHAMSVFVMTCEAAQLRKAGKVTRDTTLKRLGATHLKYGVDA 117
: . * : : : * . * .
beta_globin     ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin       KYLEFFISECITIQVLQSKHPGDFGADAQGMNKALELFRKDMASNYKELGFQG 154
neuroglobin     SFSTVGESLLYMLEKCLG-PAFTPATRAAWSQLYGAVVQAMSRGWDGE--- 151
soybean_globin QFVVVKFAALLTIKAAVG-DKWSDELSRAEVAYDELAAIKKA----- 144
rice_globin     HFEVVKFALLDTIKEEVVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
: : : : * . : : .

```

**FIGURE 6.3** Multiple sequence alignment of five distantly related globins. The output is from ClustalW using the progressive alignment algorithm of Feng and Doolittle. In stage 3, a multiple sequence alignment is created by performing progressive sequence alignments. First, the two closest sequences are aligned (soybean and rice globins). Next, further sequences are added in the order based on their position in the guide tree. An asterisk indicates positions in which the amino acid residue is 100% conserved in a column; a colon indicates conservative substitutions; a dot indicates less conservative substitutions. The proteins are human beta globin (accession NP\_000509; Protein Data Bank identifier 2hhb), human myoglobin (NP\_005359; 3RGK), human neuroglobin (NP\_067080; 1OJ6A), leghemoglobin (from the soybean Glycine max; 1FSL), and nonsymbiotic plant hemoglobin (from rice; 1D8U). Regions of alpha helices (defined in Chapter 13) based on X-ray crystallography are indicated in red letters. Three highly conserved residues are indicated by arrowheads and bold blue letters: phe44 of myoglobin; his65; and his93. Those two histidines are important in coordinating protein binding to the heme group. A green box surrounds the second histidine including five amino acids downstream (to the carboxy terminus) and 17 amino acids upstream (to the end of an alpha helical region). We discuss the alignment within this box for ClustalW in comparison to other alignment programs (**Fig. 6.6**).

Source: Kyoto University Bioinformatics Center, Courtesy of Kanehisa Laboratories.

matrix rather than a similarity matrix to describe the relatedness of the proteins. The conversion of similarity scores for each pair of sequences to distance scores is outlined in Box 6.1. The purpose of generating distance measures is to generate a guide tree (stage 2, below) to construct the alignment.

In our example, note that the best pairwise global alignment score is for rice versus soybean globin (**Fig. 6.2**, arrow 1). For a group of closely related beta globins, all have high scores (**Fig. 6.4**), even for sequences from avian and mammalian species that diverged over 300 million years ago.

2. In the second stage, a guide tree is calculated from the distance (or similarity) matrix. There are two principal ways to construct a guide tree: the unweighted pair group method of arithmetic averages (UPGMA) and the neighbor-joining method. We define these algorithms in Chapter 7. The two main features of a tree are its topology (branching order) and branch lengths (which can be drawn so that they are proportional to evolutionary distance). The tree therefore reflects the relatedness of all the proteins to be multiply aligned.

In ClustalW, the tree is described with a written syntax called the Newick format, as well as with a graphical output (**Figs. 6.2** and **6.4**, stage 2). The chicken sequence has the lowest score relative to the human, chimpanzee, dog, and mouse beta globins,

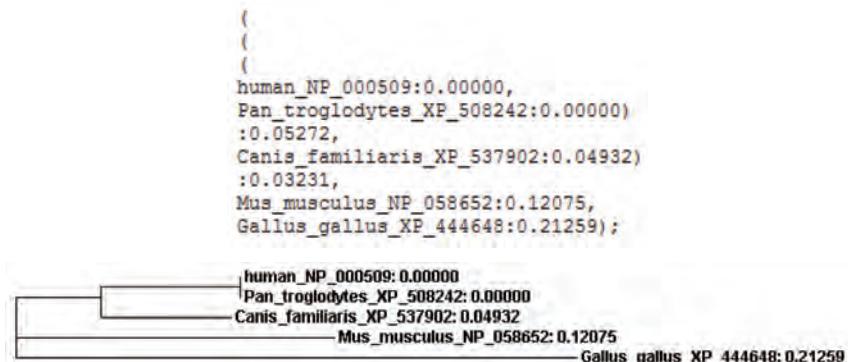
For  $N$  sequences that are multiply aligned, the number of pairwise alignments that must be calculated for the initial matrix equals  $(N - 1)/2$ . For 5 proteins, 10 pairwise alignments are made. For a multiple sequence alignment of 500 proteins,  $499 \times 500/2 = 12,250$  pairwise alignments are made; this is why the speed of an algorithm can be a concern. ClustalW is slow relative to other approaches such as MUSCLE, described below, but for most typical applications its speed is quite reasonable.

To confirm that the ClustalW scores are percent identities, perform pairwise alignments between any two of the sequences in **Figure 6.2** or **6.4** using BLASTP at NCBI (Chapter 3).

(a) Stage 1: series of pairwise alignments (closely related globin proteins)

SeqA	Name	Length	SeqB	Name	Length	Score
1	human_NP_000509	147	2	Pan_troglodytes_XP_508242	147	100.0
1	human_NP_000509	147	3	Canis_familiaris_XP_537902	147	89.8
1	human_NP_000509	147	4	Mus_musculus_NP_058652	147	80.27
1	human_NP_000509	147	5	Gallus_gallus_XP_444648	147	69.39
2	Pan_troglodytes_XP_508242	147	3	Canis_familiaris_XP_537902	147	89.8
2	Pan_troglodytes_XP_508242	147	4	Mus_musculus_NP_058652	147	80.27
2	Pan_troglodytes_XP_508242	147	5	Gallus_gallus_XP_444648	147	69.39
3	Canis_familiaris_XP_537902	147	4	Mus_musculus_NP_058652	147	78.91
3	Canis_familiaris_XP_537902	147	5	Gallus_gallus_XP_444648	147	71.43
4	Mus_musculus_NP_058652	147	5	Gallus_gallus_XP_444648	147	66.67

(b) Stage 2: create a guide tree (calculated from a distance matrix)



**FIGURE 6.4.** Example of a multiple sequence alignment of closely related globin proteins using the progressive sequence alignment method of Feng and Doolittle as implemented by ClustalW. Compare these scores to those for distantly related proteins (Fig. 6.2), and note that the pairwise alignment scores are consistently higher and the distances (reflected in branch lengths on the guide tree) are much shorter.  
Source: Kyoto University Bioinformatics Center, Courtesy of Kanehisa Laboratories.

and this is reflected in its position in the guide tree (Fig. 6.4, stages 1 and 2). A tree can also be displayed graphically at the ClustalW site by using the JalView option. Guide trees are usually not considered true phylogenetic trees, but instead are templates used in the third stage of ClustalW to define the order in which sequences are added to a multiple alignment. A guide tree is estimated from a distance matrix based on the percent identities between sequences you are aligning. In contrast, a phylogenetic tree almost always includes a model to account for multiple substitutions that commonly occur at the position of aligned amino acids (or nucleotides), as discussed in Chapter 7.

- In stage 3, the multiple sequence alignment is created in a series of steps based on the order presented in the guide tree. The algorithm first selects the two most closely related sequences from the guide tree and creates a pairwise alignment. These two sequences appear at the terminal nodes of the tree, that is, the locations of extant sequences. For example, rice globin and soybean globin are aligned. The next sequence is either added to the pairwise alignment (to generate an aligned group of three sequences, sometimes called a profile) or used in another pairwise alignment. At some point, profiles are aligned with profiles. The alignment continues progressively until the root of the tree

CLUSTAL 2.1 multiple sequence alignment

**FIGURE 6.5.** Multiple sequence of five closely related beta globin orthologs (see Fig. 6.4). The output is a screen capture from ClustalW using the progressive alignment algorithm of Feng and Doolittle. The arrowheads correspond to the human beta globin phe44, his72, and his104 residues, respectively, and the green box corresponds to the same region as in Figure 6.3. The coloring scheme (from the ClustalW program) includes groups such as acidic amino acids (blue), basic amino acids (magenta), and hydrophobic residues (red). The asterisks highlight the dozens of column positions conserved among all five globin proteins.

Source: Kyoto University Bioinformatics Center, Courtesy of Kanehisa Laboratories.

is reached, and all sequences have been aligned. At this point a full multiple sequence alignment is obtained (**Figs 6.3** and **6.5**, stage 3).

In the alignment of five distantly related globins, note that a highly conserved phenylalanine is aligned as are two histidines that coordinate heme binding in most globins (**Fig. 6.3**, arrowheads). The region of the second histidine is prone to misalignment, and we will explore how other programs treat this region. For a group of closely related globins, the level of conservation is so high that there are no gaps and therefore no ambiguities about how to perform the alignment (**Fig. 6.5**).

## **BOX 6.1 SIMILARITY VERSUS DISTANCE MEASURES**

Trees that represent protein or nucleic acid sequences usually display the differences between various sequences. One way to measure distances is to count the number of mismatches in a pairwise alignment. Another method, employed by the Feng and Doolittle progressive alignment algorithm, is to convert similarity scores to distance scores. Similarity scores are calculated from a series of pairwise alignments among all the proteins being multiply aligned. The similarity scores  $S$  between two sequences  $(i, j)$  are converted to distance scores  $D$  via:

$$D = -\ln S_{\text{eff}} \quad 6.1$$

where

$$S_{\text{eff}} = \frac{S_{\text{real}(ij)} - S_{\text{rand}(ij)}}{S_{\text{idem}(ij)} - S_{\text{rand}(ij)}} \times 100. \quad 6.2$$

Here,  $S_{\text{real}(ij)}$  describes the observed similarity score for two aligned sequences  $i$  and  $j$ ;  $S_{\text{idem}(ij)}$  is the average of the two scores for the two sequences compared to themselves (if score  $i$  compared to  $i$  receives a score of 20 and score  $j$  compared to  $j$  receives a score of 10, then  $S_{\text{idem}(ij)} = 15$ );  $S_{\text{rand}(ij)}$  is the mean alignment score derived from many (e.g., 1000) random shufflings of the sequences; and  $S_{\text{eff}}$  is a normalized score. If sequences  $i, j$  have no similarity, then  $S_{\text{eff}} = 0$  and the distance is infinite. If sequences  $i, j$  are identical, then  $S_{\text{eff}} = 1$  and the distance is 0.

The Feng–Doolittle approach includes the rule “once a gap, always a gap.” The most closely related pair of sequences is aligned first. As further sequences are added to the alignment, there are many ways that gaps could be included. The rationale for the “once a gap, always a gap” rule is that the two most closely related sequences that are initially aligned should be weighted most heavily in assigning gaps. ClustalW dynamically assigns position-specific gap penalties that increase the likelihood of having a new gap occur in the same position as a pre-existing gap. That serves to give the overall alignment a block-like structure that often appears efficient in terms of minimizing the number of gap positions.

Should an insertion be penalized by the same amount as a deletion? No, according to Löytynoja and Goldman (2005): a single deletion event is typically penalized once where it occurs, but a single insertion event that occurs once inappropriately results in multiple penalties to all the other sequences. The result of these high penalties is that many multiple sequence alignments are unrealistically aligned with too few gaps. Löytynoja and Goldman (2005) introduced a pair hidden Markov model approach that distinguishes insertions from deletions. They showed that their method creates gaps that are consistent with phylogeny, even though the alignments appear less compact than with ClustalW. Their approach applies to the alignment of protein, RNA, or DNA sequences, but it may be especially useful for the alignment of genomic DNA. There, overfitting may occur with traditional progressive alignment, for example when one sequence has long insertions. The approach of Löytynoja and Goldman (2005), reviewed in Higgins et al. (2005), provides multiple sequence alignments that have more gaps but are likely to be more accurate, based on criteria such as correct alignment of exons.

ClustalW implements a series of additional features to optimize the alignment (Thompson *et al.*, 1994). The distance of each protein (or DNA) sequence from the root of the guide tree is calculated, and those sequences that are most closely related are down-weighted by a multiplicative factor. This adjustment ensures that if an alignment includes a group of very closely related sequences as well as another group of divergent sequences, the closely related sequences will not overly dominate the final multiple sequence alignment. Other adjustments include the use of a series of scoring matrices that are applied to pairwise alignments of proteins depending on their similarity, and compensation for differences in sequence length.

Many other algorithms use variants of progressive alignment. For example, Kalign employs a string-matching algorithm to achieve speeds ten times faster than ClustalW (Lassmann and Sonnhammer, 2005). Kalign aligns 100 protein sequences of length 500 residues in less than a second.

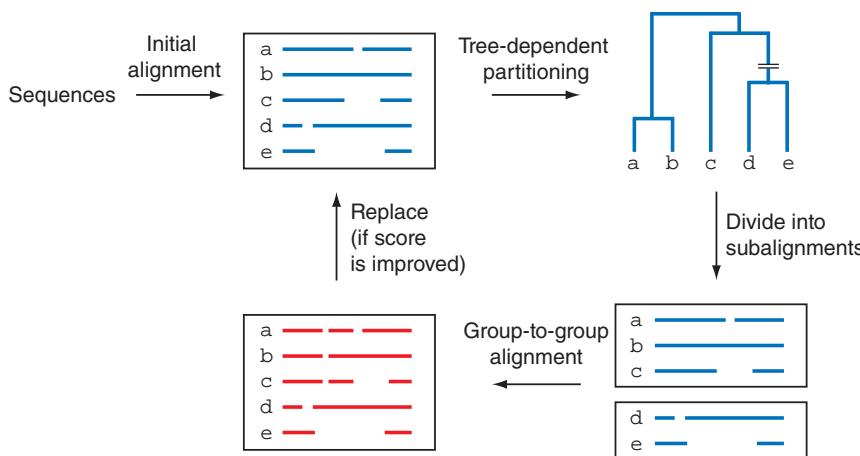
## Iterative Approaches

Iterative methods compute a suboptimal solution using a progressive alignment strategy, and then modify the alignment using dynamic programming or other methods until a solution converges. An initial tree is divided and profiles from each side are re-aligned. These methods therefore create an initial alignment and then modify it to try to improve it, using some objective function to maximize a score (Fig. 6.6).

Progressive alignment methods have the inherent limitation that once an error occurs in the alignment process it cannot be corrected; iterative approaches can overcome this limitation. In standard dynamic programming the branching order of the guide tree may be suboptimal, or the scoring parameters may cause gaps to be misplaced. Iterative refinement can search for more optimal solutions stochastically (seeking higher maximal scores according to some metric such as sum-of-pairs scores or SPS) or by systematically extracting and realigning sequences from an initial profile that is generated. Examples of programs employing iterative approaches are MAFFT (Multiple Alignment using Fast

Note that there are two different senses in which sequences are weighted by ClustalW. The “once a gap, always a gap” rule places the greatest emphasis for gap selection on the most closely related sequences (weighting their importance most heavily). Separately, a set of very closely related sequences are downweighted (reducing their impact on the alignment).

The website <http://msa.cgb.ki.se> (WebLink 6.5) includes Kalign for alignment, Kalignvu as a viewer and Mumsa to assess the quality of a multiple sequence alignment (Lassmann and Sonnhammer, 2006). Kalign is also offered through the European Bioinformatics Institute <http://www.ebi.ac.uk/kalign/> (WebLink 6.6).



**FIGURE 6.6** Iterative refinement method used by MAFFT. A progressive alignment is made then divided into subalignments by tree-dependent partitioning. Partitions are re-aligned, then subgroups are aligned. If an objective score improves, this new alignment replaces the intial one and the process may be repeated. Used with permission. Redrawn from Katoh *et al.* (2009) with permission from Springer Science and Business Media.

Fourier Transform; Katoh *et al.*, 2005), Iteralign (Karlin and Brocchieri, 1998), PRA-LINE (Profile ALIgNmEnt; Heringa, 1999; Simossis and Heringa, 2005), and MUSCLE (MULTiple Sequence Comparison by Log-Expectation; Edgar, 2004a, b).

MAFFT is an example of a multiple alignment package that is considered to be highly accurate based on recent benchmarking studies. It offers a suite of tools with choices of more speed or accuracy (Katoh *et al.*, 2009; Katoh and Standley, 2013). It includes progressive alignment, including: (1) a one-cycle progressive method (called FFT-NS-1) resembling ClustalW, but using a fast Fourier transform for the refinement step; (2) a two-cycle method (FFT-NS-2) in which a multiple alignment is created then refined distances are calculated from that alignment, and a second progressive alignment is formed; and (3) a very fast progressive aligner called PartTree that is useful to align large numbers of sequences (~50,000). The progressive alignment uses matching 6-tuples (strings of six residues) to calculate pairwise distances. This approach is called *k*-mer counting. A *k*-mer (also called a *k*-tuple or word) is a contiguous subsequence of length *k*. *k*-mer counting is extremely fast because it requires no alignment. The initial distance matrix can be recalculated once all pairwise alignments are calculated, yielding a more reliable progressive alignment. In the iterative refinement step, a weighted sum-of-pairs score is calculated and optimized. MAFFT allows options including global or local pairwise alignment.

As a practical example, we can align nine globin proteins using a variety of software tools and see that they each produce different alignments, even when using the leading programs such as MUSCLE, MAFFT, ProbCons, and T-COFFEE. While most alignment programs are available on the internet, MAFFT is designed as a command-line program (although it can be accessed via web servers). We obtain a set of nine globin proteins, put them into a text file on a computer running Linux, perform progressive alignment, and send the results to a file called msa1.txt:

```
$ home/msa$ mafft --retree 2 --maxiterate 0 betaglobins.txt > msa1.txt
```

In the resulting MAFFT alignment, the nine globins include well-aligned conserved residues (Fig. 6.7a, arrowheads 1–3) including two histidine residues that are critical for binding oxygen. The alignment includes a series of terminal and internal gaps, and the

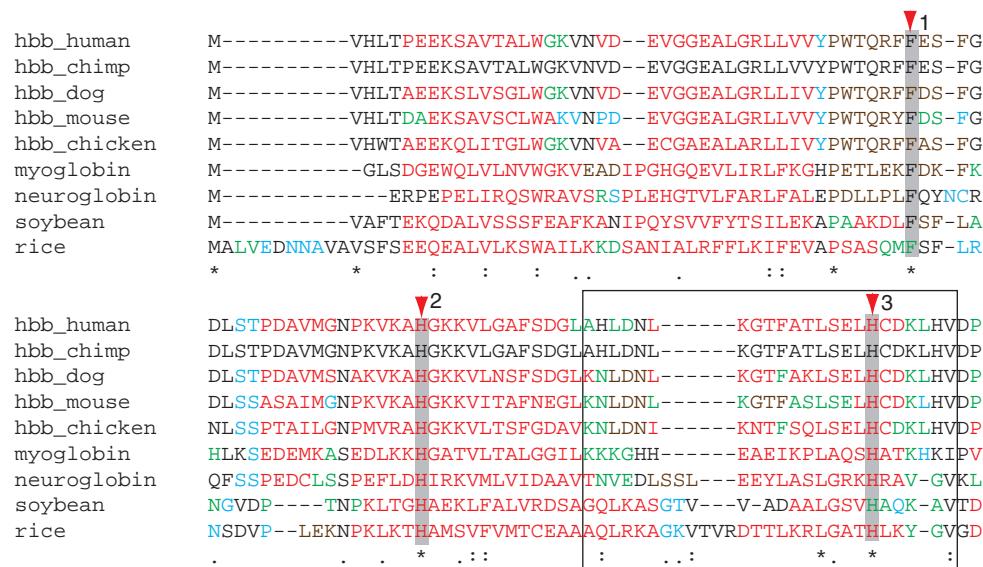
Obtain a set of sequences from NCBI, Ensembl or other sites (e.g., saving BLAST or HomoloGene results in a text file), or use Web Document 6.3 for this set of nine globin sequences. Paste the sequences into a Linux editor such as vim or nano. Instructions for downloading and installing MAFFT are available from <http://mafft.cbrc.jp/alignment/software/> (WebLink 6.7). Once you install it, type mafft -h for a list of options. MAFFT is also available (with more limited options) at the EBI website, <http://www.ebi.ac.uk/Tools/msa/mafft/> (WebLink 6.8).

(a) Alignment of nine globins by MAFFT FFT-NS-2 (v7.058b) (DSSP colors: turn, alpha helix, bend, 3/10 helix)

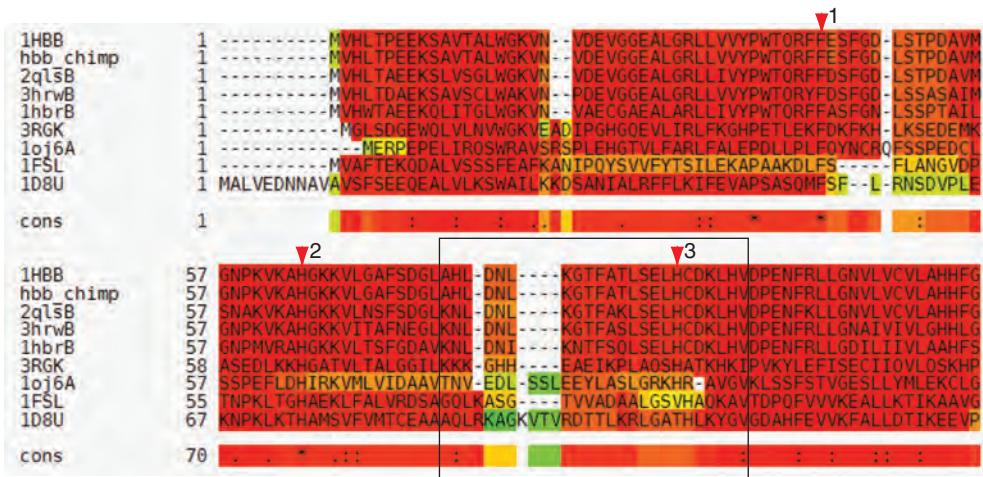
(b) Alignment of nine globins by MUSCLE (3.8)

**FIGURE 6.7** Multiple sequence alignment of nine globins using (a) MAFFT, (b) MUSCLE, (c) Prob-Cons, and (d) T-COFFEE. The sequences are color-coded according to secondary structure predictions made by DSSP (Kabsch and Sander, 1983) as provided on the Protein Data Bank website (Chapter 13). The secondary structure features are turn (shaded green), empty (no feature assigned; black), 3/10-helix (shaded brown), bend (shaded cyan) and alpha helix (shaded red). The arrowheads correspond to the human beta globin phe44, his72, and his104 residues (as in Fig. 6.5). Note that the programs differ in their abilities to align corresponding regions of alpha helical and other secondary structure; in their alignment of highly conserved residues (arrowheads 1–3); and in the number and placement of gaps (see

(c) Alignment of nine globins by ProbCons (version 1.12)



(d) Alignment of nine globins by T-COFFEE (Expresso version\_10.00)



boxed regions). The proteins used to make these alignments (given in Web Document 6.3) are as follows, including the RefSeq accession and the Protein Data Bank identifier of a structure having exactly or close to the same sequence: (1) hbb\_human (human NP\_000509.1, 1HBB); (2) hbb\_chimp (Pan\_troglodytes XP\_508242.1, no structure); (3) hbb\_dog (Canis lupus familiaris NP\_001257813.1, 2QLSIB); (4) hbb\_mouse (Mus\_musculus NP\_058652.1, 3HRWIB); (5) hbb\_chicken (Gallus\_gallus NP\_990820.1, 1HBRIB); (6) myoglobin (human NP\_005359.1, 3RGK); (7) neuroglobin (human NP\_067080.1, 1OJ6A); (8) globin\_soybean (Glycine max leghemoglobin A, NP\_001235928.1, 1FSL); and (9) globin\_rice (Oryza sativa (japonica cultivar-group) NonSymbiotic Plant Hemoglobin NP\_001049476.1, 1D8U). The first two-thirds of each alignment are shown.

secondary structure features (such as alpha helices), although not analyzed as part of this particular MAFFT module, are generally well aligned.

MAFFT and PRALINE can both incorporate information from homologous sequences that are analyzed in addition to those you submit for multiple sequence alignment. These sequences are used to improve the multiple sequence alignment; in the case of MAFFT, the extra sequences are then removed. PRALINE performs a PSI-BLAST search (Chapter 5) on the query protein sequences and then performs progressive alignment using the PSI-BLAST profiles. PRALINE also permits the incorporation of predicted secondary structure information.

Since its introduction in 2004, the MUSCLE program of Robert Edgar (2004a, 4b) has become popular because of its accuracy and its exceptional speed, especially for multiple sequence alignments involving large numbers of sequences. For example, 1000 protein sequences of average length 282 residues were aligned in 21 seconds on a desktop computer (Edgar, 2004a). MUSCLE operates in a series of three stages.

The idea of a triangular distance matrix in stage 1 is that the distance measure between sequences (A, B) equals the distance of (A, C) plus (B, C). This is a good approximation for closely related sequences, but the accuracy is further increased using the Kimura distance correction in stage 2. MUSCLE can be downloaded or accessed via web servers at <http://www.drive5.com/muscle/> (WebLink 6.10) or from the European Bioinformatics website at <http://www.ebi.ac.uk/Tools/msa/muscle/> (WebLink 6.11). As of late 2014, the Edgar (2004a, 2004b) MUSCLE papers have 12,000 literature citations.

1. A draft progressive alignment is generated. To achieve this, the algorithm calculates the similarity between each pair of sequences using either the fractional identity (calculated from a global alignment of each pair of sequences), or  $k$ -mer counting. Based on the similarities, MUSCLE calculates a triangular distance matrix, then constructs a rooted tree using UPGMA or neighbor-joining (see Chapter 7). Sequences are progressively added to the multiple sequence alignment following the branching order of the tree.
2. MUSCLE improves the tree and builds a new progressive alignment (or a new set of alignments). The similarity of each pair of sequences is assessed using the fractional identity, and a tree is constructed using a Kimura distance matrix (discussed in Chapter 7). In a comparison of two sequences there is some likelihood that multiple amino acid (or nucleotide) substitutions occurred at any given position, and the Kimura distance matrix provides a model for such changes. As each tree is constructed it is compared to the tree from stage (1), and the process results in an improved progressive alignment.
3. The guide tree is iteratively refined by systematically partitioning the tree to obtain subsets; an edge (branch) of the tree is deleted to create a bipartition. Next, MUSCLE extracts a pair of profiles (multiple sequence alignments), and realigns them (performing profile–profile alignment; see Box 6.2). The algorithm accepts or rejects the newly generated alignment based on whether the sum-of-pairs score increases. All edges of the tree are systematically visited and deleted to create bipartitions. This iterative refinement step is rapid and had earlier been shown to increase the accuracy of the multiple sequence alignment (Hirosawa et al., 1995).

In general MUSCLE is an excellent program, but in our alignment of nine globins a histidine residue that is critical for binding oxygen is not aligned among several distantly related globins (**Fig. 6.7b**, arrowhead 3).

### Consistency-Based approaches

In progressive alignments using the Feng–Doolittle approach, pairwise alignment scores are generated and used to build a tree. Consistency-based methods adopt a different approach by using information about the multiple sequence alignment as it is being generated to guide the pairwise alignments. We discuss two consistency-based multiple sequence alignment programs: ProbCons (Do *et al.*, 2005) and T-COFFEE (Notredame *et al.*, 2000). MAFFT also includes an iterative refinement approach with consistency-based scores (Katoh *et al.*, 2005), and the Ensembl program Pecan (discussed in “Analyzing Genomic DNA Alignments via Ensembl”) applies a consistency approach to aligning genomic DNA.

PRALINE can be accessed from  
<http://www.ibi.vu.nl/programs/pralinewww/> (WebLink 6.9).

## BOX 6.2 PROFILE-PROFILE ALIGNMENT WITH THE MUSCLE ALGORITHM

The name MUSCLE (multiple sequence comparison by log expectation) includes the phrase “log expectation.” Like ClustalW, MUSCLE measures the distance between sequences (Edgar, 2004a, b). In its third stage, MUSCLE iteratively refines a multiple sequence alignment by deleting the edge of the guide tree to form a bipartition, then extracting a pair of profiles and realigning them. It does this using several scoring functions to optimally align pairs of columns. For amino acid types  $i$  and  $j$ ,  $p_i$  is the background probability of  $i$ ,  $p_{ij}$  is the joint probability of  $i$  and  $j$  being aligned,  $S_{ij}$  is the score from a substitution matrix,  $f^x_i$  is the observed frequency of  $i$  in column  $x$  of the first profile,  $f^x_G$  is the observed frequency of gaps in column  $x$ , and  $\alpha^x_i$  is the estimated probability of observing residue  $i$  in position  $x$  in the family based on the observed frequencies  $f$ . (Note that  $S_{ij} = \log(p_{ij} / p_i p_j)$  as discussed in Chapter 3.) MUSCLE, ClustalW and MAFFT use a profile sum-of-pairs (PSP) scoring function:

$$\text{PSP}^{xy} = \sum_i \sum_j f_i^x f_j^y S_{ij}. \quad 6.3$$

PSP is a sequence-weighted sum of substitution matrix scores for each pair of letters (one from each column that is being aligned in a pairwise fashion). The PSP function maximizes the sum-of-pairs objective score. MUSCLE applies two PAM matrices for its PSP function. MUSCLE also employs a novel log-expectation (LE) score that is defined:

$$\text{LE}^{xy} = (1 - f_G^x)(1 - f_G^y) \log \sum_i \sum_j f_i^x f_j^y \frac{p_{ij}}{p_i p_j} \quad 6.4$$

where the factor  $(1 - f_G)$  is the occupancy of a column. This promotes the alignment of columns that are highly occupied (i.e., that have fewer gaps) while down-weighting column pairs with many gaps. Edgar (2004a) reported that this significantly improved the accuracy of the alignment.

The idea of consistency is that for sequences  $x$ ,  $y$ , and  $z$ , if residue  $x_i$  aligns with  $z_k$  and  $z_k$  aligns with  $y_j$ , then  $x_i$  should align with  $y_j$ . Consistency-based techniques score pairwise alignments in the context of information about multiple sequences, for example adjusting the score of  $x_i$  to  $y_j$  based on the knowledge that  $z_k$  aligns to both  $x_i$  and to  $y_j$ . This approach is distinctive because it incorporates evidence from multiple sequences to guide the creation of a pairwise alignment (Do *et al.*, 2005). Using the notation given in a review by Wallace *et al.* (2005), the likelihood that residue  $i$  from sequence  $x$  and residue  $j$  from sequence  $y$  are aligned, given the sequences of  $x$  and  $y$ , is given by:

$$P(x_i \sim y_j | x, y). \quad (6.8)$$

This is the posterior probability, calculated for each pair of amino acids. The consistency transformation further incorporates data from additional residues to improve the estimate of two residues aligning (i.e., given information about how  $x$  and  $y$  each align with  $z$ ):

$$P(x_i \sim y_j | x, y, z) \approx \sum_k P(x_i \sim z_k | x, z) P(y_j \sim z_k | y, z). \quad (6.9)$$

The consistency-based approach often generates final multiple sequence alignments that are more accurate than those achieved by progressive alignments, based on benchmarking studies.

The ProbCons algorithm has five steps.

1. The algorithm calculates the posterior probability matrices for each pair of sequences. This involves a pair-hidden Markov model as described in **Figure 5.10**. This HMM has three states:  $M$  (corresponding to two aligned positions of sequences  $x$  and  $y$ ),  $I_x$  (a residue in sequence  $x$  that is aligned to a gap), and  $I_y$  (a residue in  $y$  that is aligned to a gap). There is an initial probability of starting in a particular state, a transition probability from the initial state to the next residue, and an emission probability for the next residue to be aligned.

2. The expected accuracy of each pairwise alignment is computed. The expected accuracy is the number of correctly aligned pairs of residues divided by the length of the shorter sequence. The alignment is performed according to the Needleman–Wunsch dynamic programming method but, instead of using a PAM or BLOSUM scoring matrix, scores are assigned based on the posterior probability terms for the corresponding residues and gap penalties are set to zero.
3. The quality scores for each pairwise alignment are re-estimated by applying a “probabilistic consistency transformation.” This step applies information about conserved residues that were identified through all the pairwise alignments, resulting in the use of more accurate substitution scores.
4. An expected accuracy guide tree is constructed using hierarchical clustering (similar to the approach adopted by ClustalW). The guide tree is based on similarities (rather than distances).
5. The sequences are progressively aligned (as in ClustalW) by following the order specified by the guide tree. Further iterative refinements may be applied.

ProbCons is available at <http://probcons.stanford.edu/> (WebLink 6.12).

T-COFFEE was developed by Cédric Notredame, Desmond Higgins, Jaap Heringa and colleagues. It is available at <http://www.tcoffee.org> (WebLink 6.13). It is also mirrored at the European Bioinformatics Institute at <http://www.ebi.ac.uk/Tools/msa/tcoffee/> (WebLink 6.14), the Swiss Institute of Bioinformatics and the Centre National de la Recherche Scientifique (Paris).

View an output of the five distantly related globins using M-COFFEE in Web Document 6.4.

Do *et al.* (2005) reported that ProbCons outperformed six other multiple sequence alignment programs including ClustalW, DIALIGN, T-COFFEE, MAFFT, MUSCLE, and Align-m based on testing on the BAliBASE, PREFAB, and SABmark benchmark databases.

Applying ProbCons to our nine globins, note that (like MAFFT) it places the columns of key residues correctly, but all three programs handle the placement of gaps quite differently (**Fig. 6.7a–c**, boxed regions).

T-COFFEE is an acronym for tree-based consistency objective function for alignment evaluation. T-COFFEE first computes a library consisting of pairwise alignments. By default these include all possible pairwise global alignments of the input sequences (using the Needleman–Wunsch algorithm), and the 10 highest-scoring local alignments. Every pair of aligned residues is assigned a weight. These weights are recalculated to generate an “extended library” that serves as a position-specific substitution matrix. The program then computes a multiple sequence alignment by progressive alignment, creating a distance matrix, calculating a neighbor-joining guide tree, and using dynamic programming and the substitution matrix derived from the extended library.

T-COFFEE includes a suite of related alignment and evaluation tools. M-COFFEE (Meta-COFFEE) combines the output of as many as 15 different multiple sequence alignment methods (Wallace *et al.*, 2006; Moretti *et al.*, 2007). These include T-COFFEE, ClustalW, MAFFT, MUSCLE, and ProbCons. M-COFFEE employs a consistency-based approach to estimate a consensus alignment that is more accurate than any of the individual methods. By adding structural information (discussed in the next section), even more accuracy is achieved.

## Structure-Based Methods

Tertiary structures evolve more slowly than primary sequences. For example, human beta globin and myoglobin share limited sequence identity (in the “twilight zone”) yet share structures that are clearly related. It is possible to improve the accuracy of multiple sequence alignments by including information about the three-dimensional structure of one or more members of the group of proteins being aligned. Programs that enable you to incorporate structural information include PRALINE (Simossis and Heringa, 2005) and the T-COFFEE module Expresso (Armougom *et al.*, 2006b).

When you use the Expresso program at the T-COFFEE website, you submit a series of sequences (typically in the FASTA format). Each sequence is automatically searched by BLAST against the Protein Data Bank (PDB) database, and matches (sharing >60% amino acid identity) are used to provide a template to guide the creation of the multiple

sequence alignment. For our nine globin proteins the resulting alignment conserves the critical residues appropriately, and includes color-coded results showing agreement between all the pairwise structural alignments (most of which are good, indicated in red; **Fig. 6.7d**).

Structural information can also be used to assess the accuracy of a multiple sequence alignment after it has been made. This is performed in benchmarking studies (see next section) for protein families having known structures. In another approach you can incorporate structural information and assess the quality of a protein multiple sequence alignment that you make at the iRMSD-APDB (“Analyze alignments with Protein Data Bank”) server of the T-COFFEE package (O’Sullivan *et al.*, 2003; Armougom *et al.*, 2006c). It is necessary to obtain the accession numbers corresponding to the Protein Data Bank (PDB) file having the known structures of at least two of the proteins you are aligning. As an example, we can obtain the PDB accession numbers for each of the five distantly related globins described above by performing a BLASTP search at NCBI, restricting the output to PDB. Next, perform a multiple sequence alignment using T-COFFEE or any other program. Finally, input this alignment (using the PDB accession number in place of the name) to the APDB server at the T-COFFEE website. The output provides an analysis of the quality of the alignment on the basis of all pairwise comparisons of those sequences having structures, as well as an average quality assessment for each protein. The main approach to assessing how well two structures align is to measure the root mean square deviation (RMSD; see Chapter 13). The RMSD is a measure of how closely the alpha carbons of two aligned amino residues are positioned. Notredame and colleagues introduced iRMSD as an intra-molecular RMSD measure (Armougom *et al.*, 2006a).

For a set of five divergent globins analyzed with the iRMSD-APDB server, 79% of the pairwise columns could be evaluated, 51% of the columns were aligned correctly (according to APDB), and the average iRMSD over all the evaluated columns was 1.07 Ångstroms. This analysis did not depend on a reference alignment, but instead involved a calculation of the superposition of the structures in the alignment.

We described BLAST in Chapter 4, and we describe PDB in Chapter 13.

The iRMSD-APDB server is part of the T-COFFEE suite of tools (<http://www.tcoffee.org>) (WebLink 6.15). Examples of five divergent and five closely related globin sequences formatted for input to the APDB server, as well as the detailed output, are available in Web Documents 6.5 and 6.6 at <http://www.bioinfbook.org/chapter6>.

## BENCHMARKING STUDIES: APPROACHES, FINDINGS, CHALLENGES

The field of bioinformatics uses algorithms to analyze data across a huge range of applications such as pairwise alignment, database searching, measuring RNA transcript levels, or predicting protein function. Many dozens of software packages are typically available for data analysis. How do we know which to trust? Benchmarking provides an important approach. We can obtain a “gold standard” correct answer, consisting of trusted true positive relationships, then compare software programs to determine objectively which is most accurate.

Benchmark datasets may contain separate categories of multiple sequence alignments, such as those having proteins of varying length, varying divergence, insertions or deletions (indels) of various lengths, and varying motifs (such as internal repeats). McClure *et al.* (1994) performed one of the earliest benchmarking studies for multiple sequence alignment, noting the strengths of global rather than local alignment algorithms. More recently Aniba *et al.* (2010) reviewed the concept of benchmarking, particularly with respect to multiple sequence alignment. They note the following qualities of benchmark databases:

- *Relevance*: benchmarks should include tasks actually encountered by users of the software.
- *Solvability*: tasks should not be too easy (such as alignment of proteins sharing >50% amino acid identity) or too hard (such as alignment of proteins sharing limited sequence identity and limited structural data).

The website of Robert Edgar provides a downloadable collection of protein sequence alignment benchmarks (<http://www.drive5.com/bench/>, WebLink 6.16). These include BALIBASE v3, PREFAB v4, OXBENCH, and SABRE. You can directly access BAliBASE (<http://www-bio3d-igbmc.u-strasbg.fr/balibase/>, WebLink 6.17), HOMSTRAD (<http://tardis.nibio.go.jp/homstrad/>, WebLink 6.18), SABmark (<http://bioinformatics.vub.ac.be/databases/databases.html>, WebLink 6.19), and OxBenchmark (<http://www.compbio.dundee.ac.uk/downloads/oxbench/>, WebLink 6.20).

- *Scalability*: some tasks are small, while others require the analysis of large numbers of proteins.
- *Accessibility*: the benchmark should be publicly available.
- *Independence*: the methods used to construct the benchmark database should not be used to perform the sequence alignment.
- *Evolution*: the benchmark database likely needs to be expanded over time to adapt to new problems.

There are several prominent benchmark databases available for multiple sequence alignments. Existing benchmark datasets include BAliBASE (the first large-scale benchmark database; Thompson *et al.*, 2005), HOMSTRAD (Mizuguchi *et al.*, 1998), OXBench (Raghava *et al.*, 2003), PREFAB, SABmark (Van Walle *et al.*, 2005), and IRMBASE (based on synthetic datasets). The general approach is to obtain alignments based on known three-dimensional structures as established by X-ray crystallography (Chapter 13). Proteins which are by definition structurally homologous can therefore be studied. This allows an assessment of how successfully assorted multiple sequence alignment algorithms can detect distant relationships among proteins. For proteins sharing about 40% amino acid identity or more, most multiple sequence alignment programs produce closely similar results. For more distantly related proteins, the programs can produce markedly different alignments and benchmarks are useful to compare accuracy. Local alignment strategies tend to be more successful for sequences having variable lengths.

The performance of a multiple sequence alignment algorithm relative to a benchmark dataset is measured by some objective scoring function. One commonly used metric is the sum-of-pairs scores (Box 6.3). This involves counting the number of pairs of aligned residues that occur in the target and reference alignment, divided by the total number of pairs of residues in the reference. The sum-of-pairs score has been criticized because it is not practical for large numbers of sequences, it does not use an evolutionary model, and it is based on global alignment assuming that proteins in the benchmark have a similar domain organization (Edgar and Batzoglou, 2006). Several alternatives have been proposed (Edgar, 2010; Blackburne and Whelan, 2012).

Löytynoja (2012) noted that three-dimensional structures can be reliably aligned (superimposed) across their core regions, which typically are hydrophobic and evolve slowly. However, it may be misleading to base benchmarking on structural alignments outside the core. Edgar (2010) proposed improvements to benchmarks including the following: discarding sequence data having too high amino acid identity to be informative; discarding sequences for which the structure is unknown; discarding highly diverged structures for which homology is uncertain or there are ambiguous residue correspondences; assessing regions having conserved secondary structure; and taking care to identify multi-domain proteins that may be misaligned in benchmark databases. Benchmark datasets will continue to be developed to increase their utility in evaluating the performance of alignment software.

## DATABASES OF MULTIPLE SEQUENCE ALIGNMENTS

We have discussed different methods for creating multiple sequence alignments. We next examine databases of precomputed multiple sequence alignments, many of which are available. These may be searched using text (i.e., a keyword search) or using any query sequence. The query may be an already-known sequence (such as myoglobin or RBP) or any novel protein (such as the raw sequence of a new lipocalin or globin you have identified). In some databases, the query sequence you provide is incorporated into the multiple sequence alignment of a particular precomputed protein family.

### BOX 6.3 EVALUATING MULTIPLE SEQUENCE ALIGNMENTS

Thompson *et al.* (1999a, b) described two main ways to assess multiple sequence alignments. The first is the sum-of-pairs scores (SPS). This score increases as a program succeeds in aligning sequences relative to the BAliBASE or other reference alignment. The SPS assumes statistical independence of the columns. For an alignment of  $N$  sequences in  $M$  columns, the  $i$ th column is designated  $A_{i1}, A_{i2}, \dots, A_{iN}$ . For each pair of residues  $A_{ij}$  and  $A_{ik}$ , a score of 1 is assigned ( $p_{ijk} = 1$ ) if they are also aligned in the reference, and a score of 0 is assigned if they are not aligned ( $p_{ijk} = 0$ ). Then for the entire  $i$ th column, the score  $S_i$  is given by:

$$S_i = \sum_{j=1, j \neq k}^N \sum_{k=1}^N p_{ijk}. \quad (6.5)$$

For the entire multiple sequence alignment, the SPS is defined:

$$\text{SPS} = \frac{\sum_{i=1}^M S_i}{\sum_{i=1}^{Mr} S_{ri}} \quad (6.6)$$

where  $S_{ri}$  is the score  $S_i$  for the  $i$ th column in the reference alignment, and  $Mr$  corresponds to the number of columns in the reference alignment.

A second approach is to create a column score (CS). For the  $i$ th column,  $C_i = 1$  if all the residues in the column are aligned in the reference and  $C_i = 0$  if not:

$$\text{CS} = \sum_{i=1}^M \frac{C_i}{M}. \quad (6.9)$$

Sum-of-pairs scores and column scores have been used to assess the performance of multiple sequence alignment algorithms. Gotoh (1995) and others further described weighted sum-of-pairs scores that correct for biased contributions of sequences caused by divergent members of a group being aligned. Lassmann and Sonnhammer (2005) note that a column score becomes zero if even a single sequence is misaligned; it may therefore be too stringent.

A distance metric must obey three conditions:

1.  $d(x,y) = 0$  if and only if  $x = y$  (identity)
2.  $d(x,y) = d(y,x)$  (this assures symmetry)
3.  $d(x,z) \leq d(x,y) + d(y,z)$  (the triangle inequality).

Blackburne and Whelan (2012) demonstrated that SP scores are not true metrics because those dissimilarity scores violate the conditions of identity and symmetry.

### Pfam: Protein Family Database of Profile HMMs

Pfam is one of the most comprehensive databases of protein families (Punta *et al.*, 2012). It is a compilation of both multiple sequence alignments and profile HMMs of protein families. The database can be searched using text (keywords or protein names) or by entering sequence data. Its combination of HMM-based approach and expert curation makes Pfam one of the most trusted and widely used resources for protein families.

Pfam consists of two databases. Pfam-A is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. HHMER software (Chapter 5) is used to perform searches. For each family, Pfam provides features (Fig. 6.8a) including: a summary; domain organization, reflecting the architecture of protein domains (Chapter 12); a variety of alignment formats for viewing or download; a

Pfam is maintained by a consortium of researchers including Alex Bateman, Ewan Birney, Lorenzo Cerruti, Richard Durbin, Sean Eddy, and Erik Sonnhamer. Three sites host Pfam: <http://pfam.sanger.ac.uk/> (UK, WebLink 6.21), <http://pfam.janelia.org/> (US, WebLink 6.22), and <http://pfam.sbc.su.se/> (Sweden, WebLink 6.23). Version 27.0 (March 2013) contains >14,800 protein families. Pfam is based on sequences in Swiss-Prot and SP-TrEMBL (Chapter 2).

*Mycobacterium leprae* is a bacterium that causes leprosy. Its globin has accession NP\_301903.

You can also search Pfam with a DNA query, or apply many search options. Go to <http://pfam.sanger.ac.uk/search> (WebLink 6.24).

profile HMM; species data; and structural data. The full alignment can be quite large; for globins there are currently 6000 proteins in that Pfam family, and the top 20 Pfam families currently each contain from 80,000 up to 360,000 sequences in their full alignment. The seed alignments contain a smaller number of representative family members (73 in the case of the globins). Sequences in Pfam-A are grouped in families, assigned stable accession numbers (such as PF00042 for globins) and expertly curated. Additional protein sequences are automatically aligned and deposited in Pfam-B where they are not annotated or assigned permanent accession numbers. Pfam-B serves as a useful supplement that makes the database more comprehensive. For all Pfam families, the underlying HMM is accessible from the main output page.

We can see the main features of Pfam in a search for globins using the Wellcome Trust Sanger Institute site. The main ways to access the database include browsing for families, entering a protein sequence search (with a protein accession number or sequence), and entering a text search. From the front page, select a text-based search and enter “globin.” The results summary includes links to the Pfam entry and to related databases (InterPro, described below; the Protein Data Bank, introduced in Chapter 13; and clans). Each protein in Pfam can have membership in exactly one family. Some proteins, such as sperm whale myoglobin and a globin from the leprosy-causing bacterium *Mycobacterium leprae*, belong to distinct families (globins and bacterial-like globins, respectively). Those two families are distantly related and are defined as members of a larger clan.

The output includes an overview of the globin family including description of the structure of a typical member, a Pfam accession number, clan membership, and a description of the globin family from the InterPro database (discussed below). The Pfam entry further includes access to the alignment, domain organization, species distribution, and a phylogenetic tree. The alignment can be viewed for the seed set, consisting of a core group of representative members of the family (Fig. 6.8b); the full set, consisting of all known family members; or representative proteomes (Chen *et al.*, 2011). The alignment can also be retrieved in a variety of formats, including gapped alignments (useful for viewing aligned regions of the family) or ungapped alignments (useful as input into other multiple sequence alignment programs such as those discussed earlier in this chapter). One of the versatile output formats is JalView. After selecting this option, press the JalView button. A Java applet allows the multiple sequence alignment to be viewed, analyzed, and saved in a variety of ways. The applet will display a principal components analysis (PCA) on the aligned family (Fig. 6.9a). We will describe PCA, a technique to reduce highly dimensional data to two- (or three-) dimensional space, in Chapter 11 (Fig. 11.10). Here, each protein in a multiple sequence alignment is represented as a point in space based on a distance metric, and outliers are easily identified. Similar information can be represented with a phylogenetic tree (Fig. 6.9b; see Chapter 7) using the Java applet.

## SMART

The Simple Modular Architecture Research Tool (SMART) is a database of protein families implicated in cellular signaling, extracellular domains, and chromatin function (Letunic *et al.*, 2012). Like Pfam, SMART employs profile HMMs using HMMER software. SMART can be used in normal mode (providing searches against Swiss-Prot, SP-TrEMBL, and stable Ensembl proteomes) or in genomic mode (providing searches against proteomes of completely sequenced metazoan organisms from Ensembl or other organisms from Swiss-Prot including eukaryotes, bacteria, and archaea).

Also like Pfam, the SMART database is searchable by sequence or by keyword or by browsing the available domains. Domains identified in a SMART search are extensively annotated with information on functional class, tertiary structure, and taxonomy.

### (a) Pfam alignments

**Family: Globin (PF00042)**

**Summary** **Domain organisation** **Clan** **Alignments** **HMM logo** **Trees** **Curation & model** **Species** **Interactions** **Structures** **Jump to...** **enter ID/acc** **Go**

**HOME | SEARCH | BROWSE | FTP | HELP | ABOUT**

**Pfam**  
keyword search Go

**34 architectures** **6000 sequences** **5 interactions** **2886 species** **1971 structures**

## Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four [representative proteomes](#)<sup>1</sup> (RP) sets, the NCBI sequence database, and our metagenomics sequence database. [More...](#)

### View options

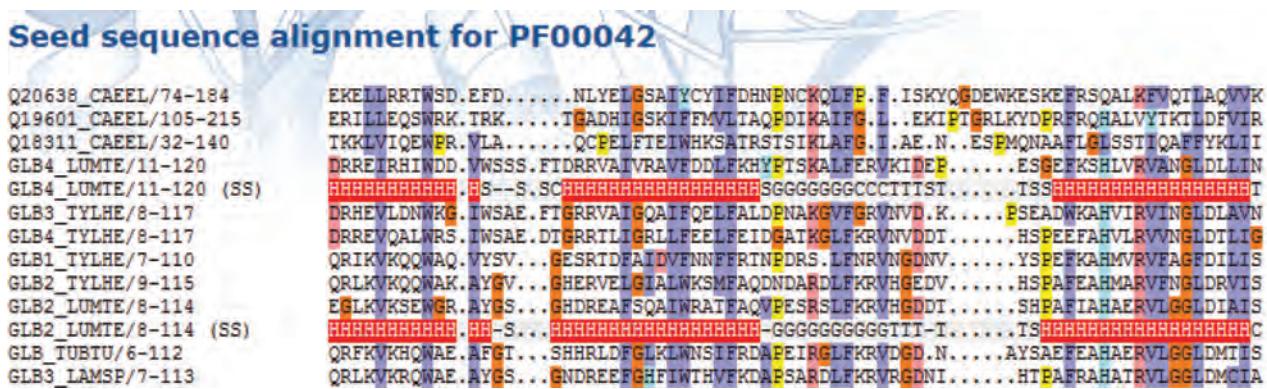
We make a range of alignments for each Pfam-A family. You can see a description of each [above](#). You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (73)	Full (6000)	Representative proteomes				NCBI (5331)	Meta (34)
			RP15 (348)	RP35 (594)	RP55 (949)	RP75 (1261)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	—	✓	✓	✓	✓	✗	✗
PP/heatmap	✗ <sup>1</sup>	—	✓	✓	✓	✓	✗	✗
Pfam viewer	✓	✓	✗	✗	✗	✗	✗	✗

<sup>1</sup>Cannot generate PP/Heatmap alignments for seeds; no PP data available

**Key:** ✓ available, ✗ not generated, — not available.

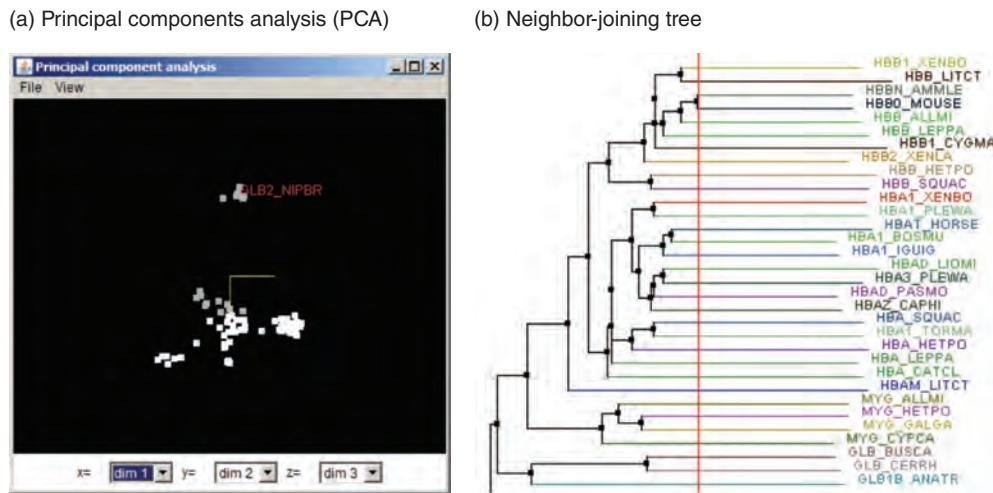
### (b) Pfam seed alignment



**FIGURE 6.8** The Pfam database is a comprehensive, authoritative resource for studying protein families. (a) A typical entry is shown for globins. The top bar shows links to protein architectures (i.e., domain organization including globins), sequences, interactions, species, and structures. The left sidebar provides links to alignments and other information. These alignments can be downloaded as seed alignments (consisting in this case of 73 representative globins), full alignments, or representative proteomes. By clicking the HTML view of the seed alignment, a multiple sequence alignment is produced; a portion is shown in (b). For those entries having known structures, secondary structure (denoted SS) is displayed (highlighted in red, with abbreviations corresponding to helix (H), turn (T), bend (S), and 3/10 helix (G) in Fig. 6.7a–c).

*Source:* PFAM. Courtesy of Dr A. Bateman.

Visualization of Pfam seed alignment of globins using JalView



**FIGURE 6.9** Pfam alignment can be retrieved in the JalView Java viewer format. The Pfam JalView applet displays a multiple sequence alignment of any Pfam protein family. The relationships of the proteins within the family can be explored using a variety of algorithms. (a) Principal components analysis (PCA), further described in Chapter 11, visualizes the relationship of the proteins based on features such as their percent identity. The view can be rotated and displays the percent of the variance that is explained along the *x*, *y*, and *z* axes. In this case, PCA shows that the group of five globins (including a rat hookworm globin, highlighted as GLB2\_NIPBR) are similar to each other but different from other globins in the seed alignment. (b) A phylogenetic tree is displayed, created with the neighbor-joining method (Chapter 7). By clicking the tree a vertical red bar is placed, with nodes to its right displayed in color. Moving this indicator bar allows you to focus analyses on a subset of the sequences in the seed alignment. JalView Java viewer is described by Waterhouse *et al.* (2009).

CDD is available at <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml> (WebLink 6.26) or through the main BLAST page (<http://www.ncbi.nlm.nih.gov/BLAST/>, WebLink 6.27). CDD can also be searched by entering a protein query sequence into the Domain Architecture Retrieval Tool (DART) at NCBI. DART is available at <http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi> (WebLink 6.28).

The InterPro project is coordinated by eight centers including EBI and The Wellcome Trust Sanger Institute. It is available at <http://www.ebi.ac.uk/interpro/> (WebLink 6.29). Release 51.0 (April 2015) contains ~27,000 entries, representing approximately 7500 domains, 18,500 families, 300 repeats, 100 active sites, 70 binding sites and 15 post-translational modification sites.

## Conserved Domain Database

The Conserved Domain Database (CDD) is an NCBI tool that allows sequence-based or text-based queries of Pfam and SMART. CDD uses reverse position-specific BLAST (RPS-BLAST) by comparing a query sequence to a set of many position-specific scoring matrices (PSSMs). RPS-BLAST is related to PSI-BLAST (Chapter 5), but is distinct because it searches against profiles generated from preselected alignments. The main purpose of CDD (and RPS-BLAST) is to identify conserved domains in the query sequence. We provided an example in Chapter 5 (Fig. 5.6). DELTA-BLAST, the most sensitive protein-protein search tool at NCBI, constructs a PSSM using the results of a CDD search and uses that to search a sequence database.

## Integrated Multiple Sequence Alignment Resources: InterPro and iProClass

A main theme of multiple sequence alignment databases is that, while each employs a unique algorithm and search format, they are well integrated with each other. Another important idea is that individual databases such as Pfam and PROSITE have evolved specific approaches to the problem of protein classification and analysis. Some databases employ HMMs; some focus on protein domains, while others assess smaller motifs. Integrated resources allow you to explore the features of a protein using several related algorithms in parallel.

At least two comprehensive resources have been developed to integrate most of the major alignment databases. The InterPro database provides an integration of PROSITE, PRINTS, ProDom, Pfam, and TIGRFAMs with cross-references to BLOCKS (Table 6.1; Hunter *et al.*, 2012).

**TABLE 6.1 Databases on which InterPro (Release 51.0) is based. Entries are rounded off to the nearest 100.**

Database	Contents (entries)
PANTHER 9.0	60,000
Pfam 27.0	14,800
PIRSF 3.01	3,300
PRINTS 42.0	2,000
ProDom 2006.1	1,900
PROSITE 20.105 patterns	1,300
PROSITE 20.105 profiles	1,100
SMART 6.2	1,000
TIGRFAMs 15.0	4,500
CATH-Gene3D 3.5.0	2,600
SUPERFAMILY 1.75	2,000
UniProtKB 2015_04	47,300,000
UniProtKB/Swiss-Prot 2015_04	531,000
UniProtKB/TrEMBL 2015_04	46,715,000
GO Classification	27,000

Source: [http://www.ebi.ac.uk/interpro/release\\_notes.html](http://www.ebi.ac.uk/interpro/release_notes.html). Accessed April 2015.

The iProClass organizes information reports for UniProtKB and NCBI protein records, with links to 170 other databases (Wu *et al.*, 2004). It provides information about protein families, domains, motifs, taxonomy, and literature. Resources such as iProClass and InterPro can be useful to identify conflicts between a variety of databases and to define the size of protein families.

You can access iProClass at  
 ☎ <http://pir.georgetown.edu/pirwww/dbinfo/iproclass.shtml>  
 (WebLink 6.30). Currently (2015)  
 it includes data on over  
 125 million proteins.

### Multiple Sequence Alignment Database Curation: Manual Versus Automated

Some databases are curated manually. This requires expert annotation; Sean Eddy and colleagues have curated Pfam, while Amos Bairoch and colleagues have curated PROSITE. BLOCKS and PRINTS are also manually annotated. Expert annotation is obviously difficult but has the great advantage of allowing judgments to be made on the protein family members. Programs such as DOMO and ProDom use automated annotation. Errors in the alignment or the addition of unrelated sequences can be problematic, as discussed for PSI-BLAST and DELTA-BLAST (Chapter 5). However, automated annotation is valuable for exhaustive analyses of large datasets such as the millions of predicted protein sequences derived from genome-sequencing projects.

## MULTIPLE SEQUENCE ALIGNMENTS OF GENOMIC REGIONS

Complete genomes are being sequenced at a rapid pace, with thousands of projects now completed or in progress. These are described in Part III of this book (Chapters 15–21). A basic problem is the alignment of entire genomes, or parts of genomes. In some cases closely related species are compared, such as humans and the chimpanzee *Pan troglodytes* (these diverged 5–7 million years ago), or different strains of the yeast *Saccharomyces cerevisiae* that diverged recently. In other cases highly divergent genomes are compared,

such as *Homo sapiens* and the monotremes (e.g., the platypus *Ornithorhynchus anatinus*) which diverged about 210 million years ago. We examine the alignment of bacterial genomes in Chapter 17 (on bacteria and archaea).

One basic motivation for performing multiple sequence alignments of genomic regions is to identify DNA sequences that are under the influence of positive selection (and are therefore changing rapidly in a given lineage) or negative selection (and are therefore highly conserved and accumulate mutations slower than the neutral rate). We will introduce the concepts of positive and negative selection in Chapter 7, and see in Part III that comparative genome analyses are used to identify highly conserved regions between genomes that are presumed to be functionally important. Practically, multiple sequence alignment of genomic regions typically uses modifications of the progressive alignment strategy we have discussed. The problem differs from that of conventional multiple sequence alignment in several ways.

- We have been considering programs that are typically used for a set of many protein or nucleic acid sequences, ranging up to hundreds or even thousands of sequences that typically have a length of no more than 1000 or 2000 residues. For genomic alignments, we typically have only a few sequences (perhaps several dozen) that may have lengths of millions or tens of millions of base pairs. The addition of sequences from multiple species improves the accuracy of multiple sequence alignments of orthologous regions, relative to pairwise alignments or to the use of a limited number of species (Margulies *et al.*, 2006).
- Aligning the genomic DNA of closely related organisms (e.g., those that diverged less than 10 million years ago) is often straightforward, but for more diverged organisms (e.g., human to mouse or human to fish) there are often islands of appreciable conservation (typically consisting of exons and conserved noncoding elements) separated by regions of extremely low conservation. This leads to the idea of “anchors” for multiple sequence alignment of genomic regions, discussed below.
- Eukaryotic genomes are riddled with repetitive DNA elements such as DNA transposons and long and short interspersed nuclear elements (LINEs, SINEs; Chapter 8). Such repeats occur in a lineage-specific fashion and can occupy a substantial portion of a genome. They must be accounted for in multiple sequence alignment.
- Chromosomal loci are subject to dynamic rearrangements such as duplications, deletions, inversions, and translocations. These often involve millions of base pairs. Such genomic changes occur frequently in individuals (serving as a major source of human disease) and as features of a species that become fixed (e.g., human chromosome 2 corresponds to two separate acrocentric chromosomes of the chimpanzee, following a chromosomal fusion event early in the hominoid lineage perhaps 5–7 million years ago). In the multiple sequence alignment of genomic regions it is common to find large stretches of apparent deletions or inversions, presenting a challenge for alignment algorithms.
- There are no benchmark datasets for genomic alignments comparable to those described above based on protein structures. However, for each algorithm it is essential to define both the sensitivity (the fraction of all truly orthologous relationships that are detected) and specificity (the fraction of predictions of an orthologous relationship that are correct). Two approaches have been adopted (Blanchette *et al.*, 2004). First, biological sequences with known features such as exons are studied, although this approach does not provide information on how to correctly align poorly conserved regions. Second, simulations have been used, although a challenge is to faithfully model varying evolutionary rates and assorted genomic features such as repetitive elements.

Human chromosome 2, the second largest chromosome, is 243 million base pairs (Mb) in size. It corresponds to chromosomes 2a and 2b of the chimpanzee *Pan troglodytes*.

## Analyzing Genomic DNA Alignments via UCSC

Consider the human beta globin locus on chromosome 11 as an illustration of the usefulness of creating and exploring multiple sequence alignments of genomic DNA. We visited this region in Chapter 5 when we introduced the BLASTZ algorithm for pairwise alignments of genomic DNA. We used the UCSC Genome Browser to visualize the extent of conservation in a region of 50,000 base pairs across multiple species relative to human (**Fig. 5.14**). This browser allows the user to select a region of interest across many scales (from single nucleotides to whole chromosomes) and across many eukaryotic organisms while displaying a user-selected set of annotation tracks. We can now revisit this region, focusing on a span of ~2400 base pairs that includes the beta globin gene (**Fig. 6.10a**). The Vertebrate Multiz Alignment and Conservation track features alignments of 46 species (a subset of which are displayed in the figure). It is based on the PHAST package consisting of two programs (Siepel *et al.*, 2005; Pollard *et al.*, 2010). (1) PhastCons is a hidden-Markov-model-based algorithm that evaluates nucleotides in both individual columns and flanking columns. Its scores (ranging from 0 to 1) represent probabilities of negative selection. (2) phyloP measures conservation at individual columns. Its scores reflect rapidly evolving, neutral, or slowly evolving positions. Multiz refers to a program that implements dynamic programming to align blocks of sequences. It is a component of the Threaded Blockset Aligner (TBA) program (Blanchette *et al.*, 2004).

The peak heights for the conservation track therefore indicate that the coding exons are highly conserved among a group of vertebrates (including mouse, dog, frog, and chicken), while much of the intergenic regions tends to be poorly conserved. Some conserved noncoding regions are apparent (e.g., **Fig. 6.10a**, arrow 1) which could represent conserved regulatory domains. By further zooming in to view just 100 base pairs, a multiple sequence alignment is displayed (**Fig. 6.10b**), in this case including the ATG codon that encodes the start methionine (arrow 2).

**Figure 6.10a, b** also shows Genomic Evolutionary Rate Profiling (GERP) scores (Davydov *et al.*, 2010). These are constraint scores for individual column positions in a multiple sequence alignment of genomic DNA. GERP++ software employs maximum likelihood evolutionary rate estimation for position-specific scoring, and has been used to identify 1.3 million constrained elements covering >7% of the human genome.

The Table Browser is always complementary to the UCSC Genome Browser. Set the Table Browser to the GRCh37/hg19 assembly, choose the Comparative Genomics group and the Conservation track, then explore the outputs of the various tables at the position of hbb on chromosome 11. For the table “Multiz Align,” the output format includes multiple alignment format (MAF). This is a standard for storing genomic alignments. Each MAF file can contain multiple blocks (these may overlap) including genomic coordinates for each species. (In contrast to BED and GFF files, coordinates on the minus strand are numbered relative to the reverse complement.)

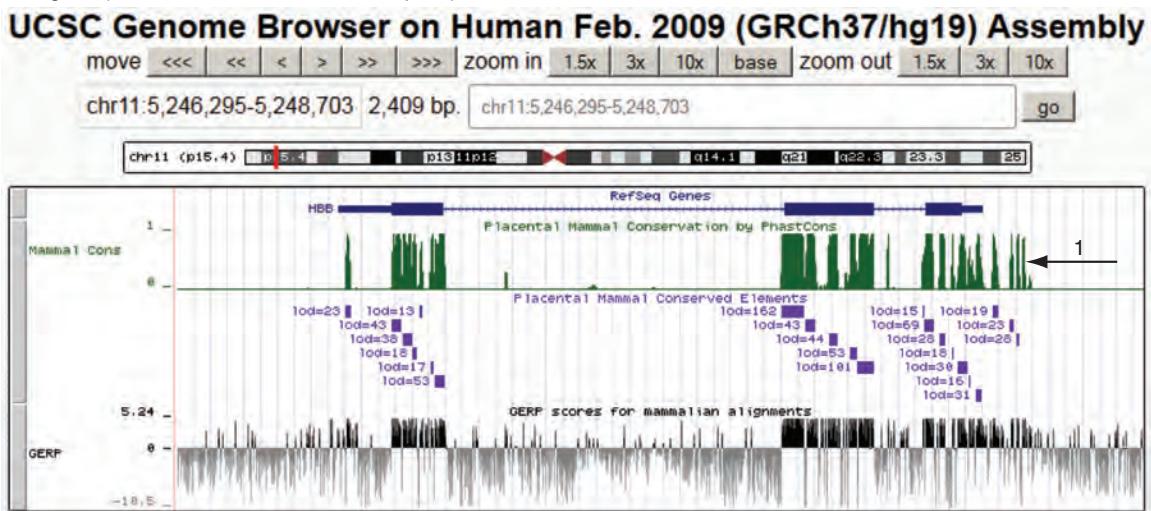
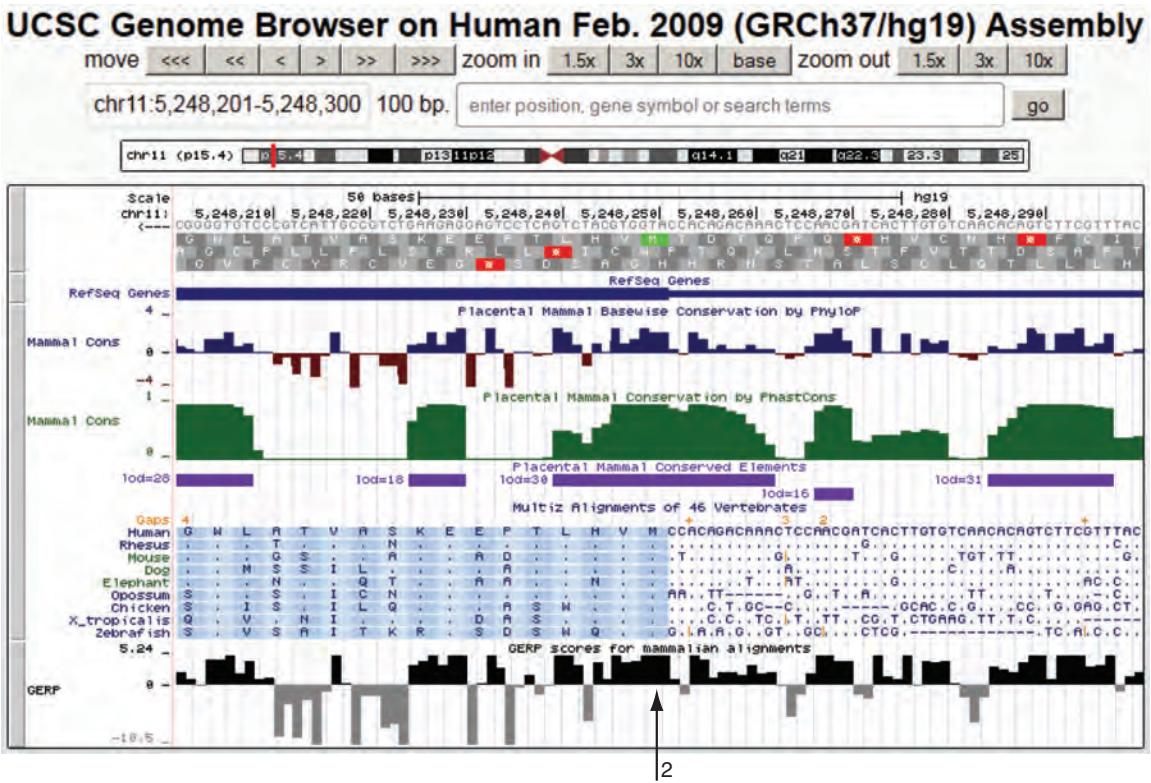
To find this genomic region, go to <http://genome.ucsc.edu> (WebLink 6.31) and select hbb (discussed further in Chapter 8). Click on the conservation track (**Fig. 6.10a**, arrow 2) to download the multiple sequence alignment.

For an example of a MAF file from the HBB region, downloaded from the UCSC Table Browser; see Web Document 6.7.

## Analyzing Genomic DNA Alignments via Galaxy

Galaxy offers useful tools for whole-genome multiple alignments (Blankenberg *et al.*, 2011).

- It includes alignment extractors. These allow you to retrieve alignments of interest, with trimming available to restrict blocks to boundaries of interest.
- Galaxy provides format converters. A MAF file can be converted to FASTA (consisting of either a single block or multiple blocks). A MAF file can also be converted to an interval format resembling a zero-based half-open BED format (refer to Box 2.5).
- It provides MAF stitchers that concatenate adjacent blocks, as well as filtering tools and calculations of alignment coverage.

(a) *HBB* gene (zoomed out 1.5x to 2,409 base pairs)(b) View of *HBB* gene (100 base pairs)

**FIGURE 6.10** Multiple sequence alignment of the human beta globin gene (*HBB*) and other vertebrate orthologs. (a) A view in the UCSC Genome Browser of the beta globin gene (zoomed out 1.5 $\times$ ). Exons are represented by blocks in the RefSeq Genes track and tend to be highly conserved among a group of vertebrate genomes. Three vertebrate conservation tracks are shown (Placental Mammal Conservation by PhastCons, Placental Mammal Conserved Elements, and GERP scores). These show that there is particular conservation in exonic regions; additional conserved noncoding regions (e.g., arrow 1) may represent regulatory elements. (b) Zoomed view of 100 base pairs shows the same tracks as well as the Multiz alignment of 46 vertebrates (nine are displayed). The aligned nucleotides are shown on the right half, while amino acids begin with the start methionine (arrow 2) and extend to the left, matching the start of protein NP\_000509.1, MVHLTPEEKS. Note that by clicking (e.g., on a peak from the PhastCons track), multiple alignment data can be downloaded.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

As an example of how to use Galaxy, begin at the UCSC Table Browser (Gene and Gene Prediction tracks, Vega genes) and select hbb (position chr11:5,246,696–5,248,301 of the GRCh37/hg19 assembly). Select the BED output format, and send the output to Galaxy (coding exons only). In Galaxy’s tools menu choose “Fetch Alignments” and then “Extract Pairwise MAF blocks” to convert the genomic interval to the corresponding blocks in another species (e.g., mouse). Choosing “Extract MAF blocks” allows you to select from 46 species. Selecting “MAF Coverage Stats” shows generally excellent coverage of these well-conserved exons across 46 species. If we then choose Tools > Convert Formats > MAF to FASTA we can output the various sequences in the FASTA format.

Web Documents 6.8 and 6.9 show the output from the pairwise MAF (human/mouse) and multiple sequence extraction (human, gorilla, mouse, rabbit, horse, dog).

### Analyzing Genomic DNA Alignments via Ensembl

Ensembl and UCSC share the same species tree for genomic alignments. However, Ensembl offers a series of innovative tools that in some cases outperform those available at UCSC. Ewan Birney and colleagues introduced several pipelines for alignment of genomic DNA. We explore Ensembl by searching for the human *HBB* gene. This leads to a series of options including Genomic Alignments (Fig. 6.11a, arrow 1). These include a series of pairwise genomic alignments based on tools such as BLASTZ (Kent *et al.*, 2003; Schwartz *et al.*, 2003) and translated BLAT.

There are also options for EPO analyses of primates, eutherian mammals, or amniota (Fig. 6.11a, arrow 2). An DNA alignment at the globin locus is shown in Figure 6.11b. The EPO pipeline is based on Enredo, Pecan, and Ortheus (Paten *et al.*, 2008a, b). First, Enredo generates colinear segments of genomes, including the ability to detect rearrangements, deletions, and duplications. This produces anchors (typically ~100 base pairs in length) between a reference genome (e.g., human) and a comparison genome of interest (e.g., mouse). Paten *et al.* (2008a) assessed its performance and reported that its coverage was better than that of Multiz (i.e., the proportion of ancient repeat bases aligned), as was its accuracy (proportion of ancient repeat bases covered by a full match). Second, Pecan builds multiple sequence alignments using a consistency approach (as described above). Third, Ortheus reconstructs ancestral sequences on a genome-wide basis. Ortheus uses a phylogenetic model to predict ancestral sequences at each node of a phylogenetic tree, improving its ability to classify insertions and deletions.

The EPO pipeline is described at  
 [http://useast.ensembl.org/info/genome/compara/epo\\_anchors\\_info.html](http://useast.ensembl.org/info/genome/compara/epo_anchors_info.html) (WebLink 6.32).

Ortheus is an example of a phylogeny-aware multiple sequence alignment technique. Examining a pairwise alignment of human and mouse beta globin DNA, it is not possible to know whether a given gap position corresponds to an insertion or a deletion in either species (Fig. 6.12a); it is also impossible to know the ancestral allele. Inference of an ancestral sequence answers these questions (Fig. 6.12b). Even more information can be gained from a multiple sequence alignment (Fig. 6.12c), but it is limited because it does not explicitly model insertion and deletion events, or complex (e.g., overlapping) indels. Ortheus produces probabilistic multiple sequence ancestor alignments (Fig. 6.12d). This facilitates both alignment and reconstruction of indels from a phylogenetic perspective.

### Alignathon Competition to Assess Whole-Genome Alignment Methods

The Alignathon competition was designed to assess the performance of a series of whole-genome sequence alignment tools (Earl *et al.*, 2014). The organizers provided the participants with three datasets: a simulated dataset modeling a great ape phylogeny, a simulation modeling a mammalian phylogeny, and a set of 20 fly genomes using real data. There were 35 submissions, including alignment software used by the major browsers (e.g., MULTIZ for the UCSC Genome Browser, EPO for Ensembl) and a variety of other tools such as VISTA-LAGAN (see Fig. 8.17) and ProgressiveMauve (Fig. 15.12).

(a) Ensembl entry for *HBB*

**Gene-based displays**

- Gene summary
- Splice variants (4)
- Transcript comparison
- Supporting evidence
- Sequence
- External references
- Regulation
- Expression
- Comparative Genomics**
  - Genomic alignments**
    - Gene tree (image)
    - Gene tree (text)
    - Gene tree (alignment)
    - Gene gain/loss tree
    - Orthologues (123)
    - Paralogues (9)
    - Protein families (1)
  - Phenotype
  - Genetic Variation
    - Variation table
    - Variation image
    - Structural variation
  - External data
    - Personal annotation
  - ID History
  - Gene history

**Gene: HBB ENSG00000244734**

**Description** hemoglobin, beta [Source: HGNC Symbol; Acc: 4827]  
**Location** Chromosome 11: 5,246,694-5,250,625 reverse strand  
**INSDC coordinates** chromosome GRCh37: CM000673.1:5246694:5250625:1  
**Transcripts** This gene has 4 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
HBB-001	ENST00000335295	754	ENSP00000333994	147	Protein coding	-	CCDS7753
HBB-004	ENST00000380315	502	ENSP00000369671	90	Protein coding	3'	-
HBB-002	ENST00000485743	680	No protein product	-	Retained intron	-	-
HBB-003	ENST00000475226	319	No protein product	-	Retained intron	-	-

**Genomic alignments**

**Alignment:** -- Select an alignment -- [Go](#)

**Key:** 6 primates EPO  
13 eutherian mammals EPO  
20 amniota vertebrates Pecan  
36 eutherian mammals EPO LOW COVERAGE

**Features**

**Human** Human AGA Human CCT Human TTG Human AGT Human GCC Human TTG Human TTT Human TAT Human AAC Human TTA

**Pairwise alignments**

Alpaca (*Vicugna pacos*) - blast  
Anole lizard (*Anolis carolinensis*) - translated blat  
Armadillo (*Dasypus novemcinctus*) - blast  
Bushbaby (*Otolemur garnettii*) - lastz  
Cat (*Felis catus*) - lastz  
Chicken (*Gallus gallus*) - lastz  
Chicken (*Gallus gallus*) - translated blat  
Chimpanzee (*Pan troglodytes*) - lastz  
Chinese softshell turtle (*Pelodiscus sinensis*) - lastz  
Ciona intestinalis - translated blat  
Ciona savignyi - translated blat  
Cod (*Gadus morhua*) - translated blat  
Coelacanth (*Latimeria chalumnae*) - translated blat

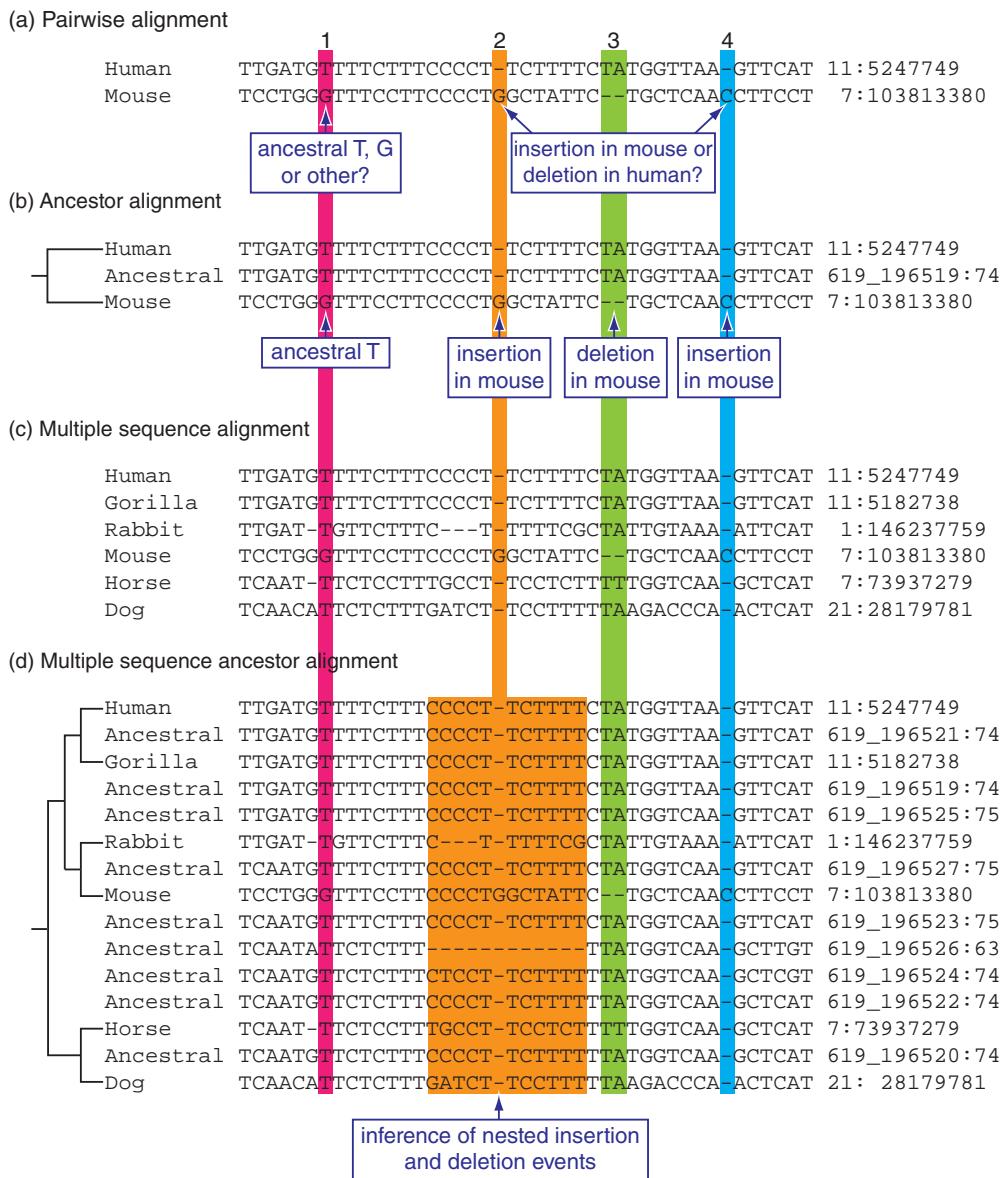
Sequence alignment table showing pairwise alignments between Human and various species. Exons are highlighted in red.

## (b) Ensembl multiple sequence alignment (Enredo/Pecan/Ortheus software)

Homo sapiens	11: 5246983	TTCATACCTCTT-ATCTCCTCCCACAG	<b>CTCCTGGGCAACGTGCTGG</b>
Gorilla gorilla gorilla	11: 5181973	TTCATACCTCTT-GTCTCCTCCCACAG	<b>CTCCTGGGCAATGTGCTGG</b>
Pongo abelii	11: 65239065	TTCATACCTCTT-GTCTCCTCCCACAG	<b>CTCCTGGGCAATGTGCTGG</b>
Oryctolagus cuniculus	1:146237264	TTCATGCCTCT--TCTCTTCCACAG	<b>CTCCTGGGCAACGTGCTGG</b>
Mus musculus	7:103812810	TTGATGGTTCTT--CCATCTTCCCACAG	<b>CTCCTGGGCAATATGATCG</b>
Bos taurus	15: 49339417	CCTTGCTTAATG-TCTTTCACACAG	<b>CTCCTGGGCAACGTGCTAG</b>
Bos taurus	15: 49074455	CCTTGCTTAATG-TCTTTCACACAG	<b>CTCCTGGGCAACGTGCTGG</b>
Sus scrofa	9: 5633260	CCTTCCCTTTTA-TCTCTCTCCCACAG	<b>CTCCTGGGCAACGTGATAG</b>
Equus caballus	7: 73936736	CCCCCTCTTT-TT-TCTTCTCCCCACAG	<b>CTCCTGGGCAACGTGCTGG</b>
Canis lupus familiaris	21: 28179266	CACATGCCTCTG-TCT--TCCCCACAG	<b>CTGCTGGGCAACGTGTTGG</b>

**FIGURE 6.11** Analyzing multiple sequence alignments at the Ensembl website. (a) Following a search for human beta globin (*HBB*) select “Genomic alignments” (arrow 1) then choose sequences from a group such as 6 primates or 36 eutherian mammals (arrow 2). The aligned sequences are displayed with exons color-coded in red (arrow 3). (b) A portion of the Enredo/Pecan/Ortheus pipeline results are shown.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.



**FIGURE 6.12** The Orpheus program of the EPO pipeline at Ensembl provides phylogeny-aware multiple sequence alignment including reconstruction of ancestral sequences. Alignments of the human beta globin region are shown. (a) In a pairwise alignment between human and mouse DNA, it is unclear whether an aligned T and G residue derive from an ancestral T, G, or other nucleotide (see column 1). It is also unclear whether gaps correspond to insertions or deletions in one or the other species. (b) By inferring an ancestral human/mouse sequence we can infer ancestral alleles and specify whether gap positions correspond to insertions or deletions in either lineage. (c) A multiple sequence alignment offers further evidence and implicitly resolves ambiguities. (d) A multiple sequence ancestor alignment includes information about every node in a phylogeny and explicitly resolves questions of the origin of insertions, deletions, and complex events such as nested insertions/deletions. Note that chromosome and position (or ancestral sequence identifier and position) are given to the right of each sequence. Multiple ancestral sequences are given because there are multiple internal nodes (introduced in Chapter 7). You can view the sequences shown in this figure by visiting the viewing HBB genomic alignments as shown in **Figure 6.11** and selecting “13 eutherian mammals EPO.” Then click “Configure this page” to add or remove species as well as inferred ancestral sequences.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

Visit the Alignathon website at  
✉ <http://compbio.soe.ucsc.edu/alignathon/> (WebLink 6.33).

For the closely related primate sequences the alignments were very similar, while results differed substantially for more divergent genome comparisons. The Alignathon project offers the community an opportunity to assess software performance, even though the particular datasets are not necessarily optimal for each software package and the statistical assessment of alignments is difficult in the absence of gold standards for true alignments.

## PERSPECTIVE

Multiple sequence alignment is the operation by which various members of a protein family (or nucleic acid family) may be grouped together. Rows correspond to sequences and columns correspond to residues, with aligned residues in a column implying shared evolutionary ancestry and/or shared positions in three-dimensional structures. Multiple sequence alignment serves many purposes, including the identification of conserved residues that are functionally important. There is tremendous enthusiasm in the bioinformatics community for the variety of novel approaches to generating accurate multiple sequence alignments, including progressive alignment, and approaches based on iterative refinement, consistency, and/or the use of structural information. A general conclusion is that most programs perform very well with sequences that are closely related (e.g., sharing approximately 40% amino acid identity or more). For more distantly related sequences, the available programs may differ considerably, particularly in where gaps are placed. For the typical user, two suggestions are: try performing multiple sequence alignments using several programs; and try using a variety of alternative parameters such as gap penalties.

The subdiscipline of multiple sequence alignment algorithms is rapidly changing. New challenges include the analysis of genomic DNA sequences. Benchmark datasets are not always available for the purpose of assessing the accuracy of a newly developed algorithm.

Databases of multiply aligned protein families such as Pfam and InterPro are rapidly expanding in size and are increasingly important tools. These databases are often accompanied by careful expert annotation. A general trend is that databases offer the integration of many alignment resources.

## PITFALLS

A very basic pitfall to avoid is the use of a group of sequences for multiple sequence alignment in which one or more sequences are not homologous to the rest. For multiple sequence alignments with relatively divergent members, it is common for different programs to give dramatically different results. A challenge is that you may not be able to assess which is the most accurate based on criteria such as structure or shared evolutionary history. Gaps are particularly hard to place, and the most compact alignment (with the fewest gaps) is not necessarily the most faithful to the evolutionary history of the sequences you are aligning. As an example of a challenge, ProbCons, MAFFT, and MUSCLE all adopt different approaches to the problem of what gap penalties to assign to terminal gaps (deletions) relative to internal gaps. There may not be a single correct approach, but this is an example of why different programs will produce different alignments.

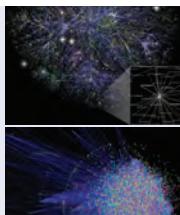
It is especially important to perform a proper multiple sequence alignment for molecular phylogeny studies. The alignment constitutes the raw data that go into making a tree (see Chapter 7). This is one of the many reasons that benchmarking studies are important to define the specificity and sensitivity of different algorithms. However, there are considerable concerns about the current state of benchmarking (Aniba *et al.*, 2010; Edgar, 2010).

## ADVICE FOR STUDENTS

It may seem that there is a bewildering number of tools available for multiple sequence alignment of proteins (and DNA sequences). In principle, benchmarking is essential to help both experts and beginners alike decide which tools perform the best. What are the criteria for deciding the specificity of these tools and their sensitivity? Read papers that perform benchmarking and those that explain and challenge the current nature of benchmark databases (e.g., Edgar, 2010; Löytynoja, 2012).

To gain a deeper understanding of multiple alignment of proteins, select a group of distantly related protein sequences. (If they are too closely related then most tools are likely to give comparable answers.) Inspect the alignments, and try to develop a feel for which alignments are better. Experts routinely make manual adjustments to alignments, using their experience and judgment. You can become expert by studying alignments, learning what objective functions are applied to evaluate which are superior, and recognizing alignments that can be improved or that contain errors.

We showed several tools for aligning genomic DNA, including PhastCons, phyloP, and GERP scores given at the UCSC Genome and Table Browser. Keep in mind that each of these tools has an associated configuration page that details what the program does and how its output may be interpreted. There are links to literature references and external resources. Actively explore these resources to learn more about the strengths and limitations of each tool.



## Discussion Questions

**[6-1]** Feng and Doolittle introduced the “once a gap, always a gap” rule, saying that the two most closely related sequences that are initially aligned should be weighted most heavily in assigning gaps. Why was it necessary to introduce this rule? How does iterative refinement overcome this rule?

**[6-2]** Could BLAST searches incorporate HMMs? How does DELTA-BLAST differ from an HMM-based search in Pfam?

**[6-3]** How would you construct a benchmark dataset for genomic DNA? What features would you need to consider (e.g., protein versus DNA, degree of conservation, chromosomal rearrangements)?

## PROBLEMS/COMPUTER LAB

**[6-1]** Practice using three NCBI resources to obtain groups of sequences in the FASTA format that you can use for multiple sequence alignment. Select a keyword such as cytochrome (other suggestions are ferritin, S100, or trypsin). In a first approach, enter this search from the home page of NCBI, and follow the link to HomoloGene. By default, the entries are displayed in the summary format. Using the pull-down menu change the display to Multiple Alignment. This allows you to scroll through a series of multiple sequence alignments. Select one for fur-

ther study. It is helpful to choose one in which there are some gaps, so that you can evaluate the performance of various software programs (see problem (6.2)). Once you identify a group of proteins of interest, click to view that HomoloGene group, and change the display to FASTA. Copy these sequences and/or save them to a text document. In a second approach, repeat this exercise beginning at the home page of NCBI, but select the link to CDD (the Conserved Domain Database). Here, there are Pfam, CDD, SMART, and/or COG identifiers. Select an entry with a CDD identifier (such as cd00904 for ferritin). Here, a multiple sequence alignment is shown. Change the format to obtain the desired number of proteins in this family (e.g., up to 5, 10, or 20) in the FASTA format; you may select the most diverse members of this group. In a third approach, perform a BLASTP search using a query such as ferritin light chain (NP\_000137) and inspect the pairwise alignments to the query. Select a group of 10 proteins by clicking on the box next to each, and click “Get selected sequences.” These ten proteins appear on an NCBI Protein page; change the display option to FASTA and use the pull-down menu option to “send to text.” The sequences are now available in the FASTA format for further study.

**[6-2]** Using the FASTA-formatted sequences from problem (6.1), perform multiple sequence alignments using programs available at the European Bioinformatics

Institute: MAFFT, MUSCLE, and T-COFFEE. Save and compare each result. How do they differ? How can you assess which is likely to be the most accurate? When applicable, try adjusting the parameters such as the scoring matrices, gap opening and extension penalties, or number of iterations to see the effects on the alignments.

**[6-3]** Use EDirect to access sets of homologous proteins from HomoloGene. These can be viewed in various formats. Retrieve the sets of protein sequences in the FASTA format containing the protein HBB. How many HomoloGene entries are there in this set?

```
$ esearch -db homologene -query "HBB" | efetch
-db homologene -format fasta
1: HomoloGene:128037. Gene conserved in
Boreoeutheria
>gi|4504351|ref|NP_000510.1| hemoglobin subunit
delta
MVHLTPEEKTAVALWGKVNVDAVGGGEALGRLLVVYPWTQRFFESFG-
DLSSPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGN-
VLVCVLRNFGKEFTPQMQAAYQKVVAGVAN
ALAHKYH
>gi|332835679|ref|XP_001162045.2| PREDICTED:
hemoglobin subunit delta isoform 1 [Pan
troglodytes]
MVHLTPEEKTAVALWGKVNVDAVGGGEALGRLLVVYPWTQRFFESFG-
DLSSPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGN-
VLVCVLRNFGKEFTPQMQAAYQKVVAGVAN
ALAHKYH
...
```

Next retrieve the alignment scores (rather than the sequences) for HomoloGene entries containing the protein HBB.

```
$ esearch -db homologene -query "HBB" | efetch
-db homologene -format alignmentscores
...
4: HomoloGene:68066. Gene conserved in
Boreoeutheria
Pairwise Alignment Scores
Gene Identity (%)
Species Symbol Protein DNA
H.sapiens HBB
vs. P.troglodytes HBB 100.0 99.8 Blast
vs. M.mulatta HBB 94.6 95.9 Blast
vs. C.lupus LOC476825 86.4 86.2 Blast
vs. C.lupus LOC480784 89.8 87.8 Blast
vs. C.lupus LOC609402 89.8 87.8 Blast
vs. M.musculus Hbb-bs 80.3 82.5 Blast
vs. M.musculus Hbb-bt 80.3 82.8 Blast
vs. R.norvegicus Hbb 81.6 82.8 Blast
vs. R.norvegicus Hbb-b1 73.5 78.7 Blast
vs. R.norvegicus LOC100134871 78.9 81.0 Blast
vs. R.norvegicus LOC689064 78.9 81.0 Blast
...
```

**[6-4]** We described how ClustalW applies a correction factor to downweight the influence of closely related proteins.

Test the performance of ClustalW: take the globins in Web Documents 6.1 and/or 6.2 and align. Then repeat the alignment with the additional input of one divergent sequence repeated a varying number of times. For example, in the closely related group of beta globins, add five copies of the chicken sequence to see its influence on the alignment.

**[6-5]** Use the T-COFFEE programs to evaluate the effect of structural information on your alignments. Follow these steps. (1) Obtain a group of five distantly related lipocalins from Web Document 6.10 (<http://www.bioinfbook.org/chapter6>). These include rat odorant-binding protein and human retinol-binding protein. (2) Align the sequences using T-COFFEE (<http://www.tcoffee.org/>, WebLink 6.13), or use another program. (3) Evaluate the alignment with the iRMSD program (<http://www.tcoffee.org/>). Include the information on two known lipocalin structures. Note the score. (4) Align the same sequences again using Expresso (<http://www.tcoffee.org/>) to incorporate structural information. Note the score. Did it improve? Do the alignments differ?

**[6-6]** MAFFT was developed as a command-line program. This problem introduces you to using MAFFT in the Linux environment. In particular, we obtain a set of alpha globin proteins and a set of beta globin proteins, align them, and then align the two profiles. (1) First obtain the globins using EDirect. Alternatively, search NCBI's HomoloGene resource with the term globin. The beta globin group (HomoloGene:68066) currently includes 15 proteins. Click the download link and save them as a text file. Repeat this for 14 proteins in the "hemoglobin, alpha 2" family (HomoloGene:469; all proteins have a length of 142 residues). These files are given as Web Documents 6.11 and 6.12, and the entries are conveniently renamed in Web Documents 6.13 and 6.14. (2) Open a Linux terminal session, and create two new documents: `vim hba.fasta` (paste your sequences into the editor, then use `:wq` to write the file and quit) and `vim hbb.fasta`. Alternatively, if you are working on a PC use WinSCP or a similar utility to transfer a text file to your working directory. (3) Perform alignments as described in this chapter for the command-line MAFFT.

**[6-7]** The purpose of this problem is to obtain mammalian DNA sequences in the beta globin region and align them. (1) Visit the UCSC Genome Browser (build GRCh37) position chr11:5,245,001–5,295,000. This 50 kilobase region of chromosome 11p15.4 includes the RefSeq genes *HBB*, *HBBP1*, *HBG1*, *HBG2*, and *HBE1*. (2) In the Comparative Genomics section click the header for the "Conservation" track. Download the sequences. (You can also download multiple alignments of 45 vertebrate genomes with human from <http://hgdownload.soe.ucsc.edu/downloads.html>. As another example to try, obtain a MAF file

from <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz46way/maf/> and browse to chrM.maf.gz (252 kB) for an example of a small set of sequences.) (3) Analyze multiple alignments in MAFFT as described above.

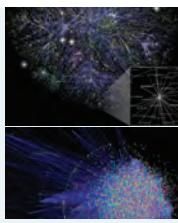
**[6-8]** This problem introduces multiple alignments in the MAF format at the Galaxy website. Go to Galaxy (either the main public server or a local instance you set up). (2) Under the “Get Data” tool (left sidebar), select UCSC Main Table Browser. Choose mammal, human, Feb. 2009 (GRCh37/hg19), Genes and Gene Prediction Tracks (group), RefSeq Genes (track), and in the position box type in hbb then “lookup” to obtain the coordinates chr11:5,246,696–5,248,301. Use the output format “BED” and send the output to Galaxy. (3) View the dataset. (4) Go to Tools > Fetch Alignments > Extract MAF blocks. For the interval, use the imported dataset from the UCSC Table Browser; for MAF source, choose “Locally Cached Alignments”; and select “46-way multiZ (hg19).” (5) Inspect the output (click the eye icon on the history panel). There are 37 blocks across this region. For each block, the line labeled “a” shows a float point score while the lines labeled “s” correspond to sequences (these are 0-based starts). (6) On the Tools panel choose “Graph/Display Data” then choose GMAJ. This is an interactive, Java-based multiple alignment viewer (Blanchette *et al.*, 2004). (7) Convert the multiple alignment to a set of FASTA files. Go to Tools > Convert Formats > “MAF to FASTA” and choose one sequence per species as the type of FASTA output. Optionally, you can download these sequences (e.g., to align them with different methods). For example, you can use Tools > Multiple Alignments > ClustalW to align these FASTA files. (8) Choose Tools > Evolution > “Neighbor Joining Tree Builder”. Use the FASTA file, and a distance model such as Kimura 2 parameter (see Chapter 7). (9) Choose Tools > Fetch Alignments > MAF Coverage Stats. Using the summarized coverage output option, the coverage is the

number of nucleotides divided by total length of the given intervals.

**[6-9]** The goal of this exercise is to understand genomic alignments available at Ensembl. We will use a Linux machine. (1) Visit the Ensembl website ([http://useast.ensembl.org/Homo\\_sapiens/Info/Index](http://useast.ensembl.org/Homo_sapiens/Info/Index)) for Human (build GRCh37). There is a comparative genomics section with information about comparative analyses as well as downloads (<ftp://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/>). Visiting the ftp site we see directories for various groups of vertebrates. We select a folder called homologies (<ftp://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/homologies/>). It has five protein files. (2) Download the first file with the `wget` command:

```
$ wget ftp://ftp.ensembl.org/pub/release-71/emf/
ensembl-compara/homologies/Compara.71.protein
.aa.fasta.gz
```

This is 226 MB in size (as shown with the `ls -lh` command) and is a `.gz` (compressed) file. We unzip using the command `gunzip Compara.71.protein.aa.fasta.gz` and the resulting uncompressed file, called `Compara.71.protein.aa.fasta`, is large (1.6 GB in size). To see how many lines are in this file, type `wc -l Compara.71.protein.aa.fasta` (there are 29,337,499 rows). These are FASTA protein records with headers beginning with `>ENSEEUP0000006240`. Explore these. (3) Use `wget ftp://ftp.ensembl.org/pub/release-71/emf/ensembl-compara/homologies/Compara.71.protein.cds.fasta.gz` to obtain the coding sequence alignment for every `protein_tree` in FASTA format (this 565 MB file uncompresses to 4.8 GB). It contains nucleotide FASTA records (beginning with `>ENSEEUP0000006240`, corresponding to the hedgehog `ZNF235` gene).



## Self-Test Quiz

**[6-1]** Benchmarking refers to:

- (a) making a set of multiple sequence alignments (MSAs) from closely related proteins that form a trusted alignment;
- (b) making a set of MSAs from proteins which have had their tertiary structure determined, allowing the MSA to be validated based on structural criteria;

- (c) making a set of MSAs with an algorithm that are subsequently employed to refine tertiary structure predictions; or
- (d) making a set of MSAs from proteins which are known, based on structural criteria, to be members of distinct protein families.

**[6-2]** Why doesn't ClustalW (a program that employs the Feng and Doolittle progressive sequence alignment algorithm) report expect values?

- (a) ClustalW *does* report expect values;
- (b) ClustalW uses global alignments for which *E* value statistics are not available;
- (c) ClustalW uses local alignments for which *E* value statistics are not available; or
- (d) ClustalW uses combined global and local alignments for which *E* value statistics are not available.

**[6-3]** The “once a gap, always a gap” rule for the Feng–Doolittle method ensures that:

- (a) gaps will not be filled in inappropriately with inserted sequences;
- (b) sequences that diverged early in evolution will be given priority in establishing the order in which a multiple sequence alignment is constructed;
- (c) gaps occurring between sequences that are most closely related in a multiple sequence alignment will be preserved; or
- (d) gaps occurring between sequences that are distantly related will be maintained in the multiple sequence alignment.

**[6-4]** How can multiple sequence alignment programs improve performance?

- (a) by performing PSI-BLAST;
- (b) by incorporating data on secondary structure;
- (c) by incorporating data on three-dimensional structures; or
- (d) all of the above.

**[6-5]** What is a main strength of consistency-based approaches (such as ProbCons)?

- (a) they include information based on position-specific scoring matrices;
- (b) they include information based on three-dimensional protein structures, typically obtained from X-ray crystallography studies;
- (c) they perform profile–profile alignments and are extremely fast algorithms; or

(d) they include information based on multiple sequence alignments to guide the determination of pairwise alignments.

**[6-6]** The main difference between Pfam-A and Pfam-B is that:

- (a) Pfam-A is manually curated while Pfam-B is automatically curated;
- (b) Pfam-A uses hidden Markov models while Pfam-B does not;
- (c) Pfam-A provides full-length protein alignments while Pfam-B aligns protein fragments; or
- (d) Pfam-A incorporates data from SMART and PROSITE while Pfam-B does not.

**[6-7]** The Enredo/Pecan/Ortheus pipeline at Ensembl includes Ortheus to reconstruct ancestral sequences. One of its strengths is that:

- (a) it is “phylogeny-aware,” meaning that it accurately infers ancestral sequences;
- (b) it is specialized to model ancestral sequences using both protein and DNA information;
- (c) it creates phylogenetic trees, breaks them, then iteratively reconstructs them; or
- (d) it uses sampling with replacement of sequences to improve accuracy.

**[6-8]** What is a feature of algorithms that align large tracts of genomic DNA, in contrast to programs such as ClustalW that align smaller blocks of DNA or protein?

- (a) they are generally unable to align DNA from organisms that are highly divergent, such as those that speciated several hundred million years ago;
- (b) they generally use progressive alignment and so are fundamentally similar;
- (c) they often employ anchors that help to align regions of conservation that are interspersed with less-conserved regions (such as those arising in noncoding regions, deleted regions, or inverted regions); or
- (d) they are specialized to accept very long inputs.

## SUGGESTED READING

Da-Fei Feng and Russell F. Doolittle's (1987) progressive alignment approach to multiple sequence alignment is an important early paper. This work stresses the relationship between multiple sequence alignment and the evolutionary relationships of proteins. It is therefore relevant to our treatment of phylogeny in Chapter 7. Doolittle (2000) also

wrote a personal account of his interest in sequence analysis, phylogeny, and bioinformatics, including mention of the historical context in which he developed his alignment algorithm.

Excellent papers that review multiple sequence alignment and explain its challenges include Löytynoja (2012), Kemeny and Notredame (2009), Pirovano and Heringa (2008), Do and Katoh (2008), and Edgar and Batzoglou (2006).

## REFERENCES

- Aniba, M.R., Poch, O., Thompson, J.D. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research* **38**(21), 7353–7363. PMID: 20639539.
- Armougom, F., Moretti, S., Kedua, V., Notredame, C. 2006a. The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics* **22**, e35–39.
- Armougom, F., Moretti, S., Poirot, O. *et al.* 2006b. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Research* **34**(Web Server issue), W604–608. PMID: 16845081.
- Armougom, F., Poirot, O., Moretti, S. *et al.* 2006c. APDB: a web server to evaluate the accuracy of sequence alignments using structural information. *Bioinformatics* **22**, 2439–2440. PMID: 17032685.
- Batzoglou, S. 2005. The many faces of sequence alignment. *Briefings in Bioinformatics* **6**(1), 6–22. PMID: 15826353.
- Blackburne, B.P., Whelan, S. 2012. Measuring the distance between multiple sequence alignments. *Bioinformatics* **28**(4), 495–502. PMID: 22199391.
- Blanchette, M., Kent, W.J., Riemer, C. *et al.* 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* **14**, 708–715. PMID: 15060014.
- Blankenberg, D., Taylor, J., Nekrutenko, A., Galaxy Team. 2011. Making whole genome multiple alignments usable for biologists. *Bioinformatics* **27**(17), 2426–2428. PMID: 21775304.
- Chen, C., Natale, D.A., Finn, R.D. *et al.* 2011. Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PLoS One* **6**(4), e18910. PubMed PMID: 21556138.
- Chothia, C., Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**, 823–826.
- Davydov, E.V., Goode, D.L., Sirota, M. *et al.* 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**(12), e1001025. PMID: 21152010.
- Do, C.B., Katoh, K. 2008. Protein multiple sequence alignment. *Methods in Molecular Biology* **484**, 379–413. PMID: 18592193.
- Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research* **15**, 330–340.
- Doolittle, R. F. 2000. On the trail of protein sequences. *Bioinformatics* **16**, 24–33.
- Earl, D., Nguyen, N.K., Hickey, G. *et al.* 2014. Alignathon: A competitive assessment of whole genome alignment methods. *Genome Research* **24**(12), 2077–2089. PMID: 25273068.
- Edgar, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.
- Edgar, R. C. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.
- Edgar, R.C. 2010. Quality measures for protein alignment benchmarks. *Nucleic Acids Research* **38**(7), 2145–2153. PMID: 20047958.
- Edgar, R.C., Batzoglou, S. 2006. Multiple sequence alignment. *Current Opinion in Structural Biology* **16**(3), 368–373. PMID: 16679011.

- Feng, D. F., Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**, 351–360.
- Feng, D. F., Doolittle, R. F. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods in Enzymology* **183**, 375–387.
- Fitch, W.M., Yasunobu, K.T. 1975. Phylogenies from amino acid sequences aligned with gaps: the problem of gap weighting. *Journal of Molecular Evolution* **5**, 1–24.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Gotoh, O. A. 1995. Weighting system and algorithm for aligning many phylogenetically related sequences. *Computer Applications in the Biosciences* **11**, 543–551.
- Heringa, J. 1999. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computational Chemistry* **23**, 341–364.
- Higgins, D.G., Blackshields, G., Wallace, I.M. 2005. Mind the gaps: progress in progressive alignment. *Proceedings of the National Academy of Science, USA* **102**, 10411–10412.
- Hirosawa, M., Totoki, Y., Hoshida, M., Ishikawa, M. 1995. Comprehensive study on iterative algorithms of multiple sequence alignment. *Computer Applications in the Biosciences* **11**(1), 13–18. PMID: 7796270.
- Hogeweg, P., Hesper, B. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of Molecular Evolution* **20**(2), 175–186. PMID: 6433036.
- Hunter, S., Jones, P., Mitchell, A. *et al.* 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* **40**(Database issue), D306–312. PMID: 22096229.
- Kabsch, W., Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**(12), 2577–2637. PMID: 6667333.
- Karlin, S., Brocchieri, L. 1998. Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. *Journal of Molecular Evolution* **47**, 565–577.
- Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**(4), 772–780. PMID: 23329690.
- Katoh, K., Kuma, K., Toh, H., Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511–518.
- Katoh, K., Asimenos, G., Toh, H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods in Molecular Biology* **537**, 39–64. PMID: 19378139.
- Kemena, C., Notredame, C. 2009. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics* **25**(19), 2455–2465. PMID: 19648142.
- Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., Haussler, D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Science, USA* **100**(20), 11484–11489. PMID: 14500911.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGgettigan, P.A. *et al.* 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21), 2947–2948. PMID: 17846036.
- Lassmann, T., Sonnhammer, E.L. 2005. Kalign: an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**, 298.
- Lassmann, T., Sonnhammer, E.L. 2006. Kalign, Kalignvu and Mumsa: web servers for multiple sequence alignment. *Nucleic Acids Research* **34**(Web Server issue), W596–599.
- Letunic, I., Doerks, T., Bork, P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research* **40**(Database issue), D302–305. PMID: 22053084.
- Löytynoja, A. 2012. Alignment methods: strategies, challenges, benchmarking, and comparative overview. *Methods in Molecular Biology* **855**, 203–235. PMID: 22407710.
- Löytynoja, A., Goldman, N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Science, USA* **102**, 10557–10562.
- Margulies, E.H., Chen, C.W., Green, E.D. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends in Genetics* **22**, 187–193.

- McClure, M. A., Vasi, T. K., Fitch, W. M. 1994. Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biology and Evolution* **11**, 571–592.
- Mizuguchi, K., Deane, C.M., Blundell, T.L., Overington, J.P. 1998. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science* **7**, 2469–2471.
- Moretti, S., Armougom, F., Wallace, I.M., Higgins, D.G., Jongeneel, C.V., Notredame, C. 2007. The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Research* **35**(Webserver issue), W645–648.
- Needleman, S. B., Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.
- Notredame C., Higgins D.G., Heringa. J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205–217.
- O’Sullivan O., Zehnder, M., Higgins, D. *et al.* 2003. APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics* **19** Suppl 1, i215–221. PMID: 12855461.
- Park, J., Karplus, K., Barrett, C. *et al.* 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology* **284**, 1201–1210. PMID: 9837738.
- Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E. 2008a. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research* **18**(11), 1814–1828. PMID: 18849524.
- Paten B., Herrero, J., Fitzgerald, S. *et al.* 2008b. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research* **18**(11), 1829–1843. PMID: 18849525.
- Pirovano, W., Heringa, J. 2008. Multiple sequence alignment. *Methods in Molecular Biology* **452**, 143–161. PMID: 18566763.
- Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**(1), 110–121. PMID: 19858363.
- Punta, M., Coggill, P.C., Eberhardt, R.Y. *et al.* 2012. The Pfam protein families database. *Nucleic Acids Research Database Issue* **40**, D290–D301.
- Raghava, G.P., Searle, S.M., Audley, P.C., Barber, J.D., Barton, G.J. 2003. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* **4**, 47.
- Schwartz, S., Kent, W.J., Smit, A. *et al.* 2003. Human-mouse alignments with BLASTZ. *Genome Research* **13**(1), 103–107. PMID: 12529312.
- Siepel, A., Bejerano, G., Pedersen, J.S. *et al.* 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**(8), 1034–1050. PMID: 16024819.
- Sievers, F., Wilm, A., Dineen, D. *et al.* 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011). PMID: 21988835.
- Simossis, V.A., Heringa, J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research* **33**(Web Server issue), W289–294.
- Thompson, J. D., Higgins, D. G., Gibson, T. J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.
- Thompson, J. D., Plewniak, F., Poch, O. 1999a. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* **27**, 2682–2690.
- Thompson, J. D., Plewniak, F., Poch, O. 1999b. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87–88.
- Thompson, J.D., Koehl, P., Ripp, R., Poch, O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* **61**, 127–136.
- Van Walle, I., Lasters, I., Wyns, L. 2005. SABmark: a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **21**(7), 1267–1268. PMID: 15333456.

- Wallace, I.M., Blackshields, G., Higgins, D.G. 2005. Multiple sequence alignments. *Current Opinion in Structural Biology* **15**(3), 261–266. PMID: 15963889.
- Wallace, I.M., O’Sullivan, O., Higgins, D.G., Notredame, C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research* **34**(6), 1692–1699. PMID: 16556910.
- Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J. 2009. Jalview Version 2: a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9), 1189–1191. PMID: 19151095.
- Watson, H.C., Kendrew, J.C. 1961. The amino-acid sequence of sperm whale myoglobin. Comparison between the amino-acid sequences of sperm whale myoglobin and of human hemoglobin. *Nature* **190**, 670–672. PMID:13783432.
- Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z., Barker, W.C. 2004. The iProClass integrated database for protein functional analysis. *Computational Biology and Chemistry* **28**, 87–96.



(a) Test tubes for precipitin reactions



(c) Findings (portion)

1. Suborder ANTHROPOIDEA	MAMMALIA																	
	Class MAMMALIA		MAMMALIA															
	1. Order PRIMATES		ANTISERA FOR.....															
25	778.	<i>Homo sapiens</i> , G. N., cut (N) 12, vi, 02	Man	+	*	-	-	-	-	-	-	-	-	-	-	-	-	-
Fam. Hominidae	26	779. " " E. G., emi (Gardner) 13, v, 02	Chimpanzee	D	120	tr	-	-	-	-	-	-	-	-	-	-	-	-
European	27	785. " " C. C., wound (N) 15, v, 02	Orang	+	240	-	-	-	-	-	-	-	-	-	-	-	-	-
Mongolian	28	786. " " B. C., cut (N) 31, v, 02	Monkey	D	tr	30	-	-	-	-	-	-	-	-	-	-	-	-
	29	327. " " Chinaman, beri-beri London (Daniels) 14, xii, 01	Hedgehog	+	240	30	-	-	-	-	-	-	-	-	-	-	-	-
	30	380. " " Chinaman, beri-beri Shanghai, China (Stanley) ca. 27, xi, 01	Cat	+	240	40	-	-	-	-	-	-	-	-	-	-	-	-
E. Indian	31	381. " " Punjab Sikh Shanghai, China (Stanley) 29, x, 01	Rabbit	+	240	tr	-	-	-	-	-	-	-	-	-	-	-	-
	32	328. " " E. Indian, Punjab London (Daniels) 13, xii, 01	Dog	+	240	40	-	-	-	-	-	-	-	-	-	-	-	-
Negro	33	819. " " Negro Lagos, Africa (Strachan) 16, ii, 02	Seal	D	tr	30	-	-	-	-	-	-	-	-	-	-	-	-
	34	820. " " 16, ii, 02	Frog	+	240	30	-	-	-	-	-	-	-	-	-	-	-	-
	35	821. " " 16, ii, 02	Wallaby	D	tr	30	-	-	-	-	-	-	-	-	-	-	-	-

(b) Description of 16,000 tests

In the following pages the results of precipitin tests with haemoserum are given in the zoological order of the antisera, the tests made by other observers being summarized in each case, the results of my tests following.

The number of tests, made by me with 30 antisera produced, is given in the following table, the total number of tests being 16,000.

Antisera for	No. of tests therewith	Antisera for	No. of tests therewith
Man	825	Ox	790
Chimpanzee	47	Sheep	701
Orang	81	Horse	790
Ceropitheque	733	Donkey	94
Hedgehog	383	Zebra	94
Cat	785	Whale	94
Hyena	378	Wallaby	691
Dog	777	Fowl	792
Seal	358	Ostrich	649
Pig	818	Fowl-egg	789
Llama	363	Emu-egg	639
Mexican Deer	749	Turtle	666
Reindeer	69	Alligator	468
Hog Deer	699	Frog	551
Antelope	686	Lobster	456
	7751		8249
		Total number of tests	16,000

For the first half of the twentieth century, the main phylogenetic analyses based on molecular data were the remarkable precipitin tests pioneered by George Nuttall and colleagues. Antisera were incubated with serum samples from a variety of species, and the time required for a precipitation reaction was recorded as well as the strength of the reaction. (a) Sample test tubes in which the reactions were conducted (Nuttall, 1904, plate I). (b) Excerpt from Nuttall (1904, p. 160) describing the 16,000 tests he performed. (c) Portion of the 92-page data summary of Nuttall (1904, p. 222–223). The 900 rows (of which 11 are shown here) represent blood samples that were tested, and the columns correspond to antisera obtained from 30 organisms (of which 18 are shown here). The values represent the time (in minutes) required for a reaction. The symbols indicate the degree of reaction (+ being greatest, and - indicating no reaction). The letter D indicates the presence of deposits in the test tube. Nuttall used these data to infer the phylogenetic relationships of assorted mammals, birds, reptiles, amphibians, and crustaceans. In the 1950s and 1960s, amino acid sequence comparisons largely replaced immunological tests for phylogenetic analysis.

Source: Nuttall (1904).

# Molecular Phylogeny and Evolution

# CHAPTER 7

*Nothing in biology makes sense except in the light of evolution.*

—Theodosius Dobzhansky (1973)

## LEARNING OBJECTIVES

Upon completing this chapter you should be able to:

- describe the molecular clock hypothesis and explain its significance;
- define positive and negative selection and test its presence in sequences of interest;
- describe the types of phylogenetic trees and their parts (branches, nodes, roots);
- create phylogenetic trees using distance-based and character-based methods; and
- explain the basis of different approaches to creating phylogenetic trees and evaluating them.

## INTRODUCTION TO MOLECULAR EVOLUTION

Evolution is the theory that groups of organisms change over time so that descendants differ structurally and functionally from their ancestors. Evolution may also be defined as the biological process by which organisms inherit morphological and physiological features that define a species. In 1859 Charles Darwin published his landmark book, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*.

As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form.

Evolution is a process of change. Heredity is generally conservative – offspring resemble their parents – and yet the structure and function of bodies changes over the course of generations. There are three main mechanisms by which changes may occur (Simpson, 1952):

- Conditions of growth affect development. Environmental factors such as accidents and disease-causing infections are not hereditary in nature (although an individual's response to disease or environmental stimuli is genetically controlled to some extent, as discussed in Chapter 21).

We explore the tree of life in Chapter 15. You can read *The Origin of Species* by Charles Darwin online at <http://www.literature.org/authors/darwin-charles/the-origin-of-species/> (WebLink 7.1).

The word phylogeny is derived from the Greek *phylon* ("race, class") and *geneia* ("origin"). Ernst Haeckel, whose tree of life is shown on the frontis to Chapter 15, coined the terms *phylogeny*, *phylum*, and *ecology*. He also wrote that "ontogenesis is a brief and rapid recapitulation of phylogenesis, determined by the physiological functions of heredity (generation) and adaptation (maintenance)" (Haeckel, 1900, p. 81). See also <http://www.ucmp.berkeley.edu/history/haeckel.html> (WebLink 7.2).

- The mechanism of sexual reproduction ensures change from one generation to the next. DNA sequences including genes are "shuffled" via recombination when an offspring inherits chromosomes from two parents.
- Mutation with selection as well as genetic drift can produce changes in genes and more generally in chromosomes.

At the molecular level, evolution is a process of mutation with selection. Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life. This discipline also uses data from present-day organisms to reconstruct the evolutionary history of species.

Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny was assessed by comparing morphological features between organisms from a variety of species (Mayr, 1982). However, molecular sequence data can also be used for phylogenetic analysis. The evolutionary relationships that are inferred, which are usually depicted in the form of a tree, can provide hypotheses of past biological events.

## PRINCIPLES OF MOLECULAR PHYLOGENY AND EVOLUTION

### Goals of Molecular Phylogeny

All life forms share a common origin and are part of the tree of life. More than 99% of all species that ever lived are extinct (Wilson, 1992). Of the extant species, closely related organisms are descended from more recent common ancestors than distantly related organisms. In principle, there may be one single tree of life that accurately describes the evolution of species. One object of phylogeny is to deduce the correct trees for all species of life. Historically, phylogenetic analyses were based upon observable phenotypic features such as the presence or absence of wings or a spinal cord. More recently, phylogenetic analyses also rely on molecular sequence data that define families of genes and proteins. Another object of phylogeny is to infer or estimate the time of divergence between organisms since the time they last shared a common ancestor.

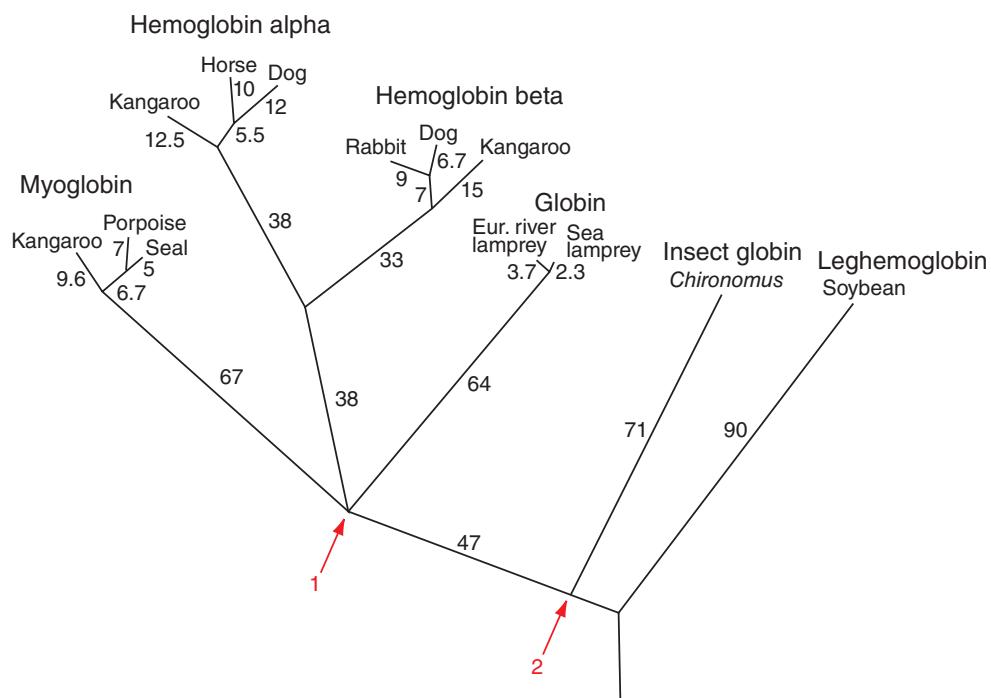
While the tree of life provides an appealing metaphor, our definition of evolution is not predicated on there necessarily being a single tree. Instead, evolution is based on a process of mutation and selection. We see in Chapter 17 that genes can be laterally transferred between species, complicating the ways in which organisms can acquire genes and traits. In many situations the tree of life has been described as a densely interconnected bush (or reticulated tree) rather than a simple tree with well-defined branches (e.g., Doolittle, 1999).

A *true tree* depicts the actual, historical events that occurred in evolution. It is essentially impossible to generate a true tree. Instead, we generate *inferred trees*, which depict a hypothesized version of the historical events. Such trees describe a series of evolutionary events that are inferred from the available data, based on some model.

The tree of life has three major branches: bacteria, archaea, and eukaryotes. We explore the global tree in Chapter 15. In this chapter we address the topic of phylogenetic trees that are used to assess the relationships of homologous proteins (or homologous nucleic acid sequences) in a family. Any group of homologous proteins (or nucleic acid sequences) can be depicted in a phylogenetic tree.

In Chapter 3 we defined two proteins as homologous if they share a common ancestor. You may perform a BLAST search and observe several proteins with high scores (low expect values) and simply view these database matches as related proteins that possibly have a related function. However, it is also useful to view orthologs and paralogs in an evolutionary context. We have applied a variety of approaches to study the relations of proteins: pairwise alignment using Dayhoff's scoring matrices (Chapter 3); BLAST searching (Chapters 4 and 5); and multiple sequence alignment (Chapter 6). We address the identification of related protein folds later in Chapters 12 and 13. All these approaches

Viruses are generally not considered to be part of the tree of life (see Chapter 16), although phylogenetic trees have been studied for all subgroups of viruses.



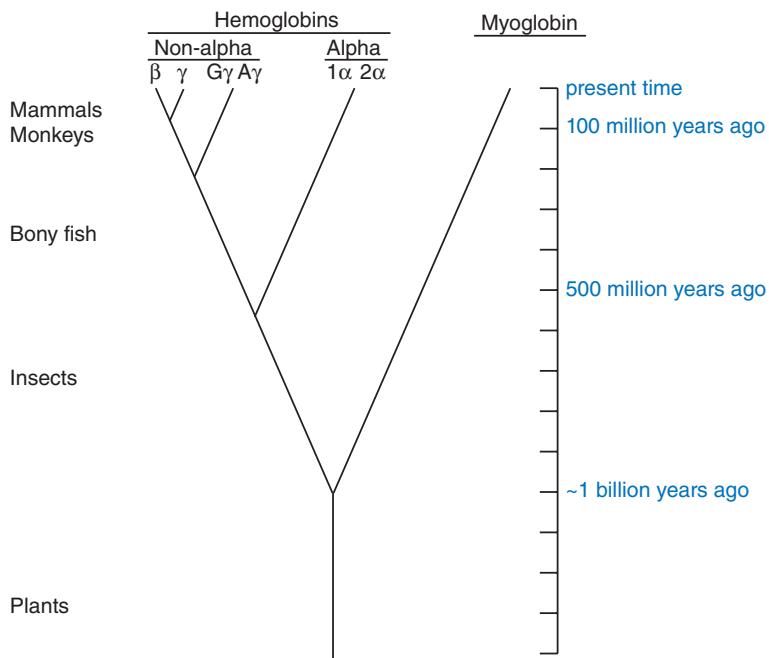
**FIGURE 7.1** In the 1960s, several groups performed pioneering studies of globin phylogeny. This tree is modified from Dayhoff et al. (1972) who used maximum parsimony analysis to infer the relationships and history of 13 globins. The observed percent difference between sequences was corrected using the data on PAM matrices in **Table 3.3**. Arrow 1 indicates a node corresponding to the last common ancestor of the group of vertebrate globins, while arrow 2 indicates the ancestor of the insect and vertebrate globins (see text for details).

Source: Dayhoff et al. (1972). Reproduced with permission from National Biomedical Research Foundation.

rely on evolutionary models to account for the observed similarities and differences between molecular sequences.

## Historical Background

Historically, the globins have been among the protein families most important to our understanding of biochemistry and molecular evolution, from the identification of hemoglobin in the 1830s and myoglobin in the 1860s to their crystallization in the nineteenth century for the purpose of comparative studies across species. (I describe this history in Web Document 7.1.) Globins were among the first proteins to be sequenced and to be analyzed using X-ray crystallography (Chapter 13). Following earlier work by Ingram (1961) and others to determine globin protein sequences, Eck and Dayhoff (1966) used parsimony analysis (defined in “Phylogenetic Inference: Maximum Parsimony” below) to generate trees of the globin family. We provided phylogenetic trees to introduce the concepts of paralogs (various human globins in Fig. 3.3) and orthologs (myoglobins in various species; Fig. 3.2). **Figure 7.1** shows a phylogenetic analysis of 13 globin proteins from various species, redrawn from Dayhoff et al. (1972). We return to these 13 sequences for phylogenetic analyses later in this chapter. **Figure 7.2** (also from Dayhoff et al., 1972) further provides a timeline of the events in which globin genes duplicated (e.g., an ancestral globin gene duplicated to form the lineages leading to modern alpha globin and beta globin), and also a timeline for speciation events (e.g., the modern fish and humans shared a common vertebrate ancestor ~400 million years ago or MYA). These studies focused on two aspects of phylogenetic trees. First, trees can depict the relatedness of particular protein subfamilies such as the alpha globins, beta globins, and myoglobins. Second,



**FIGURE 7.2** Dayhoff et al. (1972) summarized the relationship of the globin subfamilies in the context of evolutionary time. The dates of speciation events were inferred from fossil-based studies.

*Source:* Dayhoff et al. (1972). Reproduced with permission from National Biomedical Research Foundation.

Thirteen protein sequences corresponding to the proteins in Figure 7.1 are provided in Web Document 7.2 at <http://www.bioinfbook.org/chapter7>. We use these sequences as examples in this chapter. A similar phylogenetic tree was reported by Zuckerkandl and Pauling (1965).

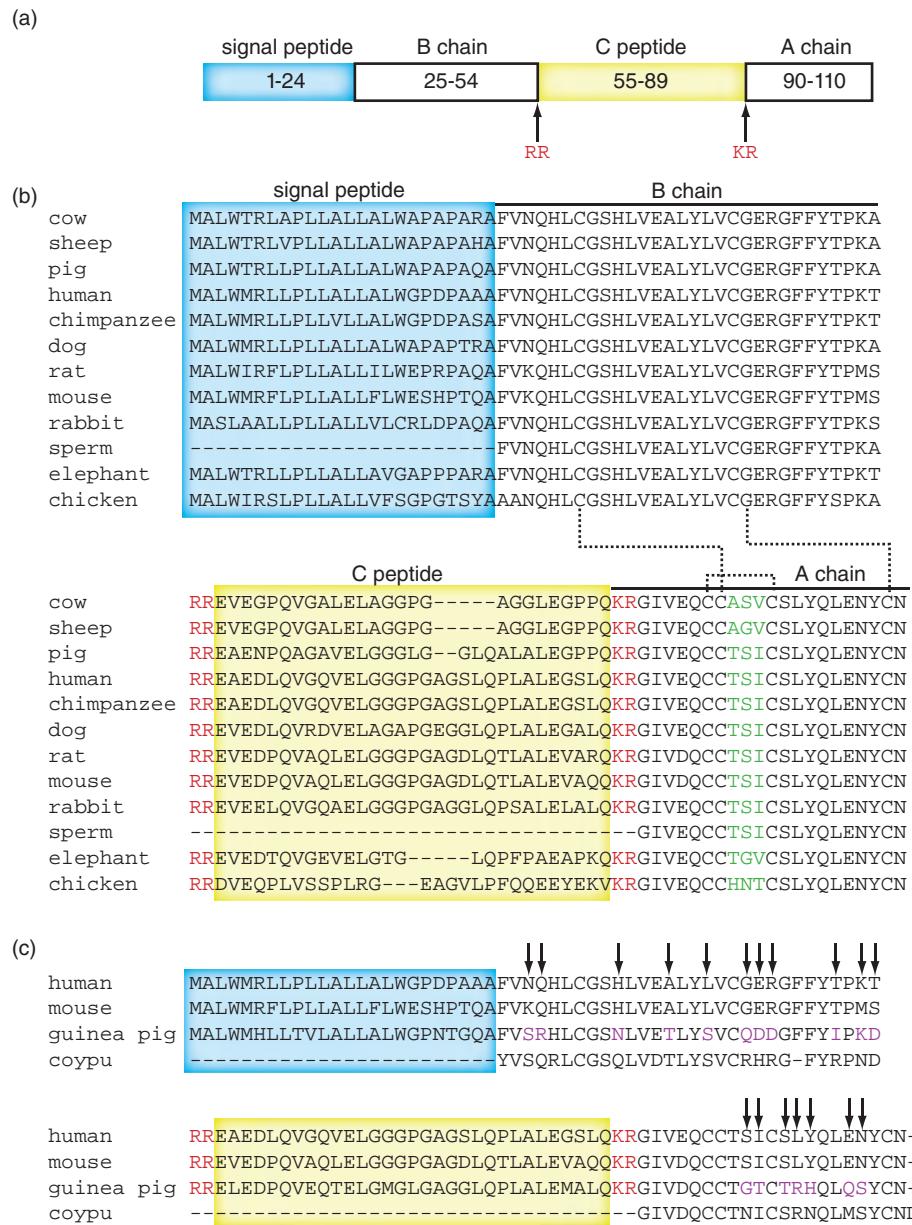
Frederick Sanger won the Nobel Prize in Chemistry (1958) "for his work on the structure of proteins, especially that of insulin" ([http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1958/](http://nobelprize.org/nobel_prizes/chemistry/laureates/1958/), WebLink 7.3). In 1980, he shared the Nobel Prize in Chemistry (with Paul Berg and Walter Gilbert) for his "contributions concerning the determination of base sequences in nucleic acids."

The protein sequences shown in Figure 7.3 are available at Web Document 7.3.

trees can depict the relatedness of species, providing inferences about the evolutionary history of life forms as well as the history of genes and gene products. We expand upon the relation of gene trees and species trees in the section "Species Trees versus Gene/Protein Trees".

Tremendous progress was also made in our understanding of molecular evolution through the study of insulin beginning in the 1950s. Insulin is a small protein secreted by pancreatic islet cells that stimulates glucose uptake upon binding to an insulin receptor on muscle and liver cells. In 1953 Frederick Sanger and colleagues determined the primary amino acid sequence of insulin, the first time this feat had been accomplished for any protein. The mature, biologically active protein consists of two subunits, the A chain and B chain, that are covalently attached through intermolecular disulfide bridges. More recently, the structure of the human preproinsulin molecule was shown to consist of a signal peptide, the B chain, an intervening sequence called the C peptide, and the A chain (Fig. 7.3a). The C peptide is flanked by dibasic residues (arg-arg or lys-arg; see Fig. 7.3a, b) at which proteolytic cleavage occurs.

Sanger and others sequenced insulin proteins from five species (cow, sheep, pig, horse, and whale). It became clear immediately that the A chain and B chain residues are highly conserved. Furthermore, amino acid differences were restricted to three residues within a disulfide "loop" region of the A chain (Fig. 7.3b, shaded turquoise). This suggested that amino acid substitutions occur nonrandomly, some changes affecting biological activity dramatically and other changes having negligible effects (Anfinsen, 1959). The differences within the disulfide loop are termed "neutral" changes (Jukes and Cantor, 1969, p. 86; Kimura, 1968). Later, when the biologically active A and B chain sequences were compared to the functionally less important C peptide, even more dramatic differences were seen. Kimura (1983) reported that the C peptide evolves at a rate of  $2.4 \times 10^{-9}$  per amino acid site per year, sixfold faster than the rate for the A and B chains ( $0.4 \times 10^{-9}$  per amino acid site per year). At the nucleotide level, the rate of evolution is similarly about sixfold faster for the DNA region encoding the C peptide (Li, 1997).



**FIGURE 7.3** Since the 1950s, studies of insulin have facilitated our understanding of molecular evolution. (a) The human insulin molecule consists of a signal peptide (required for intracellular transport; amino acid residues 1–24), the B chain, the C peptide, and the A chain. Dibasic residues (amino acids RR, KR) flank the C peptide and are the sites at which proteases cleave the protein. The A chain and B chain are then covalently linked through disulfide bridges, forming mature insulin. (b) Multiple sequence alignment of insulin from 12 species. Amino acid substitutions occur in nonrandom patterns. Note that within the A chain of insulin the amino acid residues are almost perfectly conserved between different species, except for three divergent columns of amino acids (A chain, residues in green font). However, the rate of nucleotide substitution is about six-fold higher in the region encoding the intervening C peptide than in the region encoding the B and A chains (Kimura, 1983), and gaps in the multiple sequence alignment are evident here. Disulfide bridges between cysteine residues are indicated by dashed lines. The accession numbers are NP\_000198.1 (human), NP\_001008996.1 (chimpanzee), NP\_062003.1 (rat), NP\_001123565.1 (dog), NP\_001172013.1 (mouse), NP\_001075804.1 (rabbit), NP\_001103242.1 (pig), NP\_990553.1 (chicken), NP\_001172055.1 (cow), P01318.2 (sheep), XP\_003422420.1 (elephant), and P67974.1 (sperm whale). (c) Guinea pig (*Cavia porcellus*, accession NP\_001166362.1) and coypu (*Myocastor coypus*, P01330.1) insulins evolve about seven-fold faster than insulin from other species. Human, mouse, guinea pig, and coypu insulins are aligned. Arrows indicate 18 amino acid positions at which the guinea pig sequence (purple) varies from that of human and/or mouse.

vasopressin-neurophysin 2-copeptin preproprotein [Homo sapiens]  
Sequence ID: ref|NP\_000481.2| Length: 164 Number of Matches: 1  
▶ See 5 more title(s)

Range 1: 20 to 28 GenPept Graphics			
NW Score	Identities	Positives	Gaps
47	7/9(78%)	7/9(77%)	0/9(0%)
Query 20 CYIQNCPLG 28		Oxytocin (NP_000906.1)	
CY QNCP G			
Sbjct 20 CYFQNCERG 28		Arginine vasopressin (NP_000481.2)	

**FIGURE 7.4** Human oxytocin (NP\_000906.1, residues 20–28) and arginine vasopressin (NP\_000481.2, residues 20–28) differ at only two amino acid positions, yet they have vastly different biological functions. The comparison of these peptide sequences in the 1960s led to the appreciation of the importance of primary amino acid sequences in determining protein function. The output of a pairwise alignment using BLASTP is shown.

Source: NCBI.

Perutz and Kendrew won the 1962 Nobel Prize in Chemistry “for their studies of the structures of globular proteins.” You can read about their accomplishments at <http://www.nobel.se/chemistry laureates/1962/> (WebLink 7.4). The oxytocin gene *OXT* encodes oxytocin/neurophysin I prepropeptide (NP\_000906.1, 155 amino acids); the arginine vasopressin gene *AVP* encodes vasopressin-neurophysin 2-copeptin preproprotein (NP\_000481.2, 164 amino acids including the 9 amino acid arginine vasopressing peptide 20–28).

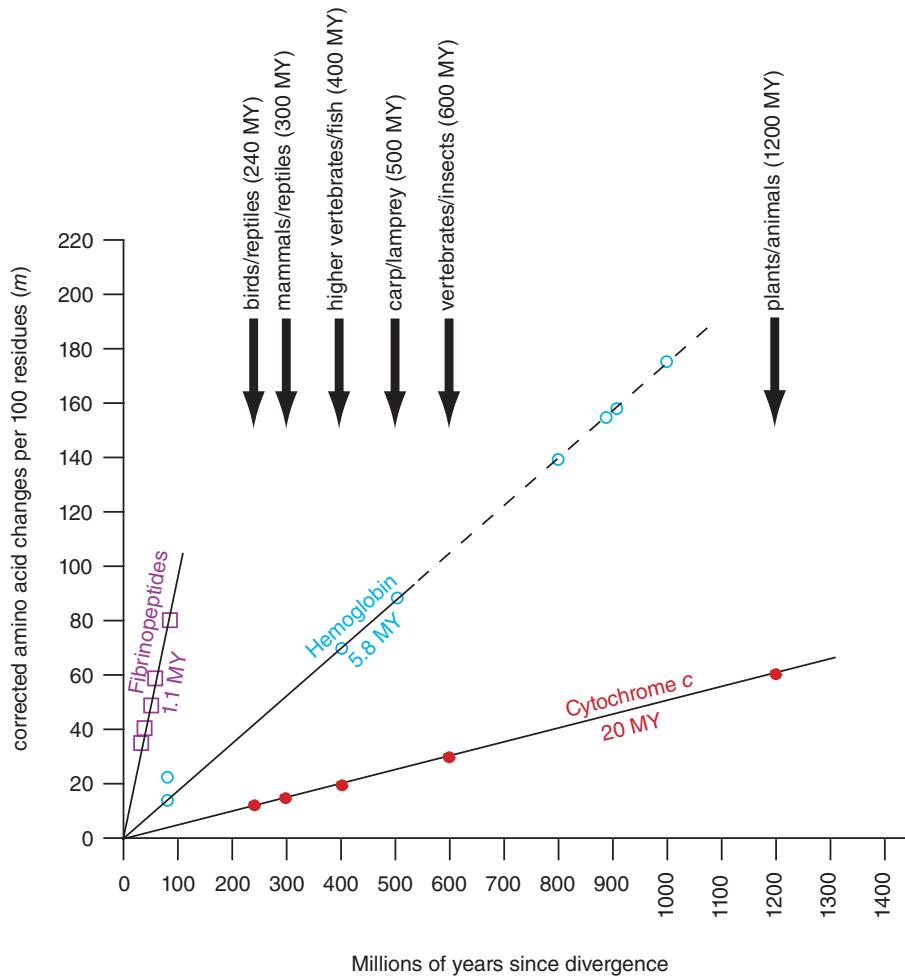
As insulin was sequenced from additional species, a surprising finding emerged. Insulin from guinea pig and a closely related species of the family *Caviidae* (the coypu) appeared to evolve seven times faster than insulin from other species. As shown in the alignment of **Figure 7.3c**, the guinea pig insulin sequence differs from human and mouse insulin at 18 different amino acid positions within the A and B chains. The explanation for this phenomenon (Jukes, 1979) is that guinea pig and coypu insulin do not bind two zinc ions, whereas insulin from all the other species do. Insulin from coypu has a very low biological potency (relative to pig or human insulin) and is primarily monomeric (Bajaj *et al.*, 1986). There is presumably a strong functional constraint on most insulin molecules to maintain amino acid residues that are able to complex zinc, but guinea pig and coypu insulin have less-selective constraint.

In the early 1950s, other laboratories sequenced vasopressin and oxytocin and found that peptides differing by only two amino acid residues have vastly different biological function (**Fig. 7.4**). In 1960 Max Perutz and John Kendrew solved the structures of hemoglobin and myoglobin. These proteins, both of which serve as oxygen carriers, are homologous and share related structures (see **Fig. 3.1**). It therefore became clear by the 1960s that there are significant structural and functional consequences to variation in primary amino acid sequence.

### Molecular Clock Hypothesis

In the 1960s, primary amino acid sequence data were accumulated for abundant, soluble proteins such as hemoglobins, cytochromes *c*, and fibrinopeptides in a variety of species. Some proteins, such as cytochromes *c* from many organisms, were found to evolve very slowly, while other protein families accumulated many substitutions. Emil Zuckerkandl and Linus Pauling (1962) as well as Emanuel Margoliash (1963) proposed the concept of a molecular clock (reviewed in Zuckerkandl, 1987). This hypothesis states that for every given gene (or protein), the rate of molecular evolution is approximately constant. In a pioneering study, Zuckerkandl and Pauling observed the number of amino acid differences between human globins including beta and delta (about 6 differences), beta and gamma (~36 differences), beta and alpha (~78 differences), and alpha and gamma (~83 differences). They could also compare human to gorilla (both alpha and beta globins), observing either 2 or 1 differences respectively, and they knew from fossil evidence that humans and gorillas diverged from a common ancestor about 11 million years ago. Using this divergence time as a calibration point, they estimated that gene duplications of the common ancestor to beta and delta occurred 44 million years ago (MYA); beta and gamma derived from a common ancestor 260 MYA; alpha and beta 565 MYA; and alpha and gamma 600 MYA.

For alignments of these globin proteins and a summary (including the correct number of differences) of the Zuckerkandl and Pauling (1962) data, see Web Document 7.4 at <http://www.bioinfbook.org/chapter7>.



**FIGURE 7.5** A comparison of the number of amino acid changes that occurs between proteins (y axis) versus the time since the species diverged (x axis) reveals that individual protein families evolve at distinct rates. Some proteins, such as cytochromes *c* from a variety of organisms, evolve very slowly; others such as hemoglobin evolve at an intermediate rate; and proteins such as fibrinopeptides undergo substitutions rapidly. This behavior is described by the molecular clock hypothesis, proposed by Zuckerkandl and Pauling (1962), Margoliash (1963), and others in the 1960s. The time of divergence of various organisms (arrows) is estimated primarily from fossil evidence. Abbreviation: MY, millions of years in the past. Adapted from Dickerson (1971) with permission from Elsevier.

A related study demonstrating the existence of a molecular clock was performed by Richard Dickerson in 1971 (Fig. 7.5). He analyzed three proteins for which a large amount of sequence data were available: cytochromes *c*, hemoglobins, and fibrinopeptides. For each, he plotted the relationship between the number of amino acid differences for a protein in two organisms versus the divergence time (in millions of years) for the organisms. These divergence times were estimated from paleontology.

When estimating the number of amino acid (or nucleic acid) differences between a group of sequences, a model is needed to explain the process by which substitutions occur; we address this subject later in this chapter. We have already encountered the idea that more mutational events occur than can be directly observed when we examined PAM matrices (Chapter 3). We saw that two proteins of length 100 that share 50% amino acid identity have sustained an average of 80 changes (Fig. 3.19). Notably, Zuckerkandl and Pauling (1962) had assumed for the purpose of their analyses that the number of observed

The correction formula of Equation (7.1) was written incorrectly in the original Margoliash and Smith (1965) article, but was used correctly by Dickerson (1971) and is further discussed by Fitch and Ayala (1994).

differences reflects the number of substitutions that have actually occurred. However, they acknowledged that the situation is more complicated because multiple substitutions may occur at any given site: “Thus the number of effective mutational events that have actually occurred since the  $\alpha$ - and  $\beta$ -chains have evolved from their common ancestor may be significantly greater than is presently apparent” (Zuckerkandl and Pauling, 1962, p. 204). Margoliash and Smith (1965, p. 233) as well as Zuckerkandl and Pauling (1965, p. 150) proposed a correction for the relationship between observed changes and actual changes. This correction was employed by Dickerson (1971; **Fig. 7.5**). The  $y$  axis of this plot consists of the corrected number of amino acid changes per 100 residues,  $m$ . The value of  $m$  is calculated via:

$$\frac{m}{100} = -\ln\left(1 - \frac{n}{100}\right). \quad (7.1)$$

This equation can be written as

$$\frac{n}{100} = 1 - e^{-(m/100)} \quad (7.2)$$

where  $m$  is the total number of amino acid changes which have occurred in a 100-amino acid segment of a protein and  $n$  is the observed number of amino acid changes per 100 residues. This correction adjusts for amino acid changes that occur but are not directly observed, such as two or more amino acid changes occurring in the same position (see “DNA, RNA, or Protein-Based Trees” below).

The results of this plot (**Fig. 7.5**) allow several conclusions (Dickerson, 1971):

- For each protein, the data lie on a straight line. This suggests that the rate of change of amino acid sequence has remained constant for each protein.
- The average rates of change are distinctly different for each protein. For example, fibrinopeptides evolve with a much higher rate of substitution. The time (in millions of years) for a 1% change in amino acid sequence to occur between two divergent lines of evolution is 20.0 MYA for cytochrome *c*, 5.8 MYA for hemoglobin, and 1.1 MYA for fibrinopeptides.
- The observed variations in rate of change between protein families reflect functional constraints imposed by natural selection.

The rate of amino acid substitution is measured by the number of substitutions per amino acid site per year,  $\lambda$ . Some values for  $\lambda$  are given in **Table 7.1**. Note that some proteins such as histones and ubiquitin undergo substitutions extraordinarily slowly.

Note that we say that histones undergo substitution very slowly, but we do not say that they *mutate* very slowly. Mutation is the biochemical process that results in a change in sequence. For example, a polymerase copies DNA (or RNA) with a particular mutation rate. Substitution is the observed change in nucleic acid or protein sequences (e.g., between various histones). The observed substitutions that are fixed in a population occur at a rate that reflects both mutation and selection, the process by which characters are selected for (or against) in evolution. If the rate of mutation of the DNA or RNA polymerases among an organism’s genes is relatively constant, then variation in substitution rates among those genes may be due primarily to positive or negative selection. In the language of Susumu Ohno (1970), some substitutions are *forbidden* because they are deleterious to the organism and are selected against. For example, substitutions in histones are almost always not tolerated, that is, they are lethal.

A significant implication of the molecular clock hypothesis is that if protein sequences evolve at constant rates, then they can be used to estimate the time that the sequences diverged. In this way phylogenetic relationships can be established between organisms.

**TABLE 7.1 Rates of amino acid substitutions per amino acid site per  $10^9$  years ( $\lambda \times 10^9$ ) in various proteins. Dayhoff (1978) expressed these rates as accepted point mutations (PAMs) per 100 amino acid residues that are estimated to have occurred in 100 million years of evolution (compare Box 3.4). The rate of mutation acceptance for serum albumin is 19 PAMs per 100 million years.**

Protein	Rate	Protein	Rate
Fibrinopeptides	9.0	Thyrotropin beta chain	0.74
Growth hormone	3.7	Parathyrin	0.73
Immunoglobulin (Ig) kappa chain C region	3.7	Parvalbumin	0.70
Kappa casein	3.3	Trypsin	0.59
Ig gamma chain C region	3.1	Melanotropin beta	0.56
Lutropin beta chain	3.0	Alpha 108asteur108108ne A chain	0.50
Ig lambda chain C region	2.7	Endorphin	0.48
Lactalbumin	2.7	Cytochrome $b_5$	0.45
Epidermal growth factor	2.6	Insulin (except Guinea pig and coypu)	0.44
Somatotropin	2.5	Calcitonin	0.43
Pancreatic ribonuclease	2.1	Neurophysin 2	0.36
Serum albumin	1.9	Plastocyanin	0.35
Phospholipase A2	1.9	Lactate dehydrogenase	0.34
Prolactin	1.7	Adenylate kinase	0.32
Carbonic anhydrase C	1.6	Cytochrome c	0.22
Hemoglobin alpha chain	1.2	Troponin C, skeletal muscle	0.15
Hemoglobin beta chain	1.2	Alpha 108asteur108108ne B chain	0.15
Gastrin	0.98	Glucagon	0.12
Lysozyme	0.98	Glutamate dehydrogenase	0.09
Myoglobin	0.89	Histone H2B	0.09
Amyloid AA	0.87	Histone H2A	0.05
Nerve growth factor	0.85	Histone H3	0.014
Acid proteases	0.84	Ubiquitin	0.010
Myelin basic protein	0.74	Histone H4	0.010

Source: Dayhoff *et al.* (1972). Reproduced with permission from National Biomedical Research Foundation and Columbia.

This is analogous to the dating of geological specimens using radioactive decay. An example of how the molecular clock may be used is given in Box 7.1.

The molecular clock hypothesis does not apply to all proteins, and a variety of exceptions and caveats have been noted:

- The rate of molecular evolution varies among different organisms. For example, some viral sequences tend to change extremely rapidly compared to other life forms.
- The clock varies among different genes (see Table 7.1) and across different parts of an individual gene (e.g., Fig. 7.3; see also the discussion on the gamma parameter in “Stage 3: Models of DNA and Amino Acid Substitution” below). The main force guiding the molecular clock is selection. Rodents tend to have a faster molecular clock than primates: this may be because their generation times are shorter and they have high metabolic rates.
- The clock is only applicable when a gene in question retains its function over evolutionary time. Genes may become nonfunctional (e.g., pseudogenes) leading to rapid

We discuss the duplication of an entire genome, followed by subsequent, rapid mutation and gene loss, in Chapters 18 (on the yeast *S. cerevisiae*) and 19 (eukaryotic genomes).

## BOX 7.1 RATE OF NUCLEOTIDE SUBSTITUTION $r$ AND TIME OF DIVERGENCE $T$

The rate of nucleotide substitution  $r$  is the number of nucleotide substitutions that occur per site per year. Similar calculations can be made for the rate of amino acid substitutions. These rates vary considerably and it is of interest to characterize whether a region evolves slowly or rapidly. The rate is defined:

$$r = \frac{K}{2T} \quad (7.3)$$

where  $T$  is the time of divergence of two extant sequences from a common ancestor.  $2T$  is used in the equation to reflect the time of divergence from a common ancestor on two separate lineages.  $T$  can sometimes be established based upon fossil (paleontological) data. As an example, the lineages leading to modern humans and rodents diverged about 90 million years ago.  $K$  is the number of substitutions per site. The  $\alpha$ -globins from rat and human differ by 0.093 nonsynonymous substitutions per site (Graur and Li, 2000); nonsynonymous changes are DNA substitutions in coding regions that result in a change in the amino acid that is specified. Given values for  $K$  and  $T$  we can estimate  $r$ :

$$r = \frac{0.093 \text{ substitutions per site}}{2(9 \times 10^7 \text{ years})}. \quad (7.4)$$

We therefore calculate that the  $\alpha$  chain of hemoglobin undergoes  $0.52 \times 10^{-9}$  nonsynonymous nucleotide substitutions per site per year. We can also use Equation (7.3) to estimate the time of divergence of two sequences given values for  $r$  and  $K$  ( $T = K/2r$ ; Graur and Li, 2000).

changes in nucleotide (and amino acid) sequence. The rate of evolution sometimes accelerates after gene duplication occurs. For example, after gene duplication generated  $\alpha$ - and  $\beta$ -hemoglobins, high rates of amino acid substitution occurred that presumably altered the function of the gene, allowing some globin proteins to be expressed at highly specific developmental stages.

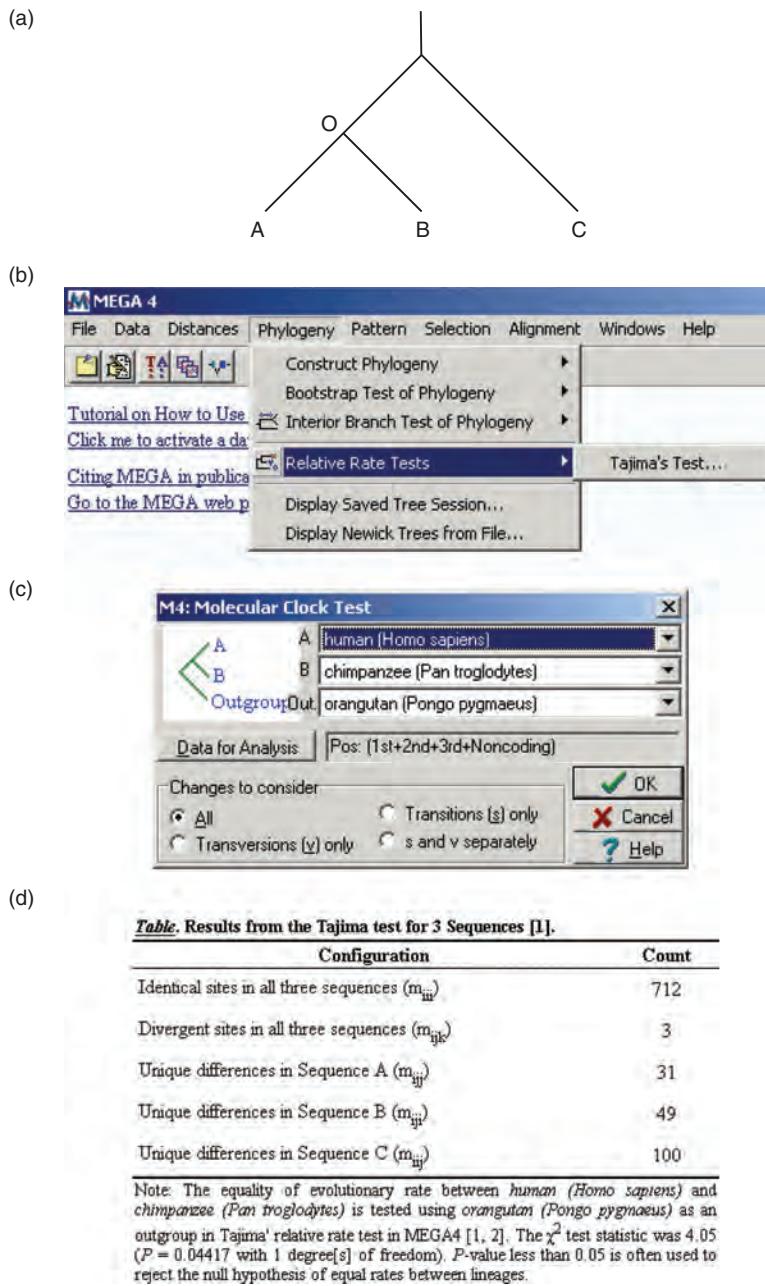
Despite these issues, the molecular clock hypothesis has proven useful and valid in many cases to which it is applied. Fitch and Ayala (1994) described a reasonably accurate molecular clock for Cu, Zn superoxide dismutase from a group of 67 protein sequences. However, obtaining correct inferences from the clock required tuning a variety of parameters.

As one practical approach to testing whether a molecular sequence has clock-like behavior, we can use the relative rate test of Tajima (1993; Box 7.2). For sequences A, B, C of the same protein or DNA/RNA from three species, let A and B be from two species from which we wish to compare the relative rates of evolution. Let C be a sequence from an outgroup, and let O be the common ancestor of A and B (Fig. 7.6a). Tajima's test determines whether there is accelerated evolution in lineage A or B, in which case we reject the null hypothesis that A and B exhibit equal evolutionary rates. Given the observed number of substitutions in sequence pairs AB, AC, and BC we can infer distances OA, OB, and OC and therefore test the null hypothesis that the relative rates OA and OB are the same (Fig. 7.6b-d). Tajima's relative rate test is implemented in MEGA software program (Tamura *et al.*, 2013). We use MEGA for phylogenetic analyses in this chapter. We provide a specific example of using the test in Problems/Computer Lab (7.1), including an explanation of how to enter the sequences into MEGA, align them, and perform Tajima's test.

MEGA is available at <http://www.megasoftware.net/> (WebLink 7.5).

### Positive and Negative Selection

Darwin's theory of evolution suggests that, at the phenotypic level, traits in a population that enhance survival are selected for (positive selection), while traits that reduce fitness are selected against (negative selection). For example, among a group of giraffes millions of years in the past, those giraffes that had longer necks were able to reach higher foliage



**FIGURE 7.6** A relative rate test to determine if two sequences follow the molecular clock hypothesis of approximately constant rates of amino acid or nucleotide substitution over evolutionary time. (a) Tajima (1993) proposed a relative rate test to determine whether protein or nucleic acid sequences from two organisms (A and B) have evolved at a similar relative rate. A and B share a common ancestor (O), and the sequence of an outgroup (C) is known. By measuring the substitution rates AB, AC, and BC it is possible to infer the rates OA and OB and to perform a chi square ( $\chi^2$ ) test to determine whether these rates are comparable (the null hypothesis) or whether one lineage has evolved at a relative accelerated or decelerated rate, thus violating the behavior of a molecular clock. Details of this test are presented in Kimura (1993) and Nei and Kumar (2000, p. 193–195). (b) Tajima's test is implemented in MEGA software. The pull-down menu for phylogenetic analysis is shown. (c) The test in MEGA allows the user to specify groups A, B, and C (outgroup). In this example mitochondrial DNA sequences from human and chimpanzee are compared using orang-utan DNA as an outgroup. (d) The output consists of a table listing the number of substitutions and an associated  $p$  value from a  $\chi^2$  test. In this example the  $p$  value is  $< 0.05$ , suggesting that the null hypothesis can be rejected and the human and chimpanzee sequences do not exhibit molecular clock-like behavior. This specific example is presented in problem (7.1) at the end of this chapter.

Source: MEGA version 5.2; Tamura et al. (2013). Courtesy of S. Kumar.

## BOX 7.2 TAJIMA'S RELATIVE RATE TEST

Tajima (1993) introduced a test for whether DNA or protein sequences in two lineages (such as human and chimpanzee) undergo evolution at equal rates. This is a test of the molecular clock: the null hypothesis is that there is an equal rate, and if we reject the null hypothesis at the 0.05 level then one of the lineages is evolving significantly faster or slower. For three protein or DNA sequences A, B, and C, let A and B be from two species we wish to compare and C is from an outgroup. For example, we can compare human and chimpanzee mitochondrial DNA using orangutan mitochondrial DNA as an outgroup. The relationships of A, B, and the outgroup C are shown in the form of a tree in **Figure 7.6a**. The observed number of sites  $n_{ijk}$  have the nucleotides  $i, j, k$  respectively. The expectation of  $n_{ijk}$  must equal that of  $n_{jik}$ , that is,

$$E(n_{ijk}) = E(n_{jik}). \quad (7.5)$$

If this equality occurs, the rate is constant per year; if it does not hold, the rate is not constant. We can measure the number of sites  $m_1$  in which residues in sequence A differ from those in B and C; similarly  $m_2$  corresponds to sites in B that are different than A and C. Given that C is an outgroup, the expectation of  $m_1$  must equal the expectation of  $m_2$ , that is,

$$E(m_1) = E(m_2). \quad (7.6)$$

This equality is tested with a chi-squared analysis:

$$\chi^2 = \frac{(m_1 - m_2)^2}{m_1 + m_2} \quad (7.7)$$

which results in a  $p$  value. If  $p < 0.05$ , the molecular evolutionary clock hypothesis is rejected at the 5% level, regardless of the substitution model. Tajima's relative rate test is implemented in Molecular Evolutionary Genetics Analysis (MEGA) software (Tamura *et al.*, 2013). For the mitochondrial sequences analyzed in **Figure 7.6**, there were 31 unique sequence differences in A (human) and 49 unique differences in B (chimpanzee), so the  $\chi^2$  test statistic was 4.05. This was obtained from:

$$\chi^2 = \frac{(31 - 49)^2}{31 + 49}. \quad (7.8)$$

This corresponds to a  $p = 0.04$  with 1 degree of freedom, suggesting that we may reject the null hypothesis of equal rates between lineages. In using Tajima's test it is important to select an outgroup that is an appropriate evolutionary distance from the two organisms you are comparing. For example, the bonobo or pygmy chimpanzee (*Pan paniscus*) may be too closely related to human and chimpanzee as all three species diverged about 5–7 million years ago; it is a problem for an intended outgroup to have the properties of an ingroup. At the other extreme, rat or mouse are too divergent as they diverged from the primate lineage about 90 million years ago. Suitable choices may include primates such as orang-utan or gorilla; select the closest true outgroup that is available.

and were more reproductively successful than their shorter-necked group members, that is, there was positive selection for height.

At the molecular level, a conventional evolutionary point of view is that positive and negative selection also operate on DNA sequences. A gene encoding an enzyme may duplicate (see Chapters 18 and 19), and then subsequent nucleotide changes may allow one of the duplicated genes to encode an enzyme with a novel function that is advantageous and hence selected for. This process of positive selection is thought to have occurred on two occasions in the evolution of lysozyme, an enzyme that breaks down bacterial peptidoglycan linkages and therefore serves as an antimicrobial protein in sources such as milk, saliva, and tears. About 25 million years ago the lysozyme gene duplicated and assumed a novel digestive function in stomach in the ancestor of goats, cows, and deer. The emergence of this novel function occurred independently in leaf-eating monkeys such as the langur some 15 million years ago (Jollès *et al.*, 1990). In each of these instances the rate of amino acid replacement increased due to positive selection as the lysozyme assumed a novel function. Other examples of positive selection include the primate ribonuclease genes (Zhang and Gu, 1998) and the MEDEA genes in plants (Spillane



**FIGURE 7.7** Phylogenetic trees can be constructed using DNA, RNA, or protein sequence data. Often, the DNA sequence is more informative than protein in phylogenetic analysis. As an example, the sequences of beta globin from three species are aligned at the 5' end of the DNA (with the corresponding amino termini of the proteins). In the 5' and 3' untranslated regions, where no protein is encoded, there is typically less selective pressure to maintain particular nucleotide residues. (Some regulatory elements may be highly conserved.) Here, just one nucleotide position varies (arrow). Within the protein-coding region, there are variant amino acid residues at amino acid positions 6, 7, and 11 (see green arrowheads). These variants may be informative in performing phylogeny. However, there is an even greater number of informative nucleotide changes, restricting our attention to the coding region. There are six positions of synonymous nucleotide changes (nucleotides shaded blue; see codons 3, 7, and 10–12) that do not cause a different amino acid to be specified. There are also six positions with nonsynonymous changes that do cause an amino acid change (red arrowheads and nucleotides). For one of these (codon 6 of the dog sequence), a single-nucleotide change of C→G, relative to the primate sequences, accounts for the amino acid change. For three other nonsynonymous codons, two nucleotides are changed relative to the primate sequences. The beta globin sequences are from human (GenBank accession NM\_000518.4), chimpanzee (*Pan troglodytes*; XM\_508242.3), mouse (*Mus musculus*; NM\_016956.3), and dog (*Canis lupus familiaris*; NM\_001270884.1).

*et al.*, 2007). In general, variants that have undergone a “selective sweep” have increased in prevalence through positive selection (Cutter and Payseur, 2013).

There are several ways to assess whether selection has occurred in sequence data. One approach relies on the fact that the portion of DNA that codes for a protein can have both synonymous and nonsynonymous substitutions. For a nucleotide change in a given codon, a synonymous substitution does not result in a change in the amino acid that is specified. For example, consider an alignment of human, chimpanzee, mouse, and dog beta globin DNA sequences at their 5' ends (amino termini of the proteins; Fig. 7.7). In the third codon the nucleotides CAT in the human and dog sequences encode a histidine. Changing the third position to yield CAC in the chimpanzee and mouse sequences does not alter the amino acid that is encoded. Other synonymous changes are evident (Fig. 7.7, red-colored nucleotides). A nonsynonymous substitution does change the amino acid that is specified. For example, human and chimpanzee beta globin have a CCT codon that specifies a proline, but the corresponding canine sequence has a single substitution resulting in a codon (GCT) that specifies an alanine (Fig. 7.7, codon 6).

Comparison of the rates of nonsynonymous substitution per nonsynonymous site ( $\hat{d}_N$ ) versus synonymous substitution per synonymous site ( $\hat{d}_S$ ) may reveal evidence of positive or negative selection. If  $\hat{d}_S$  is greater than  $\hat{d}_N$ , this suggests that the DNA sequence is under negative or purifying selection. Negative selection limits change in a corresponding amino acid sequence; this occurs when some aspect of the structure and/or function of a protein is critical and cannot tolerate substitutions. When  $\hat{d}_N$  is greater than  $\hat{d}_S$ , this

Refer to the genetic code in  
Box 3.6.

SNAP is available at the HIV sequence database website (<http://www.hiv.lanl.gov/>, WebLink 7.6) in the tools menu. Web Document 7.5 introduces 12 globin DNA coding sequences (11 myoglobin orthologs plus one cytoglobin sequence as an outgroup); see <http://www.bioinfbook.org/chapter7>. That file includes multiple sequence alignments of those sequences. We use these sequences as examples later in this chapter. Web Document 7.6 provides an example of how to use four of those globin-coding sequences to test for selection using SNAP software, while Web Document 7.7 shows an example of tests for selection in MEGA software. Datammonkey is available at <http://www.datammonkey.org/> (WebLink 7.7).

suggests that positive selection occurs. An example of positive selection is a duplicated gene that is under pressure to evolve new functions.

A variety of computer programs assess the ratio of synonymous to nonsynonymous substitutions. One is Synonymous Nonsynonymous Analysis Program (SNAP), which requires codon-aligned nucleotide sequences as its input (Korber, 2000). Datammonkey is a suite of tools including robust maximum likelihood approaches for determining positive or negative selection (Delport *et al.*, 2010). MEGA employs the Nei and Gojobori (1986) method to test the null hypothesis that the sequences are under either positive, negative, or neutral selection (Tamura *et al.*, 2013).

There is considerable interest in measuring positive or negative selection on a genome-wide basis. Many approaches have been adopted (Nielsen, 2005, Sabeti *et al.*, 2006). For example, Bustamante *et al.* (2005) studied the DNA sequence 11,000 genes in 39 individuals and reported rapid amino acid evolution at 9% of the informative loci. For many of the genomes that have recently been sequenced (e.g., human, chimpanzee, dog, chicken, rat), a description of those genes that are under positive selection is a basic part of the genome analysis (see Chapter 19).

Positive and negative can also be studied on a highly compressed time scale in viruses. In 1978, 500 women were inadvertently infected with hepatitis C virus (HCV). Stuart Ray and colleagues (2005) sequenced a 5.2 kilobase portion of the HCV genome from the original inoculum and from 22 women about 20 years after the infection. They showed loci with both positive and negative selection, reflecting the evolution of the virus to optimize its fitness in each host. For example, amino acid substitutions in known epitopes diverged from the consensus sequence in individuals having the human leukocyte antigen (HLA) allele for that epitope, indicating a mechanism of immune selection. In another study, Cox *et al.* (2005) studied sequence variation of HCV both before, during, and after HCV infection. They showed that amino acid substitutions reflect escape from T cell recognition; in those individuals with persistent infection, there were selection pressures on epitopes that resulted in nonsynonymous changes. The Ray *et al.* (2005) and Cox *et al.* (2005) results provide examples of the usefulness of longitudinal studies in phylogeny, and they reveal mechanisms through which positive and natural selection shape the fitness of viruses.

## Neutral Theory of Molecular Evolution

There is a tremendous amount of DNA polymorphism in all species that is difficult to account for by conventional natural selection. We examine this throughout the tree of life in Part III. In Chapter 8, we examine single-nucleotide polymorphisms (SNPs), an extremely common form of polymorphism that does not appear to be under selection in most instances. Similarly, many chromosomal copy number variants occur in apparently normal individuals (Chapter 8). These involve multiple regions of up to millions of base pairs of DNA that are deleted or duplicated. The majority of copy number variants appear to be sporadic, benign and not under positive or negative selective pressure.

In the decades up to the 1960s the prevailing model of molecular evolution was that most changes in genes are selected for or against in a Darwinian sense. Motoo Kimura (1968, 1983) proposed a different model to explain evolution at the DNA level. Kimura (1968) noted that the rate of amino acid substitution averages approximately one change per  $28 \times 10^6$  years for proteins of 100 residues. He further estimated that the corresponding rate of nucleotide substitution must be extremely high (one base pair of DNA replaced in the genome of a population every 2 years on average).

Kimura's conclusion was that most observed DNA substitutions must be neutral or nearly neutral, and that the main cause of evolutionary change (or variability) at

Kimura (1968) based his calculations on substitution rates measured within the families of alpha and beta globin, cytochrome c, and triosephosphate dehydrogenase proteins.

the molecular level is random drift of mutant alleles. Most nonsynonymous mutations are deleterious, and are therefore not observed as substitutions in the population. Under this model, called the neutral theory of evolution, positive Darwinian selection plays an extremely limited role. Indeed, the existence of a molecular clock makes sense in the context of the neutral hypothesis because most amino acid substitutions are neutral. (Substitutions are therefore tolerated by natural selection to change in a manner that has clock-like properties. If substitutions occurred primarily in the context of positive or negative selection, it is unlikely that they could account for clock-like evolution.) In the decades since his 1983 publication, the neutral theory continues to be tested in a variety of organisms. We explore some of these studies when we consider the eukaryotic chromosome in Chapter 8.

A typical human genome has ~3.5 million SNPs, most of which are intergenic (residing outside genes). Of the SNPs in exons, about 11,000 are synonymous and ~11,000 are nonsynonymous. See Chapter 20.

## MOLECULAR PHYLOGENY: PROPERTIES OF TREES

Molecular phylogeny is the study of the evolutionary relationships among organisms or among molecules using the techniques of molecular biology. Many other techniques are used to study evolution, including morphology, anatomy, paleontology, and physiology. We focus on phylogenetic trees using molecular sequence data, and begin with an explanation of the nomenclature used to describe trees.

### Topologies and Branch Lengths of Trees

There are two main kinds of information inherent in any phylogenetic tree: the topology and the branch lengths. The topology of a tree defines the relationships of the proteins (or other objects) that are represented in the tree. For example, the topology shows the common ancestor of two homologous protein sequences. The branch lengths sometimes (but not always) reflect the degree of relatedness of the objects in the tree.

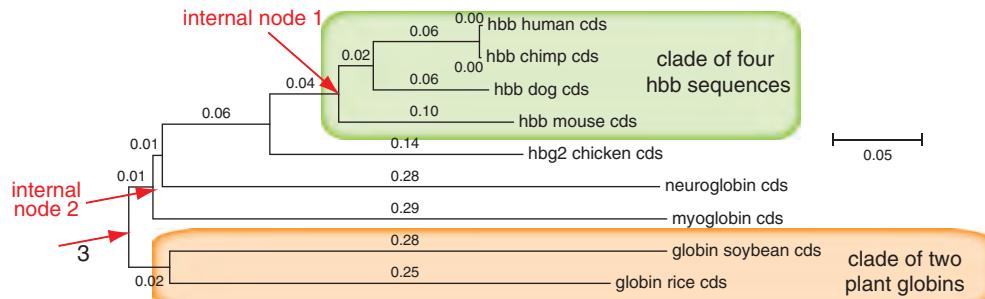
We define the main parts of a tree and the main types of trees, using nine globin coding sequences as an example. The four trees shown in **Figure 7.8a–d** are all made from the same input dataset (a multiple sequence alignment), and present alternative ways to view and analyze the data. You can make these same trees in MEGA (Tamura *et al.*, 2013) by following these steps. (1) Install MEGA software. (2) Copy the set of nine DNA sequences. (3) In MEGA select Align > Edit/Build Alignment > Create a New Alignment > DNA (see **Fig. 7.9a, b**) and paste the sequences into the Alignment Explorer. (4) Select the sequences and choose Alignment > Align by MUSCLE (codons). Keeping the default options, select Compute. (5) Save the alignment (via Data > Save Session to create a .mas file suffix) and choose Data > Export Alignment > MEGA format to save the alignment as 9globin\_cds.meg. (6) Choose Phylogeny > Construct/Test Neighbor-Joining Tree (**Fig. 7.9a**). A dialog box opens with many options (**Fig. 7.9c**); choose Compute and a phylogenetic tree is generated.

A phylogenetic tree is a graph composed of branches and nodes. Only one branch (also called an edge) connects any two nodes. The nodes represent the taxonomic units (taxa or taxons); the node (from the Latin for “knot”) is the intersection or terminating point of two or more branches. For us, taxa will typically be DNA or protein sequences. An operational taxonomic unit (OTU) is an extant taxon present at an external node or leaf; the OTUs are the available nucleic acid or protein sequences that we are analyzing in a tree. The internal nodes represent ancestral sequences that we can infer but can only very rarely observe (as in the case of sequencing DNA from extinct organisms, discussed in Chapter 15).

Consider the trees in **Figure 7.8**. Each tree includes nine OTUs (globins) which define the external nodes. In addition, there are internal nodes, each of which represents an inferred ancestor of the OTUs. For example, internal node 1 (**Fig. 7.8a**) corresponds to the globin DNA sequence of the ancestral sequence of mouse, dog, chimpanzee,

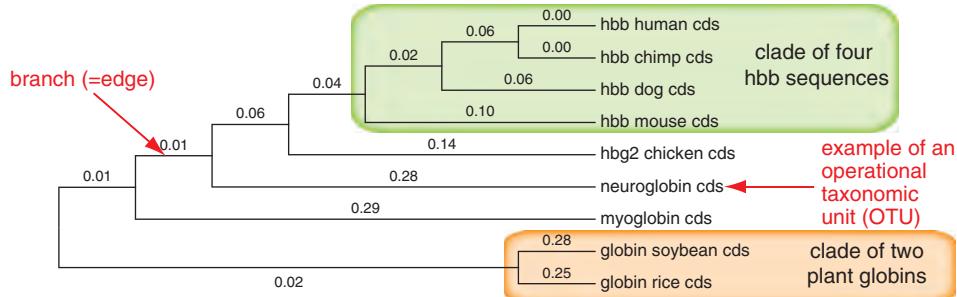
You can download MEGA on any platform from <http://www.megasoftware.net/>. The nine globin coding sequences (and their accession numbers) are available from Web Document 7.8 at <http://bioinfbook.org> (WebLink 7.8). 9globin\_cds.meg is available as Web Document 7.9; upon starting MEGA you can import this file directly. These nine globin DNAs correspond to the nine globin proteins we studied in Chapter 6. (DNA sequences are not yet available for some of the 13 globin proteins we viewed in **Fig. 7.1**.)

(a) Nine globin coding sequences: neighbor-joining tree (rectangular tree style)

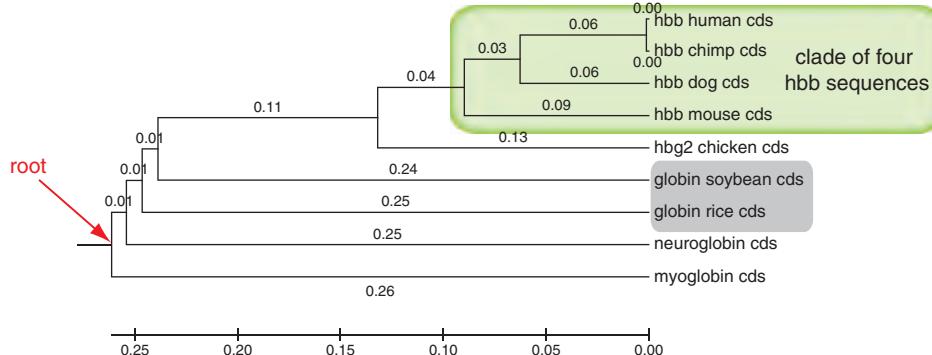


**FIGURE 7.8** Phylogenetic trees contain nodes and branches, and are defined by the branch lengths and topology. These trees were created with MEGA software by importing nine globin DNA coding sequences, aligning them using MUSCLE, and creating or displaying the phylogenetic trees four different ways. (a) Neighbor-joining (NJ) tree. Two clades are highlighted, and two internal nodes are indicated. Branch lengths were calculated by the *p*-distance correction in the units of number of nucleotide differences per site. (b) The tree was made as in (a) with the option to show topology only. Branch length values are shown (as in (a)) but the lengths of each branch are not proportional. Note that the operational taxonomic units (OTUs, i.e., the nine extant sequences) are now neatly aligned at the right side of the tree. An example of a typical branch is indicated. (c) The same dataset is used to create a tree by the UPGMA rather than neighbor-joining method. Note that the beta globin (hbb) clade maintains the same topology, but the two plant globins (highlighted with gray background) have a different topology (unrealistically sharing a common ancestor with vertebrate globins other than neuroglobin and myoglobin). The UPGMA tree is rooted. (d) The neighbor-joining tree of (a) is plotted as a radial tree.

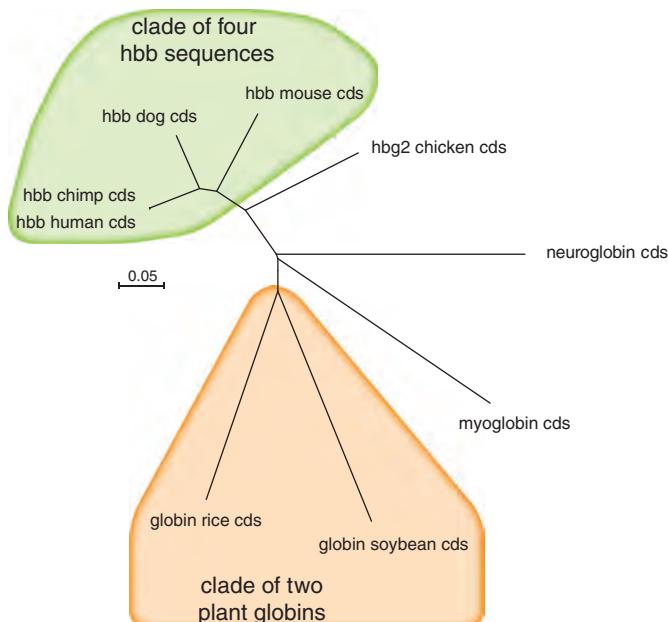
(b) Nine globin coding sequences: neighbor-joining tree ("topology only" tree style)



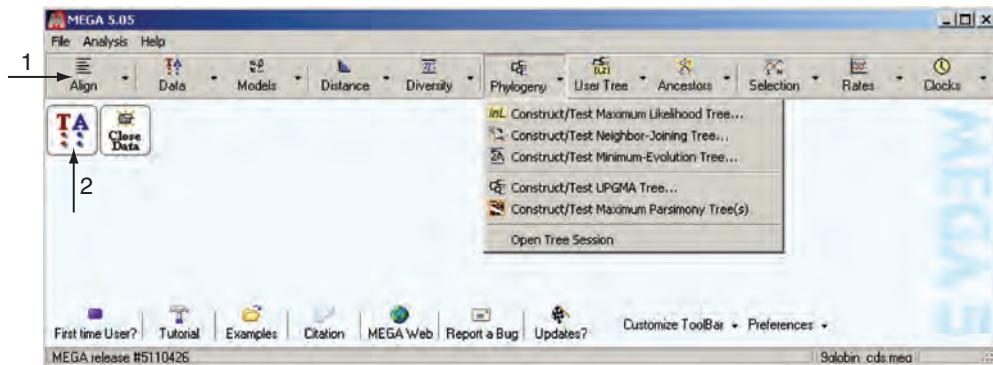
(c) Nine globin coding sequences: UPGMA tree



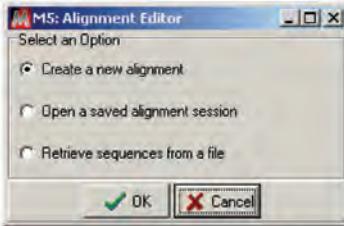
(d) Nine globin coding sequences: neighbor-joining tree (radial tree style)



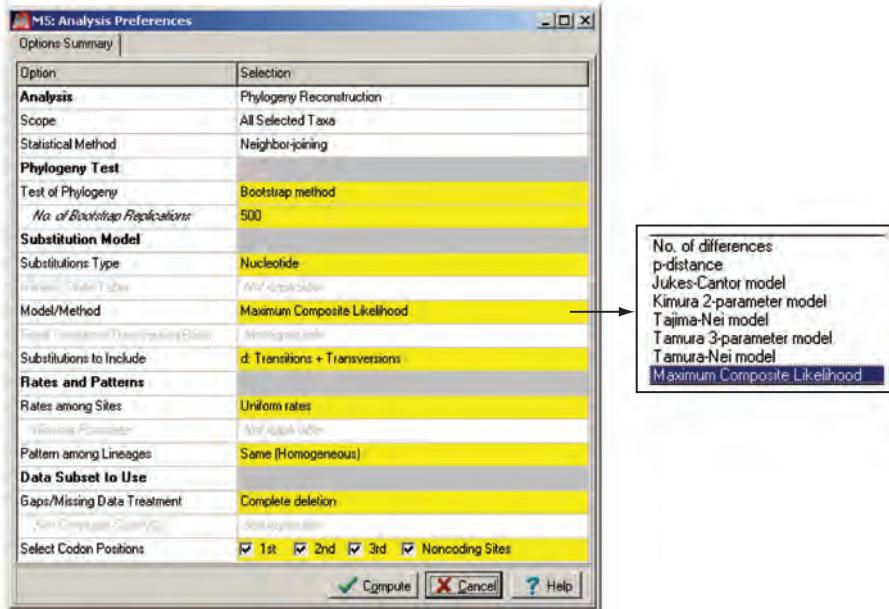
(a) MEGA main dialog box



(b) Alignment editor



(c) Analysis preferences (for making a neighbor-joining tree)



**FIGURE 7.9** Using MEGA to make and analyze phylogenetic trees. (a) The main dialog box includes an option to import and align DNA or protein sequences (arrow 1). Once entered the data can be viewed (arrow 2) and manipulated (e.g., you may include/exclude particular taxa or sequence positions). The pull-down menu for phylogenetic analysis is shown. (b) Alignment editor. It is good practice to save sequences for analysis as a text file so they can be studied later. (c) Analysis options for making a neighbor-joining tree. Similar preference boxes are available for other tree-making and analysis methods. Options for the Model/Method are shown to the right.

Source: MEGA version 5.2; Tamura et al. (2013). Courtesy of S. Kumar.

and human beta globin. (I like to imagine a small, furry creature that roamed the world about 100 MYA.) Node 1 represents the beta globin sequence of that creature. Following a series of speciation events the dog, rodent, and primate lineages emerged, each with its beta globin sequence that underwent changes to the forms we observe today. Internal node 2 represents an ancestral sequence that existed in an organism that predated the divergence of metazoans (animals) and plants some 1500 MYA (1.5 billion years ago).

Some OTUs can be swapped (that is, rotated or exchanged) without altering the topology of the tree. For example, the soybean globin is depicted above the rice globin in **Figure 7.8a**, but if it is swapped (rotated about the node so that the rice is now on top) the topology has not been changed. In general OTUs or clades that share an immediate ancestor node can be rotated on that node. Others cannot be swapped however, such as soybean globin and any of the other OTUs. (When you view a tree in MEGA, a tool allows you to swap branches.)

Branches define the topology of the tree, that is, the relationships among the taxa in terms of ancestry. In the trees of **Figure 7.8**, the branches leading to each of the nine OTUs are called external branches (or peripheral branches). The others are called internal branches.

Branch lengths should be defined for every tree. In some trees, the branch length represents the number of nucleotide or amino acid changes that have occurred in that branch. In **Figure 7.8a, c, d** scale bars are given, and the branch lengths are in units of base differences per site. This format (called a phylogram) has the helpful feature of conveying a clear visual idea of the relatedness of different proteins within the tree. In **Figure 7.8b**, the branches are unscaled. This implies that they are not proportional to the number of changes. This form of presenting a tree (called a cladogram) has the advantage of aligning the OTUs neatly in a vertical column. This may be especially useful if the tree has many dozens of OTUs. Note however the branches leading to soybean and rice: they have lengths of 0.28 and 0.25 and are drawn to scale in **Figure 7.8a** but not **7.8b**.

An internal node is bifurcating if it has only two immediate descendant lineages (branches). Bifurcating trees are also called binary or dichotomous; any branch that divides splits into two daughter branches. A tree is multifurcating if it has a node with more than two immediate descendants. It is not uncommon to see such trees in the literature, particular when it is challenging to resolve the relationships between closely related species or sequences.

A clade is a group of all the taxa that have been derived from a common ancestor plus the common ancestor itself. A clade is also called a monophyletic group. In our context, a clade is a set of sequences that form a group within a tree. In the example of any of the trees in **Figure 7.8a-d** a clade of four beta globin sequences is highlighted, including three internal nodes. Chicken *HBG2* is not a member of this clade. Another clade is highlighted, including plant globins (rice and soybean) and their common ancestor.

## Tree Roots

A phylogenetic tree can have a root representing the most recent common ancestor of all the sequences. Our set of nine globin DNA sequences is represented as a rooted tree in **Figure 7.8c**. If one assumes a constant molecular clock, then time and distance are proportional: the direction of time moves from oldest (at the root) to newest (at the OTUs). Often the root is not known today, and some tree-making algorithms do not provide conjectures about placement of a root. The alternative to a rooted tree is an unrooted tree (shown in **Fig. 7.8a, b, d**). An unrooted tree specifies the relationships among the OTUs. However, it does not define the evolutionary path completely or make assumptions about common ancestors.

A multifurcation is also called a polytomy. Multifurcating trees are by definition nonbinary. For an example of a multifurcating tree, see Rokas *et al.* (2005). They reported that many metazoan (animal) phyla are unresolved, reflecting a temporal compression due to the rapid radiation of many animal groups. Philippe and colleagues (Baurain *et al.*, 2007) suggested that such multifurcations occur in a phylogenetic tree because of insufficient sampling. For an example of multifurcation see **Figure 7.29** below.

If a tree is unrooted you may choose to add a root. The two main ways to do this are by specifying an outgroup and by midpoint rooting. To specify an outgroup, include one or more sequences that are known to have diverged earlier than the rest. Consider **Figure 7.8a, b, d**. Since plants diverged from the vertebrate lineage about 1500 MYA, the two plant globins could be selected to define the position of the root (see arrow 3 of **Fig. 7.8a** for the position at which the root could be placed). A second way to place a root is through midpoint rooting. Here, the longest branch is determined and presumed to be the most reasonable site for a root.

In Web Document 7.5 we include human cytoglobin as an outgroup for 11 closely related myoglobin DNA sequences.

## Enumerating Trees and Selecting Search Strategies

The number of possible trees to describe the relationships of a dozen protein sequences is staggeringly large. It is important to know the number of possible trees for any tree you are making. There is only one “true” tree representing the evolutionary path by which molecular sequences (or even species) evolved. The number of potential trees is useful in deciding which tree-making algorithms to apply.

The number of possible rooted and unrooted trees is described in Box 7.3. For two OTUs, there is only one tree possible. For three taxa, it is possible to construct either one

Some phylogeny projects involve the generation of trees for thousands of taxa. See the Deep Green plant project at <http://ucjeps.berkeley.edu/bryolab/GPphylo/> (WebLink 7.9). The Ribosomal Database at <http://rdp.cme.msu.edu/> (WebLink 7.10) includes an analysis of over 2.7 million ribosomal RNA sequences (see Chapter 10). For typical analyses, you may analyze several dozen taxa. If you want to make a phylogenetic tree with the globins that are currently in Pfam (version 27.0), you could use the 73 proteins available in the seed alignment or all 6000 proteins available in the full alignment.

### BOX 7.3 NUMBER OF ROOTED AND UNROOTED TREES

The number of bifurcating unrooted trees ( $N_U$ ) for  $n$  OTUs ( $n \geq 3$ ) is given by Cavalli-Sforza and Edwards (1967):

$$N_U = \frac{(2n-5)!}{2^{n-3} (n-3)!} \quad (7.9)$$

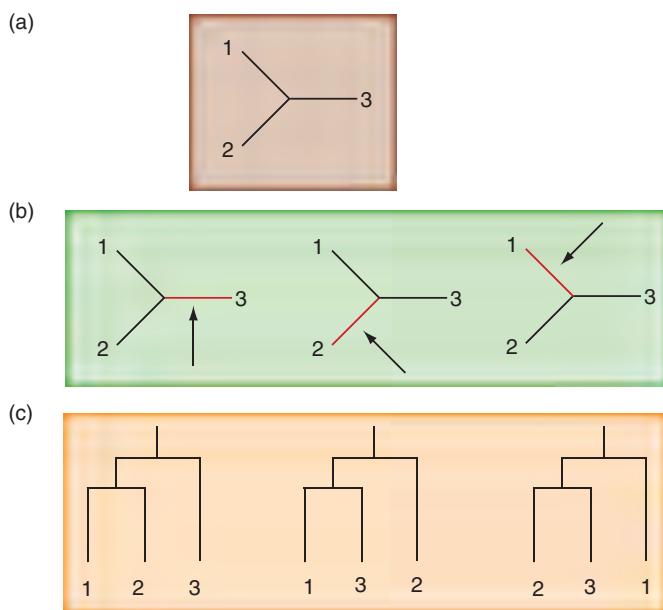
The number of bifurcating rooted trees ( $N_R$ ) for  $n$  OTUs ( $n \geq 2$ ) is:

$$N_R = \frac{(2n-3)!}{2^{n-2} (n-2)!} \quad (7.10)$$

For example, for four OTUs,  $N_R$  equals  $(8-3)!/(2^2)(2)! = 5!/8 = 15$ . The number of possible rooted and unrooted trees (up to 50 OTUs) is as follows. The values were calculated using MatLab software (MathWorks).

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,489,707	2,027,025
15	213,458,046,676,875	$8 \times 10^{12}$
20	$8 \times 10^{21}$	$2 \times 10^{20}$
50	$2.8 \times 10^{76}$	$3 \times 10^{74}$

To give a sense of the immense number of possible trees corresponding to just a few dozen taxa, there are on the order of  $10^{79}$  protons in the universe.



**FIGURE 7.10** For three operational taxonomic units (such as three aligned protein sequences 1–3), there is (a) one possible unrooted tree. (b) Any of these edges may be used to select a root (see arrows), from which (c) three corresponding rooted trees are possible.

Source: MEGA version 5.2; Tamura et al. (2013). Use of software courtesy of S. Kumar.

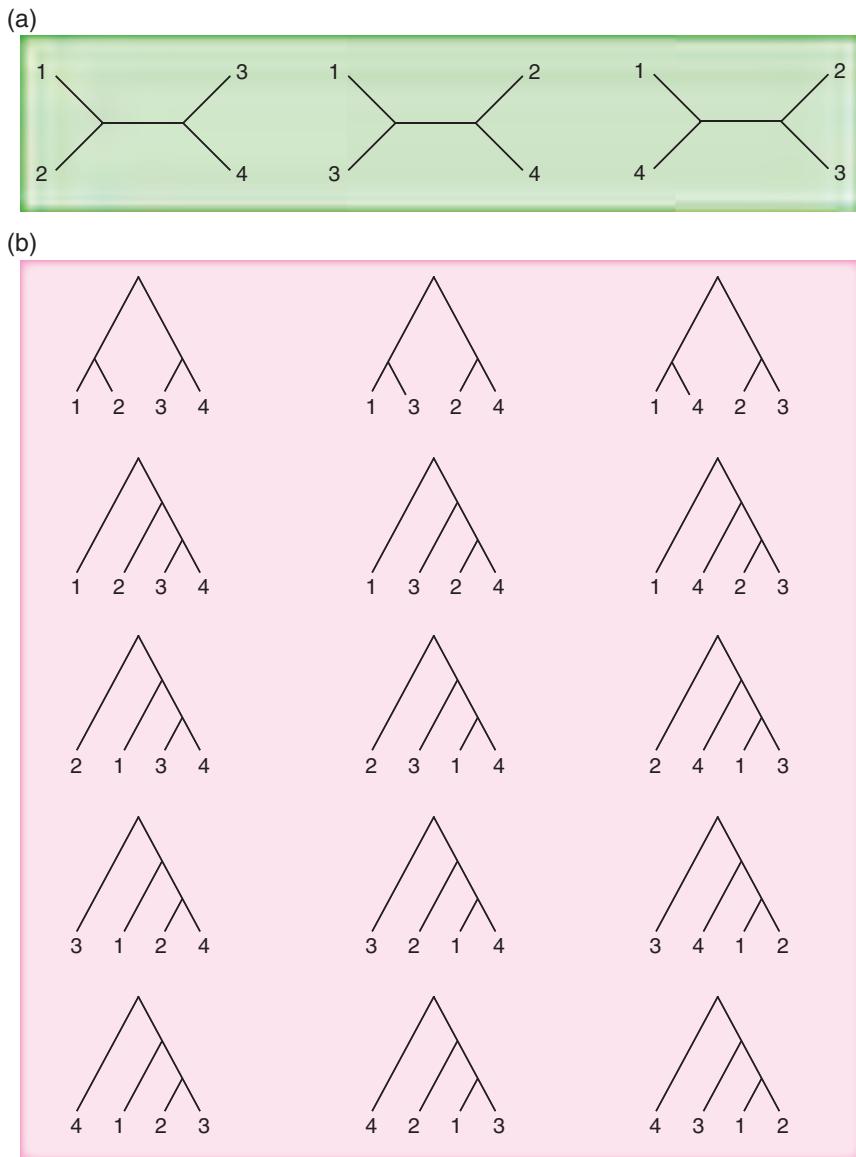
unrooted tree or three different rooted trees (Fig. 7.10). For four taxa, the number of possible trees rises to three unrooted trees or 15 rooted trees (Fig. 7.11).

An exhaustive search examines all possible trees and selects the one with the most optimal features such as the shortest overall sum of the branch lengths. An important practical limit is reached at around 12 sequences, for which there are over  $6.5 \times 10^8$  possible unrooted trees and  $1.3 \times 10^{10}$  rooted trees. For about 12 taxa (or fewer) it is possible for a standard desktop computer to perform exhaustive searches for which all possible trees are evaluated.

The branch-and-bound method provides an exact algorithm for identifying the optimal tree (or trees) without performing an exhaustive search (Penny *et al.*, 1982; reviewed in Felsenstein, 2004). In one variant of this approach three taxa are used to make a tree; only one unrooted tree is possible. A fourth taxon is added, creating three possible trees. Upon addition of a fifth taxon there are three times five (i.e., 15) possible trees. By considering the tree in each group having the shortest branch lengths, it is possible to efficiently identify candidates for the optimal tree(s). This allows a strategy of not performing exhaustive searches for trees (or subtrees) having a worse score than the potential optimal tree. The name of this method refers to a boundary that is reached once the search process has identified a subtree with a suboptimal score.

For more than a dozen sequences it is generally necessary to use a heuristic algorithm to identify an optimal tree (or trees). A heuristic algorithm explores a subset of all possible trees, discarding vast numbers of trees that have a topology that is unlikely to be useful. In this way it is possible to create phylogenetic trees having hundreds or even thousands of protein (or DNA) sequences. As an example of how a heuristic algorithm works, consider a dataset in which the algorithm seeks a tree with the shortest total branch lengths (i.e., the most parsimonious tree). This search occurs without evaluating all possible trees, but instead by performing a series of rearrangements of the topology. Once a tree with a particular score is obtained, the algorithm can establish that score

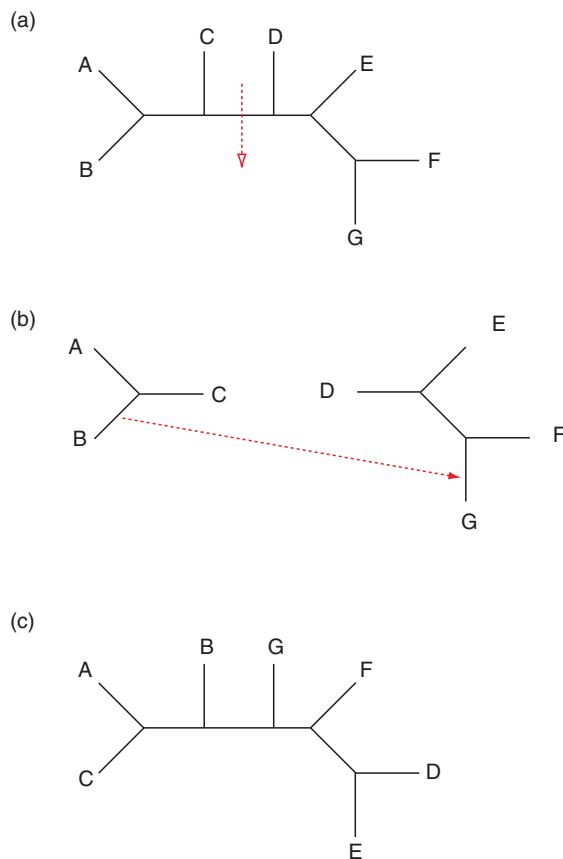
By analogy to branch-and-bound approach, the Needleman–Wunsch method identifies the optimal subpaths in a pairwise alignment without exhaustively evaluating all possible subpaths (Chapter 3).



**FIGURE 7.11** For four operational taxonomic units (such as four aligned protein sequences 1–4), there are (a) 3 possible unrooted trees and (b) 15 possible rooted trees. Only one of these is a true tree in which the topology accurately describes the evolutionary process by which these sequences evolved.

as an upper limit and discard all trees for which rearrangements are unlikely to yield a shorter tree.

A variety of heuristic approaches are available. Stepwise addition involves the addition of taxa (as described for branch-and-bound) with subsequent branch swapping on the shortest tree(s). The choice of which three taxa are joined initially may be determined arbitrarily (e.g., by the order in which the sequence are input), randomly, or based on which three taxa are most closely related. Another heuristic algorithm is branch swapping. In the “tree bisection and reconnection” version, a tree is bisected along a branch, generating two subtrees. These are reconnected by systematically joining all possible pairs of branches with one branch originating from each subtree (Fig. 7.12). Heuristic algorithms have an inherent tradeoff between search time and confidence in the search result. One can assume that they provide an approximation of the “best” tree.



**FIGURE 7.12** Branch swapping using the tree bisection reconnection (TBR) approach. (a) After a tree is made it is bisected along a branch to form two subtrees. (b, c) These are reconnected by joining one branch from each subtree. All possible bisections are evaluated, as well as all possible reconnection patterns. The goal is to identify the most optimal tree(s).

Source: Redrawn from PAUP User's Guide. Courtesy of D. Swofford.

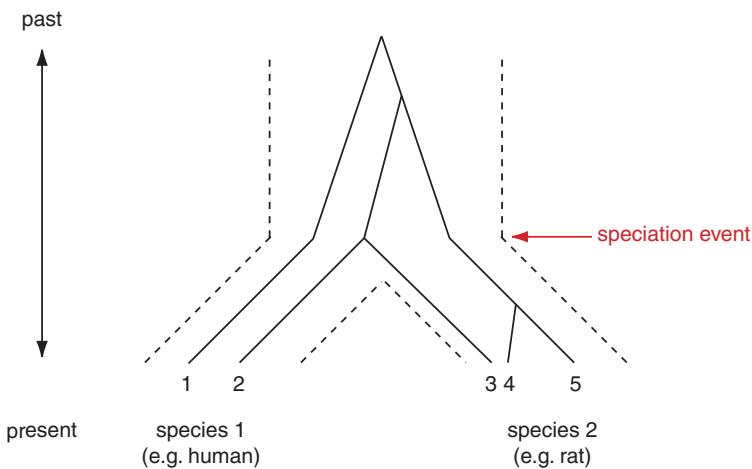
## TYPE OF TREES

### Species Trees versus Gene/Protein Trees

Species evolve and genes (and proteins) evolve. The analysis of molecular evolution can be complicated by the time that two species diverged. Speciation, the process by which two new species are created from a single ancestral species, occurs when the species become reproductively isolated (Fig. 7.13). In a species tree, an internal node represents a speciation event. For example, for a species tree containing human and mouse taxa connected by a node, that node corresponds to the last common ancestor of humans and mice, a creature that lived some 90 MYA. In a gene tree (or protein tree), an internal node represents the divergence of an ancestral gene into two new genes (or proteins) with distinct sequences. Phylogeny software such as MEGA can reconstruct ancestral DNA or protein sequences that are present at an inferred node. An example is shown for a group of globin sequences (Fig. 7.14). For a tree containing rat and mouse myoglobin sequences, the node connecting those two taxa represents the sequence of an ancestral rodent that existed at the time of the rat–mouse speciation (~25 MYA). In almost all cases this ancestral sequence is not known but is inferred. Reconstructions of ancestral states are subject to a variety of artifacts, especially when rates of evolution are rapid in some branches of the tree (Cunningham *et al.*, 1998).

For estimates of species' divergence times see <http://www.timetree.org> (WebLink 7.11).

This interpretation of a phylogenetic tree should be in terms of historical events (Baum *et al.*, 2005). Consider the tree of globins shown in Figure 7.8a. Is a globin from



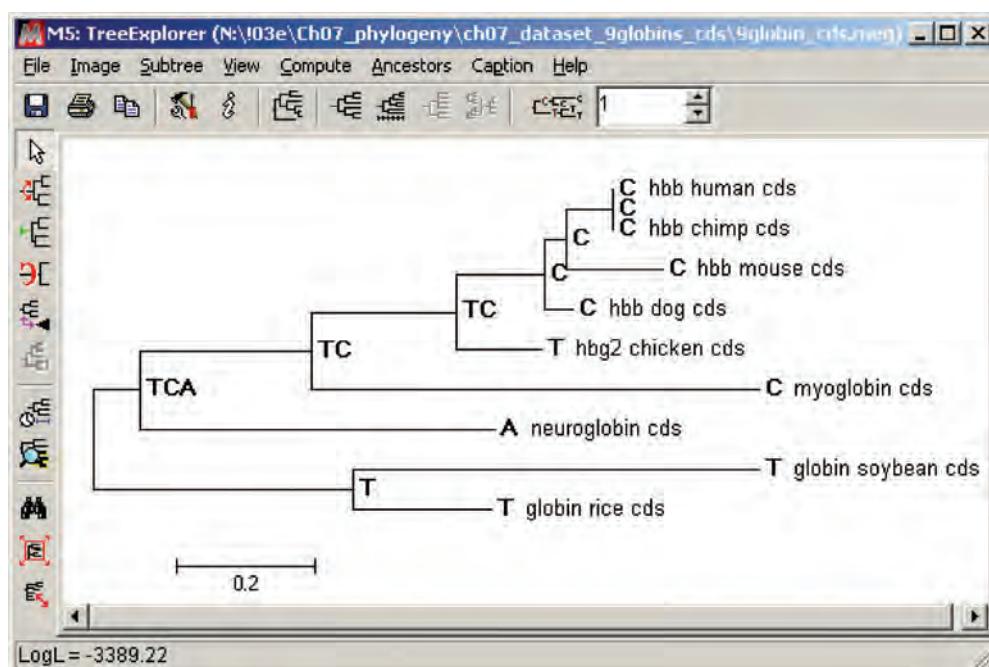
**FIGURE 7.13** A species tree and a protein (or gene) tree can have a complex relationship. A speciation event, such as the divergence of the lineage that generated modern humans and rodents, can be dated to a specific time (e.g., 90 MYA). When speciation occurs, the species become reproductively isolated from one another. This event is represented by dotted lines (see horizontal arrow). Phylogenetic analysis of a specific group of homologous proteins is complicated by the fact that a gene duplication could have preceded or followed the speciation event. In essentially all phylogenetic analyses, the extant proteins (OTUs) are sequences from organisms that are alive today. It is necessary to reconstruct the history of the protein family as well as the history of each species. In the above example, there are two human paralogs and three rat paralogs. Proteins 1 and 5 diverged at a time that greatly predates the divergence of the two species. Proteins 2 and 3 diverged at a time that matches the date of species divergence. Proteins 4 and 5 diverged recently, after the time of species divergence. It is possible to reconstruct both species trees and protein (or gene) trees. Adapted from Graur and Li (2000), based upon Nei (1987). Reproduced with permission from Sinauer Associates and Columbia University Press.

chicken more closely related to mouse beta globin than to human beta globin? No, it is not: mouse and human globin are members of a clade that share a common ancestor (see internal node 1), and that ancestor is the descendant of the last common ancestor of mammalian and chicken globin. Interpreting trees in phylogeny contrasts with the analysis of trees in other areas of biology such as microarray data analysis (Chapter 11). There the nodes connecting samples or genes do not have a historical meaning.

In a genetically polymorphic population, gene duplication events may occur before or after speciation. A protein (or gene) tree differs from a species tree in two ways (Graur and Li, 2000): (1) the divergence of two genes from two species may have predated the speciation event, which may cause overestimation of branch lengths in a phylogenetic analysis; or (2) the topology of the gene tree may differ from that of the species tree. In particular, it may be difficult to reconstruct a species tree from a gene tree. A molecular clock may be applied to a gene tree in order to date the time of gene divergence, but it cannot be assumed that this is also the time that speciation occurred.

Reconstructing a phylogenetic tree based upon a single protein (or gene) can therefore give complicated results. For this reason, many researchers construct trees from a variety of distinct protein (or gene) families in order to assess the relationships of different species. Another strategy that has been adopted is to generate concatenated protein (or DNA) sequences. For example, Baldauf *et al.* (2000) used four concatenated protein sequences to create a comprehensive phylogenetic tree of eukaryotes (Fig. 19.1). Such a strategy produces a tree that is weighted by the average protein length; the choice of which sequences are included will impact the outcome.

In looking at phylogenetic trees, it is important to be aware of the type of data that were used to generate the tree. It is also important to inspect the scale bar (if present)



**FIGURE 7.14** Reconstruction of ancestral sequences using MEGA. A maximum likelihood tree was constructed with nine globin DNA sequences using the Tamura–Nei model with uniform rates among sites. The resulting tree was saved (in the Newick format) and used as input to the Ancestors tab (see top of Fig. 7.9a) and the tool “Infer Ancestral Sequences (ML).” The tree shows inferred ancestral states at a single position (from most likely to least likely). The full dataset can be exported as a spreadsheet.

Source: MEGA version 5.2; Tamura et al. (2013). Use of software courtesy of S. Kumar.

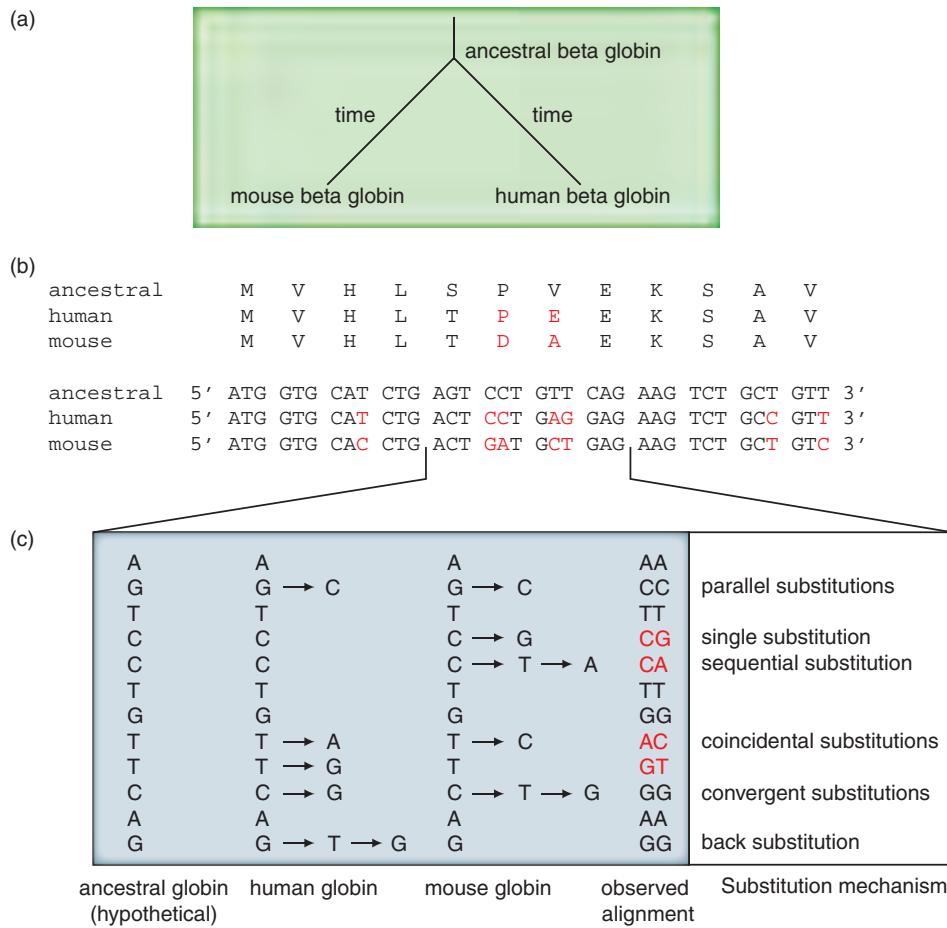
which describes whether the units are number of substitutions per site, number of substitutions per branch, elapsed time, or some other measure.

### DNA, RNA, or Protein-Based Trees

When you generate a phylogenetic tree using molecular sequence data, you can use DNA, RNA, or protein sequences. In one common scenario, you may want to evaluate the relationship of a group of molecules such as globins. The choice of whether to study protein or DNA depends in part on the question you are asking. In some cases, protein studies are preferable; you may prefer to study a multiple sequence alignment of proteins, or the lower rate of substitutions in protein relative to DNA may make protein studies more appropriate for comparisons across widely divergent species. In many other cases, studying DNA is more informative than protein. There are several reasons for this.

- DNA allows the study of synonymous and nonsynonymous mutation rates, as discussed above (Fig. 7.7).
- Substitutions in DNA include those that are directly observed in an alignment, such as single-nucleotide substitutions, sequential substitutions, and coincidental substitutions (depicted in Fig. 7.15). By analyzing two sequences with reference to an ancestral sequence (Fig. 7.15a, b), it is possible to infer a great deal of information about mutations that do not appear in a direct comparison of two (or more) sequences. These mutational processes include parallel substitutions, convergent substitutions, and back substitutions (Fig. 7.15c).
- Noncoding regions (such as the 5' and 3' untranslated regions of genes, or introns; see Fig. 7.7) may be analyzed using molecular phylogeny. For some portions of noncoding DNA, there is little evolutionary pressure to conserve the nucleotide sequence, and these

Synapomorphy is defined as a character state that is shared by several taxa. Homoplasy is defined as a character state that arises independently (e.g., through convergent substitutions or back substitutions) but is not derived from a common ancestor (i.e., is not homologous). See Graur and Li (2000).



**FIGURE 7.15** Multiple types of mutations occur in sequences. (a) There is a hypothetical, ancestral globin sequence from which human and murine beta globin diverged in the past at time T when these organisms last shared a common ancestor. We can infer the nucleotide and amino acid sequences of the ancestor. (b) Consider a portion of the coding sequence of human and murine beta globin (the data are from Fig. 7.7). There are two observed mismatches at the amino acid level, and seven observed mismatches at the nucleotide level. Many more than seven mutations may have occurred in this region. Hypothetical ancestral protein and DNA sequences are shown, selected for the purpose of illustration. (c) Comparison of 12 nucleotides of the hypothetical ancestral sequence with the observed human and murine sequences illustrates several mutational mechanisms. Single-nucleotide substitution, sequential substitution, and coincidental substitution could all account for observed mutations (red-colored nucleotides). Parallel, convergent, and back substitutions could all occur without producing an observed mismatch. In this example, four mutations are observed (nucleotides colored red) while 13 mutations actually occurred. (a, c) Data from Graur and Li (2000).

regions may vary greatly. That is, the nucleotide substitution rate equals the neutral mutation rate. In other cases there is tremendous nucleotide conservation, perhaps because of the presence of a regulatory element such as a transcription factor binding motif.

- Pseudogenes have been studied using molecular phylogeny, for example to estimate the neutral rate of evolution. By definition, pseudogenes do not encode functional proteins (see Chapter 8). Similarly, inactive DNA transposons and other repetitive DNA elements have been analyzed as “molecular fossils” to explore speciation events and the evolution of chromosomes.
- The rate of transitions and transversions can be evaluated (Box 7.4). In a comparison of mitochondrial DNA among a group of primate species (human, chimpanzee, and gorilla), 92% of the differences were transitions (Brown *et al.*, 1982). Transitions commonly occur far more frequently than transversions in nuclear DNA as well, and this is reflected in various models of nucleotide substitution (see below).

We describe several ribosomal RNA databases in Chapter 10. These serve as important sources of sequences for phylogenetic analyses.

## BOX 7.4 TRANSITIONS AND TRANSVERSIONS

A transition is a nucleotide substitution between two purines (A to G or G to A) or between two pyrimidines (C to T or T to C). A transversion is the substitution between a purine and a pyrimidine (e.g., A to C, C to A, G to T; there are eight possible transversions). The International Union of Pure and Applied Chemistry (IUPAC; <http://www.iupac.org>) defines many symbols commonly used in science. The abbreviations of the four nucleotides are adenine (A), cytosine (C), guanine (G), and thymine (T). Additional abbreviations are for an unspecified or unknown nucleotide (N), an unspecified purine nucleotide (R), and an unspecified pyrimidine nucleotide (Y).

You can assess the rate of transitions and transversions using the MEGA package. Open a protein-coding DNA alignment file in MEGA. Visit the Sequence Data Editor and, under the Statistics pull-down menu, choose Nucleotide Pair Frequencies (Directional). The output tabulates the number of identical pairs of nucleotides, the transitional and transversional pairs, and their ratio. Alternatively, use the Pattern pull-down menu, and choose Computer Transition/Transversion Bias.

We will show how the entire genome of a fungus duplicated in Chapter 18. The evidence for this consisted of BLASTP searches of all *Saccharomyces cerevisiae* proteins against each other, resulting in the detection of conserved blocks of sequence from various chromosomes (see Fig. 18.10). Here, BLASTN searches would not have been sensitive enough to reveal the homology between different chromosomes.

While the analysis of DNA can offer many advantages, it is sometimes preferable to study proteins for phylogenetic analysis. The evolutionary distance between two organisms may be so great that any DNA sequences are saturated. That is, at many sites all the possible nucleotide changes may occur (even multiple times), so that phylogenetic signal is lost. Proteins have 20 states (amino acids) instead of only four states for DNA, so there is a stronger phylogenetic signal. We saw that BLASTP searches of human globins against plants were more sensitive than BLASTN searches (Chapter 4). For closely related sequences, such as mouse versus rat beta globin, DNA-based phylogeny can be more appropriate than protein studies because of the advantages of DNA discussed above.

Whether nucleotides or amino acids are selected for phylogenetic analysis, the effects of character changes can be defined. An unordered character is a nucleotide or amino acid that changes to another character in one step. An ordered character is one that must pass through one or more intermediate states before it changes to a different character. Partially ordered characters have a variable or indeterminate number of states between the starting value and the ending value. Nucleotides are unordered characters: any one nucleotide can change to any other in one step (Fig. 7.16a). Amino acids are partially ordered. If you inspect the genetic code, you will see that some amino acids can change to a different amino acid in a single step of one nucleotide substitution, while other amino acid changes require two or even three nucleotide mutations (Fig. 7.16b).

## FIVE STAGES OF PHYLOGENETIC ANALYSIS

Molecular phylogenetic analyses can be divided into five stages: (1) selection of sequences for analysis; (2) multiple sequence alignment of homologous protein or nucleic acid sequences; (3) specification of a statistical model of nucleotide or amino acid evolution; (4) tree building; and (5) tree evaluation. These stages are discussed in the following sections.

### Stage 1: Sequence Acquisition

We have discussed some issues regarding the choice of DNA, RNA, or protein sequences for molecular phylogeny. You can acquire the sequences from many sources, including the following.

- HomoloGene at NCBI includes thousands of eukaryotic protein families. HomoloGene entries can be viewed as sequences in the FASTA format (or as a multiple sequence alignment).

(a)

	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

(b)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C	0	2	3	1	1	2	2	3	2	3	2	2	2	3	1	1	2	2	1	1
D	0	1	2	1	1	2	2	2	2	3	1	2	2	2	2	2	2	1	3	1
E	0	3	1	2	2	1	2	2	2	2	2	2	1	2	2	2	2	1	2	2
F	0	2	2	1	3	1	2	2	2	2	3	2	1	2	1	2	1	2	1	1
G	0	2	2	2	2	2	2	2	2	2	2	2	1	1	2	1	1	1	2	2
H	0	2	2	1	3	1	1	1	1	1	1	1	1	2	2	2	2	3	1	1
I	0	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	3	2	2
K	0	2	1	1	2	1	1	2	1	1	2	1	1	2	1	2	2	2	2	2
L	0	1	2	1	1	1	1	1	1	2	1	1	1	2	1	1	2	1	1	2
M							0	2	2	2	1	2	1	1	1	2	3			
N							0	2	2	2	1	1	1	2	3	1				
P								0	1	1	1	1	1	2	2	2	2			
Q								0	1	2	2	2	2	2	2	2				
R									0	1	1	2	1	2	2	2				
S									0	1	2	1	1	1	2	2				
T										0	2	2	2	2	2					
V											0	2	2	2	2					
W												0	2	2						
Y													0							

**FIGURE 7.16** Step matrices for (a) nucleotides or (b) amino acids describe the number of steps required to change from one character to another. For the amino acids, between one and three nucleotide mutations are required to change any one residue to another. Adapted from Graur and Li (2000). Used with permission.

- Results from the BLAST family of proteins can be selected, viewed in NCBI Protein or NCBI Nucleotide, and formatted in the FASTA format. Sequences may be obtained from the European Bioinformatics Institute or Ensembl.
- Sequences from a large variety of databases can be output in the FASTA format (or as multiple sequence alignments). For RNA, these databases include Rfam and the Ribosomal Database (Chapter 10). For proteins, these databases include Pfam and InterPro (Chapter 6). For viruses, examples include reference databases for human immunodeficiency virus and hepatitis C virus.

You can access the HIV Sequence Database at <http://www.hiv.lanl.gov/> (WebLink 7.12), and the HCV database at <http://hcv.lanl.gov/> (WebLink 7.13).

## Stage 2: Multiple Sequence Alignment

Multiple sequence alignment (Chapter 6) is a critical step of phylogenetic analysis. In many cases, the alignment of nucleotide or amino acid residues in a column implies that they share a common ancestor. If you misalign a group of sequences, you will still be able to produce a tree. However, it is not likely that the tree will be biologically meaningful. If you create a multiple alignment of sequences and include a nonhomologous sequence, it may still be incorporated into the phylogenetic tree.

In preparing a multiple sequence alignment for phylogenetic analysis, there are several important considerations in creating and editing the alignment. Let us introduce these ideas by referring to a specific example of 13 globins. We presented a phylogenetic tree

Web Document 7.10 (at <http://www.bioinfbook.org/chapter7>) includes 13 quasi-randomly selected protein sequences. If you import these into MEGA you can align them using ClustalW and generate a tree. Can you distinguish that tree from one generated using a group of homologous proteins?

of these proteins in **Figure 7.1**. The multiple sequence alignment from which this tree was generated is shown in **Figure 7.17**. There are several notable features:

1. Carefully inspect the alignment to be sure that all sequences are homologous. It is sometimes possible to identify a sequence that is so distantly related that it is not homologous. You can further test this possibility by performing pairwise alignments (is the expect value significant?), BLAST searches, or checking whether the proteins are members of a Pfam family. If a sequence is not apparently homologous to the others, it should be removed from the multiple sequence alignment.
2. Some multiple sequence alignment programs may treat distantly related sequences by aligning them outside the block of other sequences. If necessary, lower the gap creation and/or gap extension penalties to accommodate the distantly related homolog(s) into the multiple sequence alignment. As discussed in Chapter 6, include methods that incorporate structural information into the alignment of proteins when possible. In some cases, a group of proteins share a domain (defined in Chapter 12) but are unrelated outside the domain; you can restrict your analyses to just the region of the homologous domain using software such as MEGA. These programs allow you to select any specific residues for inclusion or exclusion in the phylogenetic analysis.
3. The complete sequence is not known for many genes. Whenever possible, the multiple sequence alignment data used for phylogenetic analyses should be restricted to portions of the proteins (or nucleic acids) that are available for all the taxa being studied.
4. There are both terminal and internal gaps in this alignment (**Fig. 7.17**, arrowheads). A gap could represent an insertion in some of the sequences or a deletion in the others. Most phylogeny algorithms are not equipped to evaluate insertions or deletions (also called indels). Many experts recommend that any column of a multiple sequence alignment that includes a gap in any position should be deleted, and software programs typically delete columns with incomplete data as a default option.
5. In this example, note that the sequences include three myoglobins, three alpha globins, three beta globins, and four other globins. Intuitively, we expect these globin sequences to be distinguished in a phylogenetic tree, and this is the case (**Figs. 7.1** and **7.2**). Indeed, we can see such differences by inspecting the multiple sequence alignment. There are positions in which the amino acid in a particular position differs between the myoglobins, alpha globins, and beta globins (**Fig. 7.17**, columns with open circles and red lettering). Other positions are highly conserved among all these proteins (columns indicated with diamonds), as expected for a family of proteins having closely related structures. The phylogenetic tree (**Fig. 7.1**) visualizes these various relationships. Any time you inspect a multiple sequence alignment and a tree, you are looking at related information from different perspectives.

A variety of tree-building programs accept a multiple sequence alignment as input. ReadSeq is a convenient program that translates multiple sequence alignments into formats compatible with most commonly used phylogeny packages.

First released in 1993, ReadSeq was written by Don Gilbert. Many ReadSeq servers are available online, such as ones at EBI (<http://www.ebi.ac.uk/Tools/sfc/readseq/>, WebLink 7.14) and the NIH (<http://www-bimas.cit.nih.gov/molbio/readseq/>, WebLink 7.15). It can be downloaded from SourceForge (<http://sourceforge.net/projects/readseq/>, WebLink 7.16).

### Stage 3: Models of DNA and Amino Acid Substitution

Phylogenetic analyses rely on models of DNA or amino acid substitution. These models may be implicit or explicit. For distance-based methods, statistical models are employed to estimate the number of DNA or amino acid changes that occurred in a series of pairwise comparisons of sequences. For maximum likelihood and Bayesian approaches, statistical models are applied to individual characters (residues) in order to assess the most likely topology as well as other features such as substitution rates along individual

myoglobin\_kanga - - - - - MGLSDGEWQLVLNWIWGKVETDEGGHKGKDVIRLFKGHPETLEKFDKF  
 myoglobin\_harbo - - - - - MGLSEGEWQLVLNVGKVEADLAGHGDVIRLFKGHPETLEKFDKF  
 myoglobin\_gray\_ - - - - - MGLSDGEWHLVLNVWGKVETDLAGHGQEVLIRLFKSHPETLEKFDKF  
 alpha\_globin\_ho - - - - - MV-LSAADKTNVKAAWSVKVGGHAGEYGAEALERMFLGFPITTKTYFPHF  
 alpha\_globin\_ka - - - - - V-LSAADKGHVKAIGWKVGGHAGEYGAEGLERTFHSFPPTTKTYFPHF  
 alpha\_globin\_do - - - - - V-LSPADKTNIKSTWDKIGGHAGDGYGEALDRFTQSFPPTTKTYFPHF  
 beta\_globin\_dog - - - - - MVHLTAEEKSLVSGLGKVV-NVDEVGGAEALGRLLIVYPWTQRFFDSF  
 beta\_globin\_rab - - - - - MVHLSSFEKSATLWGKV-NVVEVGGEALGRLLVVYPWTQRFFESF  
 beta\_globin\_kan - - - - - VHLTAEEKNAITSLWGKV-AIEQTGGEALGRLLIVYPWTQRFFDHF  
 globin\_riverlam - PIVDS---GSPAVLSAAEKTKIRSAWAPVSYNYESGVDILVKFFTSTPAQAEFFPKF  
 globin\_sealampr - MPIVDT---GSVAPLSAAEKTKIRSAWAPVYSTYESGVDILVKFFTSTPAQAEFFPKF  
 globin\_soybean - - - - - VAFTEKQDALVSSSFEAFKANI PQYSVVFYTSILEKAPAAKDLFSL  
 globin\_insect MKFLILALCFAAASALSADQISTVQASFDFDKVKGD---PVGILYAVFKADPSIMAKTQF  
  
 myoglobin\_kanga KHLKSEDEMKAEDLKKHGITVLTALGNILKKKGHHAEELKPLAQ---HATKHKIPVQF  
 myoglobin\_harbo KHLKTEAEMKAEDLKKHGNTVLTALGGILKKKGHHDAELKPLAQ---HATKHKIPYK  
 myoglobin\_gray\_ KHLKSEDDMRRSEDLRKHGNTVLTALGGILKKKGHHAEELKPLAQ---HATKHKIPYK  
 alpha\_globin\_ho - DLSHGS-----QVKAHGKKVGDALTAVGHLDLPGALSNSLSD---HAHKLRVDPVN  
 alpha\_globin\_ka - DLSHGS-----QIQAHQGKKIADALGQAVEHIDDLPGTLSKLSLD---HAHKLRVDPVN  
 alpha\_globin\_do - DLSPGS-----QVKAHGKKVADALTAVAHLDLPGALSALSDL---HAYKLRVDPVN  
 beta\_globin\_dog GDLSTPDAMVSNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSEL---HCDKLHVDPEN  
 beta\_globin\_rab GDLSSANAVMNNPKVKAHGKKVLAASFSEGLSHLDNLKGTFAKLSEL---HCDKLHVDPEN  
 beta\_globin\_kan GDLSNAKAVMANPKVLAHGAKVLVAFGDAIKNLDNLKGTFAKLSEL---HCDKLHVDPEN  
 globin\_riverlam KGMTSADELKKSADVVRWAERIINAVNDAVASMDTEKMSMK---DLGSKHAKSFQVDPQY  
 globin\_sealampr KGLTTADQLIKKSADVVRWAERIINAVNDAVASMDTEKMSMLRDLGSKHAKSFQVDPQY  
 globin\_soybean ANPTDG---VNPKLTGHAEKLFALVRDSAGQL-KASGTVVADAALGSVHAQKAVTNPEF  
 globin\_insect AG-KDLESIKGTAPFEIHNARIVGFFSKIIGELPNIEADVNFTVAS---HKPRGVTHDQ  
  
 myoglobin\_kanga LEFISDAIIQVIQSKHAGNFGADAQAAAMKKALELFRHDMAAKYKEFGFQG  
 myoglobin\_harbo LEFISEAIIHVLHRSRHPAEGFADAQGAMNKALELFRKDIATKYKELGFHG  
 myoglobin\_gray\_ LEFISEAIIHVLHSKHPAEGFADAQAAAMKKALELFRNDIAAKYKELGFHG  
 alpha\_globin\_ho FKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLLSVSTVLTSKYR----  
 alpha\_globin\_ka FKLLSHCLLVFAAHLGDAFTPEVHASLDKFLLAVSTVLTSKYR----  
 alpha\_globin\_do FKLLSHCLLVTLACHHPTTEFTPAVHASLDKFLLFAAVSTVLTSKYR----  
 beta\_globin\_dog FKLLGNVLVCVLAAHFGKEFTPQVQAAQKVVAGVANALAHKYH----  
 beta\_globin\_rab FRLGNVLVVLSHFGKEFTPQVQAAQKVVAGVANALAHKYH----  
 beta\_globin\_kan FKLLGNIIIVICLAEHFGKEFTIDTQVAWQKLVAGVANALAHKYH----  
 globin\_riverlam FKVL-AVIADTVAAAG-----DAGFEKLSMCIIILMLRSAY----  
 globin\_sealampr FKVLAAVIADTVAAAG-----DAGFEKLSMSMCILLRSAY----  
 globin\_soybean -VVKEALLKTIKAAGDKWSDELSRAWEVAYDELAAAIKAK----  
 globin\_insect ---LNNFRAGFVSYMKAHHTDFAGAEAAWGTALDTFFGMIFSKM---

**FIGURE 7.17** We introduce tree-making approaches with a multiple sequence alignment of 13 globin proteins, made using MAFFT (FFT-NS-1 v5.861). The sequences correspond to those in **Figure 7.1**. There are three myoglobins (red kangaroo *Macropus rufus*, P02194; harbor porpoise *Phocoena phocoena*, P68278; gray seal *Halichoerus grypus*, P68081); three alpha globins (horse *Equis caballus*, P01958; eastern gray kangaroo *Macropus giganteus*, P01975; dog *Canis lupus familiaris*, P60529); three beta globins (dog *Canis lupus familiaris*, XP\_537902; rabbit *Oryctolagus cuniculus*, NP\_001075729; eastern gray kangaroo *Macropus giganteus*, P02106); two fish globins (European river lamprey *Lampetra fluviatilis*, 690951A; sea lamprey *Petromyzon marinus*, P02208); an insect globin (midge larva *Chironomus thummi thummi*, P02229); and a plant leghemoglobin (soybean *Glycine max* 711674A). Gaps in the alignment (solid arrowheads) are not easily interpretable by phylogenetic algorithms and could represent either insertions or deletions. Four positions are 100% conserved (open diamonds). Amino acids in many other positions distinguish the groups of myoglobins, alpha globins, beta globins, and other globins (examples are shown in columns with open circles; in some cases the groups are perfectly distinguishable in an aligned column). A phylogenetic tree provides a visualization of these relationships (**Fig. 3.2** and this chapter).

Source: MAAFT. Software used with permission from K. Katoh.

branches. For maximum parsimony, the criterion for finding the best tree is based on the shortest branch lengths and, while individual characters are also evaluated, most of these statistical models are not applicable.

The simplest approach to defining the relatedness of a group of nucleotide (or amino acid) sequences is to align pairs of sequences and count the number of differences. The degree of divergence is sometimes called the Hamming distance. For an alignment of length  $N$  with  $n$  sites at which there are differences, the degree of divergence  $d$  is defined as:

$$d = \frac{n}{N} \times 100. \quad (7.11)$$

Earlier in this chapter we discussed an example of this type of calculation by Zuckerkandl and Pauling (1962) who counted the number of amino acid differences between human beta globin and delta, gamma, and alpha globin. The Hamming distance is simple to calculate, but it ignores a large amount of information about the evolutionary relationships among the sequences. The main reason is that character *differences* are not the same as *distances*: the differences between two sequences are easy to measure, but the genetic distance involves many mutations that cannot be observed directly. As shown in **Figure 7.15**, there are many kinds of mutations that occur but are not detected in an estimate of divergence based on counting differences. We also discussed a correction implemented by Dickerson (1971) that was proposed by Margoliash and Smith (1965) and by Zuckerkandl and Pauling (1965); see Equations (7.1) and (7.2). In MEGA software this is referred to as the Poisson correction (see Nei and Kumar, 2000, p. 20). The Poisson correction for distance  $d$  assumes equal substitution rates across sites and equal amino acid frequencies. It uses the following formula to correct for multiple substitutions at a single site:

$$d = -\ln(1 - p) \quad (7.12)$$

where  $d$  is the distance and  $p$  is the proportion of residues that differ. We make the following assumptions (Uzzell and Corbin, 1971). First, the probability of observing a change is small but nearly identical across the genome. This probability is proportional to the length of the time interval  $\lambda\Delta t$  for some constant  $\lambda$ . The probability of observing no changes is therefore  $1 - \lambda\Delta t$ . Second, we assume the number of nucleotide or amino acid changes is constant over the time interval  $t$ . When a mutation does occur, this does not alter the probability of another mutation occurring at this same position. Third, we assume that changes occur independently. Equation (7.12) is derived from the Poisson distribution which describes the random occurrence of events when that probability of occurrence is small. The Poisson distribution is used to model a variety of phenomena, such as the decay of radioactivity over time. It is given by the formula:

$$P(X) = \frac{e^{-\mu} \mu^X}{X!} \quad (7.13)$$

where  $P(X)$  is the probability of  $X$  occurrences per unit of time,  $\mu$  represents the population mean number of changes over time, and  $e$  is  $\sim 2.718$  (Zar, 1999).

Let us consider a practical example of how different substitution models affect the distances that are measured in a set of 13 globin proteins. We enter the proteins into MEGA and select the Distances menu (**Fig. 7.9a**) to compute pairwise distances between the 13 proteins. We can view the number of amino acid differences per sequence (**Fig. 7.18a**), highlighting several pairwise comparisons that are relatively closely or distantly related. Next we estimate the differences based on the Hamming distance (Equation (7.11); called the *p*-distance in MEGA; **Fig. 7.18b**). When we next use the Poisson correction, the distance values are comparable (relative to the Hamming distance) for closely related sequences such as globins from two lampreys (**Fig. 7.18c**, dashed red boxes). However, the estimated evolutionary divergence for distantly related sequences is dramatically different using

(a) Number of differences

	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbharbor porpoise P68278 <i>Phocoena pho...</i>	19											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	16	12										
4. alphahorse P01958 <i>Equus caballus</i>	84	84	84									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	85	87	84	24								
6. alphadog P60529 <i>Canis lupus familiaris</i>	88	88	86	22	27							
7. betadog XP 537902 <i>Canis lupus familia...</i>	80	79	78	66	69	67						
8. betarabbit NP 001075729 <i>Oryctolagus c...</i>	80	81	78	64	67	65	16					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	83	82	80	68	69	66	25	28				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	88	92	88	77	77	76	83	83	81			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	89	91	89	76	77	76	83	85	81	85	83	8
12. globinsoybean 711674A <i>Glycine max</i> (so...	98	97	97	93	93	93	87	90	90	90	93	94
13. globininsect P02229 <i>Chironomus thummi...</i>	87	88	86	92	93	97	92	90	94	88	89	91

(b) p-distance

	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbharbor porpoise P68278 <i>Phocoena pho...</i>	0.17											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	0.14	0.11										
4. alphahorse P01958 <i>Equus caballus</i>	0.74	0.74	0.74									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	0.75	0.77	0.74	0.21								
6. alphadog P60529 <i>Canis lupus familiaris</i>	0.78	0.78	0.76	0.19	0.24							
7. betadog XP 537902 <i>Canis lupus familia...</i>	0.71	0.70	0.69	0.58	0.61	0.59						
8. betarabbit NP 001075729 <i>Oryctolagus c...</i>	0.71	0.72	0.69	0.57	0.59	0.58	0.14					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	0.73	0.73	0.71	0.60	0.61	0.58	0.22	0.25				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	0.78	0.81	0.78	0.68	0.68	0.67	0.73	0.73	0.72			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	0.79	0.81	0.79	0.67	0.68	0.67	0.73	0.75	0.72	0.07		
12. globinsoybean 711674A <i>Glycine max</i> (so...	0.87	0.86	0.86	0.82	0.82	0.82	0.77	0.80	0.80	0.82	0.82	0.83
13. globininsect P02229 <i>Chironomus thummi...</i>	0.77	0.78	0.76	0.81	0.82	0.86	0.81	0.80	0.83	0.78	0.79	0.81

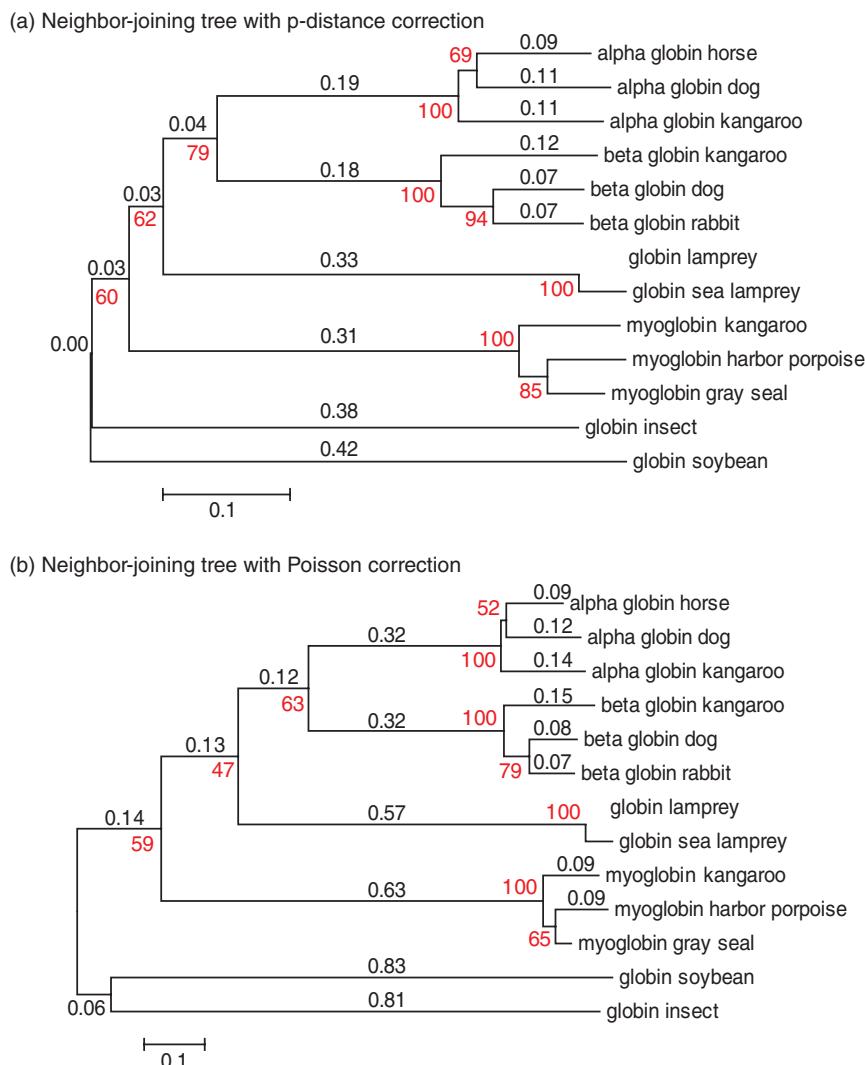
(c) Poisson correction

	1	2	3	4	5	6	7	8	9	10	11	12
1. mbkangaroo P02194 <i>Macropus rufus</i> (red...)												
2. mbharbor porpoise P68278 <i>Phocoena pho...</i>	0.18											
3. mbgray seal P68081 <i>Halichoerus grypus</i>	0.15	0.11										
4. alphahorse P01958 <i>Equus caballus</i>	1.36	1.36	1.36									
5. alphakangaroo P01975 <i>Macropus gigante...</i>	1.40	1.47	1.36	0.24								
6. alphadog P60529 <i>Canis lupus familiaris</i>	1.51	1.51	1.43	0.22	0.27							
7. betadog XP 537902 <i>Canis lupus familia...</i>	1.23	1.20	1.17	0.88	0.94	0.90						
8. betarabbit NP 001075729 <i>Oryctolagus c...</i>	1.23	1.26	1.17	0.84	0.90	0.86	0.15					
9. betakangaroo P02106 <i>Macropus giganteu...</i>	1.33	1.29	1.23	0.92	0.94	0.88	0.25	0.28				
10. globinlamprey 690951A <i>Lampetra fluvia...</i>	1.51	1.68	1.51	1.14	1.14	1.12	1.33	1.33	1.26			
11. globinsealamprey P02208 <i>Petromyzon ma...</i>	1.55	1.64	1.55	1.12	1.14	1.12	1.33	1.40	1.26	0.07		
12. globinsoybean 711674A <i>Glycine max</i> (so...	2.02	1.95	1.95	1.73	1.73	1.73	1.47	1.59	1.59	1.73	1.78	
13. globininsect P02229 <i>Chironomus thummi...</i>	1.47	1.51	1.43	1.68	1.73	1.95	1.68	1.59	1.78	1.51	1.55	1.64

**FIGURE 7.18** Estimating the evolutionary divergence between sequences. The MEGA software package includes a menu for choosing models of nucleotide or amino acid substitution. Similar options are available in other software packages such as PHYLIP. (a) The number of amino acid differences per sequence is displayed below the diagonal, based on pairwise analyses of 13 globins (see Fig. 7.17 legend for their accession numbers). Two closely related globins (with few differences) are highlighted in dashed green boxes, while two divergent globins (with many differences) are highlighted in solid red boxes. (Standard error estimates can be displayed above the diagonal.) (b) Evolutionary divergence was estimated using the p-distance option to calculate the number of amino acid differences per site. Note that each cell (below the diagonal) represents the number of observed differences divided by the total number of positions in the dataset (113 in this case, with all columns containing gaps eliminated from the final data matrix). For example, the value of 0.87 for a comparison of taxa 1 (myoglobin from kangaroo) and 12 (soybean globin), shown in a red box, is obtained by dividing 98 by 113. (c) Evolutionary divergence was estimated using the Poisson correction. Note that this introduces a substantial increase in the estimated distance for the more divergent sequences. Such larger estimates are likely to be more realistic than simple Hamming distances, and will lead to the creation of trees with different branch lengths and topologies.

Source: MEGA version 5.2; Tamura et al. (2013). Courtesy of S. Kumar.

the Poisson correction (Fig. 7.18c, solid red boxes). The consequence of choosing among these models is that entirely different phylogenetic trees may be constructed. We can use this dataset of globin proteins to construct a neighbor-joining tree (defined in “Making Trees by Distance-Based Methods: Neighbor-Joining” below) using either the *p*-distance (Fig. 7.19a) or the Poisson correction (Fig. 7.19b). Note that the topologies of the two trees are the



**FIGURE 7.19** The effect of differing models of amino acid substitution on phylogenetic trees. Phylogenetic trees of 13 globin proteins were made using the neighbor-joining method which uses the distance information that is presented in Figure 7.18. The trees were made using (a) the *p*-distance or (b) the Poisson correction. Branch lengths are in the units of evolutionary distances used to infer each tree. The sum of the branch lengths was 2.81 in (a) and 4.93 in (b). Trees were created using MEGA software. Bootstrapping was performed using 500 bootstrap replicates to identify the percent of instances (indicated in red) in which bootstrap trees support each clade in the inferred tree. For example, in panel (b) in 100% of the bootstrap trials, horse, dog, and kangaroo alpha globin were supported as being in a clade. However, the clade containing horse and dog alpha globin proteins was supported in only 52% of the bootstrap replicates. This means that in 48% of the bootstrap trees kangaroo alpha globin joined that group of proteins, and we can infer that there is not strong support for a distinct, closely related horse/dog group that shared an ancestor with the kangaroo protein. In general the bootstrap can provide a measure of how well supported an inferred tree topology is upon repeated samplings of the dataset.

Source: MEGA version 5.2; Tamura et al. (2013). Courtesy of S. Kumar.

same in this example, but the branch lengths differ. For the optimal tree using the  $p$ -distance correction the sum of the branch lengths is 2.81, while for the tree made with the Poisson correction the sum of the branch lengths is 4.93. Such differences can have large effects on the interpretation of a phylogenetic tree.

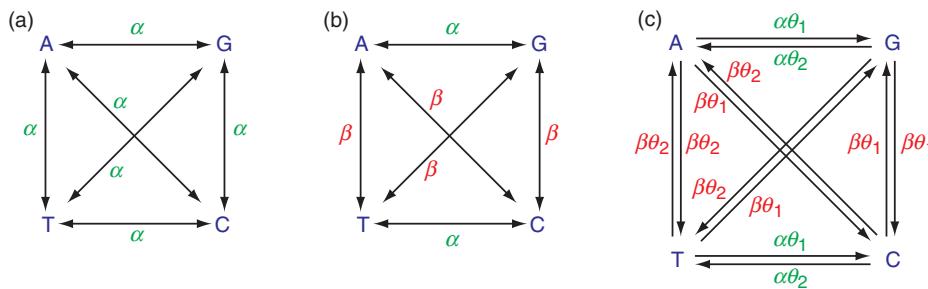
In order to model substitutions that occur in DNA sequences, Jukes and Cantor (1969, p. 100) proposed a fundamentally useful corrective formula:

$$D = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} p \right). \quad (7.14)$$

As an example of how to use Equation (7.14), consider an alignment where 3 nucleotides out of 60 aligned residues differ. The normalized Hamming distance is  $3/60 = 0.05$ . The Jukes–Cantor correction  $d = -(3/4) \ln[1 - (4 \times 0.05/3)] = 0.052$ . In this case, applying the correction causes only a small effect. When 30/60 nucleotides differ, the Jukes–Cantor correction is  $-(3/4) \ln [1 - (4 \times 0.5/3)] = 0.82$ , a far more substantial adjustment.

The Jukes–Cantor one-parameter model describes the probability that each nucleotide will mutate to another (Fig. 7.20a). It makes the simplifying assumption that each residue is equally likely to change to any of the other three residues and that the four bases are present in equal frequencies. This model therefore assumes that the rate of transitions equals the rate of transversions. The corrections are minimal for very closely related sequences, but can be substantial for more distantly related sequences. Beyond about 70% differences, the corrected distances are difficult to estimate. This approaches the percent differences found in randomly aligned sequences.

Dozens of models have been developed that are more sophisticated than Jukes–Cantor. Usually, the transition rate is greater than the transversion rate; for eukaryotic nuclear DNA it is typically two-fold higher. The Kimura (1980) two-parameter model adjusts the transition and transversion ratios by giving more weight to transversions to account for their likelihood of causing nonsynonymous changes in protein-coding regions (Fig. 7.20b). In any region of DNA (including noncoding sequence), the transition/transversion ratio corrects for the biophysical threshold for creating a purine–purine or pyrimidine–pyrimidine pair in the double helix. For example, Tamura (1992) extended the two-parameter model to adjust for the guanine and cytosine (GC) content of the DNA sequences (Fig. 7.20c). We see in Part III of this book that the GC content varies greatly among different organisms and different chromosomal regions within an organism’s genome.



**FIGURE 7.20** Models of nucleotide substitution. (a) The Jukes–Cantor model of evolution corrects for superimposed changes in an alignment. The model assumes that each nucleotide residue is equally likely to change to any of the other three residues and that the four bases are present in equal proportions. The rate of transitions  $\alpha$  equals the rate of transversions  $\beta$ . (b) In the Kimura two-parameter model,  $\alpha \neq \beta$ . Typically, transversions are given more weight. (c) Tamura’s model, which accounts for variations in GC content. This is an example of a more complex model of nucleotide substitution. Note that there are distinct parameters for nucleotide substitutions, and that many of these parameters are directional (e.g., the rate of changing from nucleotides T to C differs from the rate for C to T).

Changes in nucleotide substitution at a given position of an alignment represent one kind of DNA variation, and we have discussed several ways to correct for changes that occur. Substitution rates are often variable across the length of a group of sequences. This represents a second distinct category of DNA variation, and we can also model these changes. Some sites (columns of aligned residues) are relatively invariant, while others undergo substitutions readily.

- Because of the degeneracy of the genetic code, the third position of a codon almost always has a higher substitution rate than the first and second codon position.
- Some regions of a protein have conserved domains. We saw an example of this with the insulin orthologs in **Figure 7.3**. Viruses, immunoglobulin genes, and mitochondrial genomes often display hypervariable regions of mutation.
- Noncoding RNAs (Chapter 10) often have functional constraints such as stem and loop structures that include highly conserved positions with low substitution rates.

A gamma ( $\Gamma$ ) model accounts for unequal substitution rates across variable sites (Box 7.5). The gamma family of distributions can be plotted with the substitution rate ( $x$  axis) versus the frequency (y axis, **Fig. 7.21**). This shape of the distribution varies as determined by the gamma shape parameter  $\alpha$ . Zhang and Gu (1998) measured  $\alpha$  for protein sequences from 51 vertebrate nuclear genes and 13 mammalian mitochondrial genes. They reported a range of values from 0.17 to 3.45 (median value 0.71) for the 51 nuclear genes. There was a negative correlation between the extent of among-site rate variation and the mean substitution rate. Genes with a high level of rate variation among sites (large  $\alpha$ ) have a low mean substitution rate and are therefore slowly evolving. Rapidly evolving proteins have a low level of rate variation among sites.

When we create a phylogenetic tree using 13 globin protein sequences using MEGA or other software, we can specify that there is a uniform rate of variation among sites (thus not invoking the gamma distribution), or we can set the shape parameter of  $\alpha$  to any positive value. For a group of globin proteins, there are dramatic differences in the branch lengths and the topologies of trees created using the same neighbor-joining method and the Poisson correction with varying gamma distributions and shape parameters  $\alpha = 0.25$ ,  $\alpha = 1$ , or  $\alpha = 5$  (**Fig. 7.22a–c**).

It is now routine for users to evaluate dozens or even  $>100$  different models of nucleotide or amino acid substitutions, and to apply criteria to select the best one for

## BOX 7.5 THE GAMMA DISTRIBUTION

In mathematics, the gamma distribution  $\Gamma$  is commonly used to model continuous variables that have skewed distributions. The gamma distribution has been used to model the among-site rate variation of proteins. Given a substitution rate  $r$  at a site, the  $\Gamma$  distribution has the following probability density function (Zhang and Gu, 1998):

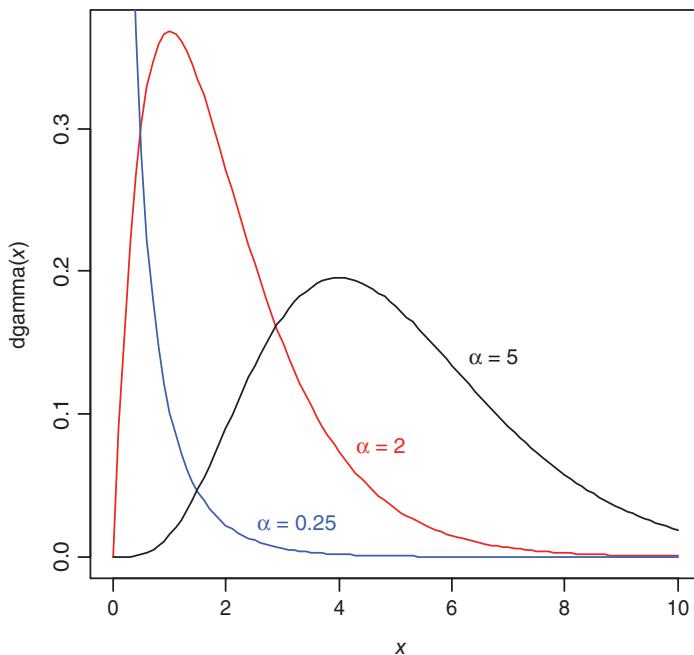
$$g(r) = \frac{(\alpha / \mu)^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-(\alpha/\mu)r}. \quad (7.15)$$

The two parameters in this equation are the mean rate  $\mu = E(r)$  and the shape parameter  $\alpha$ . Here  $E(r)$  is the mean substitution rate (or the expectation of  $r$ ). Small values of  $\alpha$  correspond to a high degree of rate variation among sites. In a study by Zhang and Gu (1998), genes with a high value of  $\alpha$  included the C-kit proto-oncogene ( $\alpha = 3.45$ ) and  $\alpha$ -globin ( $\alpha = 1.93$ ), while genes with a low  $\alpha$  value included histone H2A.X ( $\alpha = 0.19$ ) and  $\beta$ 2 thyroid hormone receptor ( $\alpha = 0.21$ ).

In the R programming language, you can invoke the gamma distribution from the `stats` package with the commands

```
> x=seq(0,10,length=101)
> plot(x,dgamma(x,shape=2)
> lines(x,dgamma(x,shape=0.25))
```

You can also display the gamma distribution using Microsoft Excel with the function gammadist.



**FIGURE 7.21** The gamma distribution describes the substitution rate ( $x$  axis; from low to high) with a frequency distribution (y axis) that is dependent on shape parameter  $\alpha$ . For small values of  $\alpha$  (e.g.,  $\alpha = 0.25$ ), most of the nucleotides undergo substitutions at slow rates and most of the observed variation is attributed to relatively few nucleotide sites that evolve rapidly. For large values of  $\alpha$  (e.g.,  $\alpha = 5$ ) few nucleotide sites undergo very fast or very slow evolution, and there is minimal among-site rate variation. For intermediate values of  $\alpha$  (e.g.,  $\alpha = 2$ ) some nucleotides evolve with high substitution rates. This figure was generated in the R programming language using the `dgamma` function of the stats package.

Source: R Foundation, from <http://www.r-project.org>.

a particular analysis. For example, the ModelTest program implements a log likelihood ratio test to compare models (Posada and Crandall, 1998; Posada, 2006). The log likelihood ratio test is a statistical test of the goodness-of-fit between two models. ModelTest systematically tests up to 56 models of variation. The likelihood scores of a null model ( $L_0$ ) and an alternative model ( $L_1$ ) are calculated for comparisons of a relatively simple model and a relatively complex model. A likelihood ratio test statistic is obtained:

$$\delta = -2 \log \Lambda \quad (7.16)$$

where

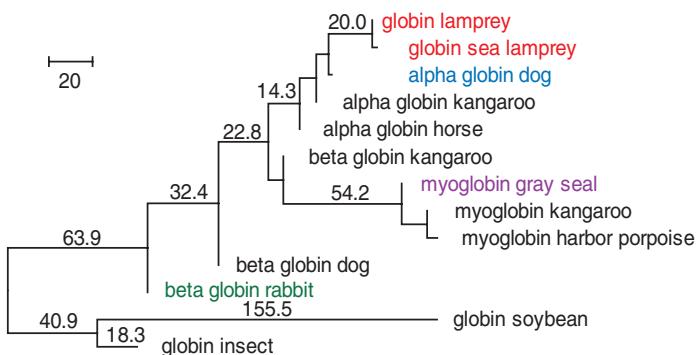
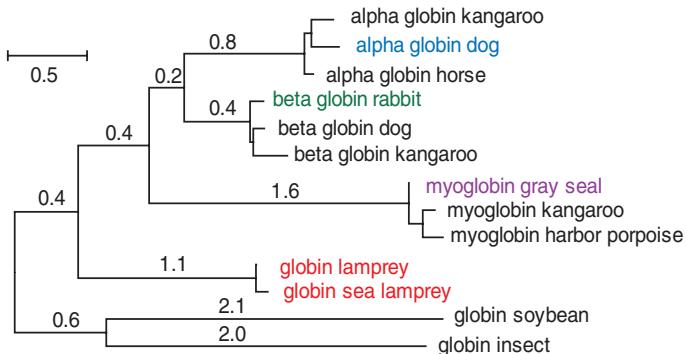
$$\Lambda = \frac{\max[L_0(\text{Null Model} | \text{Data})]}{\max[L_1(\text{Alternative Model} | \text{Data})]} \quad (7.17)$$

This test statistic follows a  $\chi^2$  distribution and, given the number of degrees of freedom (equal to the number of additional parameters in the more complex model), a probability value is obtained. As an alternative to log likelihood ratio tests, ModelTest also uses the Akaike information criterion (AIC; Posada and Buckley, 2004). This measures the best-fitting model as that having the smallest AIC value:

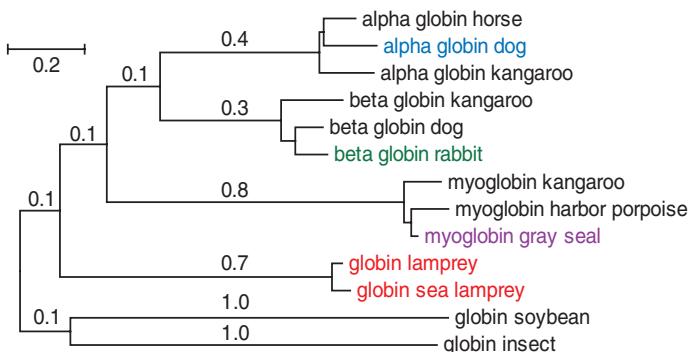
$$\text{AIC} = -2 \ln L + 2N \quad (7.18)$$

where  $L$  is the maximum likelihood for a model using  $N$  independently adjusted parameters for that model. In this way good maximum likelihood scores are rewarded, while using too many parameters is penalized.

$\Lambda$  is the Greek letter corresponding to  $L$ .

(a) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter  $\alpha=0.25$ (b) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter  $\alpha=1$ 

ModelTest and ProtTest as well as jModelTest2 (Darriba *et al.*, 2012) were developed by David Posada and colleagues, and are available from <http://darwin.uvigo.es/our-software/> (WebLink 7.17). The site includes servers for online analyses. An example of an output file from ModelTest, showing the results of analyzing 56 substitution models from 11 myoglobin coding sequences, is shown in Web Document 7.11 at <http://www.bioinfbook.org/chapter7>. ProtTest output for 13 globins is shown as Web Document 7.12. The Hepatitis C Virus (HCV) sequence database at the Los Alamos National Laboratories (<http://hcv.lanl.gov/>, WebLink 7.18) offers Findmodel, a web-based implementation of ModelTest that accepts DNA sequences as input. It displays over two dozen models at <http://hcv.lanl.gov/content/sequence/findmodel/findmodel.html> (WebLink 7.19).

(c) Neighbor-joining tree with Poisson correction and gamma distribution shape parameter  $\alpha=5$ 

**FIGURE 7.22** Effect of changing the  $\alpha$  parameter of the  $\Gamma$  distribution on phylogenetic trees. A dataset consisting of 13 globin proteins (see Fig. 7.1) was aligned and trees were generated in MEGA software using the neighbor-joining technique, the Poisson correction, and  $\alpha$  parameters of (a) 0.25, (b) 1, or (c) 5. Note the dramatic effects on the estimated branch lengths. Also note that the topologies differ within the alpha globin, beta globin, and myoglobin clades. The scale bars are in units of number of substitutions.

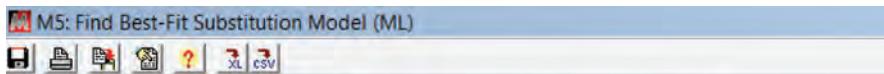
Source: MEGA version 5.2; Tamura *et al.* (2013). Use of software courtesy of S. Kumar.

For protein sequences, we can use the ProtTest software to evaluate an alignment (made by MAFFT) of 13 globin protein sequences. ProtTest evaluates a dozen different amino acid substitution matrices (Abascal *et al.*, 2005; Darriba *et al.*, 2011). For 13 globins it selects the LG+G+F (Fig. 7.23a). This refers to the LG matrix of Le and Gascuel (2008) (+G assesses the probability of the rate of change, and +F uses the

(a) ProtTest lowest- (best-)scoring of 112 models for amino acid substitution (13 globins)

Model	deltaAIC*	AIC	AICw	-lnL
<hr/>				
LG+G+F	0.00	5883.41	0.52	-2898.71
LG+I+G+F	0.37	5883.78	0.43	-2897.89
LG+I+F	5.04	5888.45	0.04	-2901.23
LG+F	10.15	5893.56	0.00	-2904.78
RtREV+I+G+F	23.23	5906.65	0.00	-2909.32
RtREV+G+F	23.90	5907.31	0.00	-2910.65
Dayhoff+G+F	26.95	5910.37	0.00	-2912.18
RtREV+I+F	26.99	5910.40	0.00	-2912.20
DCMut+G+F	27.28	5910.69	0.00	-2912.34
Dayhoff+I+G+F	28.08	5911.49	0.00	-2911.75

(b) MEGA models for amino acid substitution (13 globins)



**Table. Maximum Likelihood fits of 48 different amino acid substitution models**

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)	f(A)	f(R)	f(N)
WAG+G	24	4964.195	4836.725	-2393.968	n/a	5.07	0.087	0.044	0.039
WAG+I	24	4965.561	4838.092	-2394.652	0.03	n/a	0.087	0.044	0.039
WAG	23	4967.943	4845.755	-2399.515	n/a	n/a	0.087	0.044	0.039
WAG+G+I	25	4968.408	4835.661	-2392.403	0.02	7.77	0.087	0.044	0.039
Dayhoff+G+I	25	4970.283	4837.535	-2393.340	0.02	6.81	0.087	0.041	0.040
Dayhoff+G	24	4990.568	4863.098	-2407.155	n/a	4.99	0.087	0.041	0.040
JTT+G	24	5003.961	4876.492	-2413.852	n/a	5.25	0.077	0.051	0.043
JTT+I	24	5004.353	4876.884	-2414.048	0.03	n/a	0.077	0.051	0.043
JTT+G+I	25	5005.191	4872.444	-2410.795	0.03	6.48	0.077	0.051	0.043
Dayhoff+I	24	5013.028	4885.559	-2418.385	0.03	n/a	0.087	0.041	0.040

**FIGURE 7.23** Evaluation of evolutionary models. (a) ProtTest software evaluates dozens of amino acid substitution matrices and many models of evolution to select a model that best fits a given sequence alignment. (Here, 13 globin proteins aligned by MAFFT are evaluated.) (b) MEGA also evaluates multiple models of evolution. For both software packages the Akaike information criterion (AIC) is used to identify the optimal model(s).

Source: (a) ProtTest software. (b) MEGA version 5.2; Tamura et al. (2013). Use of software courtesy of S. Kumar.

amino acid frequencies observed in the dataset). A comparable analysis can be performed in MEGA, which ranks the model of amino acid substitution with the lowest AIC value as best (Fig. 7.23b).

## Stage 4: Tree-Building Methods

There are many ways to build a phylogenetic tree, and these have been reviewed in both books (Durbin *et al.*, 1998; Nei and Kumar, 2000; Felsenstein, 2004; Yang, 2006; Baxevanis and Ouellette, 2009; Lemey *et al.*, 2009; Hall, 2011) and articles (Felsenstein, 1988, 1996; Nei, 1996; Thornton and DeSalle, 2000; Bos and Posada, 2005; Whelan, 2008; Yang and Rannala, 2012).

We consider four principal methods of making trees: distance-based, maximum parsimony, maximum likelihood, and Bayesian inference. Distance-based methods begin by analyzing pairwise alignments of the sequences and using those distances to infer the relatedness between all the taxa. Maximum parsimony is a character-based method in which columns of residues are analyzed in a multiple sequence alignment to identify the tree with the shortest overall branch lengths that can account for the observed character differences. Maximum likelihood and Bayesian inference are model-based statistical approaches in which the best tree that can account for the observed data is inferred.

Molecular phylogeny captures and visualizes the sequence variation that occurs in homologous DNA, RNA, or protein molecules. The most popular software tools for phylogeny include the following. All are extremely versatile and offer a broad range of approaches to making trees.

Phylib is available from <http://evolution.genetics.washington.edu/phylib/general.html> (WebLink 7.20). MEGA can be downloaded from <http://www.megasoftware.net/> (WebLink 7.21). The TREEPUZZLE site is <http://www.tree-puzzle.de/> (WebLink 7.22). MrBayes is available from <http://mrbayes.sourceforge.net/index.php> (WebLink 7.23). Joseph Felsenstein offers a web page with about 200 phylogeny software links at <http://evolution.genetics.washington.edu/phylib/software.html> (WebLink 7.24).

- PAUP (Phylogenetic Analysis Using Parsimony) was developed by David Swofford *et al.* (1996).
- MEGA (Molecular Genetic Evolutionary Analysis) was written by Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei (Tamura *et al.*, 2013). Many of its concepts are explained in an excellent textbook by Nei and Kumar (2000), *Molecular Evolution and Phylogenetics*.
- PHYLIP (the PHYLogeny Inference Package), developed by Joseph Felsenstein, is one of the most widely used phylogeny programs. Felsenstein has written an outstanding book, *Inferring Phylogenies* (2004).
- TREE-PUZZLE was developed by Korbinian Strimmer, Arndt von Haeseler, and Heiko Schmidt. It implements a maximum likelihood method, which is a model-based approach to phylogeny.
- MrBayes was developed by John Huelsenbeck and Fredrik Ronquist. It implements Bayesian estimation of phylogeny, another model-based approach. MrBayes estimates a quantity called the posterior probability distribution, which is the probability of a tree conditioned on the observed data.

### *Distance-Based*

Distance-based methods begin the construction of a tree by calculating the pairwise distances between molecular sequences (Felsenstein, 1984; Desper and Gascuel, 2006). A matrix of pairwise scores for all the aligned proteins (or nucleic acid sequences) is used to generate a tree. The goal is to find a tree in which the branch lengths correspond as closely as possible to the observed distances. The main distance-based methods include the unweighted-pair group method with arithmetic mean (UPGMA) and neighbor-joining (NJ). Distance-based methods of phylogeny are computationally fast, so they are particularly useful for analyses of a larger number of sequences (e.g., >50 or even hundreds or thousands).

These methods use some distance metric, such as the number of amino acid changes between the sequences, or a distance score (see Box 6.3). A distance metric is distinguished by three properties: (1) the distance from a point to itself must be zero, that is,  $D(x, x) = 0$ ; (2) the distance from point  $x$  to  $y$  must equal the distance from  $y$  to  $x$ , that is,  $D(x, y) = D(y, x)$ ; and (3) the triangle inequality must apply in that  $D(x, y) \leq D(x, z) + D(z, y)$ . While similarities are also useful, distances (which differ from differences when they obey the above properties) offer appealing properties for describing the relationships between objects (Sneath and Sokal, 1973).

The observed distances between any two sequences  $i, j$  can be denoted  $d_{ij}$ . The sum of the branch lengths of the tree from taxa  $i$  and  $j$  can be denoted  $d'_{ij}$ . These two distance measures are ideally the same, but phenomena such as the occurrence of multiple

substitutions at a single position typically cause  $d_{ij}$  and  $d'_{ij}$  to differ. The goodness-of-fit of the distances based on the observed data and the branch lengths can be estimated as follows (see Felsenstein, 1984):

$$\sum_i \sum_j w_{ij} (d_{ij} - d'_{ij})^2 \quad (7.19)$$

The goal is to minimize this value; it is zero when the branch lengths of a tree match the distance matrix exactly. For Cavalli-Sforza and Edwards (1967)  $w_{ij} = 1$  while for Fitch and Margoliash (1967)  $w_{ij} = 1/d_{ij}^2$ .

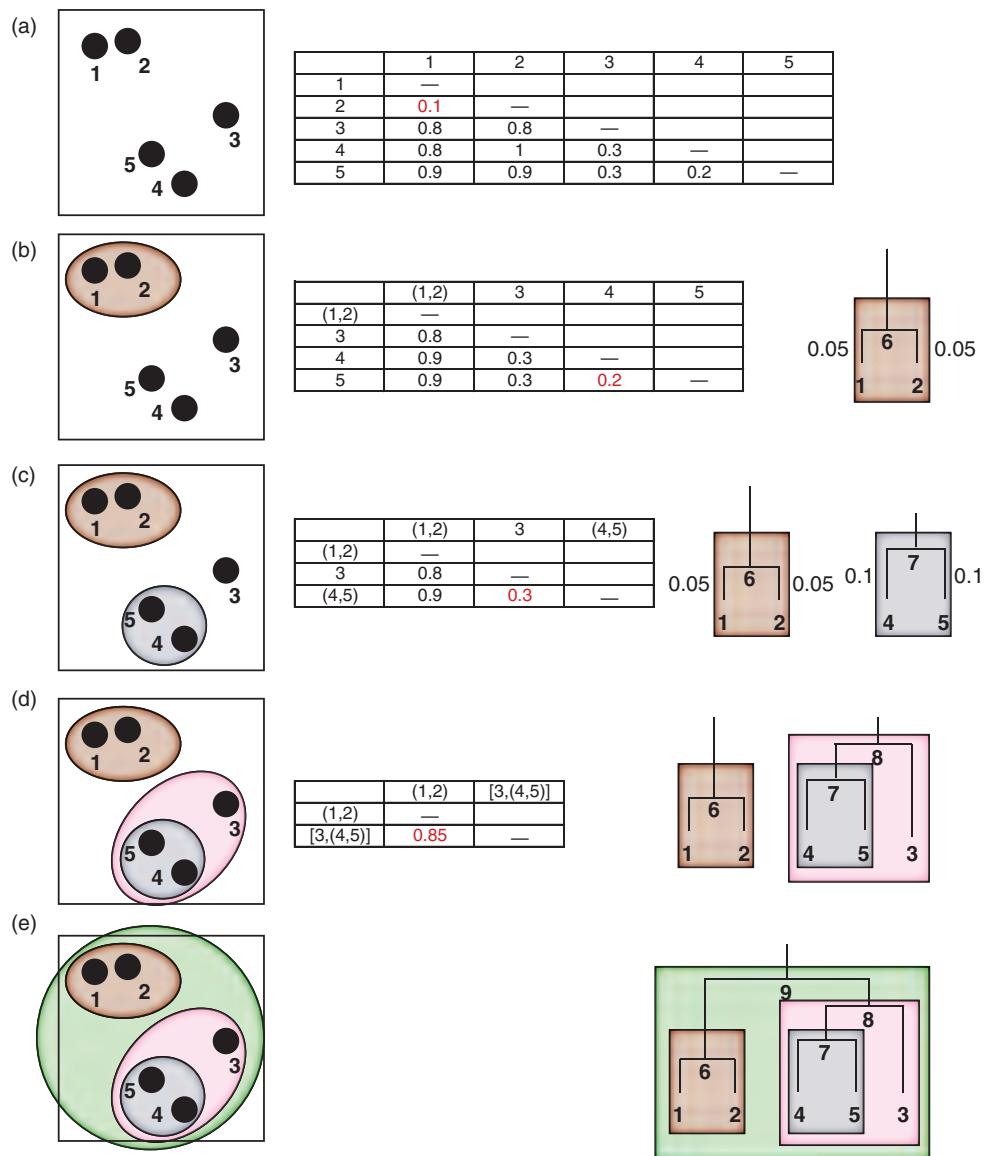
We can inspect the multiple sequence alignment in **Figure 7.17** as well as the tree in **Figure 7.1** to think about the essence of distance-based molecular phylogeny. In this approach, the percent amino acid similarity between each pair of proteins in the multiple sequence alignment can be calculated. Some pairs, such as dog and rabbit beta globins, are very closely related and will be placed close together in the tree. Others, such as insect globin and soybean globin, are more distant than the other sequences and will be placed farther away on the tree. In a sense, we can look at the sequences in **Figure 7.17** horizontally, calculating distance measurements between the entire sequences. This approach discards a large amount of information about the characters (i.e., the aligned columns of residues), instead summarizing information about the overall relatedness of sequences. In contrast, character information is evaluated in maximum parsimony, maximum likelihood, and Bayesian approaches. All strategies for inferring phylogenies must make some simplifying assumptions, but nonetheless the simpler approaches of distance-based methods very often produce phylogenetic trees that closely resemble those derived by character-based methods.

**The UPGMA Distance-Based Method** We introduce UPGMA here because the tree-building process is relatively intuitive and UPGMA trees are broadly used in the field of bioinformatics. However, the algorithm most phylogeny experts employ to build distance-based trees is neighbor-joining (described in the following section). We can make an UPGMA tree in MEGA using the phylogeny menu (**Fig. 7.9a**). UPGMA clusters sequences based on a distance matrix. As the clusters grow, a tree is assembled. A tree of 9 globins using UPGMA is shown in **Figure 7.8c**. As we would expect, the alpha globins, beta globins, lamprey globins, and myoglobin are clustered in distinct clades. The two most closely related protein (lamprey globins) are clustered most closely together.

We described the use of a distance matrix to create a guide tree in Chapter 6.

The UPGMA algorithm was introduced by Sokal and Michener (1958) and works as follows. Consider five sequences whose distances can be represented as points in a plane (**Fig. 7.24a**). We also represent them in a distance matrix. Some protein sequences, such as 1 and 2, are closely similar while others (such as 1 and 3) are far less related. UPGMA clusters the sequences as follows (adapted from Sneath and Sokal, 1973, p. 230):

1. We begin with a distance matrix. We identify the least dissimilar groups (i.e., the two OTUs  $i$  and  $j$  that are most closely related). All OTUs are given equal weights. If there are several equidistant minimal pairs, one is picked randomly. In **Figure 7.24a** we see that OTUs 1 and 2 have the smallest distance.
2. Combine  $i$  and  $j$  to form a new group  $ij$ . In our example, groups 1 and 2 have the smallest distance (0.1) and are combined to form cluster (1, 2; see **Fig. 7.24b**). This results in the formation of a new, clustered distance matrix having one fewer row and column than the initial matrix. Dissimilarities that are not involved in the formation of the new cluster remain unchanged; for example, in the distance matrix of **Figure 7.23b**, taxa 3 and 4 still maintain a distance of 0.3. The values for the clustered taxa (1, 2) reflect the average of OTUs 1 and 2 to each of the other OTUs. The distance of OTU 1 to OTU 4 was initially 0.8, 1.0 for OTU 2 to OTU 4, and 0.9 for OTU (1, 2) to OTU 4.



**FIGURE 7.24** Explanation of the UPGMA method. This is a simple, fast algorithm for making trees. It is based on clustering sequences. (a) Each sequence is assigned to its own cluster. A distance matrix, based on some metric, quantitates the distance between each object. The circles in the figure represent these sequences. (b) The taxa with the closest distance (sequences 1 and 2) are identified and connected. This allows us to name an internal node (right, node 6 in (b)). The distance matrix is reconstructed counting taxa 1 and 2 as a group. We can also identify the next closest sequences (4 and 5; distance is in red). (c) These next closest sequences (4 and 5) are combined into a cluster, and the matrix is again redrawn. In the tree (right side) taxa 4 and 5 are now connected by a new node, 7. We can further identify the next smallest distance (value 0.3, red font) corresponding to the union of taxon 3 to cluster (4, 5). (d) The newly formed group (cluster 4, 5 joined with sequence 3) is represented on the emerging tree with new node 8. Finally, (e) all sequences are connected in a rooted tree.

3. Connect  $i$  and  $j$  through a new node on the nascent tree. This node corresponds to group  $ij$ . The branches connecting  $i$  to  $ij$  and  $j$  to  $ij$  each have a length  $D_{ij}/2$ . In our example, OTUs 1 and 2 are connected through node 6, and the distance between OTU1 and node 6 is 0.05 (Fig. 7.24b, right side). We label the internal node 6 because we reserve the numbers 1–5 on the  $x$  axis as the terminal nodes of the tree.

4. Identify the next smallest dissimilarity (between OTUs 4 and 5 in **Fig. 7.24b**), and combine those taxa to generate a second clustered dissimilarity matrix (**Fig. 7.24c**). In this step it is possible that two OTUs will be joined (if they share the least dissimilarity), or a single OTU (denoted *i*) will be joined with a cluster (denoted *jk*), or two clusters will be joined (*ij*, *kl*). The dissimilarity of a single OTU *i* with a cluster *jk* is computed simply by taking the average dissimilarity of *ij* and *ik*. In this process a new distance matrix is formed, and the tree continues to be constructed. In **Figure 7.24c** the smallest distance in the matrix is 0.3, corresponding to the relation of OTU 3 to the combined OTU 4, 5. These are joined in **Figure 7.24d** in the graphic representation, in the distance matrix, and in the tree.

5. Continue until there are only two remaining groups, and join these.

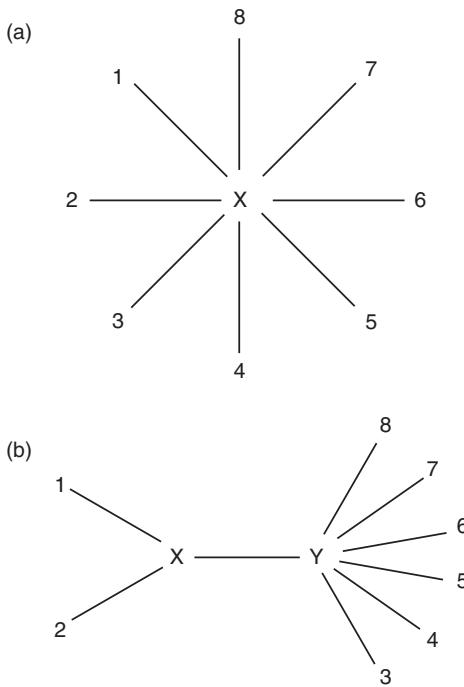
The tree shown in **Figure 7.24** was made by the UPGMA approach using the sequences of 13 globin proteins, for which the Poisson-corrected distances are shown in **Figure 7.18c**. We demonstrate how to perform UPGMA calculations on this dataset in a series of 12 tables available on the supplementary website. Compare **Figures 7.18c** and **7.24** and note that the two closest OTUs of the distance matrix (globin from lamprey and sea lamprey) have the shortest branch lengths on the UPGMA tree. The second closest group (myoglobin from harbor porpoise and gray seal) has the next shortest branch lengths. That group of two OTUs collectively has a short branch length to kangaroo myoglobin. These relationships are visualized in the phylogenetic tree.

A critical assumption of the UPGMA approach is that rate of nucleotide or amino acid substitution is constant for all the branches in the tree, that is, the molecular clock applies to all evolutionary lineages. If this assumption is true, branch lengths can be used to estimate the dates of divergence, and the sequence-based tree mimics a species tree. An UPGMA tree is rooted because of its assumption of a molecular clock. If it is violated and there are unequal substitution rates along different branches of the tree, the method can produce an incorrect tree. Note that other methods (including neighbor-joining) do not automatically produce a root, but a root can be placed by choosing an outgroup or by applying midpoint rooting.

The UPGMA method is a commonly used distance-based method in a variety of applications including microarray data analysis (see Chapter 11). In phylogenetic analyses using molecular sequence data, its simplifying assumptions tend to make it significantly less accurate than other distance-based methods such as neighbor-joining. We used UPGMA to make a rooted tree of nine globin DNA sequences (**Fig. 7.8c**). In contrast to the neighbor-joining tree it placed two plant globins in a position that is biologically implausible.

**Making Trees by Distance-Based Methods: Neighbor-Joining** The neighbor-joining method is used for building trees by distance methods (Saitou and Nei, 1987). It produces both a topology and branch lengths. We begin by defining a neighbor as a pair of OTUs connected through a single interior node *X* in an unrooted, bifurcating tree. In the tree of globins shown in **Figure 7.1**, porpoise and seal myoglobins are neighbors while kangaroo myoglobin is not a neighbor because it is separated from those two proteins by two nodes. In general, the number of neighbor pairs in a tree depends on the particular topology. For a bifurcating tree with *N* OTUs,  $N-2$  pairs of neighbors can potentially occur. The neighbor-joining method first generates a full tree with all the OTUs in a star-like pattern with no hierarchical structure (**Fig. 7.25a**). All  $N(N-1)/2$  pairwise comparisons are made to identify the two most closely related sequences. These OTUs give the smallest sum of branch lengths (see taxa 1 and 2 in **Fig. 7.25b**). OTUs 1 and 2 are now treated as a single OTU, and the method identifies the next pair of OTUs that gives the smallest sum of branch lengths. This could be two OTUs such as 4 and 6, or a single OTU such as 4 paired with the newly formed clade that includes OTUs 1 and 2. The tree has  $N-3$  interior

See Web Document 7.13 at  
 ↗ <http://www.bioinfbook.org/chapter7/> for a detailed UPGMA analysis.



**FIGURE 7.25** The NJ method is a distance-based algorithm. (a) The OTUs are first clustered in a star-like tree. “Neighbors” are defined as OTUs that are connected by a single, interior node in an unrooted, multifurcating tree. (b) The two closest OTUs are identified, such as OTUs 1 and 2. These neighbors are connected to the other OTUs via the internal branch XY. The OTUs that are selected as neighbors in (b) are chosen as those that yield the smallest sum of branch lengths. This process is repeated until the entire tree is generated. Adapted from Saitou and Nei (1987) with permission from Oxford University Press.

branches, and the neighbor-joining method continues to successively identify nearest neighbors until all  $N-3$  branches are identified.

The process of starting with a star-like tree and finding and joining neighbors is continued until the topology of the tree is completed. We describe how branch lengths are calculated in Box 7.6. The neighbor-joining algorithm minimizes the sum of branch lengths at each stage of clustering OTUs, but the final tree is not necessarily the one with the

### BOX 7.6 BRANCH LENGTHS IN A NEIGHBOR-JOINING TREE

Saitou and Nei (1987) defined the sum of the branch lengths as follows. Let  $D_{ij}$  equal the distance between OTUs  $i$  and  $j$ , and let  $L_{ab}$  equal the branch lengths between nodes  $a$  and  $b$ . The sum of the branch lengths  $S$  for the tree in **Figure 7.25a** is:

$$S = \sum_{i=1}^N L_{iX} = \frac{1}{N-1} \sum_{i < j} D_{ij}. \quad (7.20)$$

This result follows from the fact that in computing the total distance each branch is counted  $N-1$  times. For the tree in **Figure 7.25b** the branch length between nodes X and Y (given by  $L_{XY}$ ) is:

$$L_{XY} = \frac{1}{2(N-2)} \left[ \sum_{k=3}^N (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^N L_{iY} \right]. \quad (7.21)$$

In Equation (7.21) the first term in the brackets is the sum of all distances that include  $L_{XY}$ , and the other terms exclude irrelevant branch lengths. Saitou and Nei (1987) provide further detailed analyses of the total branch lengths of the tree.

shortest overall branch lengths. Its results may therefore differ from minimum evolution strategies or maximum parsimony (discussed in the following section). Neighbor-joining produces an unrooted tree topology (because it does not assume a constant rate of evolution), unless an outgroup is specified or midpoint rooting is applied.

We have shown several examples of neighbor-joining trees for 13 globins (**Figs. 7.8, 7.19, 7.22**). This algorithm is especially useful when studying large numbers of taxa. There are many examples of its use in the literature such as studies of the 1918 influenza virus (Taubenberger *et al.*, 2005). There are many alternative distance-based approaches, some of which have been systematically compared (Hollich *et al.*, 2005; Desper and Gascuel, 2006).

### ***Phylogenetic Inference: Maximum Parsimony***

The main idea behind maximum parsimony is that the best tree is that with the shortest branch lengths possible (Czelusniak *et al.*, 1990). Parsimony-based phylogeny based on morphological characters was described by Hennig (1966), and Eck and Dayhoff (1966) used a parsimony-based approach to generating phylogenetic trees such as that in **Figure 7.1**. According to maximum parsimony theory, having fewer changes to account for the way a group of sequences evolved is preferable to more complicated explanations of molecular evolution. We therefore seek the most parsimonious explanations for the observed data. The assumption of phylogenetic systematics is that genes exist in a nested hierarchy of relatedness, and this is reflected in a hierarchical distribution of shared characters in the sequences. The most parsimonious tree is supposed to best describe the relationships of proteins (or genes) that are derived from common ancestors.

The steps are as follows:

- Identify informative sites. If a site is constant (e.g., **Fig. 7.17**, diamonds), then it is not informative. MEGA software includes an option to view parsimony-informative sites (**Fig. 7.26a**, arrow). Noninformative sites include constant sites (**Fig. 7.26a**, closed arrowheads). Also, informative sites are column positions in which there are at least two states (e.g., two different amino acid residues) with at least two taxa having each state. Non-informative sites are indicated in **Figure 7.26a**, open arrowheads.
- Construct trees. Every tree is assigned a cost, and the tree with the lowest cost is sought. When a reasonable number of taxa are evaluated, such as about a dozen or fewer, all possible trees are evaluated and the one with the shortest branch length is chosen. When necessary, a heuristic search is performed to reduce the complexity of the search by ignoring large families of trees that are unlikely to contain the shortest tree.
- Count the number of changes and select the shortest tree (or trees).

Parsimony analysis assumes that characters are independent of each other. The length  $L$  of a full tree is computed as the sum of the lengths  $l_j$  of the individual characters:

$$L = \sum_{j=1}^C w_j l_j \quad (7.22)$$

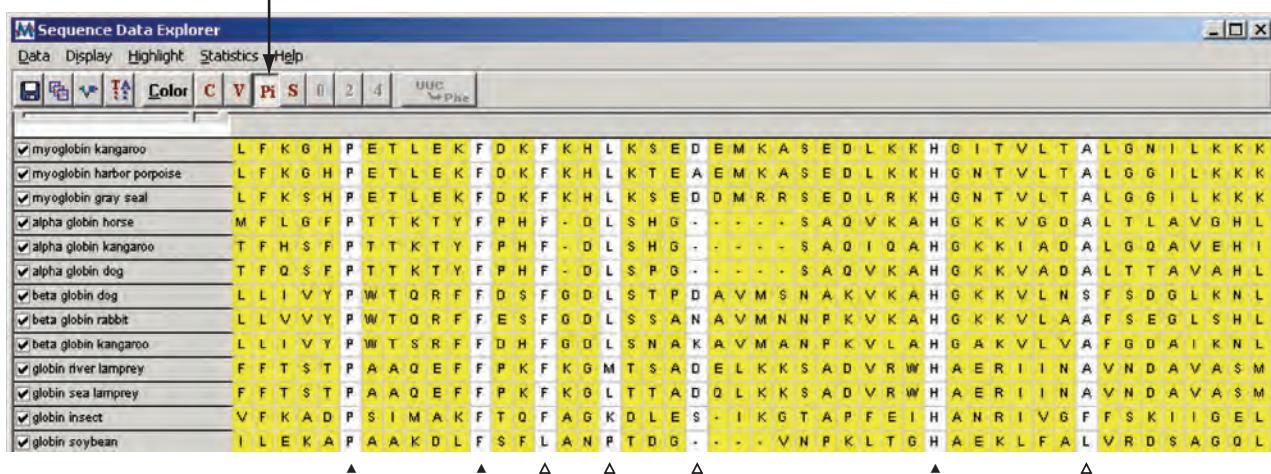
where  $C$  is the total number of characters, and the weight  $w_j$  assigned to each character is typically 1. A different weight might be assigned if, for example, nucleotide transversions incur a greater penalty than transitions.

As an example of how maximum parsimony works, consider five aligned amino acid sequences (**Fig. 7.26b**, taken from the upper left of **Fig. 7.26a**). Two possible trees describe these sequences (**Fig. 7.26c, d**); each tree has hypothetical sequences assigned to ancestral nodes. One of the trees (**Fig. 7.26c**) requires fewer changes to explain how the observed sequences evolved from a hypothetical common ancestor. In this example, each site is treated independently.

An artifact called long-branch attraction sometimes occurs in phylogenetic inference, and parsimony approaches may be particularly susceptible. In a phylogenetic

The word parsimony (from the Latin parcere, "to spare") refers to simplicity of assumptions in a logical formulation.

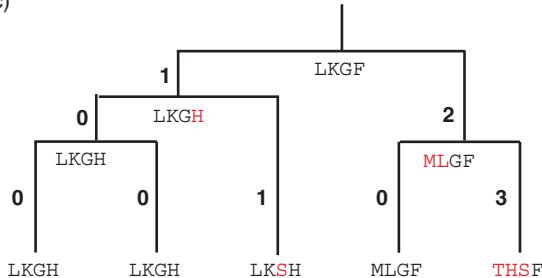
(a)



(b)

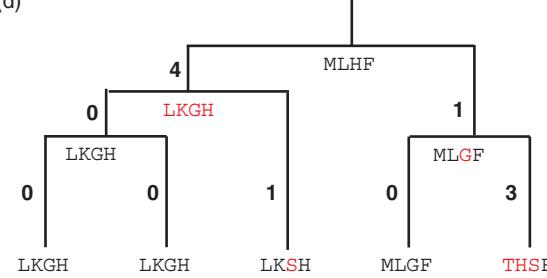
kangaroo LKGH  
porpoise LKGH  
gray seal LKSH  
horse  $\alpha$  MLGF  
kangaroo  $\alpha$  THSF

(c)



Total cost: 7

(d)

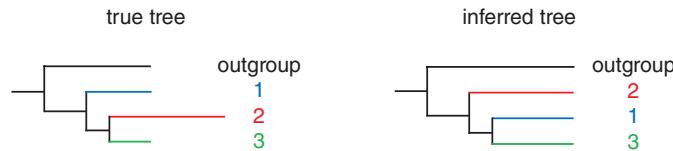


Total cost: 9

**FIGURE 7.26** Principle of maximum parsimony. (a) Many columns of aligned residues are informative for parsimony analysis. However, columns with entirely conserved residues (filled arrowheads) are not informative, nor are columns in which there are at least two different residues that occur at least two times (open arrowheads). This alignment of 13 globin proteins was viewed in MEGA software, with the option to display parsimony-informative characters selected (see arrow); other options include viewing conserved or variable positions. (b) Example of four amino acid residues from five different species (taken from the top left of (a)). Maximum parsimony identifies the simplest (most parsimonious) evolutionary path by which those sequences might have evolved from ancestral sequences. (c, d) Two trees showing possible ancestral sequences. The tree in (c) requires 7 changes from its common ancestor, while the tree in (d) requires 9 changes. Maximum parsimony would therefore select the tree in (c).

Source: MEGA version 5.2; Tamura et al. (2011). Use of MEGA software courtesy of S. Kumar.

reconstruction of protein or DNA sequences, a branch length indicates the number of substitutions that occur between two taxa. Parsimony algorithms assume that all taxa evolve at the same rate and that all characters contribute the same amount of information. Long-branch attraction is a phenomenon in which rapidly evolving taxa are placed together on a tree, not because they are closely related, but artifactually because they both have many mutations. Consider the true tree in **Figure 7.27**, in which taxon 2 represents a DNA or protein that changes rapidly relative to taxa 1 and 3. The outgroup is (by definition) more distantly related than taxa 1, 2, and 3 are to each other. A maximum parsimony algorithm may generate an inferred tree (**Fig. 7.27**) in which taxon 2



**FIGURE 7.27** Long-branch-chain attraction. The true tree includes a taxon (labeled 2) that evolves more quickly than the other taxa. It shares a common ancestor with taxon 3. However, in the inferred tree taxon 2 is placed separately from the other taxa because it is attracted by the long branch of the outgroup. Adapted from Philippe and Laurent (1998), with permission from Elsevier.

is “attracted” toward another long branch (the outgroup) because these two taxa have a large number of substitutions. Whenever two long branches are present, they may be attracted. This may also account for the apparent artifact involving the plant globins in our UPGMA tree (Fig. 7.8c).

#### Model-Based Phylogenetic Inference: Maximum Likelihood

Maximum likelihood is an approach that is designed to determine the tree topology and branch lengths that have the greatest likelihood of producing the observed dataset. A likelihood is calculated for each residue in an alignment, including some model of the nucleotide or amino acid substitution process. It is one of the most computationally intensive but most flexible methods available (Felsenstein, 1981). Maximum parsimony methods sometimes fail when there are large amounts of evolutionary change in different branches of a tree. In contrast, maximum likelihood provides a statistical model for evolutionary change that varies across branches. For example, maximum likelihood can be used to estimate positive and negative selection across individual branches of a tree.

A computationally tractable maximum likelihood method is implemented in the TREE-PUZZLE program (Strimmer and von Haeseler, 1996; Schmidt *et al.*, 2002). The program allows you to specify various models of nucleotide or amino acid substitution and rate heterogeneity (e.g., the  $\Gamma$  distribution). There are three steps. First, TREE-PUZZLE reduces the problem of tree reconstruction to a series of quartets of sequences. For quartet A, B, C, D there are three possible topologies (Fig. 7.11a). In the maximum likelihood step, the program reconstructs all quartet trees. For N sequences there are

$$\binom{N}{4}$$

possible quartets; for example, for 12 myoglobin DNA sequences there are

$$\binom{12}{4}$$

or 495 possible quartets. The three quartet topologies are weighted by their posterior probabilities.

$$\binom{n}{k}$$

is a binomial coefficient that is pronounced “n choose k.” It describes the number of combinations, that is, how many ways there are to choose  $k$  things out of  $n$  possible choices. Given the factorial functions  $n!$  and  $k!$  we can write the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Try the following in R:

```
> choose(12, 4)
[1] 495
```

A file showing how to format 13 globin proteins for input into the TREE-PUZZLE program is provided in Web Document 7.14 at <http://www.bioinfbook.org/chapter7>. Web Document 7.15 shows the TREE-PUZZLE output file.

The TREE-PUZZLE program of Heiko Schmidt, Korbinian Strimmer and Arndt von Haeseler is available at <http://www.tree-puzzle.de/>. You can also perform maximum likelihood (and quartet puzzling) using DNAML (Phylip), PAUP, and MEGA. PhyML (Guindon *et al.*, 2010) is available online at <http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html> (WebLink 7.25).

Nguyen *et al.* (2015; PMID 25371430) introduced IQ-TREE, a fast method for producing phylogenetic trees by maximum likelihood. IQ-Tree implements a series of strategies (both “uphill” moves that perform rearrangements to increase the tree likelihood and “downhill” moves to sample trees avoiding local optima). Thus it is both fast and effective (finding trees with higher likelihoods). These investigators also developed UFboot, an ultrafast bootstrap approximation method (Minh *et al.*, 2015; PMID 23418397). Both IQ-TREE and UFboot are available online and for download at <http://www.cibiv.at/software/iqtree>.

For

$$\binom{12}{4}$$

this corresponds to

$$\frac{12!}{4!(8!)} = \frac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = 495.$$

In the second step, called the quartet puzzling step, a large group of intermediate trees is obtained. The program begins with one quartet tree. Since that tree has four sequences,  $N-4$  sequences remain. These are systematically added to the branches that are most likely based on the quartet results from the first step. Puzzling allows estimates of the support to each internal branch of the tree that is constructed; such estimates are not available for distance- or parsimony-based trees. In the third step, the program generates a majority consensus tree. The branch lengths and maximum likelihood value are estimated.

To use TREE-PUZZLE we can download it from its website, install it (e.g., in a folder named TREE-PUZZLE), and place a set of aligned protein sequences in the Phylip format into the folder. You can try this with the 13 aligned globins available on the website of this textbook, and follow the protocols suggested by Schmidt and von Haeseler (2007). Run the command:

```
$ puzzle 13globins.phy
```

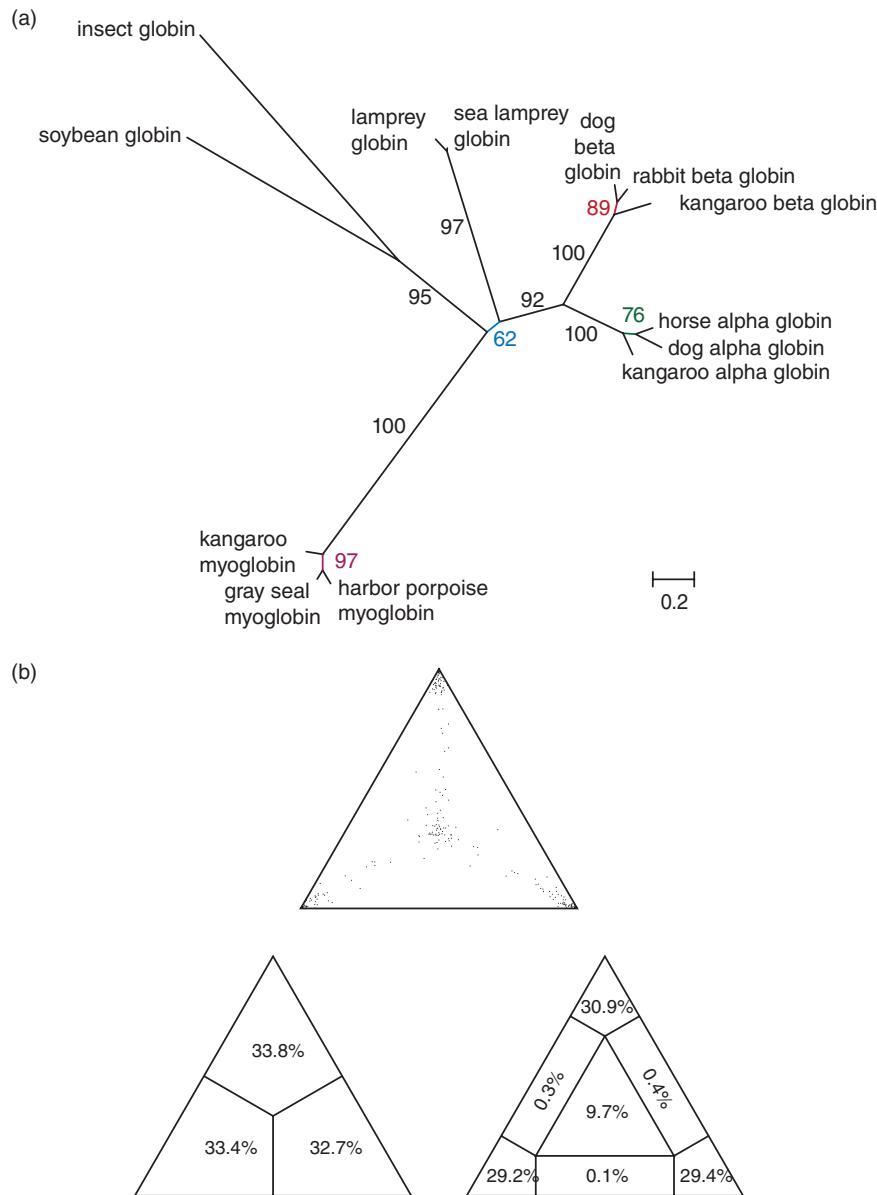
There are many options for the analysis type, choice of outgroup, evolutionary model,  $\Gamma$ -distributed or other models of rate heterogeneity, and other features. There are three output files, including a distance matrix, a puzzle report file, and a tree file that can be drawn with independent programs such as Tree-View or the DrawTree program of Phylip (Fig. 7.28a). Note that TREE-PUZZLE provides support values for the branches of the tree, and these values differ from bootstraps (section “Stage 5: Evaluating Trees” below).

The TREE-PUZZLE program also allows an option called likelihood mapping which describes the support of an internal branch as well as a way to visualize the phylogenetic content of a multiple sequence alignment (Strimmer and von Haeseler, 1996). The quartet topology weights sum to 1, and likelihood mapping plots them on a triangular surface. In this plot, each dot corresponds to a quartet that is positioned spatially according to its three posterior weights (Fig. 7.28b). For 13 globin protein sequences, 9.7% of the quartets were unresolved (as indicated in the center of the triangle). An additional 0.3% + 0.4% + 0.1% of the quartets were partially resolved. For 12 myoglobin DNA coding sequences, only 3% of the quartets were unresolved (not shown). Likelihood mapping summarizes the strength (or conversely the ambiguity) inherent in a dataset for which you perform tree puzzling.

### ***Tree Inference: Bayesian Methods***

Bayesian inference is a statistical approach to modeling uncertainty in complex models. Conventionally we calculate the probability of observing some data (such as the result of a coin toss) given some probability model. This probability is denoted  $P(\text{data}|\text{model})$ , that is, the probability of the data given the model. (This is also read as “the probability of the data conditional upon the model.”) Bayesian inference instead seeks the probability of a tree conditional on the data (that is, based on the observations such as a given multiple sequence alignment). This assumes the form  $P(\text{model}|\text{data})$ ,  $P(\text{hypothesis}|\text{data})$ , or in our case  $P(\text{tree}|\text{data})$ . According to Bayes’s theorem (Huelsenbeck *et al.*, 2002),

$$\Pr[\text{Tree} | \text{Data}] = \frac{\Pr[\text{Data} | \text{Tree}] \times \Pr[\text{Tree}]}{\Pr[\text{Data}]} \quad (7.23)$$



**FIGURE 7.28** Maximum likelihood inference of phylogenetic trees using quartet puzzling. The taxa in any tree with four or more sequences can be represented as quartets of sequences (A, B, C, D) as shown in **Figure 7.11b**. These can be placed in a tree with three possible topologies. Quartet puzzling applies maximum likelihood criteria to identify the most likely tree. (a) This tree of 13 globin proteins was constructed using the TREE-PUZZLE program. Support values for the branches are shown. (b) Likelihood mapping (in TREE-PUZZLE) indicates the frequency with which quartets are successfully resolved. In the top triangle, there are 495 points corresponding to all possible quartets. Each quartet has three posterior weights which are mapped in triangles. For the analysis of 13 globins, only 9.7% of the quartets were unresolved. Likelihood mapping provides an estimate of the ability of a given dataset to be successfully analyzed in quartets.

Source: TREE-PUZZLE.

Bayesian estimation of phylogeny is focused on a quantity called the posterior probability distribution of trees,  $\text{Pr}[\text{Tree}|\text{Data}]$ . (This is read as “the probability of observing a tree given the data.”) For a given tree, the posterior probability is the probability that the tree is correct, and our goal is to identify the tree with the maximum probability. On the right-hand side of Equation (7.23), the denominator  $\text{Pr}[\text{Data}]$  is a normalizing constant

over all possible trees. The numerator consists of the prior probability of a phylogeny  $\text{Pr}[\text{Tree}]$  and the likelihood  $\text{Pr}[\text{Data}|\text{Tree}]$ . These terms represent a distinctive feature of Bayesian inference of phylogeny: the user specifies a prior probability distribution of trees (although it is allowable for all possible trees to be given equal weight).

Practically, we can apply a Bayesian inference approach using the MrBayes software program (Ronquist *et al.*, 2012). There are four steps. First, read in a Nexus data file. This can be accomplished by performing a multiple sequence alignment of interest, then converting it into the Nexus format with a tool such as ReadSeq. We use an example of 13 globin protein-coding DNA sequences. Install the program (on the PC, Linux, or Mac O/S), open a terminal window, and enter `mb` (or in some cases `./mb`) on the command line to begin the program.

```
MrBayes > execute globins.nex
```

The procedure for performing this analysis is described in Web Document 7.16 at <http://www.bioinfbook.org/chapter7>.

MrBayes is available from <http://mrbayes.sourceforge.net/> (WebLink 7.26). It was developed by Fredrik Ronquist and John Huelsenbeck. By late 2014, the 2001 and 2003 papers on MrBayes had been cited in over 28,000 publications. Web Document 7.17 shows how to format 13 globin proteins for input into MrBayes, and Web Document 7.18 shows the output. See <http://www.bioinfbook.org/chapter7>.

The `execute` command reads in the globin alignment.

Second, specify the evolutionary model and the parameters of your tree construction. This can be considered either a strength of the Bayesian approach (because your judgment may help you to select optimal parameters) or a weakness (because there is a subjective element to the procedure). All priors do not have to be informative; conservative settings can be selected.

Options are available for data comprising DNA (whether coding or not), ribosomal DNA (for the analysis of paired stem regions; see Chapter 10), and protein. Before performing the analysis, a prior probability distribution is specified for the parameters of the likelihood model. There are six types of parameters that are set as the priors for the model in the case of the analysis of nucleotide sequences: (1) the topology of the trees (e.g., some nodes can be constrained to always be present); (2) the branch lengths; (3) the stationary frequencies of the four nucleotides; (4) the six nucleotide substitution rates (for  $A \leftrightarrow C$ ,  $A \leftrightarrow G$ ,  $A \leftrightarrow T$ ,  $C \leftrightarrow G$ ,  $C \leftrightarrow T$ , and  $G \leftrightarrow T$ ); (5) the proportion of invariant sites; and (6) the shape parameter of the gamma distribution of rate variation. Your decisions on how to specify these parameters may be subjective.

As an example of specifying a parameter for protein studies, use the `prset` command to set the priors for the probabilities. The mixed option of the `aamodelpr` parameter samples across amino acid rate matrices and is one method of averaging across ten models:

```
MrBayes > prset aamodelpr=mixed
```

As an alternative, specify a particular matrix (such as Dayhoff's) with the argument `fixed`.

```
MrBayes > prset aamodelpr=fixed(dayhoff)
```

Third, run the analysis. This is invoked with the `mcmc` (Monte Carlo Markov Chain) command. You may include dozens of optional arguments such as the number of cycles (`ngen`) or chains (`nchains`) used in MCMC analysis.

```
MrBayes > mcmc nchains=4 ngen=300000
```

The posterior probability of the possible phylogenetic trees is ideally calculated as a summation over all possible trees and, for each tree, all combinations of branch lengths and substitution model parameters are evaluated. In practice this probability cannot be determined analytically, but it can be approximated using MCMC. This is done by drawing many samples from the posterior distribution (Huelsenbeck *et al.*, 2002). MrBayes

runs two simultaneous, independent analyses beginning with distinct, randomly initiated trees. This helps to ensure that your analysis includes a good sampling from the posterior probability distribution. Eventually the two runs should reach convergence. An MCMC analysis is performed in three steps: first, a Markov chain is started with a tree that may be randomly chosen. Second, a new tree is proposed. Third, the new tree is accepted with some probability. Typically tens to hundreds of thousands of MCMC iterations are performed. The proportion of time that the Markov chain visits a particular tree is an approximation of the posterior probability of that tree. Some authors have cautioned that MCMC algorithms can give misleading results, especially when data have conflicting phylogenetic signals (Mossel and Vigoda, 2005). During the MCMC analysis you can observe a progressive decline of the average standard deviation of split frequencies, in our case declining from 0.12 to ~0.005.

Fourth, summarize the parameters of the run with `sump`, and summarize the trees from MCMC analysis with `sumt`:

```
MrBayes > sump
MrBayes > sumt
```

The summary of parameters includes convergence diagnostics indicating whether parameters have been undersampled and which models are favored. MrBayes provides a variety of additional output files providing information such as a list of trees found during the MCMC search (sorted by posterior probability values for the best trees), phylogram, branch lengths (in units of the number of expected substitutions per site), and clade credibility values. The summary statistics for a Bayesian analysis are provided. An example for 13 globin proteins is shown as a phylogram (**Fig. 7.29a**) or as a radial tree (**Fig. 7.29b**). This radial tree represents a consensus tree and includes support values for interior branches.

Bayesian inference of phylogeny resembles maximum likelihood because each method seeks to identify a quantity called the likelihood which is proportional to observing the data conditional on a tree. The methods differ in that Bayesian inference includes the specification of prior information and uses MCMC to estimate the posterior probability distribution. Although they were introduced relatively recently, Bayesian approaches to phylogeny have become increasingly commonplace.

The tree shown in **Figure 7.29a** is from the `sumt` output of MrBayes. The tree shown in **Figure 7.29b** is made by processing one of the MrBayes output files (`globins.nex.con.tre`) using the FigTree graphical viewer of Andrew Rambaut and colleagues. FigTree is available at <http://tree.bio.ed.ac.uk/software/figtree/> (WebLink 7.27).

## Stage 5: Evaluating Trees

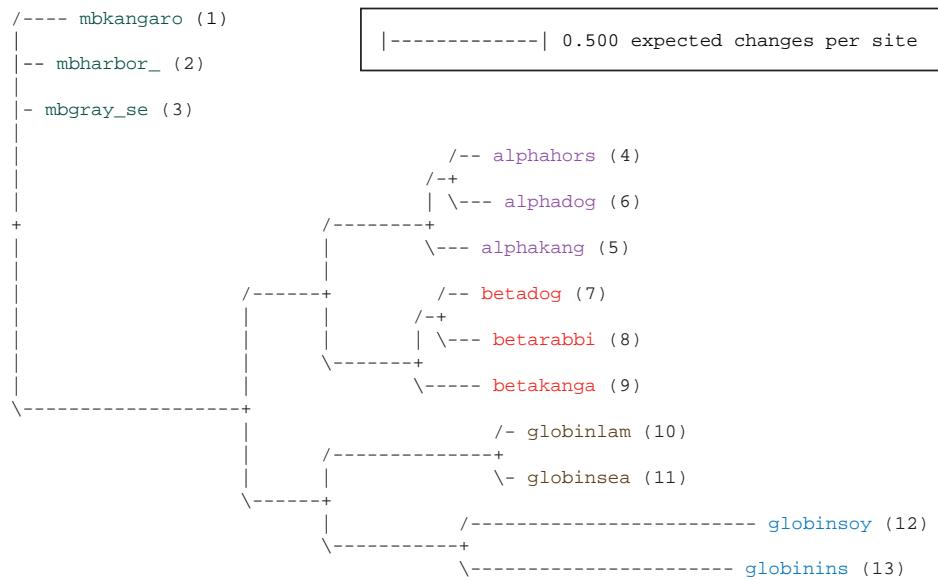
After you have constructed a phylogenetic tree, how can you assess its accuracy? The main criteria by which accuracy may be assessed are consistency, efficiency, and robustness (Hillis and Huelsenbeck, 1992; Hillis, 1995). The accuracy of a tree-building approach or the accuracy of a particular tree can be studied. The most common approach is bootstrap analysis (Felsenstein, 1985; Hillis and Bull, 1993). Bootstrapping describes the robustness of the tree topology: given a particular branching order, how consistently does a tree-building algorithm find that branching order using a randomly permuted version of the original dataset? Bootstrapping allows an inference of the variability in an unknown distribution from which the data were drawn (Felsenstein, 1985).

Nonparametric bootstrapping is performed as follows. A multiple sequence alignment is used as the input data to generate a tree using some tree-building method. The program then makes an artificial dataset of the same size as the original dataset by randomly picking columns from the multiple sequence alignment. This is usually performed with replacement, meaning that any individual column may appear multiple times (or not at all). A tree is generated from the randomized dataset. A large number of bootstrap replicates are then generated; typically, between 100 and 1000 new trees are made by this process. The bootstrap trees are compared to the original, inferred tree(s). The

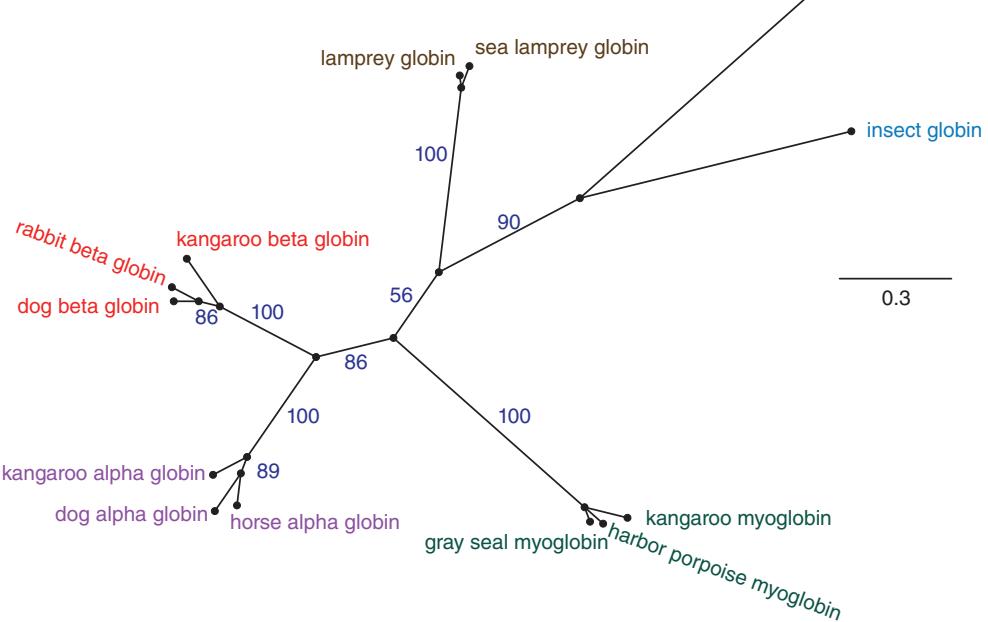
Accuracy refers to the degree to which a tree approximates the true tree. We define and discuss precision and accuracy in Chapter 11 in the context of microarray data analysis.

Parametric bootstrapping refers to repeated random sampling without replacement from the original sample. It is not used as often as nonparametric bootstrapping.

(a) Phylogram (MrBayes output)



(b) Radial tree with clade credibility values (MrBayes output)



**FIGURE 7.29** Bayesian inference of phylogeny for 13 globin proteins using MrBayes software (version 3.2.2). The input sequences were aligned using MAFFT at EBI (see Chapter 6). The amino acid model (using default settings) was Poisson, with 20 states corresponding to the amino acids and equal rates of substitution. Prior parameters included equal, fixed frequencies for the states, an equal probability for all topologies, and unconstrained branch lengths. Monte Carlo Markov Chain estimation of the posterior distribution was achieved using a run of 1,000,000 trials. (a) Phylogram output shows clades containing various globin subtypes. Note that the myoglobin group is unresolved. (b) Tree files can be exported from MrBayes and viewed using FigTree software. An unrooted radial tree is shown here. Nodes are indicated as closed circles. Clade credibility values (values along branches) give 100% support for the separation of clades containing myoglobins, alpha globins, beta globins, and lamprey globins as well as 90% support for the insect globin and soybean globin clade. The node connecting the three unresolved myoglobins is multifurcating. The scale bar is 0.3 expected changes per amino acid site.

Source: MrBayes. Courtesy of J. Huelsenbeck and F. Ronquist.

information you get from bootstrapping is the frequency with which each clade in the original tree is observed.

An example of the bootstrap procedure using MEGA is shown in **Figure 7.19**. The percentage of times that a given clade is supported in the original tree is provided based on how often the bootstraps supported the original tree topology. Bootstrap values above 70% are sometimes considered to provide support for the clade designations. Hillis and Bull (1993) have estimated that such values provide statistical significance at the  $p < 0.05$  level. This approach measures the effect of random weighting of characters in the original data matrix, giving insight into how strongly the phylogenetic signal that produces a tree is distributed through the multiple sequence alignment. In **Figure 7.19a, b**, the clade containing three alpha globins has 100% bootstrap support, indicating that in all 500 bootstrap replicates that clade maintained its integrity (with none of the three alpha globins assigned to a different clade, and no non-alpha globin joining that clade). However, the clade containing horse and dog alpha globin received only 52% bootstrap support (**Fig. 7.19b**), suggesting that about half the time kangaroo alpha globin was in a clade with the dog or horse orthologs. This example shows how viewing the bootstrap percentages can be useful to estimate the robustness of each clade in a tree. Note that bootstrapping supports a model in which alpha globins, beta globins, myoglobins, and lamprey globins each are assigned to a unique clade.

Maximum likelihood approaches report the tree with the greatest likelihood, and they also report the likelihood for internal branches. For Bayesian inference of phylogeny, the result is typically the most probable tree (called a maximum *a posteriori* probability estimate). The results are often summarized using a majority rule consensus tree in which the values represent the posterior probability that each clade is true. The confidence estimates may sometimes be too liberal (Suzuki *et al.*, 2002). For example, Mar *et al.* (2005) found that Bayesian posterior probabilities reached 100% at bootstrap percentages of 80%.

## PERSPECTIVE

Molecular phylogeny is a fundamental tool for understanding the evolution and relationships of protein (and nucleic acid) sequences. The main output of this analysis is a phylogenetic tree, which is a graphical representation of a multiple sequence alignment. The recent rapid growth of DNA and protein sequence data, along with the visual impact of phylogenetic trees, has made phylogeny increasingly important and widely applied. We will show examples of trees in Chapters 15–19 as we explore genomes across the tree of life.

The field of molecular phylogeny includes conceptually distinct approaches including those outlined in this chapter (distance, maximum parsimony, maximum likelihood, and Bayesian methods). For each of these approaches software tools continue to evolve. It is therefore quite reasonable for you to obtain a multiple sequence alignment and perform phylogenetic analyses with all four tree-making approaches and with a variety of substitution models. The relative merits of these maximum parsimony versus model-based approaches continue to be debated (e.g., Kolaczkowski and Thornton, 2004; Steel, 2005).

## PITFALLS

The quality of a phylogenetic tree based on molecular sequence data depends upon the quality of the sequence data and the multiple sequence alignment. It is also necessary to choose appropriate models of nucleotide or amino acid substitution for the

phylogeny. There is an active debate within the field concerning the importance of selecting models without too few or too many parameters. Furthermore, the choice of tree-making approaches (from distance to maximum parsimony, maximum likelihood, and Bayesian frameworks) may produce an optimal tree having different topologies and branch lengths. In contrast to multiple sequence alignments of proteins having known structures, there are few “gold standard” benchmark datasets that allow objective definitions of the true trees.

In practice, for many published phylogenetic trees the underlying multiple sequence alignments are not available and it is challenging to assess the quality of published trees. A group of 28 phylogeny experts has begun to define reporting standards for phylogenetic analysis (called “Minimum Information about a Phylogenetic Analysis” or MIAPA; Leebens-Mack *et al.*, 2006). Such standards may someday require those who report trees to include the underlying data as well as descriptions of the models used to construct trees.

Each approach has potential pitfalls. Neighbor-joining trees may introduce errors for distantly related sequences because they do not adequately account for the high variances. Bayesian and maximum likelihood approaches are both dependent on appropriate prior choices of parameters as well as proper estimation. Maximum parsimony essentially offers no statistical model for phylogeny.

Finally, your understanding of the output of the phylogenetic analysis is critical. Each of the methods used to reconstruct phylogenetic trees involves many assumptions and suffers from potential weaknesses. It is also important to learn how to interpret trees as graphs that reflect the historical relationships of taxa; in a tree of protein sequences, for example, the internal nodes correspond to inferred ancestral sequences.

## ADVICE FOR STUDENTS

While there are dozens of leading phylogeny software packages, I suggest that you immerse yourself in two: MEGA and MrBayes. Both are extraordinarily popular. As of 2015 MEGA has been downloaded over 1.1 million times, while the MrBayes papers from 2001 and 2003 have together been cited over 25,000 times. Using MEGA provides an excellent introduction to phylogeny with the ability to compare distance matrices to trees and the ability to create trees using a wide range of methods. MrBayes provides a good method of becoming familiar with the thinking behind Bayesian analyses which have applications across the field of bioinformatics.

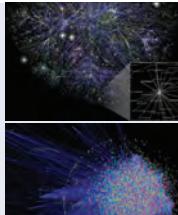
For MEGA and MrBayes, try these four approaches. (1) Read the documentation accompanying each tool. (2) Read the papers by the authors of the software. (3) Make and analyze as many trees as possible. Try taking some globin sequences (from the Web Documents given in this chapter), or choose your own data. (4) Find papers of interest in the literature – perhaps beginning with papers written by the authors of MEGA or MrBayes software – and find several examples of published phylogenetic trees. Read those papers and understand how phylogenetic trees were made and interpreted. The raw data that go into making a tree are almost never provided with publications (although they should always be required), but in many cases you should be able to find the DNA, RNA, or protein sequences in public databases and then try to make the same trees that were reported in the publications. All this will help you become familiar with the methods and scope of phylogeny. More broadly, you will learn the broad range of biological principles that can be elucidated through phylogenetic analyses.

## WEB RESOURCES

An informative starting point for phylogeny resources on the World Wide Web is the site of Joseph Felsenstein (<http://evolution.genetics.washington.edu/phylip/software.html>; WebLink 7.24). About 400 packages and 50 web servers are listed, organized by categories such as phylogenetic methods, computer platforms, and assorted types of data. All of the major software tools listed in this chapter have websites that we have listed, and most of these sites include detailed documentation and examples that further illustrate both the practical use of the software and the conceptual issues addressed by the authors' particular approach to phylogeny.

The HIV Sequence Database at the Los Alamos National Laboratory (discussed in Chapter 16 on viruses) offers a brief online guide to making and interpreting phylogenetic trees.

The Los Alamos tutorial is available at [http://www.hiv.lanl.gov/content/sequence/TUTORIALS/TREE\\_TUTORIAL/Tree-tutorial.html](http://www.hiv.lanl.gov/content/sequence/TUTORIALS/TREE_TUTORIAL/Tree-tutorial.html) (WebLink 7.28).



## Discussion Questions

**[7.1]** Consider a multiple sequence alignment containing a grossly incorrect region. What is the likely consequence of using this alignment to infer a phylogenetic tree using a distance-based or character-based method?

**[7.2]** Are there gene (or protein) families for which you expect distance-based tree-building methods to give substantially different results than character-based methods?

**[7.3]** How would you test whether a particular human gene (or protein) is under positive selection? What species would you select for comparison to the human sequence?

**[7.4]** We showed in Chapter 3 that two proteins sharing 50% identity have undergone an average of about 80 changes per 100 aligned residues. Does the Jukes–Cantor correction show the same phenomenon for DNA sequences?

### PROBLEMS/COMPUTER LAB

**[7.1]** Determine whether human and chimpanzee mitochondrial DNA sequences have equal evolutionary rates between lineages. To do this, use Tajima's relative rate test as implemented in MEGA.

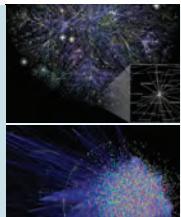
- (1) Obtain MEGA software.
- (2) Obtain mitochondrial DNA sequences from human, chimpanzee, bonobo, orangutan, gorilla, and gibbon from Web Document 7.19 at <http://www.bioinfbook.org/chapter7>.
- (3) Apply Tajima's test using an appropriate outgroup. Is the probability value significant (< 0.05)?

**[7.2]** Perform phylogenetic analyses using MEGA software.

- (1) Go to the conserved domain database (<http://www.ncbi.nlm.nih.gov/cdd>) at NCBI.
- (2) Enter lipocalins (or another family of your choice; you can also begin at Ensembl, HomoloGene, or Pfam).
- (3) Select the mFasta format then click "Reformat." The result is a multiple sequence alignment. Copy this into a text editor (such as NotePad++), and simplify the names of the sequences.
- (4) Import the file (or paste the sequences) into MEGA as shown in **Figure 7.9**. Align the sequences and save in the .mas and .meg formats.
- (5) Choose Phylogeny > Construct/Test to create neighbor-joining, maximum likelihood, or other trees.
- (6) For each tree you create, read the caption. Try the tree tools (e.g., placing a root, flipping nodes, showing or hiding branch lengths, interconverting display formats).
- (7) Perform bootstrapping. Identify clades having low levels of support. Why does this occur?

**[7.3]** Perform Bayesian inference of phylogeny using MrBayes software. A detailed analysis for 13 globin proteins is provided in Web Documents 7.17 and 7.18. Use a group of proteins, and also perform an analysis for DNA coding sequences from a group of myoglobins (and cyto-globin as an outgroup; provided in Web Document 7.5).

**[7.4]** For students interested in Python, explore the ETE programming toolkit for the automated manipulation, analysis, and visualization of phylogenetic trees. The website <http://pythonhosted.org/ete2/> (WebLink 7.29) includes documentation, access to ETE, and a tutorial.



## Self-Test Quiz

**[7.1]** According to the molecular clock hypothesis:

- (a) all proteins evolve at the same, constant rate;
- (b) all proteins evolve at a rate that matches the fossil record;
- (c) for every given protein, the rate of molecular evolution gradually slows down like a clock that runs down; or
- (d) for every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages.

**[7.2]** The two main features of any phylogenetic tree are:

- (a) the clades and the nodes;
- (b) the topology and the branch lengths;
- (c) the clades and the root; or
- (d) the alignment and the bootstrap.

**[7.3]** Which one of the following is a character-based phylogenetic algorithm?

- (a) neighbor-joining;
- (b) Kimura;
- (c) maximum likelihood; or
- (d) PAUP.

**[7.4]** Two basic ways to make a phylogenetic tree are distance-based and character-based. A fundamental difference between them is:

- (a) distance-based methods essentially summarize relatedness across the length of protein or DNA sequences while character-based methods do not;
- (b) distance-based methods are only used for DNA data while character-based methods are used for DNA or protein data;
- (c) distance-based methods use parsimony while character-based methods do not; or

(d) distance-based methods have branches that are proportional to time while character-based methods do not.

**[7.5]** An example of an operational taxonomic unit (OTU) is:

- (a) multiple sequence alignment;
- (b) protein sequence;
- (c) clade; or
- (d) node.

**[7.6]** For a given pair of OTUs, which of the following is true?

- (a) the corrected genetic distance is greater than or equal to the proportion of substitutions; or
- (b) the proportion of substitutions is greater than or equal to the corrected genetic distance.

**[7.7]** Transitions are almost always weighted more heavily than transversions:

- (a) true; or
- (b) false.

**[7.8]** One of the most common errors in making and analyzing a phylogenetic tree is:

- (a) using a bad multiple sequence alignment as input;
- (b) trying to infer the evolutionary relationships of genes (or proteins) in the tree;
- (c) trying to infer the age at which genes (or proteins) diverged from each other; or
- (d) assuming that clades are monophyletic.

**[7.9]** You have 1000 viral DNA sequences of 500 residues each, and you want to know if there are any pairs that are identical (or nearly identical). Which of the following is the most efficient method to use?

- (a) BLAST;
- (b) maximum-likelihood phylogenetic analysis;
- (c) neighbor-joining phylogenetic analysis; or
- (d) Popset.

## SUGGESTED READING

For an excellent review of phylogeny see Yang and Rannala (2012) as well as books by Yang (2006) and Felsenstein (2004). For an overview of trends in phylogeny see Blair and Murphy (2011). Sudhir Kumar, Koichi Tamura and colleagues discuss statistical issues in phylogenomics (Kumar *et al.*, 2012).

For Bayesian inference of phylogeny, excellent articles have been published by Ronquist (2004) and Huelsenbeck *et al.* (2002).

## REFERENCES

- Abascal, F., Zardoya, R., Posada, D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**(9), 2104–2105. PMID: 15647292.
- Anfinsen, C. B. 1959. *The Molecular Basis of Evolution*. John Wiley and Sons, New York.
- Bajaj, M., Blundell, T.L., Horuk, R. et al. 1986. Coypu insulin. Primary structure, conformation and biological properties of a hystricomorph rodent insulin. *Biochemical Journal* **238**(2), 345–351. PMID: 3541911.
- Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., Doolittle, W. F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972–977.
- Baum, D.A., Smith, S.D., Donovan, S.S. Evolution. 2005. The tree-thinking challenge. *Science* **310**, 979–980.
- Baurain, D., Brinkmann, H., Philippe, H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Molecular Biology and Evolution* **24**, 6–9. PMID: 17012374.
- Baxevanis, A. D., Ouellette, B. F. 2009. *Bioinformatics*, 3rd edition. Wiley-Interscience, New York.
- Blair, C., Murphy, R.W. 2011. Recent trends in molecular phylogenetic analysis: where to next? *Journal of Heredity* **102**(1), 130–138. PMID: 20696667.
- Bos, D.H., Posada, D. 2005. Using models of nucleotide evolution to build phylogenetic trees. *Developmental and Comparative Immunology* **29**, 211–227.
- Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* **18**, 225–239.
- Bustamante, C.D., Fledel-Alon, A., Williamson, S. et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157. PMID: 16237444.
- Cavalli-Sforza, L. L., Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* **19**, 233–257.
- Cox, A.L., Mosbruger, T., Mao, Q. et al. 2005. Cellular immune selection with hepatitis C virus persistence in humans. *Journal of Experimental Medicine* **201**, 1741–1752. PMID: 15939790.
- Cunningham, C.W., Omland, K.E., Oakley, T.H. 1998. Reconstructing ancestral character states: a critical reappraisal. *Tree* **13**, 361–366.
- Cutter A.D., Payseur B.A. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* **14**(4), 262–274. PMID: 23478346.
- Czelusniak, J., Goodman, M., Moncrief, N. D., Kehoe, S. M. 1990. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods in Enzymology* **183**, 601–615.
- Darriba, D., Taboada, G.L., Doallo, R., Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**(8), 1164–1165. PMID: 21335321.
- Darriba, D., Taboada, G.L., Doallo, R., Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**(8), 772. PMID: 22847109.
- Darwin, C. 1859. *The Origin of Species by Means of Natural Selection*. John Murray, London.
- Dayhoff, M. O. 1978. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
- Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J., Jones, D.D. 1972. Gene duplications in evolution: the globins. In: *Atlas of Protein Sequence and Structure 1972* (ed. Dayhoff, M.O.), National Biomedical Research Foundation, Washington, DC, Vol. 5.
- Delport, W., Poon, A.F., Frost, S.D., Kosakovsky Pond, S.L. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**(19), 2455–2457. PMID: 20671151.
- Desper, R., Gascuel, O. 2006. Getting a tree fast: Neighbor Joining, FastME, and distance-based methods. *Current Protocols in Bioinformatics Chapter 6*, Unit 6.3. PMID: 18428768.
- Dickerson, R. E. 1971. Sequence and structure homologies in bacterial and mammalian-type cytochromes. *Journal of Molecular Biology* **57**, 1–15.

- Dobzhansky, T. 1973. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher* **35**, 125–129.
- Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129.
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
- Eck, R.V., Dayhoff, M.O. 1966. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, MD.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.
- Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* **38**, 16–24.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791.
- Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22**, 521–565.
- Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology* **266**, 418–427.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland MA.
- Fitch, W.M., Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**(3760), 279–284. PMID: 5334057.
- Fitch, W.M., Ayala, F.J. 1994. The superoxide dismutase molecular clock revisited. *Proceedings of the National Academy of Science, USA* **91**(15), 6802–6807. PMID: 8041700.
- Graur, D., Li, W.-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Guindon S., Dufayard, J.F., Lefort, V. et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3), 307–321. PMID: 20525638.
- Haeckel, E. 1900. *The Riddle of the Universe*. Harper and Brothers, New York.
- Hall, B. G. 2011. *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists*. Sinauer Associates, Sunderland, MA.
- Hennig, W. 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana.
- Hillis, D. M. 1995. Approaches for assessing phylogenetic accuracy. *Systematic Biology* **44**, 3–16.
- Hillis, D. M., Huelsenbeck, J. P. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity* **83**, 189–195.
- Hillis, D. M., Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**, 182–192.
- Hollich, V., Milchert, L., Arvestad, L., Sonnhammer, E.L. 2005. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Molecular Biology and Evolution* **22**, 2257–6422.
- Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* **51**, 673–688. PMID: 12396583.
- Ingram, V.M. 1961. *Hemoglobin and its Abnormalities*. Charles C. Thomas, Springfield, IL.
- Jollès, J., Prager, E.M., Alnemri, E.S. et al. 1990. Amino acid sequences of stomach and nonstomach lysozymes of ruminants. *Journal of Molecular Evolution* **30**, 370–382. PMID: 2111849.
- Jukes, T.H. 1979. Dr. Best, insulin, and molecular evolution. *Canadian Journal of Biochemistry* **57**(6), 455–458. PMID: 383230.
- Jukes, T. H., Cantor, C. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*. (eds H. N. Munro, J. B. Allison). Academic Press, New York, pp. 21–132.
- Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**, 624–626.
- Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**, 111–120.

- Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Kolaczkowski, B., Thornton, J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**(7011), 980–984. PMID: 15496922.
- Korber B. (2000). HIV signature and sequence variation analysis. In *Computational Analysis of HIV Molecular Sequences* (eds A. G.Rodrigo, G. H.Learn), pp. 55–72. Kluwer Academic Publishers, Dordrecht, Netherlands.
- Kumar S., Filipski, A.J., Battistuzzi, F.U. *et al.* 2012. Statistics and truth in phylogenomics. *Molecular Biology and Evolution* **29**(2), 457–472. PMID: 21873298.
- Le, S.Q., Gascuel, O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**(7), 1307–1320. PMID:18367465.
- Leebens-Mack, J., Vision, T., Brenner, E. *et al.* 2006. Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA). *OMICS* **10**, 231–237. PMID: 16901231.
- Lemey, P., Salemi, M., Vandamme, A.-M. 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition. Cambridge University Press, Cambridge.
- Li, W.-H. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.
- Mar, J.C., Harlow, T.J., Ragan, M.A. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. *BMC Evolutionary Biology* **5**, 1–20.
- Margoliash, E. 1963. Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Science, USA* **50**, 672–679.
- Margoliash, E., Smith, E.L. 1965. Structural and functional aspects of cytochrome c in relation to evolution. In *Evolving Genes and Proteins* (eds V.Bryson and H.J.Vogel), pp. 221–242. Academic Press, Inc., New York.
- Mayr, E. 1982. *The Growth of Biological Thought. Diversity, Evolution, and Inheritance*. Belknap Harvard, Cambridge, MA.
- Mossel, E., Vigoda, E. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**, 2207–2209.
- Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics* **30**, 371–403.
- Nei, M., Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.
- Nei, M., Kumar, S. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.
- Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197–218.
- Nuttall, G. H. F. 1904. *Blood Immunity and Blood Relationship*. Cambridge University Press, Cambridge.
- Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.
- Penny, D., Foulds, L.R., Hendy, M.D. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature* **297**, 197–200.
- Philippe, H., Laurent, J. 1998. How good are deep phylogenetic trees? *Current Opinion in Genetics and Development* **8**, 616–623.
- Posada, D. 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research* **34**(Web Server issue), W700–W703.
- Posada, D., Crandall, K. A. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.
- Posada, D., Buckley, T.R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**, 793–808. PMID: 15545256.

- Ray, S.C., Fanning, L., Wang, X.H. *et al.* 2005. Divergent and convergent evolution after a common-source outbreak of hepatitis C virus. *Journal of Experimental Medicine* **201**, 1753–1759. PMID: 15939791.
- Rokas, A., Kruger, D., Carroll, S.B. 2005. Animal evolution and the molecular signature of radiations compressed in time. *Science* **310**, 1933–1938.
- Ronquist, F. 2004. Bayesian inference of character evolution. *Trends in Ecology and Evolution* **19**, 475–481.
- Ronquist F., Teslenko, M., van der Mark, P. *et al.* 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**(3), 539–542. PMID: 22357727.
- Sabeti, P.C., Schaffner, S.F., Fry, B. *et al.* 2006. Positive natural selection in the human lineage. *Science* **312**, 1614–1620. PMID: 16778047.
- Saitou, N., Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.
- Schmidt, H.A., von Haeseler, A. 2007. Maximum-likelihood analysis using TREE-PUZZLE. *Current Protocols in Bioinformatics Chapter 6*, Unit 6.6. PMID: 18428792.
- Schmidt, H. A., Strimmer, K., Vingron, M., von Haeseler, A. 2002. Tree-Puzzle: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.
- Simpson, G. G. 1952. *The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man*. Yale University Press, New Haven.
- Sneath, P.H.A., Sokal, R.R. 1973. *Numerical Taxonomy*. W.H. Freeman & Co., San Francisco.
- Sokal, R.R., Michener, C.D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**, 1409–1437.
- Spillane, C., Schmid, K.J., Laouelli-Duprat, S. *et al.* 2007. Positive darwinian selection at the imprinted MEDEA locus in plants. *Nature* **448**, 349–352. PMID: 17637669.
- Steel, M. 2005. Should phylogenetic models be trying to “fit an elephant”? *Trends in Genetics* **21**, 307–309.
- Strimmer, K., von Haeseler, A. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* **13**, 964–969.
- Suzuki, Y., Glazko, G.V., Nei, M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of the National Academy of Science, USA* **99**, 16138–16143.
- Swofford, D. L., Olsen, G. J., Waddell, P. J., Hillis, D. M. 1996. Molecular systematics: Context and controversies. In *Molecular Systematics*, 2nd edition (eds D. M. Hillis, C. Moritz, B. K. Mable). Sinauer Associates, Sunderland, MA.
- Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599–607.
- Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* **9**, 678–687.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**(12), 2725–2729. PMID: 24132122.
- Taubenberger, J.K., Reid, A.H., Lourens, R.M. *et al.* 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**, 889–893. PMID: 16208372.
- Thornton, J. W., DeSalle, R. 2000. Gene family evolution and homology: Genomics meets phylogenetics. *Annual Review in Genomics and Human Genetics* **1**, 41–73.
- Uzzell, T., Corbin, K.W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089–1096.
- Whelan, S. 2008. Inferring trees. *Methods in Molecular Biology* **452**, 287–309. PMID: 18566770
- Wilson, E. O. 1992. *The Diversity of Life*. W. W. Norton, New York.
- Yang, Z. 2006. *Computational Molecular Evolution* (Oxford Series in Ecology and Evolution). Oxford University Press, New York.

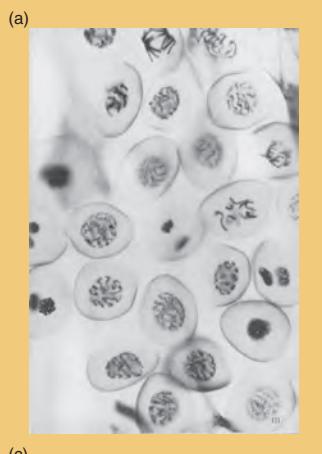
- Yang, Z., Rannala, B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**(5), 303–314. PMID: 22456349
- Zar, J. H. 1999. *Biostatistical Analysis*. Fourth edition. Prentice Hall, Upper Saddle River, NJ.
- Zhang, J., Gu, X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* **149**, 1615–1625.
- Zuckerkandl, E. 1987. On the molecular evolutionary clock. *Journal of Molecular Evolution* **26**(1–2), 34–46. PMID: 3125336
- Zuckerkandl, E., Pauling, L. 1962. Molecular disease, evolution, and genic heterogeneity. In *Horizons In Biochemistry* (eds M.Kasha, B.Pullman). Albert Szent-Gyorgyi Dedicatory Volume. Academic Press, New York.
- Zuckerkandl, E., Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins* (eds B.Bryson, H.Vogel). Academic Press, New York, pp. 97–166.



# Genomewide Analysis of DNA, RNA, and Protein

PART

|||



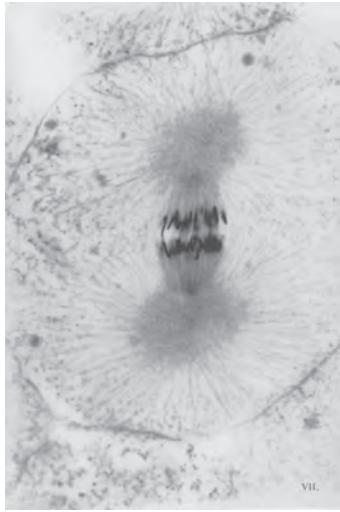
(a)



(b)



(c)



(d)

In this second part of the book we follow the flow of central dogma of molecular biology by moving from DNA (Chapters 8 and 9) to RNA (Chapters 10 and 11) to protein (Chapters 12 and 13). We then discuss functional genomics (Chapter 14), which is the genome-wide study of the function of genes and gene products. We focus on next-generation sequencing in our studies of DNA and RNA. This builds on Part I, in which we described how to find sequences (and other information) in databases, how to align DNA or protein sequences in a pairwise fashion or in a multiple sequence alignment, and how to perform evolutionary studies through molecular phylogeny.

Through the first half of the twentieth century, Charles Darlington performed brilliant studies of the chromosomes. (a) First PG mitosis in *Paris quadrifolia*, Liliaceae, showing all stages from prophase to telophase;  $n = 10$ ; 800 $\times$  magnification. (b) First PG mitosis in polar view. *Tradescantia virginiana*, Commelinaceae;  $n = 9$  (from aberrant plant with 22 chromosomes). 1200 $\times$  magnification. (c) Root tip squashes showing anaphase separation in *Fritillaria pudica*, 3 $\times$  = 39. Note the spiral structure of chromatids (daughter chromosomes). 3000 $\times$  magnification. (d) Cleavage mitosis in the morula of the teleostean fish, *Coregonus clupeoides*, in the middle of anaphase. Spindle structure revealed by slow fixation. 4000 $\times$  magnification.

Source: Darlington (1932). Reproduced with permission from Taylor & Francis.

# DNA: The Eukaryotic Chromosome

# CHAPTER 8

*Science is about building causal relations between natural phenomena (for instance, between a mutation in a gene and a disease). The development of instruments to increase our capacity to observe natural phenomena has, therefore, played a crucial role in the development of science – the microscope being the paradigmatic example in biology. With the human genome, the natural world takes an unprecedented turn: it is better described as a sequence of symbols. Besides high-throughput machines such as sequencers and DNA chip readers, the computer and the associated software becomes the instrument to observe it, and the discipline of bioinformatics flourishes. However, as the separation between us (the observers) and the phenomena observed increases (from organism to cell to genome, for instance), instruments may capture phenomena only indirectly, through the footprints they leave. Instruments therefore need to be calibrated: the distance between the reality and the observation (through the instrument) needs to be accounted for. [We are] calibrating instruments to observe gene sequences; more specifically, computer programs to identify human genes in the sequence of the human genome.*

—Martin Reese and Roderic Guigó (2006, p. S1.1), introducing EGASP, the Encyclopedia of DNA Elements (ENCODE) Genome Annotation Assessment Project

*Inasmuch as the only requirement to be qualified as partitioning sequences [i.e., intergenic DNA including pseudogenes] is to be untranscribable and/or untranslatable, it is not likely that these sequences came into being as a result of positive selection. Our view is that they are the remains of nature's experiments which failed. The earth is strewn with fossil remains of extinct species; is it a wonder that our genome too is filled with the remains of extinct genes?*

—Susumu Ohno, *So Much “Junk” DNA in our Genome* (1972, p. 368)

## LEARNING OBJECTIVES

Upon reading this chapter you should be able to:

- define features of eukaryotic genomes such as the C value;
- define five major types of repetitive DNA and bioinformatics resources to study them;
- describe eukaryotic genes;
- explain several categories of regulatory regions;
- use bioinformatics tools to compare eukaryotic DNA;
- define single-nucleotide polymorphisms (SNPs) and analyze SNP data; and
- compare and contrast methods to measure chromosomal change.

## INTRODUCTION

Synonyms of eukaryotes include eucaryotae, eucarya, eukarya, and eukaryotae. The word derives from the Greek *eu-* ("true") and *karutos* ("having nuts"; this refers to the nucleus). There is currently a debate as to whether the word prokaryote should no longer be used. In its place we refer to bacteria and archaea. See the discussion in Chapter 15.

Eukaryotes are single-celled or multicellular organisms that are characterized by the presence of a membrane-bound nucleus and a cytoskeleton. Genomic DNA is organized into chromosomes, a topic we explore in this chapter from a bioinformatics perspective. Later we examine specific eukaryotic genomes beginning with fungi (Chapter 18), including *Saccharomyces cerevisiae*. We then broadly survey eukaryotes (Chapter 19), from the simplest primitive single-celled organisms to plants and metazoans (animals).

At the start of Chapter 15 we address five basic perspectives on the field of genomics. With respect to the topic of eukaryotic chromosomes, these five perspectives are as follows.

- *Perspective 1: Catalog genomic information.* We examine genome sizes, noncoding DNA (e.g., repetitive DNA), and coding DNA (genes). For a given segment of genomic DNA, we address the problem of annotation: how much repetitive DNA is present and of what type? How many protein-coding genes or RNA genes are present?
- *Perspective 2: Catalog comparative genomic information.* How can comparative genomics help us to understand chromosomal rearrangements that have occurred over time?
- *Perspective 3: Biological principles.* What are the mechanisms underlying chromosomal functions and chromosomal variations such as duplications, inversions, and translocations? More broadly, as we examine genomic DNA, we want to address the molecular basis of how species evolve.
- *Perspective 4: Human disease relevance.* In what ways are chromosomal variants associated with disease?
- *Perspective 5: Bioinformatics aspects.* What tools are available to understand chromosomes, from genome browsers to gene-finding algorithms?

In Chapter 15 we suggest that students in genomics classes complete a project in which they select either one genome of a favorite organism or one gene and analyze it according to these five principles. These topics are consistent with a vision for the future of genomics research that was outlined by Eric Green and his colleagues at the National Human Genome Research Institute (Green and Guyer, 2011). Their five sequential, overlapping domains of genomics research are: (1) understanding the structure of genomes; (2) understanding the biology of genomes; (3) understanding the biology of disease; (4) advancing the science of medicine; and (5) improving the effectiveness of healthcare. The material in this chapter is essential for the understanding of the structure and biology of genomes from which further advances in our understanding of disease may be made.

### Major Differences between Eukaryotes and Bacteria and Archaea

Eukaryotes share a common ancestry with bacteria and archaea but, when we compare them, we find several outstanding differences (Vellai and Vida, 1999; Watt and Dean, 2000; Cavalier-Smith, 2002; Katz, 2012). Some of these genomic features are highlighted in **Table 8.1**.

- There is a tremendous diversity of bacterial, archaeal, and eukaryotic life forms. Very few bacterial or archaeal life forms are visible to the human eye and indeed many eukaryotes are also single-celled, microscopic organisms. Most life forms that we can see are multicellular eukaryotes (e.g., plants and metazoans).

Learn more about the NHGRI strategic plan for genomics at <http://www.genome.gov/sp2011/> (WebLink 8.1).

**TABLE 8.1 Features of several sequenced bacterial and eukaryotic genomes.**  
**Adapted from Gardner *et al.* (2002), Blattner *et al.* (1997), International Human Genome Sequencing Consortium (2001, 2004), and <http://www.ensembl.org/>.**

Feature	E. coli K-12	Parasite <sup>a</sup>	Yeast <sup>b</sup>	Slime Mold <sup>c</sup>	Plant <sup>d</sup>	Human <sup>e</sup>
Genome size (Mb)	4.64	22.8	12.5	8.1	115	3324
GC content (%)	50.8	19.4	38.3	22.2	34.9	41
Number of coding genes	4288	5268	5770	2799	25,498	20,774
Gene density (kb per gene)	0.95	4.34	2.09	2.60	4.53	27
Percent coding	87.8	52.6	70.5	56.3	28.8	1.3
Number of introns	0	7406	272	3578	107,784	53,295
Repeat (%)	<1	<1	2.4	<1	14	46

<sup>a</sup>*Plasmodium falciparum*; <sup>b</sup>*Saccharomyces cerevisiae*; <sup>c</sup>*Dictyostelium discoideum*; <sup>d</sup>*Arabidopsis thaliana*;

<sup>e</sup>*Homo sapiens*.

- Eukaryotic cells have three cellular features that are lacking in bacteria and archaea: (1) a membrane-bound nucleus; (2) an extensive system of organelles bound by intracellular membranes; and (3) a cytoskeleton, including elements such as actin and tubulin, and molecular motors. Notably, bacteria and archaea lack energy-producing organelles and are incapable of endocytosis, the process by which extracellular cargo is internalized (Vellai and Vida, 1999).
- Most eukaryotes undergo sexual reproduction, although some (such as Bdelloid rotifers) are asexual. Bacteria lack gamete fusion and do not exchange DNA by sex.
- The genome size of eukaryotes varies widely, spanning five orders of magnitude (**Table 8.2**). In contrast, most archaeal and bacterial genomes are between about 0.2 and 13 Mb (million base pairs or megabases) in size (see Chapters 15 and 17).
- Bacterial and archaeal genomes tend to have a relatively high density of protein-coding genes and little repetitive or other noncoding DNA. For example, 0.7% of the *Escherichia coli* genome consists of noncoding repeats (Blattner *et al.*, 1997). In contrast, many eukaryotic genomes include large tracts of noncoding DNA. Several examples are provided in **Table 8.1**.
- Bacteria and archaea are haploid, that is, the organism has one set of chromosomes. Eukaryotes may be haploid or diploid (2x; having two sets of chromosomes) or have other ploidy states (such as triploid; 3x). This higher level of ploidy offers eukaryotes a variety of evolutionary mechanisms such as heterozygous advantage (Watt and Dean, 2000).
- The genomes are organized differently. The majority of bacterial and archaeal genomes are organized in circular chromosomes, often with small accompanying plasmids (see **Fig. 17.1**). Eukaryotic nuclear genomes are organized primarily into linear chromosomes. These eukaryotic chromosomes are typically numerous (ranging from a few to over one hundred) and each has a centromere (defined below) as well as telomeres at either end. These features are absent from bacterial and archaeal chromosomes, although centromere-like elements have been described (Hazan and Ben-Yehuda, 2006). The mechanisms by which bacteria segregate DNA are relatively obscure.

Sexual reproduction is called syngamy, the process by which the haploid chromosomes of the male and female gametes combine to form the zygote (i.e., the fertilized ovum).

**TABLE 8.2** Genome size of selected phyla or classes eukaryotes. Note that 0.001 Gb (gigabases) equals 1 Mb. Values in picograms were multiplied by  $0.9869 \times 10^9$  to obtain Gb. Adapted from Graur and Li (2000) with permission from Sinauer Associates, Animal Genome Size Database of T. R. Gregory (<http://www.genomesize.com>) and the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>).

Taxon	Phylum, class, or division	Genome size range (Gb)	Ratio of genome sizes (highest/lowest)
All eukaryotes	—	0.003–686	228,667
Alveolata	—	—	22,333
	Apicomplexians	0.009–201	22,333
	Ciliates	0.024–8.62	359
	Dinoflagellates	1.37–98	72
Diatoms		0.035–24.5	700
Amoebae		0.035–686	19,600
Euglenozoa		0.098–2.35	24
Fungi/microsporidia		0.003–1.47	490
Animals	—	—	3,325
	Sponges	0.059–1.78	30
	Cnidarians	0.227–1.83	8
	Insects	0.089–9.47	106
	Elasmobranchs	1.47–15.8	11
	Bony fishes	0.345–133	386
	Amphibians	0.93–84.3	91
	Reptiles	1.23–5.34	4
	Birds	1.67–2.25	1
	Mammals	1.7–6.7	4
	Placozoa	0.04	—
Plants	—	—	6,140
	Algae	0.080–30	375
	Pteridophytes	0.098–307	3,133
	Gymnosperms	4.12–76.9	19
	Angiosperms	0.050–125	2,500

## GENERAL FEATURES OF EUKARYOTIC GENOMES AND CHROMOSOMES

### C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly

The C value is measured in base pairs or in picograms (pg) of DNA. One picogram of DNA corresponds to approximately 1 Gb.

In eukaryotic genomes, the haploid genome size (C value) varies enormously. This is shown in Table 8.2 for various taxa of eukaryotes, and in Table 8.3 for specific eukaryotic species. Some genomes are relatively quite small, such as the microsporidian *Encephalitozoon cuniculi* (2.9 Mb; Chapter 18). Others have genome sizes in the range of hundreds of billions of base pairs. Tremendous variation in C values occurs among the unicellular protists such as amoebae, with a 20,000-fold range. Within the animal kingdom, the range is about 3000-fold.

**TABLE 8.3** Genome size (*C* value) for various eukaryotic species. Adapted from Graur and Li (2000) with permission from Sinauer Associates; NCBI (<http://www.ncbi.nlm.nih.gov>); Cameron *et al.* (2000); and the Database of Genome Sizes (<http://www.cbs.dtu.dk/databases/DOGS/>).

Species	Common name	<i>C</i> value (Gb)
<i>Saccharomyces cerevisiae</i>	Yeast	0.012
<i>Neurospora crassa</i>	Fungus	0.043
<i>Dysidea crawlshagi</i>	Sponge	0.054
<i>Caenorhabditis elegans</i>	Nematode	0.097
<i>Drosophila melanogaster</i>	Fruit fly	0.12
<i>Paramecium aurelia</i>	Ciliate	0.19
<i>Oryza sativa</i>	Rice	0.47
<i>Strongylocentrotus purpuratus</i>	Sea urchin	0.80
<i>Gallus domesticus</i>	Chicken	1.23
<i>Erysiphe cichoracearum</i>	Powdery mildew	1.5
<i>Boa constrictor</i>	Snake	2.1
<i>Parascaris equorum</i>	Roundworm	2.5
<i>Carcharias obscurus</i>	Sand-tiger shark	2.7
<i>Canis familiaris</i>	Dog	2.9
<i>Rattus norvegicus</i>	Rat	2.9
<i>Xenopus laevis</i>	African clawed frog	3.1
<b><i>Homo sapiens</i></b>	<b>Human</b>	<b>3.3</b>
<i>Nicotiana tabacum</i>	Tobacco plant	3.8
<i>Locusta migratoria</i>	Migratory locust	6.6
<i>Paramecium caudatum</i>	Ciliate	8.6
<i>Allium cepa</i>	Onion	15
<i>Truturus cristatus</i>	Warty newt	19
<i>Thuja occidentalis</i>	Western giant cedar	19
<i>Coscinodiscus asteromphalus</i>	Centric diatom	25
<i>Lilium formosanum</i>	Lily	36
<i>Amphiuma means</i>	Two-toed salamander	84
<i>Pinus resinosa</i>	Canadian red pine	68
<i>Protopterus aethiopicus</i>	Marbled lungfish	140
<i>Amoeba proteus</i>	Amoeba	290
<i>Amoeba dubia</i>	Amoeba	690

Remarkably, the range in *C* values does not correlate well with the complexity of organisms. This is true for different species, such as an onion (*Allium cepa*) compared to a human. It is also true for species that are phenotypically very similar. Some organisms such as *Fugu rubripes* (a pufferfish with a 400 megabase genome) have extremely compact genomes while closely related organisms of similar biological complexity have

An online database of plant *C* values is available at <http://data.kew.org/cvalues/> (WebLink 8.2). Currently (February 2015) it lists data for over 8500 species. The Animal Genome Size Database (from T. Ryan Gregory) is online at <http://www.genomesize.com/> (WebLink 8.3).

genomes that are orders of magnitude larger (for example, the lungfish *Protopterus aethiopicus* genome size is ~130,000 megabases). This lack of correlation is called the *C* value paradox (Hartl, 2000; Hancock, 2002; Kidwell, 2002; Knight, 2002).

The genomes of many eukaryotes were sequenced over a decade ago, including *Caenorhabditis elegans* (1998), *Drosophila melanogaster* (2000), *Homo sapiens* (2001), and *Mus musculus* (2002) (see Chapters 15 and 19). These whole-genome studies provide one clear answer to the *C* value paradox: genomes are filled with large tracts of noncoding DNA sequences in varying amounts, which accounts for the variation in genome size. A major hypothesis is that this “extra” DNA, first called “junk” DNA by Susumu Ohno (1972), has little or no adaptive advantage for the organism (or for the species). Of course, some intergenic DNA contains elements having critical roles such as regulating gene expression. One way to assess the functional importance of large amounts of genomic DNA is to determine the extent to which DNA loci are under selective constraint; approximately 10% of the human genome has been estimated to be selective pressure to be conserved across species.

This explanation of the *C* value paradox has been reasonably well accepted in recent decades (Eddy, 2012), until it was challenged by the ENCODE project in 2012. Below we discuss the ENCODE explanation and compare it to the traditional model.

### Organization of Eukaryotic Genomes into Chromosomes

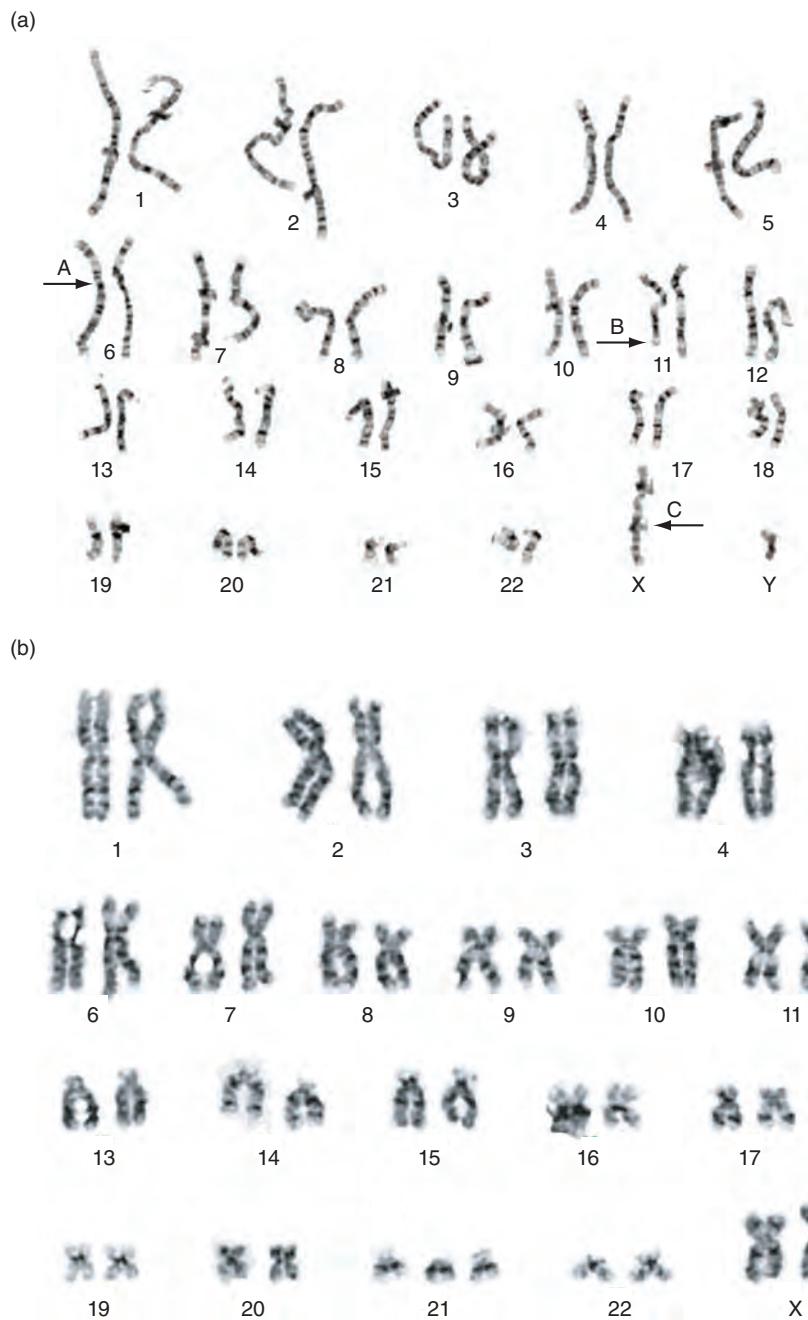
Genomic DNA is organized in chromosomes. Originally, chromosomes were defined morphologically as the bodies into which the nucleus resolves itself at the beginning of mitosis and from which it is derived at the end of mitosis (Waldeyer, 1888; Darlington, 1932). It was clear by the 1880s that the nucleus is the cellular organelle that directs the cell division process, and that mitosis occurs in both plants and animals (Lima-de-Faria, 2003). Visualizing chromosomes cytogenetically was challenging. Reports from the 1920s that there are 48 human chromosomes were not corrected until Joe Hin Tjio and Albert Levan (1956) reported that the diploid number of chromosomes is 46, that is, there are 23 pairs of human chromosomes.

Chromosomes are often studied at metaphase, when they are thickest and most condensed. For human studies, a sample is typically collected from blood cells or amniotic fluid. Chromosomes are most often visualized using dyes or using specific DNA probes by fluorescence *in situ* hybridization (FISH).

As we explore a variety of eukaryotic genomes that have been completely sequenced, it is helpful to describe the structure and content of chromosomes. We refer to a karyotype of human metaphase chromosomes visualized with Wright’s stain (Fig. 8.1a). A variety of stains produce banding patterns on chromosomes. These include Q bands (based on stains using quinacrine mustard or derivatives) and G bands (based on the Giemsa dye; Wright’s stain is an example of such a dye). These dyes stain the entire length of each chromosome and produce a characteristic banding pattern. A band is defined as a portion of a chromosome that is distinguishable from adjacent segments by appearing lighter or darker.

There are several major features of eukaryotic chromosomes. The most apparent landmarks are the two telomeres (the chromosome ends) and the centromere. Telomeres are structures characterized by tandem arrays of repetitive sequences found at the chromosome ends. They provide stability to chromosomes by preventing the degradation of the chromosome end and by blocking the fusion of chromosome ends. The centromere, a region that remains unstained with many dyes, appears as a constriction. Centromeres may be metacentric (located near the middle of the chromosome) or acrocentric (located close to a telomere). In humans, the five acrocentric chromosomes are 13, 14, 15, 21, and 22. In some species such as the mouse (*Mus musculus*), all chromosomes are acrocentric.

The human autosomes consist of chromosomes 1–22, while X and Y are the sex chromosomes. In the particular karyotype shown in Figure 8.1a there is a hemizygous deletion of the terminus of chromosome 11q. In a euploid (apparently normal) individual



**FIGURE 8.1** Example of human karyotypes. (a) The chromosomes are visualized with Wright's stain. Centromeres are visible as an indentation in the chromosome (e.g., see arrows A and C). This karyotype is of a person with a hemizygous deletion of a telomeric portion of chromosome 11q, resulting in a loss of several million base pairs of DNA (arrow B). (b) Karyotype of a female with trisomy 21 (Down syndrome). Note that there are three copies of chromosome 21.

there are two copies of each autosome per nucleus; in a hemizygous deletion there is only one copy; and in a homozygous deletion there are zero copies. Using conventional karyotyping, deletions or duplications as small as several million base pairs can be observed by inspection of the banding patterns. **Figure 8.1b** shows a trisomy of chromosome 21 in which the entire chromosome (~48 Mb) is present in three copies.

Deletion 11q syndrome results in trigonencephaly (a triangle-shaped head), a carp-shaped mouth, and cardiac defects (Jones, 1997).

An ideogram is a diagram of a karyotype. A karyotype is an image (often a photograph) of the chromosomes from a cell during metaphase, when each chromosome is a pair of sister chromatids. Karyotypes display the chromosomes in numerical order, with the short arm (p arm) oriented upward. For humans, the short arm is called "p" for *petit* (French for "small"), while the q arm (long arm) is named as the letter following p. Online databases of chromosomes (and karyotypes) include the Ensembl genome browser (<http://www.ensembl.org/>, WebLink 8.4), the Ideogram Album (<http://www.pathology.washington.edu/research/cytopages/>, WebLink 8.5), and KaryotypeDB (<http://www.nenno.it/karyotypedb/>, WebLink 8.6).

The Map Viewer is available at <http://www.ncbi.nlm.nih.gov/mapview/> (WebLink 8.7). dbVar is at <http://www.ncbi.nlm.nih.gov/dbvar> (WebLink 8.8); enter a query for HBB and when the browser opens click the wheel-shaped Configure link to select from a vast number of annotation options. We previously introduced Genome Workbench (<http://www.ncbi.nlm.nih.gov/tools/gbench/>, WebLink 8.9) in Chapter 2, and we show how to view next-generation sequence reads by uploading BAM files to that viewer in Chapter 9.

Ensembl (<http://www.ensembl.org/>, WebLink 8.10) is a joint project between EMBL-EBI and the Sanger Institute.

Inside a nucleus, chromosomes tend to be unraveled structures that occupy restricted spaces called chromosome territories. Meaburn and Misteli (2007) provide an overview of the spatial organization of chromosomes and genomes, including visualization with chromosome-specific fluorescent probes. Trask (2002), Speicher and Carter (2005), Dolan (2011), and South (2011) have written overviews of the field of human cytogenetics.

## Analysis of Chromosomes Using Genome Browsers

The diploid number of chromosomes is constant in each species, although there may be individual variation. We explore the 16 *S. cerevisiae* chromosomes in Chapter 18, including a variety of databases such as NCBI, MIPS, and SGD that provide graphic displays. In humans, the diploid number is 46 (i.e., there are 23 pairs of chromosomes in somatic cells). Ideograms of karyotypes for some other organisms are available online.

DNA databases (Chapter 2) store billions or trillions of base pairs of DNA from various organisms. For a particular organism of interest, whether a fungus, plant or animal, genome browsers represent an essential tool to store, centralize, process, and display both raw sequence data and analyses based on annotation of the data. Annotation consists of adding information about features such as the experimentally determined or computationally predicted repetitive elements or genes or sites of variation.

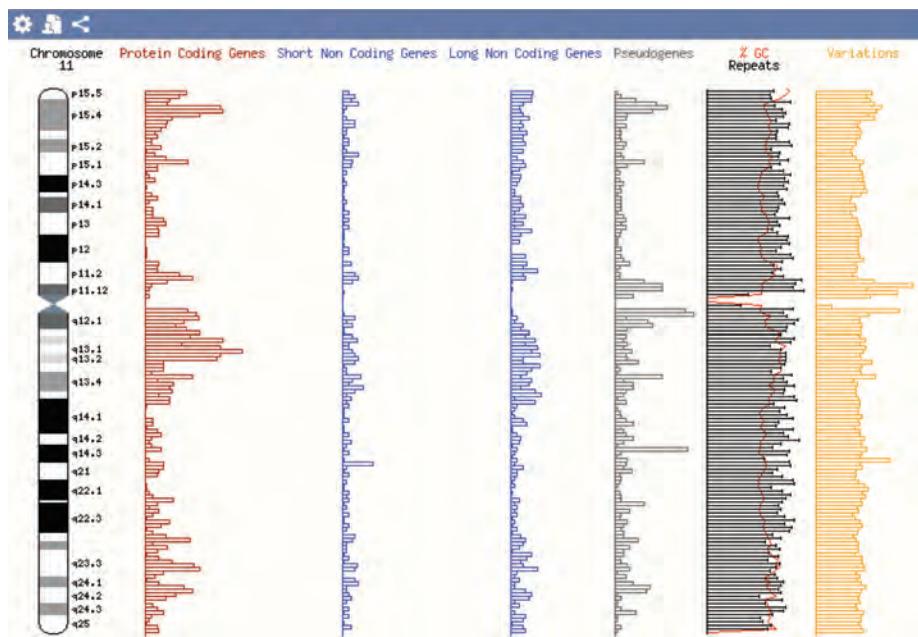
There are several genome browsers that provide broad and deep coverage of a variety of eukaryotic genomes, as listed below; we focus on Ensembl and UCSC resources. We introduce many other genome browsers for other organisms (e.g., protozoans, plants, or fungi) in Part III of this book.

1. NCBI offers a Map Viewer for dozens of species (Wolfsberg, 2011). It includes an ideogram, tracks that can be added or removed, and links for each gene to a variety of database resources. Additional browsers include one at the database of genomic structural variation (dbVar) and Genome Workbench.
2. The Ensembl project offers a map viewer filled with annotation data (Flicek *et al.*, 2014). A view of human chromosome 11 includes summaries of the genomic features such as GC content, single-nucleotide polymorphisms (SNPs), and coding and noncoding gene content (Fig. 8.2a). A link to the location-based viewer (Fig. 8.2.b) provides access to hundreds of additional tracks of features that can be viewed or downloaded for detailed chromosomal analyses.
3. The UCSC Genome Browser includes a gateway to select a genome and chromosomal region of interest (Kuhn *et al.*, 2013; Meyer *et al.*, 2013). The main genome browser page depicts the chromosome of interest along with a series of user-selected annotation tracks.
4. The vertebrate genome annotation (Vega) database offers a genome browser reflecting high-quality manual annotation of selected vertebrate genomes.

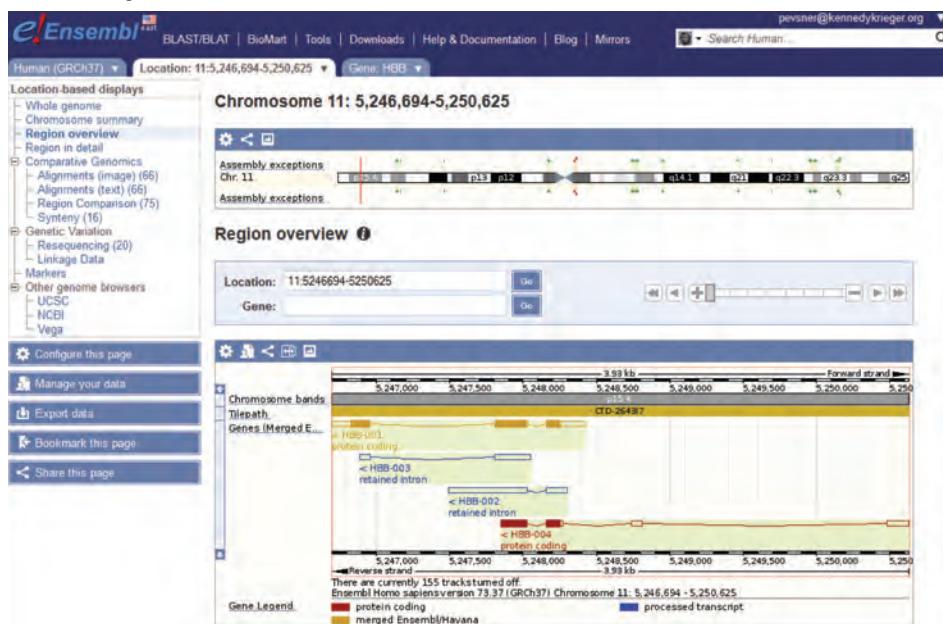
## Analysis of Chromosomes Using BioMart and biomaRt

Data on chromosomes can also be obtained in tabular form. One resource is the UCSC Table Browser which provides access to the tables underlying the UCSC Genome Browser. Another resource is BioMart, a web service accessed through the home page of Ensembl. You can use BioMart to answer thousands of queries. For example, given five globin genes in the form of HGNC symbols (e.g., *HBB*, *HBD*, *HBE1*, *HBG1*, *MB*), what is their GC content, and what are the associated Protein Data Bank (PDB) identifiers? Go to the BioMart home page. (1) Choose a database (we select Ensembl Genes 73). (2) Choose a dataset (we select *Homo Sapiens* genes, GRCh37.p12 for patch 12 of the current GRC human genome assembly). (3) Choose filters. This specifies the input information. In the Gene section, choose "ID list limit" and enter the five globin gene symbols (Fig. 8.3a). You can also upload a text file containing entries of interest. Note that in the pulldown

(a) Ensembl: chromosome summary



(b) Ensembl: Region overview



**FIGURE 8.2** View of human chromosome 11 using Ensembl (release 73, September 2013). This is one of the key genome browsers (with UCSC and NCBI) and offers an exceptionally wide range of viewing and analysis options. (a) Summary of chromosome 11. The gear-shaped “configure” icon (upper left corner) can be clicked to change settings (e.g., adding or reformatting datasets). (b) The horizontal ideogram (upper portion) shows chromosome 11, with a red bar indicating the location of *HBB* (the gene that was searched for in this instance). The region overview (lower right graphic) includes *HBB* exons, and notes that 155 tracks are currently turned off. These can be accessed with the gear-shaped link. The left sidebar includes a large number of additional display options. Data can be imported or exported, the page can be configured with a rich set of options, and there are links to other browsers.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

The UCSC Genome Bioinformatics site is <http://genome.ucsc.edu/> (WebLink 8.11). We explore it further in this chapter, and have seen examples of the Genome and Table Browser and BLAT (Figs. 2.12, 5.14, 5.16, 6.10).

Vega is available at <http://vega.sanger.ac.uk/> (WebLink 8.12).

(a) BioMart at Ensembl: specify filters (input for which you want to apply queries)

(b) BioMart output

Ensembl Gene ID	% GC content	HGNC symbol	PDB ID
ENSG00000198125	50.32	MB	
ENSG00000198125	50.32	MB	3RGK
ENSG00000244734	37.64	HBB	3W4U
ENSG00000244734	37.64	HBB	1A00
ENSG00000244734	37.64	HBB	1A01
ENSG00000244734	37.64	HBB	1A0U
ENSG00000244734	37.64	HBB	1A0Z
ENSG00000244734	37.64	HBB	1A3N
ENSG00000244734	37.64	HBB	1A3Q
ENSG00000244734	37.64	HBB	1ABW

**FIGURE 8.3** Using BioMart service at Ensembl. (a) The user selects a dataset (e.g., human genes; other options include variation and regulation databases as well as Vega high-quality annotations), filters (including many options from genes to microarray elements to chromosomal regions), and attributes (of which thousands are available). (b) The result may be sent to a spreadsheet. BioMart may also be accessed within Galaxy using the Get Data tool.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

GRC refers to the Genome Research Consortium (see Chapters 15 and 20).

menu you must specify HGNC symbols. (4) Choose attributes. These are the features we wish to search for. Under Gene > Ensembl choose Ensembl Gene ID and % GC content. Under External > External References select HGNC gene symbol and PDB ID. (5) Click count; as expected, results are available for all five genes. Click the results, and the data we requested can be viewed (Fig. 8.3b) or downloaded in several tabular formats.

The R package `biomaRt` also provides powerful access to BioMart. We demonstrate its versatility in the following examples.

### Example 1

Given a set of NCBI gene identifiers, we can find the official (HGNC) gene symbols and the GC content for five globins (Fig. 8.4). We first install R and RStudio, and then specify the package we wish to use.

**getBM function:**  
 --used to perform a query  
 --has four main arguments  
 (attributes, filters, values, mart)  
 --returns a data.frame

Attributes: a vector specifying the output you request

```
> ens_att <- listAttributes(ensembl)
> ens_att[1:10,]
  name
1   ensembl_gene_id
2   ensembl_transcript_id
3   ensembl_peptide_id
4   ensembl_exon_id
5   description
6   chromosome_name
7   start_position
8   end_position
9   strand
10  band
# currently the full list has 1,720
# attributes you can choose from!
```

```
> mydata = getBM(attributes=c("entrezgene", "hgnc_symbol",
  "percentage_gc_content"), filters="entrezgene",
  values=myentrez, mart=ensembl)
```

	entrezgene	hgnc_symbol	percentage_gc_content
1	3043	HBB	37.64
2	3045	HBD	37.91
3	3046	HBE1	38.96
4	3047	HBG1	45.86
5	4151	MB	50.32

Values refers to vector of values for the filters

```
> myentrez = c("3043", "3045",
  "3046", "3047", "4151")
> myentrez
[1] "3043" "3045" "3046" "3047" "4151"
```

**mart**  
 --object of the class Mart  
 --to invoke (e.g. for mouse):  

```
> UseMart
> mouse=useMart("ensembl",
  dataset="mmusculus_gene_ensembl")
```

Filters: a vector that defines a restriction on your query. To see your options:  

```
>filters = listFilters(ensembl)
>filters[1:10,]
  1   chromosome_name
  2   start
  3   end
  4   band_start
  5   band_end
  6   marker_start
  7   marker_end
  8   type
  9   encode_region
10  strand
# Currently ~350 filters!
```

It is a surprisingly common error to analyze genes using obsolete or incorrect gene symbols.

When we choose gene symbols such as *HBG1* and *MB*, we can confirm that they are official HGNC symbols by entering them individually into <http://www.genenames.org> (WebLink 8.13). Alternatively, we can add query them in BioMart or `biomaRt` as described in this section, even analyzing a text file with thousands of gene symbols to confirm that all are indeed correct.

**FIGURE 8.4** Using the R package `biomaRt` to obtain information about a chromosome. This figure shows two central, boxed panels. The top one defines the output filename (we call it `mydata`) and the command `getBM`. We supply attributes (a vector of information we request), filters (restrictions on the query), values (defining the query) and mart (the database we wish to query). The next boxed panel shows the result, given by typing `mydata` (then pressing enter). `biomaRt` is extremely versatile, allowing many thousands of queries across many species and databases. Note that a command such as `filters` (followed by pressing the enter key) displays that particular file, while `filters[1:10, ]` specifies rows, columns (within the brackets); here rows 1:10 are displayed, and since no column value is entered, by default all columns are displayed.

Source: R Foundation, from <http://www.r-project.org>.

To begin, visit the `biomaRt` page at <http://www.bioconductor.org> (WebLink 8.14). It includes instructions for downloading `biomaRt` as well as vignettes. You can also install packages using RStudio. The examples below are adapted from the `biomaRt` website. The R code below is available as a text file (so you can paste the commands into your own R session) at Web Document 8.1 at <http://bioinfbook.org/chapter8>.

To perform a query we'll use the `getBM` function. This requires four main arguments that we assemble: (1) attributes (the output of a query that we want to produce); (2) filters (a vector of filters that restrict the input, such as searching only on a particular chromosome or database or organism); (3) values which apply to the filters, such as a vector of gene identifiers; and (4) mart, which is the mart we'll create with `useMart` (e.g., searching Ensembl or Vega). The R commands below start with a prompt in red (`>`); commands are in blue; and comments are in green preceded by a hash mark.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("biomaRt")
> library("biomaRt")
# We need to choose a BioMart database.
> listMarts()
# Choices include ensembl, vega, unimart, or many others.
> ensembl <- useMart("ensembl")
> listDatasets(ensembl)
# We can browse the datasets and select human
> ensembl = useDataset("hsapiens_gene_ensembl", mart=ensembl)
```

We need to select a filter (or filters) to restrict our query to some area of interest such as a chromosome, a region, or a list of identifiers. First we look at available filters.

```
> filters = listFilters(ensembl)
# Look at the first seven rows of filters,
# then at the last few rows with the tail function.
> filters[1:7,]
   name           description
1  chromosome_name Chromosome name
2      start        Gene Start (bp)
3       end        Gene End (bp)
4  band_start      Band Start
5  band_end        Band End
6 marker_start     Marker Start
7 marker_end      Marker End
> tail(filters)
   name           description
296 with_transmembrane_domain Transmembrane domains
297 with_signal_domain        Signal domains
298 germ_line_variation_source limit to genes with germline variation
                                 data sources
299 somatic_variation_source  limit to genes with somatic variation
                                 data sources
300 with_validated_snp        Associated with validated SNPs
301 so_parent_name            Parent term name
```

There are therefore about 300 filters to choose from. We next review the available attributes for our output:

```
> attributes = listAttributes(ensembl)
> attributes[1:5,]
   name           description
1  ensembl_gene_id Ensembl Gene ID
2  ensembl_transcript_id Ensembl Transcript ID
3  ensembl_peptide_id Ensembl Protein ID
4  ensembl_exon_id Ensembl Exon ID
5  description      Description
> tail(attributes)
   name           description
1144 phase          phase
1145 cdna_coding_start cdNA coding start
1146 cdna_coding_end  cdNA coding end
1147 genomic_coding_start Genomic coding start
1148 genomic_coding_end  Genomic coding end
1149 is_constitutive Constitutive Exon
```

The values for the analysis in **Figure 8.4** are obtained by creating a file with a list of Entrez gene identifiers. We are then ready to perform the search:

```
> mydata = getBM(attributes=c("entrezgene", "hgnc_symbol",
"percentage_gc_content"), filters="entrezgene", values=myentrez,
mart=ensembl)
```

This script will create a file called `mydata` (you can assign it any name). The `getBM` function performs the query. The attributes specify what output you seek; currently, for the human Ensembl dataset we use in this example, there are 1720 different types of information you can request. The values for our query in this particular example are the NCBI Entrez identifiers for our five genes.

### *Example 2*

What are the HGNC gene symbols for genes on human chromosome 21?

```
> chrom=21
# You could use chrom=c(21,22) to specify two chromosomes
> getBM(attributes="hgnc_symbol", filters="chromosome_name",
values=chrom, mart=ensembl)
  hgnc_symbol
1   MIR548X
2   PPIAP22
3   SLC6A6P1
# We truncate this output of HGNC symbols from chromosome 21.
```

### *Example 3*

What Ensembl genes are in a 100,000 base pair region of chromosome 11 surrounding *HBB*? What chromosome band are they on, what strand, and what type of genes are they?

```
> getBM(c("hgnc_symbol", "band", "strand", "gene_biotype"),
filters=c("chromosome_name", "start", "end"),
values=list(11,5200000,5300000), mart=ensembl)
  hgnc_symbol    band      strand   gene_biotype
1                p15.4       1      antisense
2                p15.4      -1      misc_RNA
3   HBBP1        p15.4      -1     pseudogene
4                p15.4      -1 sense_overlapping
5   OR52A1        p15.4      -1 protein_coding
6   OR51V1        p15.4      -1 protein_coding
7   HBB          p15.4      -1 protein_coding
8   HBD           p15.4      -1 protein_coding
9   HBG1          p15.4      -1 protein_coding
10  HBG2          p15.4      -1 protein_coding
11  HBE1          p15.4      -1 protein_coding
```

Note that we can expand the attributes (e.g., adding “`start_position`”, “`end_position`” after “`band`”) for more information.

### *Example 4*

What are the rat homologs of the genes in a 100 kilobase region of human chromosome 11?

```
> getBM(c("rnorvegicus_homolog_ensembl_gene"),
filters=c("chromosome_name", "start", "end"),
values=list(11,5200000,5300000), mart=ensembl)
[1] "ENSRNOG00000299978" "ENSRNOG00000015940"
"ENSRNOG0000049424" "ENSRNOG0000047098"
[5] "ENSRNOG0000048955" "ENSRNOG00000031230"
"ENSRNOG0000048992" "ENSRNOG0000030879"
[9] "ENSRNOG0000030784" "ENSRNOG00000029286"
```

***Example 5***

What are the paralogs of the genes in a 50 kb region of human chromosome 11? Since this region includes beta globin genes, we might expect the result to include alpha globin gene loci on chromosome 16.

```
> getBM(attributes=c("hsapiens_paralog_chromosome",
+ "hsapiens_paralog_chrom_start","hsapiens_paralog_chrom_end"),
  filters=c("chromosome_name","start","end"),
  values=list(11,5250000,5300000), mart=ensembl)
      hs_paralog_chromosome hs_paralog_chrom_start hs_paralog_chrom_end
1           NA             NA                 NA
2            16            202686            204502
3            16            222846            223709
4            16            230452            231180
5            16            226679            227521
6            16            203891            216767
7            11            5253908            5256600
8            11            5289582            5526847
9            11            5274420            5667019
10           11            5269313            5271122
11           11            5246694            5250625
# The + sign indicates a line break in the R code
# For clarity the column titles hsapiens... are truncated to hs...
```

For some queries, using BioMart (or other web-based software) is adequate. In many other cases, especially when we have a list of many genes (or any other features) of interest, it is easier to use a program such as `biomaRt`. By using a script, your search is less likely to be error-prone, your results are saved to a file, your methods are more likely to be reproducible, and your data can easily be plotted in R (e.g., computer lab exercises 8.3, 8.4 and 8.11).

### Analysis of Chromosomes by the ENCODE Project

An initial version of the human genome sequence was reported by a public consortium (International Human Genome Sequencing Consortium, 2001) and by Venter *et al.* (2001). It was clear that the annotation of the functional elements embedded in the genomic DNA is extraordinarily complex. The Encyclopedia of DNA Elements (ENCODE) project was initiated to investigate the properties of the human and other genomes (ENCODE Project Consortium, 2004). The ENCODE Project Consortium *et al.* (2007) released its findings on 1% of the genome in a paper with over 250 coauthors. This represented the generation of over 200 datasets by 35 groups. A total of 44 regions of the human genome were selected, spanning 30 megabases. In the production phase, the ENCODE Project Consortium *et al.* (2012) reported its findings across the entire human genome; a set of 30 papers were published at the same time.

We next describe the scope of the ENCODE project, their main conclusions, and where you can find, analyze, and explore ENCODE data and literature.

1. *Scope of the ENCODE project.* The goal of the ENCODE project is to generate a comprehensive catalog of all functional elements in the human genome (as well as the genomes of model organisms such as fly, worm, and mouse). As of 2015 nearly 4600 experiments have been performed, with an emphasis on using a wide range of assays on a large number of different cell types. A key aspect of this project has been its focus on functional elements which are defined as genomic segments encoding an RNA or protein product or a biochemical signature such as a chromatin modification. This emphasis on function may be contrasted with other approaches that focus on generating DNA sequences. Both fundamental approaches (sequencing/annotating genomic DNA and defining functional elements in DNA) can serve to build a rich catalog of DNA elements (such as gene structures or repetitive DNA elements). The

- scope of the ENCODE project included redefining the meaning of the gene (see the end of this chapter) and the transcript (see Chapter 10 on the topic of RNA).
2. *Main conclusions of ENCODE.* Stamatoyannopoulos (2012) provides an overview of the main conclusions of the ENCODE project, as well as its significance and future directions. Conclusions include the following (from ENCODE Project Consortium *et al.* 2007, 2012):
    - a. The human genome is pervasively transcribed; we discuss this in Chapter 10. While exons span less than 3% of the genome, RNA transcripts are generated from 62% of the genome.
    - b. 80.4% of the human genome is functionally active, defined as participating in at least one RNA and/or chromatin-associated event in at least one cell type. While this is presented as perhaps the major single finding of the ENCODE project, objections have been raised (see the following section). Previous estimates indicate that ~10% of the human genome is functionally active (e.g., Smith *et al.*, 2004), rather than 80.4% as championed by the ENCODE project. Indeed, the ENCODE consortium argues that the 80.4% estimate is conservative because not all cell types or physiological conditions were assayed.
    - c. Many novel noncoding transcripts were identified, sometimes overlapping protein-coding genes. A set of long noncoding RNAs was characterized (Derrien *et al.*, 2012).
    - d. Novel transcriptional start sites were identified and characterized in detail. Regulatory sequences surrounding transcription start sites are symmetrically distributed. Previously, it had been thought that there is a bias toward the location of regulatory sequences upstream of genes.
    - e. Histone modification and chromatin accessibility predict the presence and activity of transcription start sites. 56.1% of the genome was shown to be enriched for histone modifications.
    - f. Of the 80.4% of the human genome spanned by elements defined by ENCODE as functional, if we exclude the RNA elements and histone elements 44.2% of the genome is covered. These regions include sites of hypersensitivity to digestion by the endonuclease DNase I; transcription factor binding sites; or DNA binding sites defined by chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq). The consortium reported finding 4.1 million DNase I hypersensitive sites and estimated this is half of the true total.
    - g. 5% of the nucleotides in the human genome are under evolutionary constraint in mammals. Of these constrained bases, there is experimental evidence of some function for about 60%. Not all bases that are under evolutionary constraint have been experimentally shown to have function.
- ENCODE has been extended to the mouse genome (Mouse ENCODE Consortium *et al.*, 2012). It also includes the model organism (modENCODE) project which generated detailed functional annotation of a fruit fly genome (*Drosophila melanogaster*) and a nematode (*Caenorhabditis elegans*). We describe these project in Chapter 19.
3. *Where to find ENCODE data.* The UCSC Genome Bioinformatics site serves as a central repository to browse and mine ENCODE data (Rosenbloom *et al.*, 2013). It also enables downloads of raw and processed data files. Many of the datasets are also available at the Gene Expression Omnibus (GEO) database at the National Center for Biotechnology Information (NCBI) as well as the Sequence Read Archive at NCBI. The journal *Nature* also offers an extensive ENCODE explorer with many resources including links to publications. The ENCODE Project Consortium (2011) also provides a useful user guide to the project.

The modENCODE project website is <http://www.genome.gov/modencode/> (WebLink 8.15).

Visit the ENCODE site at UCSC: <http://genome.ucsc.edu/ENCODE/> (WebLink 8.16). The National Human Genome Research Institute which sponsored the ENCODE project offers information at <http://www.genome.gov/10005107> (WebLink 8.17). See the *Nature* site at <http://www.nature.com/encode/> (WebLink 8.18).

## Critiques of ENCODE: the C Value Paradox Revisited and the Definition of Function

Several challenges have been made to the ENCODE Project Consortium *et al.* (2012) claims that 80.4% of the human is functional and that the concept of junk DNA is obsolete. This debate highlights the different meanings of function.

One objection is that junk DNA may have biochemical activity (as described by the ENCODE project) without having function in an evolutionary sense (Niu and Jiang, 2013). The ENCODE definition of function does not distinguish between biologically important activities (such as a globin gene encoding a globin protein) and the activity found in “junk” DNA such as the activity of transposable elements that comprise half the human genome.

Sean Eddy (2012) notes that transposons that fill much of the human genome are expected to have a biochemical function according the ENCODE definitions of biochemical activity: many transposable sequences are actively transcribed and regulated, and they are also actively repressed by some host-mediated chromatin modifications. Eddy (2013) contrasts two views: (1) a specific and reproducible biochemical phenomenon must have a biologically meaningful function; and (2) biology is “noisy” and background biochemical activity is tolerated. To test these models he proposes a Random Genome Project in which a million bases of random synthetic DNA are introduced into a cell. He predicts it will display functional biochemical properties (e.g., transcription, transcription factor binding, histone modifications) such as those reported by the ENCODE project, although such function would not be biologically meaningful. The effects, he predicts, would even be cell-type specific since each cell has its own regulatory machinery. In essence, Eddy proposes a negative control experiment that is likely to expose false positive results in ENCODE. Niu and Jiang (2013) and Graur *et al.* (2013) also distinguish biochemical function as defined by ENCODE from evolutionarily meaningful function.

Ford Doolittle (2013) proposes a different thought experiment. Suppose the ENCODE project were extended to a set of compact genomes (e.g., *Takifugu rubripes*; 400 Mb) and large genomes (e.g., a lungfish or various giant plant or protist genomes). He predicts two possible outcomes. First, functional elements could be constant in number, regardless of C value (similarly, the number of protein-coding genes typically does not scale with genome size). In this outcome, the density of functional elements per kilobase would be dramatically smaller in such large genomes, yet would be surrounded by vast amounts of junk DNA. A second outcome is that functional elements as defined by ENCODE increase in proportion to C value (independent of organismal complexity). Would lungfish having 300-fold larger genome size and 300-fold more functional elements then be expected to display more organismal complexity than related *Takifugu* having compact genomes?

These concerns lead Doolittle (2013) as well as Niu and Jiang (2013), Graur *et al.* (2013), and Eddy (2013) to consider the definition of function. One definition implies an evolutionarily selected effect: the function of a trait (or genomic feature) reflects its effects for which it was (or is) under positive natural selection. *FOXP2* is a gene that, in the human lineage, facilitates speech (Lai *et al.*, 2001). While it is conserved among other vertebrates, in humans it has a selective function in speech that is selected for. We do not call some effects “functions” if they are incidental (e.g., lower back pain in primates that walk upright), and in the context of sequences we often infer selection by examining evolutionary conservation.

A second way to define function is based on causal roles. Elements are often studied by ablation: if a DNA region is deleted (or blocked from performing its inherent activity such as expression) and some effect then goes away, we may assign a causal role for that DNA element. Doolittle suggests that most biologists see experiments that indicate such causal roles as providing indirect evidence for a selected effect. Graur *et al.* (2013) cite an example of the DNA sequence TATAAA. This has a well-known selected effect function,

maintained by natural selection, to bind a transcription factor. If another sequence arises by mutation and happens to closely resemble TATAAA it might bind the transcription factor (having a function based on causal role), but without any adaptive or maladaptive effect (there is therefore no selected effect function).

A third definition of function is based on existence (in the language used by critics). A structure or element exists and so must be functional whether it is an intron, *Alu* element, or endogenous retrovirus. These various elements may be actively transcribed. Doolittle suggests that this definition of function based on mere existence is the principal sense in which the ENCODE consortium defines over 80% of the genome as having function. To Graur *et al.* (2013), the human genome could be said to consist of 100% functional DNA according to a definition of function based on activity, since 100% of DNA is transcribed by DNA polymerases. This would extend the ENCODE conclusions to an absurd extent. We return to the subject of defining function in Chapters 12 and 14.

TATAAA commonly occurs upstream of polyadenylation sites.

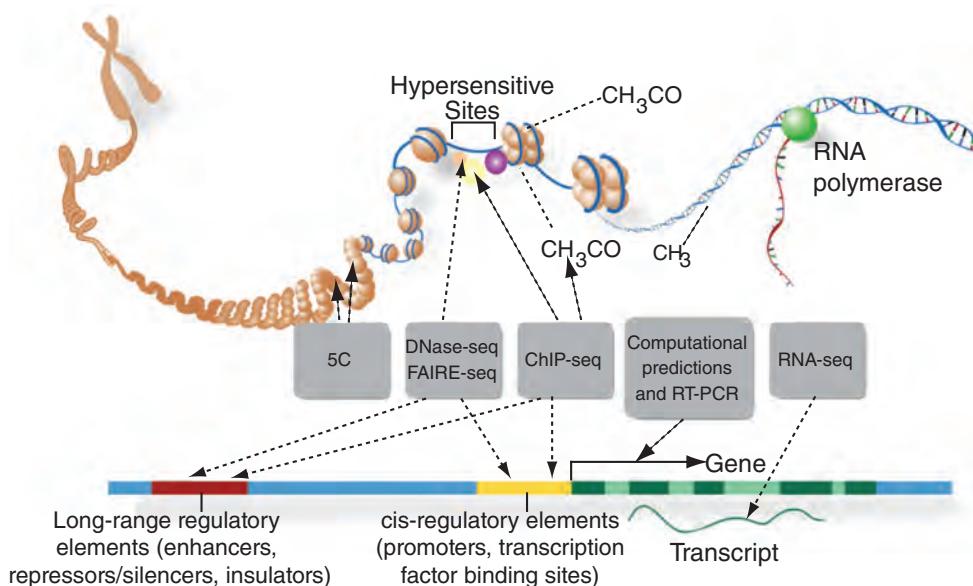
We return to the subject of defining function in Chapter 12 (in the context of protein function; see Fig. 12.18). In Chapter 14 we discuss definitions of function and approaches to function (involving the ENCODE project and other sources); see Fig. 14.17.

## REPETITIVE DNA CONTENT OF EUKARYOTIC CHROMOSOMES

### Eukaryotic Genomes Include Noncoding and Repetitive DNA Sequences

Bacterial and archaeal genomes have both genes and additional, relatively small intergenic regions. Typically, these genomes are circular, and there is almost one gene per kilobase of genomic DNA (Chapter 17; Table 8.1). In contrast, eukaryotic genomes contain a smaller proportion of protein-coding genes and large amounts of noncoding DNA. This noncoding material includes repetitive DNA, genes encoding RNAs that have assorted functions, introns that interrupt exons and are spliced from mature RNA transcripts, and intergenic regions.

Repetitive DNA sequences can occupy vast proportions of eukaryotic genomes (Richard *et al.*, 2008). These sequences consist of repeated nucleotides of various

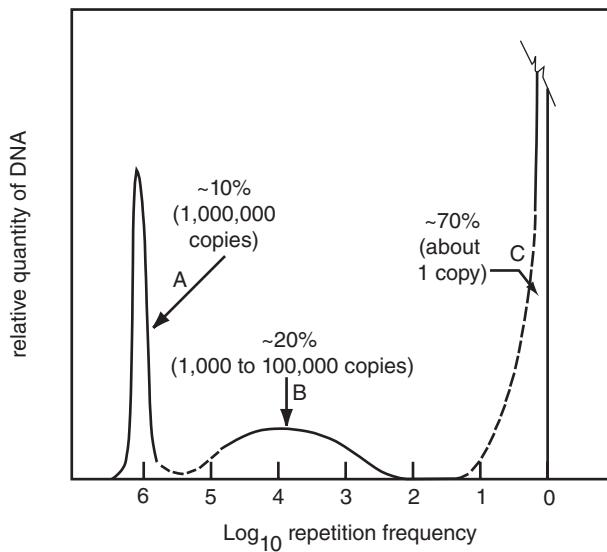


**FIGURE 8.5** The ENCODE project is intended to catalog functional elements in the human genome (as well as the genomes of mouse, fruit fly, and a nematode). This catalog includes regulatory DNA elements that control processes such as transcription, and elements that function at the RNA and protein levels. The image shows a chromosome (upper left) with the DNA unwound to show the types of technologies used by the ENCODE project: 5C (chromosome conformation capture carbon copy); DNA-seq; FAIRE-seq; ChIP-seq; polymerase chain reaction with reverse transcription (RT-PCR); and RNA-seq. Source: ENCODE, courtesy of UCSC.

lengths (Jurka, 1998). We also discuss these repeats in our analysis of the human genome (Chapter 20). In mammals, up to 60% of genomic DNA is repetitive; in some yeasts, 20% is repetitive. Identifying repetitive DNA elements in eukaryotic DNA is essential in genome analysis. Such repeats can powerfully influence the structure of the genome, including the capacity of chromosomes to rearrange and to regulate transcription. They are often important in disease, serving as substrates for recombination events that delete or duplicate chromosomal segments. Repeats are also useful as “molecular fossils” in evolutionary studies based on comparative analysis of genomes from different species (Chapter 20).

Britten and Kohne (1968) performed some of the earliest experiments that defined the repetitive nature of eukaryotic DNA. They purified genomic DNA from a wide variety of species, sheared it, and dissociated the DNA strands. Under appropriate conditions of salt, temperature, and time, the DNA strands reanneal. They measured the rate at which the DNA reassociates and found that for dozens of eukaryotes (but not for several viruses or bacteria) DNA reassociates in several distinct fractions. Large amounts of eukaryotic DNA reassociate extremely rapidly. For the mouse genome, about 10% of genomic DNA reassociates rapidly and consists of about one million copies (Fig. 8.6, arrow A). This highly repetitive DNA is localized to the highly condensed portion of chromosomes referred to as heterochromatin (Redi *et al.*, 2001; Avramova, 2002). A further 20% of the DNA reassociates in a fraction containing from 1000 to 100,000 distinct DNA species (arrow B). Finally, about 70% of the DNA is unique, consisting of only a single copy (arrow C). This DNA forms the euchromatin, a portion of the chromosome that is not

Britten and Kohne (1968) used several techniques to distinguish single-stranded from double-stranded DNA such as hydroxyapatite chromatography (a calcium phosphate column), binding of radiolabeled DNA fragments to immobilized DNA on filters, and spectrophotometry. The rate of DNA reassociation is a function of the incubation time  $t$  and the DNA concentration  $C_0$ . The  $C_0 t$  plot displays the fraction of DNA that remains single-stranded versus the  $C_0 t$  value, and it is the basis for the data shown in Figure 8.6.



**FIGURE 8.6** The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a  $C_0 t_{1/2}$  curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large  $C_0 t_{1/2}$  value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow A), and represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.

Source: Britten and Kohne (1968). Reproduced with permission from AAAS.

condensed and is therefore accessible for the transcription of genes. The banding pattern of chromosomes (Fig. 8.1) corresponds to regions of heterochromatin and euchromatin. Heterochromatic regions lack (or actively inhibit) gene expression, although some expressed genes have been identified in the heterochromatin of a variety of species from *Drosophila* to human (Yasuhara and Wakimoto, 2006).

RepeatMasker software, introduced in the following section, has been the most widely used tool for characterizing repetitive DNA. More recently, the leading researchers in this area (including Robert Finn, Arian Smit, Jerzy Jurka, and Sean Eddy) introduced Dfam, a database of repetitive DNA that relies on hidden Markov models (Wheeler *et al.*, 2013). They report coverage of 54.5% of the human genome by repetitive elements.

The origin of these repeats and their function present fascinating questions. What different kinds of repeats occur? From where did they originate and when? Is there a logic to their promiscuous growth, or do they multiply without purpose? We are beginning to understand the extent and nature of the repeat content of eukaryotic genomes, including the human genome. Repetitive DNA has in the past been called “junk DNA” or “selfish DNA,” reflecting its propensity to expand throughout genomes. However, it is likely that repetitive DNA has important roles in chromosome structure, recombination events, and the function of some genes (Makalowski, 2000; see also the following section).

There are five main classes of repetitive DNA in eukaryotes (Jurka, 1998; Makalowski, 2000; IHGSC, 2001; Kidwell, 2002; Jurka *et al.*, 2007) as described in the following sections.

#### *Interspersed Repeats (Transposon-Derived Repeats)*

Together, interspersed repeats constitute about 45% of the human genome (see Chapter 20; reviewed by Jurka *et al.*, 2007; Rebollo *et al.*, 2012). These repeats can be generated by elements that copy RNA intermediates (retroelements) or DNA intermediates (DNA transposons) (Table 8.4). Genes may be copied by retrotransposition when an mRNA is reverse-transcribed and then integrated into the genome. Such genes can be identified because they usually lack introns, while they do have short direct flanking repeats. Examples of some mammalian retrotransposed genes are presented in Table 8.5.

Interspersed repeats can be divided into four categories (Ostertag and Kazazian, 2001; Kidwell, 2002; see also Figures 20.9, 20.10, Table 20.5):

- Long-terminal-repeat (LTR) transposons, which are RNA-mediated elements. These are also called retrovirus-like elements. LTR transposons have LTRs of several hundred base pairs at either end of the element.

Dfam is available at <http://dfam.janelia.org/> (WebLink 8.19).

A retrotransposon (also called a retroposon or retroelement) is a transposable element that copies itself to genomic locations through a process of reverse transcription with an RNA intermediate. This process is similar to that of a retrovirus.

Barbara McClintock was awarded a Nobel Prize in 1983 for her discovery of mobile genetic elements in maize (*Zea mays*). You can read more about this pioneering work at <http://www.nobel.se/medicine/laureates/1983/> (WebLink 8.20).

A search of NCBI Nucleotide with the term “retropseudogene” yields ~70 hits (February 2015), while “retrotransposed” yields 120 hits. A search with the term “retrotransposon” however yields >325,000 core nucleotide matches, >21,000 expressed sequence tags, and >100,000 genome survey sequences.

**TABLE 8.4 Examples of repeat classes and transposable elements. Adapted from Kidwell (2002) with permission from Springer Science and Business Media.**

Class	Subclass	Superfamily	Examples of family	Approximate size range (bp)
Retroelements (RNA-mediated elements)	LTR retrotransposons	Ty1-copia	Opie-1 (maize)	3000–12,000
	Non-LTR retrotransposons	LINEs	LINE-1 (human)	1000–7000
		SINEs	Alu (human)	100–500
DNA transposons	Cut-and-paste transposition	Mariner-Tc1	Tc1 in <i>C. elegans</i>	1000–2000
		P	P in <i>Drosophila</i>	500–4600
	Rolling circle transposition	Helitrons	Helitrons in <i>A. thaliana</i> , <i>O. sativa</i> , and <i>C. elegans</i>	5500–17,500

**TABLE 8.5 Examples of mammalian genes generated by retrotransposition. Retrotransposed genes lack introns, and they often have flanking direct repeats and a polyadenine tail. Chr, chromosome; ADAM, a disintegrin and metalloproteinase; Cetn, centrin, EF-hand protein; Glud, glutamate dehydrogenase; Pdha2, pyruvate dehydrogenase (lipoamide) alpha 2; Supt4h, suppressor of Ty 4 homolog (*S. cerevisiae*). Adapted from Betrán and Long (2002) and from a search of Entrez (NCBI) with the term *retropseudogene*.**

Retrotransposed gene			Original gene				
Name	RefSeq	Chr	Name	RefSeq	Chr	Distribution	Age (Ma)
ADAM20	NM_003814	14q	ADAM9	NM_003816	8p	Human, not macaque	<20
Cetn1	NM_004066	18p	Cetn2	NM_004344	Xq28	Mammals	>75
Glud2	NM_012084	Xq	Glud1	NM_005271	10q	Human, not mouse	<70
Pdha2	NM_005390	4q	Pdha1	NM_000284	Xp	Placentals	~70
SRP46	NM_032102	11q	PR264/SC35	NM_003016	17q	Human, simians	~89
Supt4h2	NM_011509	10	Supt4h	NM_009296	11	Mouse	<70

SINEBase is a database of SINES (Vassetzky and Kramerov, 2013) available at <http://sines.eimb.ru/> (WebLink 8.21).

RepBase Update has been developed from 1990 by Jerzy Jurka and colleagues. RepeatMasker was written by Arian Smit and Phil Green, and is available at <http://www.repeatmasker.org/> (WebLink 8.22). The Censor Server at the Genetic Information Research Institute (GIRI) is available at <http://www.girinst.org/censor/index.php> (WebLink 8.23). RepeatMasker servers are available online at the Institute for Systems Biology (<http://www.repeatmasker.org/>, WebLink 8.24) and the NCKU Bioinformatics Center (Taiwan) (<http://www.binfo.ncku.edu.tw/RM/RepeatMasker.php>, WebLink 8.25).

The 50,000 bases of genomic DNA we used are available as Web Document 8.2 at <http://www.bioinfbook.org/chapter8>.

- Long interspersed elements (LINEs), which encode an enzyme with reverse transcriptase activity (and possibly additional proteins). In mammals, LINE1 and LINE2 families are most prevalent.
- Short interspersed elements (SINEs), which are also RNA-mediated elements. *Alu* repeats, found in primates, are well-known examples of SINEs. We see an example of an *Alu* repeat sequence below.
- DNA transposons comprise about 3% of the human genome.

We can illustrate interspersed repeats using the UCSC Genome Browser. A region of 13,000 base pairs including the beta globin (*HBB*) gene is shown in **Figure 8.7a**. Information on repeats that is precomputed using the RepeatMasker software package shows SINE, LINE, LTR, and DNA transposon elements as well as several other categories of repetitive DNA (simple repeats, low-complexity DNA, satellite DNA). By clicking the Table link on the top sidebar of the UCSC Genome Browser, you can access the Table Browser (**Fig. 8.7b**). You can select the BED output format (**Fig. 8.7c**). By clicking “get output” you can obtain a tab-delimited file listing all the elements detected by RepeatMasker as well as their genomic coordinates. By selecting the “Sequence” output option, you are directed to a dialog box (**Fig. 8.8.a**) from which you obtain the RepeatMasked sequences (a portion of which are shown in **Fig. 8.1.b**). These include repetitive regions such as A-rich, AT-rich, and those with a repeating pattern such as (TA)<sub>n</sub>, (CA)<sub>n</sub>, or (TAAAA)<sub>n</sub>, where n denotes the number of occurrences of each pattern. Additional patterns such as L1 elements (**Fig. 8.8.b**, bottom) are not readily discernible by eye as repetitive.

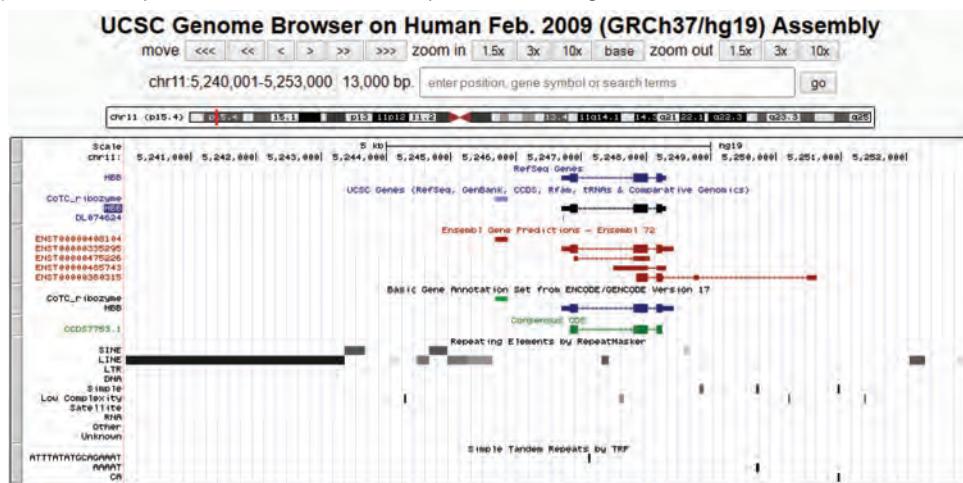
RepeatMasker searches a DNA query of interest against RepBase, a database of known repeats and low-complexity regions in eukaryotic DNA. Several programs, including RepeatMasker and the Censor Server at GIRI, effectively allow searches of DNA query sequences against this database (Smit, 1999; Jurka, 2000).

We can explore repetitive DNA by using a RepeatMasker server to analyze 50,000 bp of genomic DNA from human chromosome 11 in the beta globin locus. The output includes a summary (**Table 8.6**) as well as detailed tables of scores using the Smith-Waterman algorithm, the position of the repeat, and information on the type of repeat (e.g., SINE/*Alu*, LTR, or simple repeat).

### Processed Pseudogenes

These genes are not actively transcribed or translated (Echols *et al.*, 2002; Harrison and Gerstein, 2002). They represent genes that were once functional, but are defined by their lack of protein product. They can be recognized because of the presence of a stop

(a) Genes and repeat elements in 13,000 base pairs of the beta globin locus



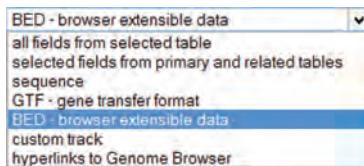
(b) Access to tabular data on repeat elements using the UCSC Table Browser

**Table Browser**

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal   genome: Human   assembly: Feb. 2009 (GRCh37/hg19)   group: Variation and Repeats   track: RepeatMasker   add custom tracks   track hubs  
table: rmsk   describe table schema  
region:  genome  ENCODE Pilot regions  position chr11:5,240,001-5,253,000   lookup   define regions  
identifiers (names/accessions):    
filter:   
intersection:   
output format: BED - browser extensible data   Send output to  Galaxy  GREAT  
output file:  (leave blank to keep output in browser)  
file type returned:  plain text  gzip compressed

(c) Options for Table Browser output formats



**FIGURE 8.7** Interspersed and other repetitive DNA elements are visualized and tabulated using the UCSC Genome Browser and Table Browser. (a) A region of 13,000 bases in the beta globin region of chromosome 11 is shown (chr11:5,240,001–5,253,000). The RepeatMasker track is set to “full,” displaying the location of several repetitive DNA elements such as SINE, LINE, LTR, and DNA transposons. Gene tracks are also displayed (for RefSeq, UCSC Genes, Ensembl Gene Predictions, GENCODE, and the Consensus Coding Sequence or CCDS project). Note that the gene models differ slightly. (b) A link from the Genome Browser to the Table Browser allows you to access this (or other) information as a tabular output. (c) Table Browser output formats include browser extensible data (BED) files which are defined in detail at the UCSC site. See also **Figure 2.13**.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

codon or frameshift that interrupts an open reading frame. There are two main classes of pseudogenes. Processed pseudogenes arise through retrotransposition events (i.e., random insertion events mediated by LINEs having reverse transcriptase activity) via an RNA intermediate. Nonprocessed pseudogenes are remnants of duplicated genes.

Recall that a BLAST search uses the SEG and/or DUST programs to define and mask repetitive DNA sequences, and also to detect and mask low-complexity protein sequences (Chapter 4).

The mouse genome contains one functional gene encoding glyceraldehyde 3-phosphate dehydrogenase (*Gapdh*; NM\_008084.2) and at least 400 pseudogenes distributed across 19 chromosomes (Mouse Genome Sequencing Consortium *et al.*, 2002). The functional *Gapdh* gene was listed as assigned to mouse chromosome 7 (Mouse Genome Sequencing Consortium *et al.*, 2002), but currently (February 2015) it is assigned to chromosome 6 by Gene at NCBI and by Ensembl. The presence of many pseudogenes contributes to the difficulty of assigning correct chromosomal loci.

(a) UCSC Table Browser sequence output

**Sequence Retrieval Region Options:**

Add  extra bases upstream (5') and  extra downstream (3')

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending past the edge of the chromosome.

**Sequence Formatting Options:**

- All upper case.
- All lower case.
- Mask repeats:  to lower case  to N

(b) Repeat sequences identified by RepeatMasker

**FIGURE 8.8** RepeatMasker output. (a) The sequence retrieval option includes several formatting options such as masking repeats to lower case. (b) The sequences identified by RepeatMasker include nucleotide-rich repeats, repeating motifs such as  $(TAAAAA)_n$ , and SINE and LINE elements.

*Source:* RepeatMasker.

Mark Gerstein's laboratory provides a website in pseudogenes (<http://www.pseudogene.org/>, WebLink 8.26). This includes a browser and descriptions of pseudogenes in human, worm, fly, yeast, and plant. psiDR is available at <http://www.pseudogenes.org/psidr/> (WebLink 8.27).

While pseudogenes are defined as nonfunctional, many studies have emphasized their possible functional roles (Balakirev and Ayala, 2003; Castillo-Davis, 2005; Pavlicek *et al.*, 2006). These include gene expression, the regulation of gene function, and roles in recombination. Evolutionary studies suggest that some pseudogenes do not evolve at the neutral rate (e.g., compared to extinct repeat elements), consistent with some functional role.

As part of ENCODE, the GENCODE project defined the expression levels of human pseudogenes, transcription factor binding, RNA polymerase II binding, and chromatin marks

**TABLE 8.6 RepeatMasker analysis of 50,000 base pairs of genomic DNA in the human *HBB* locus. The sequence (given at Web Document 8.2) was entered into the search engine at <http://www.repeatmasker.org> (version open 4.0.3, default mode).**

Elements	Type	Number of elements*	Length occupied (bp)	Percentage of sequence
SINEs		8	2093	4.19
	ALUs	7	2011	4.02
	MIRs	1	82	0.16
LINEs		16	12,279	24.56
	LINE1	12	11,419	22.84
	LINE2	4	860	1.72
LTR elements	L3/CR1	0	0	0
		5	1556	3.11
	ERVL	1	513	1.03
	ERVL-MaLRs	2	669	1.34
	ERV_classI	2	374	0.75
DNA elements	ERV_classII	0	0	0
		1	248	0.5
	hAT-Charlie	1	248	0.5
Unclassified	TcMar-Tigger	0	0	0
		0	0	0
	Total interspersed repeats		16,176	32.35
Small RNA		0	0	0
Satellites		0	0	0
Simple repeats	repeats:	18	824	1.65
Low complexity		3	363	0.73

\*Most repeats fragmented by insertions or deletions were counted as one element.

Source: RepeatMasker.

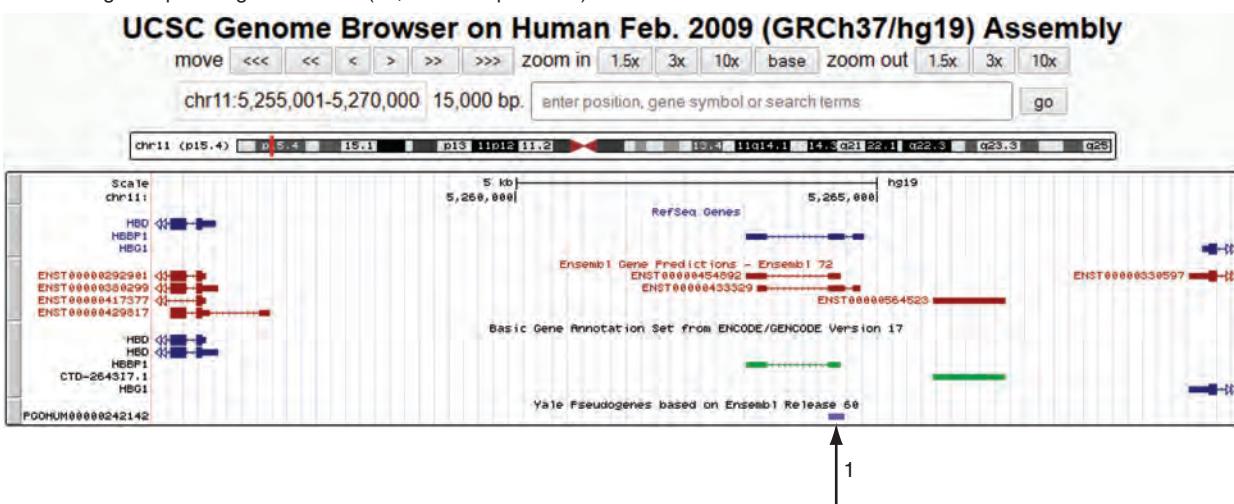
(Pei *et al.*, 2012). They concluded that some pseudogenes retain partial activity, for example having regulatory functions as noncoding RNAs. They also produced the pseudogene Decoration Resource (psiDR) to annotate pseudogenes. We describe mechanisms for the origin of pseudogenes later in this chapter, and in Chapter 18 we discuss the duplication of entire yeast genomes followed by rapid, subsequent gene loss to generate pseudogenes.

The number of pseudogenes in the human genome is remarkably close to the number of predicted protein-coding genes. For example, chromosome 1 has 3141 protein-coding genes and 991 pseudogenes (Gregory *et al.*, 2006); chromosome 2 has 1346 genes and 1239 pseudogenes (Hillier *et al.*, 2005); and the smallest autosome, chromosome 21, has 225 known and predicted genes and 59 pseudogenes (Hattori *et al.*, 2000).

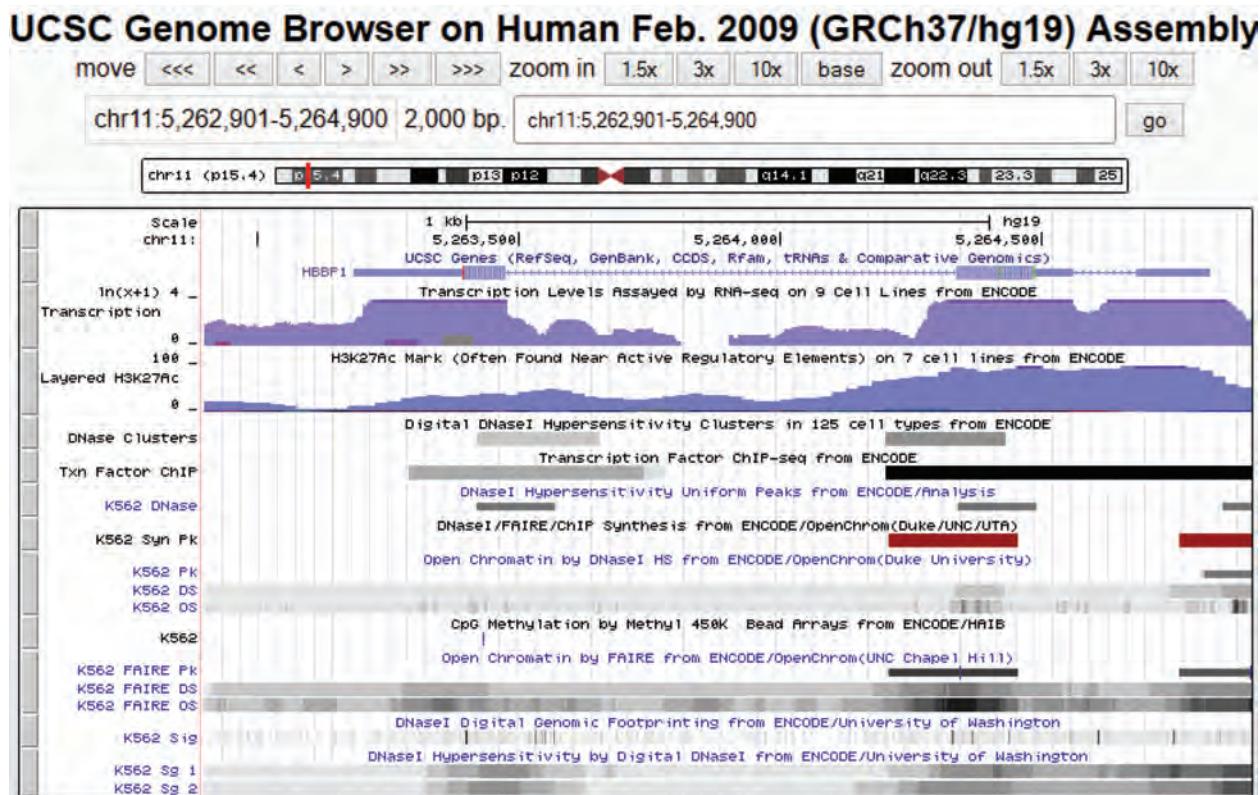
You can activate the pseudogene track at the UCSC Genome Browser. This is shown for a segment of 15,000 base pairs within the beta globin locus (Fig. 8.9a). The RefSeq track shows three genes (the globin genes *HBD* and *HBG1* flanking the pseudogene *HBBP1*), while the Ensembl and GENCODE genes tracks show several additional gene models. A pseudogenes track is also displayed, showing *HBBP1*. As for any UCSC Genome Browser track, you can click on the title “pseudogenes” above the pull-down menu to

According to Ensembl (assembly GRCh38) there are ~20,300 protein-coding genes and ~14,200 pseudogenes in the human genome.

(a) Region of beta globin pseudogene *HBBP1* (15,000 base pair view)



(b) *HBBP1* (2,000 base pair view)



**FIGURE 8.9** Viewing pseudogenes at the UCSC Genome Browser. (a) 15,000 base pair view of the beta globin region (chr11:5,255,001–5,270,000). Note that a consensus annotation track for pseudogenes is activated as well as a RefSeq, Ensembl, and GENCODE gene tracks. A pseudogene is evident, beta globin pseudogene 1 (*HBBP1*). It is flanked by *HBD* (to the 5' side) and *HBG1* (at the 3' edge). The Yale Pseudogenes track indicates the pseudogene (arrow 1); clicking on its entry displays more information, including that the status of this entry is “ambiguous” (not shown). (b) Detailed view of *HBBP1* (showing 2 kilobases at chr11:5,262,901–5,264,900) includes extensive ENCODE annotation of the regulation and expression of *HBBP1*. For example, histone modifications (such as H3K27AC) and DNaseI hypersensitivity are shown.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

access more details on the methodology as well as literature citations. This pseudogene corresponds to accession NR\_001589.1 and is annotated in NCBI Gene as beta hemoglobin pseudogene 1 (official symbol *HBBP1*). Viewing *HBBP1* at higher magnification (a 2000 base pair view; **Fig. 8.9b**) reveals more details of the structure of the pseudogene; by clicking it, details of a model of the gene and information on its expression and RNA folding properties are obtained.

It is also straightforward to analyze repeats (or any other genomic features of interest) in R. Install `biomaRt`, specify the genome of interest (e.g., human), the region, and the features.

### Simple Sequence Repeats

These microsatellites (typically from 1 to 6 base pairs in length) and minisatellites (typically from a dozen to 500 bp repeats) include short sequences such as  $(A)_n$ ,  $(CA)_n$ , or  $(CGG)_n$ . We saw examples of these repeats from our RepeatMasker analysis of human genomic DNA (**Fig. 8.8**). Replication slippage is a mechanism by which simple sequence repeats may occur (Richard *et al.*, 2008). Many functions have been ascribed to simple sequence repeats, from influencing transcription factor binding to influencing morphological traits in dogs and yeast (reviewed in Kashi and King, 2006).

Simple sequence repeats of particular length and composition occur preferentially in different species. For example,  $(AT)_n$  is especially common in *A. thaliana* and  $(CT/GA)_n$  occurs preferentially in *C. elegans* (Schlöterer and Harr, 2000). In *Drosophila virilis*, the density and length of microsatellites are considerably greater than in *D. melanogaster* or *H. sapiens* (Schlöterer and Harr, 2000). In humans, simple sequence repeats are of particular interest because they are highly polymorphic between individuals and therefore serve as useful genetic markers. The expansion of triplet repeats such as CAG is also associated with over a dozen diseases including Huntington disease (Cummings and Zoghbi, 2000). We discuss these issues in Chapter 21 (on human disease). A disease characterized by cerebellar ataxia and seizures (spinocerebellar atrophy type 10; SCA10) is caused by the expansion of the sequence ATTCT in intron 9 of the ataxin 10 gene on chromosome 22q13.31 (Matsuura *et al.*, 2000). While there are 10–29 repeats in apparently normal individuals, those with SCA10 have from several hundred to as many as 4500 repeats.

### Segmental Duplications

Segmental duplications are often defined as two genomic regions sharing at least 90% nucleotide identity over a span of one kilobase, although they sometimes consist of blocks of 200 or 300 kilobases (kb) in length (Bailey *et al.*, 2001). These duplications occur both within and between chromosomes (intra- and interchromosomally). The euchromatic portion of the human genome consists of about 5.3% duplicated regions (She *et al.*, 2004). This includes about 150 megabases. In “Mechanisms of Creating Duplications, Deletions, and Inversions” we discuss mechanisms by which segmental duplications (also called low-copy repeats) may cause genes to become deleted, duplicated, or inverted. A practical consideration is that after whole-genome shotgun sequencing, the assembly of segmentally duplicated regions (especially those >15 kilobases in length and sharing >97% sequence identity) is problematic (She *et al.*, 2004). As a consequence, assemblies based on whole-genome shotgun assembly may underestimate the extent of duplications (including duplicated genes), underestimate the length of euchromatin, and underrepresent duplication-rich regions including pericentromeric and subtelomeric areas.

We can view segmental duplications using the UCSC Genome Browser for both the beta globin locus (**Fig. 8.10a**) and the alpha globin locus (**Fig. 8.10b**). At the beta globin locus, the immediately adjacent *HBG1* and *HBG2* genes represent a segmental

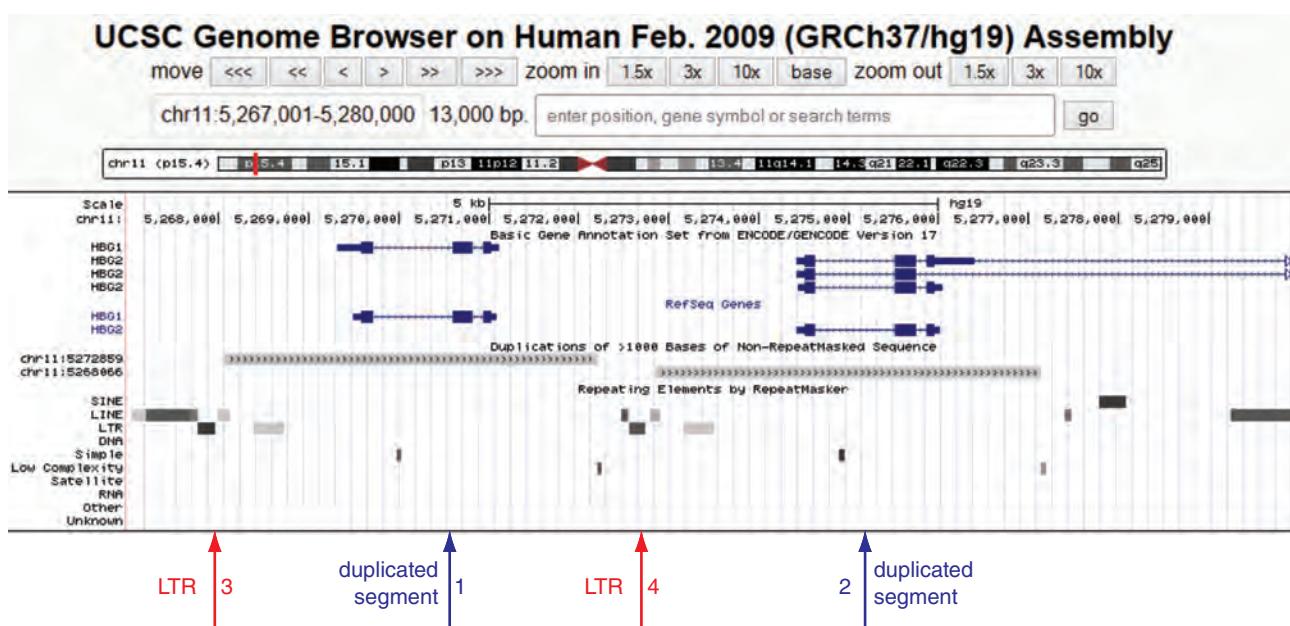
Web Document 8.3 shows a pairwise alignment between *HBB* and its pseudogene *HBBP1* (<http://www.bioinfbook.org/chapter8>).

Some authors define microsatellites as having a length of 1–6 bp, while others suggest 1–12 bp.

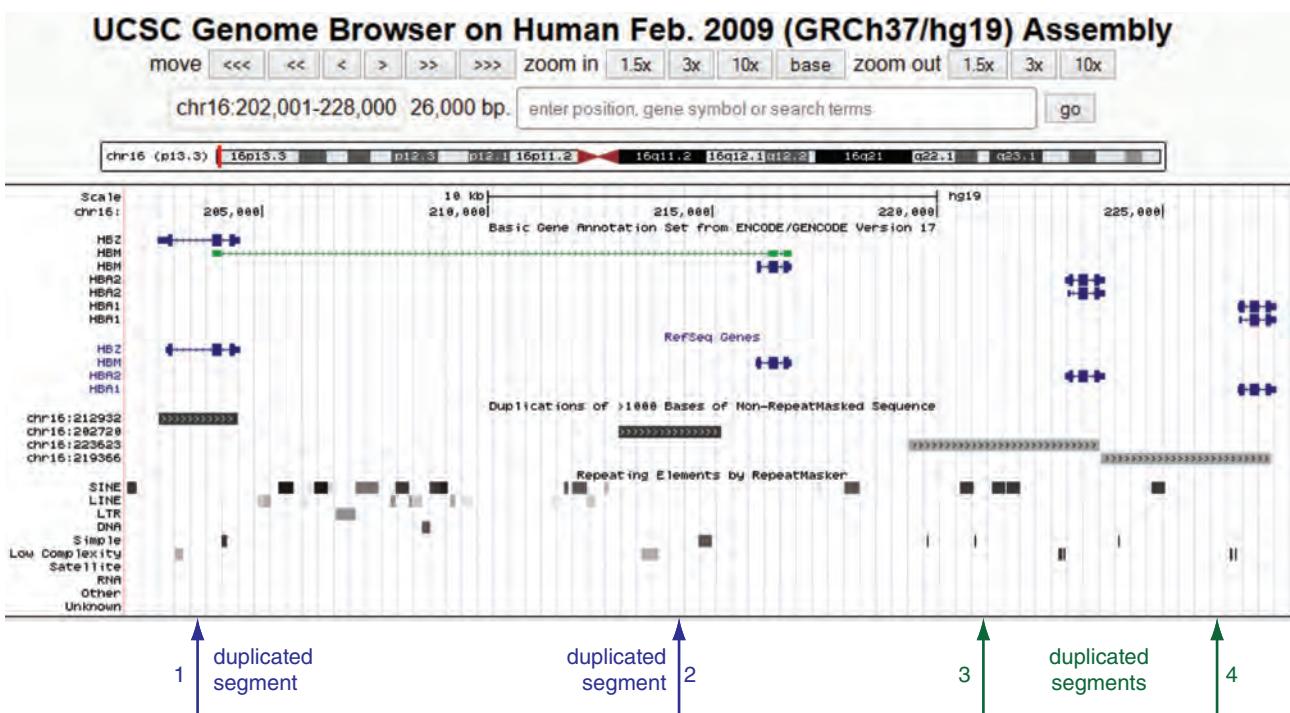
To see specific examples of simple sequence repeats, go to NCBI Nucleotide and enter “microsatellite.” There are over 700,000 entries from which to choose. The Tandem Repeats Finder is an online tool that allows you to search a sequence for tandem repeats of up to 2000 bp (<http://tandem.bu.edu/trf/trf.html>, WebLink 8.28; Benson, 1999). Its output is available at UCSC in the Variation and Repeats group for the track Microsatellite. Currently (February 2015) there are >41,000 entries for the human genome (GRCh37). For more information on SCA10 and its repeats, enter the query SCA10 at NCBI and see the Online Mendelian Inheritance in Man (OMIM) entry #603516.

You can download a table of segmentally duplicated regions from the UCSC Genome Browser. Currently (February 2015) they span 164 megabases (5.7% of the genome). To see this for the GRCh37 human genome assembly, set the group to Variation and Repeats, set the region to the genome, and click “summary/statistics.”

(a) Segmental duplication at the beta globin locus on chromosome 11



(b) Segmental duplications at the alpha globin locus on chromosome 16



**FIGURE 8.10** Segmental duplications visualized at the UCSC Genome Browser. (a) The beta globin region includes a segmentally duplicated region (13,000 base pairs on chromosome 11p15.4 are shown; chr11:5,267,001–5,280,000). These two regions (arrows 1 and 2) are flanked at the 5' end by long terminal repeats (LTRs; arrows 3 and 4) that could mediate tandem duplication. (b) A region of 26 kilobases is shown (chr16:202,001–228,000). Two pairs of segmentally duplicated regions are evident. For the first pair, the region encompasses the *HBZ* gene (arrow 1) but the duplicated region has no annotated genes (arrow 2), although expressed sequence tags (see Chapter 10) are localized to it (not shown). A second segmentally duplicated pair (arrows 3 and 4) includes the alpha globin genes *HBA2* and *HBA1*; these encoded proteins share 100% amino acid identity.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

**TABLE 8.7** Telomeric repeat sequences from several eukaryotic organisms.

Organism	Telomeric repeat	Reference
<i>Arabidopsis thaliana</i> , other plants	TTTAGGG	McKnight et al., 1997
<i>Ascaris suum</i> (nematode)	TTAGGC	Jentsch et al., 2002
<i>Euplotes aediculatus</i> , <i>Euplotes crassus</i> , <i>Oxytricha nova</i> (ciliates)	TTTTGGGG	Jarstfer and Cech, 2002; Shippen-Lentz and Blackburn, 1989; Melek et al., 1994
<i>Giardia duodenalis</i> , <i>Giardia lamblia</i>	TAGGG	Upcroft et al., 1997; Hou et al., 1995
<i>Guillardia theta</i> (cryptomonad nucleomorph)	[AG] <sub>n</sub> AAG <sub>n</sub> A	Douglas et al., 2001
<i>Homo sapiens</i> , other vertebrates	TTAGGG	Nanda et al., 2002
<i>Hymenoptera</i> , <i>Formicidae</i> (ants)	TTAGG	Lorite et al., 2002
<i>Paramecium</i> , <i>Tetrahymena</i>	TTGGGG, TTTGGG	McCormick-Graham and Romero, 1996
<i>Plasmodium falciparum</i>	AACCCTA	Gardner et al., 2002
<i>Plasmodium yoelii yoelii</i>	AACCCTG	Carlton et al., 2002

See Web Document 8.4 for a detailed description of segmental duplication of lipocalins. Lipocalins are proteins that transport hydrophobic ligands such as odorants (Pevsner et al., 1988). These genes offer further examples of segmental duplication that enable the diversification of gene function.

Web document 8.5 shows a global pairwise alignment between the two segmentally duplicated blocks at the beta globin locus. See <http://www.bioinfbook.org/chapter8>.

Telomeric repeats are synthesized by telomerase, a ribonucleoprotein that has specialized reverse transcriptase activity.

duplication. For the alpha globin locus on chromosome 16, the *HBZ* gene (zeta globin) is tandemly duplicated to generate a pseudogene less than 10,000 base pairs apart. By clicking on the segmental duplication block on the Genome Browser output, you can access the exact genomic coordinates of the duplicated blocks as well as a global pairwise alignment of the two.

## *Blocks of Tandemly Repeated Sequences*

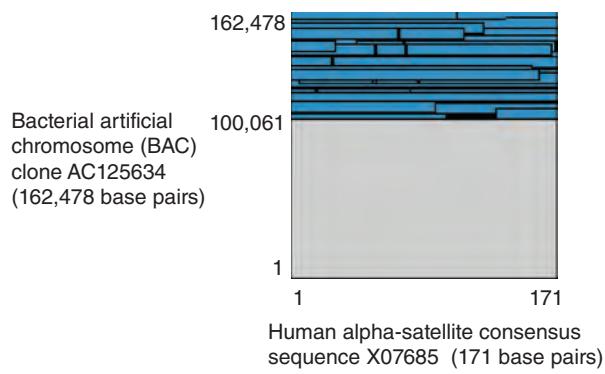
Tandem repeats occur at loci such as telomeres, centromeres, and ribosomal gene clusters. Several telomere repeat sequences are listed in **Table 8.7**. In human telomeres, the short sequence TTAGGG is repeated thousands of times. These repeats span up to 20 kilobases (while in mice they span 25–150 kb). Try a BLASTN search using TTAGGG TTAGGG TTAGGG as a query, restricting the output to human, and remove the filter for low complexity. The result is several thousand BLAST hits, most from telomeric sequences such as that shown in **Figure 8.11**.

The centromere is a constricted site of a chromosome that serves as an attachment point for spindle microtubules, allowing chromosomal segregation during mitotic and

```
>gi|224514922|ref|NT_024477.14| Homo sapiens chromosome 12 genomic
contig, GRCh37.p13 Primary Assembly (displaying 3' end)
CGGGAAATCAAAGCCCTCTGAATCCTGCGCACCGAGATTCTCCCCAGCCAAGGTGAGGCCGCAGCAGT
GGGAGATCCACACCGTAGCATTGGAACACAATGCAGCATTACAATGCAGACATGACACCGAAAATATA
ACACACCCCATTGCTCATGTAACAAGCACCTGTAATGCTAATGCACTGCCTCAAAACAAAATTTAAAT
AAGATCGCAATCCGCACACTGCCGTGCAAGTGCTAACAGCAATGAAAATAGTCACACATAATAACCCCTA
ATAGTGTAGGGTTAGGGTCAAGGTCCCGGTCGGGTCCGGGTCCGGGGTCCGGGTCAAGGTGAGGGTGA
GGGTTAGGGTTAGGGTCAAGGTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
```

**FIGURE 8.11** A BLASTN search of the human genome (all assemblies) database was performed at the NCBI website using TTAGGGTTAGGGTTAGGG as query (i.e., three TTAGGG repeats). There were matches to hundreds of genomic scaffolds. This figure shows an example (NT\_024477.14) assigned to the telomere of chromosome 12q having many dozens of TTAGGG repeats. These occurred at the 3' end of the genomic contig sequence.

*Source:* BLASTN, NCBI.



**FIGURE 8.12** The repetitive nature of  $\alpha$ -satellite DNA. A consensus sequence for human  $\alpha$ -satellite DNA (X07685) was compared to a BAC clone (AC125634) assigned to a pericentromeric region of chromosome 9q. BLASTN at NCBI was used, and the dotplot is shown. Note that a consecutive 60 kilobases of the BAC clone (y axis) matches the satellite consensus sequence repeatedly.

Source: BLASTN, NCBI.

meiotic cell divisions (Choo, 2001). All eukaryotic chromosomes have a functional centromere, although the primary nucleotide sequence is not well conserved between species. In humans, this DNA consists largely of a 171 base pair repeat of  $\alpha$ -satellite DNA extending for 1–4 Mb. Almost all eukaryotic centromeres are able to bind a histone H3-related protein (called CENP-A in vertebrates). This protein–DNA complex forms a building block of centromeric chromatin that is essential for the function of the kinetochore, the site of attachment of the spindle fiber.

The GenBank accession number for a human  $\alpha$ -satellite consensus sequence is X07685. An alignment of this sequence (171 base pairs) with a typical bacterial artificial chromosome (BAC) clone from a pericentromeric region dramatically shows how often the satellite sequence is repeated (Fig. 8.12). A BLASTN search of the nonredundant database, using this as a query and turning off filtering, results in thousands of database hits. If you exclude human entries from the output of your search, you find that the human  $\alpha$ -satellite sequence matches other primates. However, the human sequence has only very little conservation to nonprimate sequences, with nonsignificant expect values.

Satellite DNA is a feature of every known eukaryotic centromere, with only two documented exceptions. In the yeast *S. cerevisiae*, the entire centromere sequence extends only several hundred base pairs. A second exception is the neocentromere, an ectopic centromere that assembles a functional kinetochore, is stable in mitosis but lacks  $\alpha$ -satellite DNA (Amor and Choo, 2002; Marshall *et al.*, 2008). Over 90 human neocentromeres have been described, many involving trisomy or tetrasomy (extra chromosomal copies). As part of the analysis of the genome of the rhesus macaque *Macaca mulatta*, Ventura *et al.* (2007) described evolutionarily new centromeres that appeared while the conventional centromere was inactivated. They reported that in the 25 million years since macaque and human lineages diverged, 14 evolutionarily new centromeres have emerged and become fixed in one or the other species.

We described Expect values in Chapter 4. To perform this search try Net BLAST at the NCBI Genome Workbench (installation instructions are available at the NCBI website). Use the “Search Tool” (set for NCBI Nucleotide) to find X07685. Right-click that entry to “Add to Project” and create a new project with this sequence (it appears in a data folder on the left sidebar). Right-click to “Run Tools” and select megaBLAST search. View the results (with Alignment Summary View).

## GENE CONTENT OF EUKARYOTIC CHROMOSOMES

### Definition of Gene

We have begun our analysis of eukaryotic genomes by considering noncoding and repetitive DNA. The coding portions of a genome are of particular interest, as they largely determine the phenotype of all organisms. Two of the biggest challenges in understanding

any eukaryotic genome are defining what a gene is and identifying genes within genomic DNA. We first define the variety of genes and then provide the criteria for identifying them:

- Protein-coding genes form a major category of genes. Several criteria are applied to the assignment of a DNA sequence as a protein-coding gene. The principal requirement is that there must be an open reading frame (ORF) of at least some minimum length such as 90 base pairs (corresponding to 30 codons encoding amino acids, or an 3 kD protein). Frith *et al.* (2006) identified large numbers of short proteins (less than 100 amino acids). Of the 3701 proteins they identified, only 232 matched a mouse International Protein Index or Swiss-Prot database.
- Pseudogenes do not encode functional gene products although, as discussed above, some important exceptions have been reported.
- Many kinds of noncoding genes do not encode protein, but instead encode functional RNA molecules (Eddy, 2001, 2002). These include transfer RNA (tRNA) genes that translate information from the triplet codons in mRNA to amino acids. tRNAscan-SE software identifies 99–100% of tRNA genes in genomic DNA sequence with an error rate of one false positive per 15 Gb (Lowe and Eddy, 1997). We show an example of the tRNAscan-SE server in **Figure 10.5**.
- We discuss a variety of other noncoding genes in Chapter 10. These include ribosomal RNA (rRNA) genes that function in translation; small nucleolar RNAs (snoRNAs) that function in the nucleolus; small nuclear RNAs that function in spliceosomes to remove introns from primary RNA transcripts; and microRNAs (miRNAs) of about 21–25 nucleotides in length that are widely conserved among species and may serve as antisense regulators of other RNAs (Ambros, 2001; Ruvkun, 2001).

In annotating genomic DNA, an emphasis is often placed on describing the protein-coding genes. However, it is now clear that noncoding genes encoding various types of RNA products have diverse and important functions. Furthermore, it is not as straightforward to identify noncoding RNAs (Eddy, 2002), although the ENCODE Project Consortium *et al.* (2012) applied methods to characterize them. Their full size might be extremely small, as in the case of miRNAs. There is no ORF to help define the boundaries of noncoding genes. Database searches may be less sensitive than is possible for protein-coding genes, because the scoring matrices for amino acids are more sensitive and specific. We discuss databases of noncoding RNAs such as Rfam (Griffiths-Jones *et al.*, 2003) in Chapter 10.

Classically, a gene was defined as a unit of hereditary information localized to a particular chromosome position and encoding one protein. More recently, we have become aware of alternative splicing to produce multiple transcripts from one gene locus, we have identified large numbers of noncoding RNAs, and we have observed pervasive transcription throughout the genome (including transcriptionally active regions that have not been annotated as genes). Given the insights of the ENCODE project (ENCODE Project Consortium *et al.*, 2007, 2012) as well as the analysis of completed genome sequences, the historical definitions of a gene have been challenged. Gerstein *et al.* (2007, p. 677) proposed that “The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.” (This definition is vague and suffers from the terms “coherent” and “potentially.”) Djebali *et al.* (2012) of the ENCODE project wrote “...We would propose that the transcript be considered as the basic atomic unit of inheritance. Concomitantly, the term gene would then denote a higher-order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait.” Stamatoyannopoulos (2012) wrote: “Although the gene has conventionally been viewed as the fundamental unit of genomic organization, on the basis of ENCODE data it is now compellingly argued that this unit is not the gene but rather the transcript.”

A gene is a DNA sequence that codes for an RNA product that may further be translated into a protein product. The genomic sequence constitutes the genotype that is related to the phenotype of a cell or ultimately of an organism. The ENCODE project has expanded our understanding of the complexity of transcription. This includes cataloging a large amount of the genome that is transcribed, and finding many RNA transcripts derived from multiple genomic loci. By calling the transcript the “basic atomic unit of inheritance,” the ENCODE authors place greater weight on the product of a gene than on the gene itself. The use of the term “atomic” represents a metaphor; in an earlier metaphor, DNA is like the blueprint of a house, specifying products that assume various functions in the house (plumbing, trash removal, making compartments). The ENCODE project might see the blueprints as being so complex that only when they are interpreted (as transcripts) do they become functional blueprints.

## Finding Genes in Eukaryotic Genomes

RNA polymerase I synthesizes most ribosomal RNAs; RNA polymerase II synthesizes messenger RNAs and small nuclear RNAs (snRNAs); and RNA polymerase III synthesizes 5S rRNA and transfer RNAs (Chapter 10).

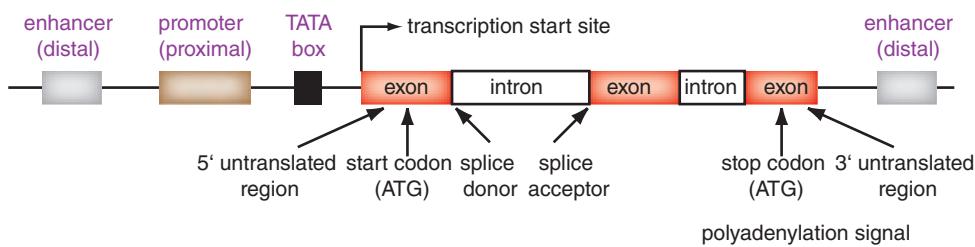
Finding protein-coding genes in eukaryotic genomes is a far more complex problem than for bacteria and archaea (Picardi and Pesole, 2010; Alioto, 2012). While bacterial genes typically correspond to long open reading frames (ORFs), most eukaryotic genes have exons and introns. The structure of a typical eukaryotic gene that is transcribed by RNA polymerase II is summarized in **Figure 8.13a**. Distal upstream and/or downstream enhancers and silencers as well as proximal (more neighboring) promoter elements regulate transcription. CCAAT box and a TATA box are promoter elements, with the TATA-box typically located 20–30 base pairs upstream of the transcription start site and the CCAAT box further to the 5' side. There are several kinds of exons:

1. Noncoding exons correspond to the untranslated 5' or 3' region of DNA.
2. Initial coding exons include the start methionine and continue to the first 5' splice junction.
3. Internal exons begin with a 3' splice site and continue to a 5' splice site.
4. Terminal exons proceed from a 3' splice site to a termination codon.
5. Single-exon genes are intronless, beginning with a start codon and ending with a stop codon (**Fig. 8.13b**; **Table 8.5**).

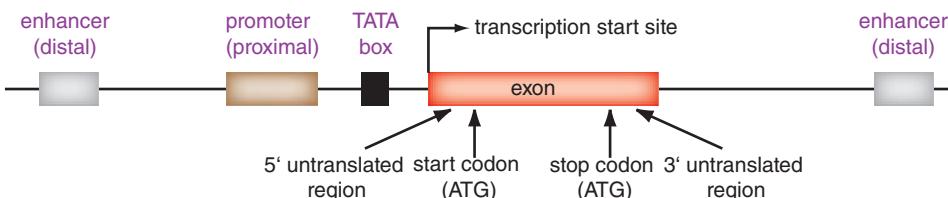
Introns have been categorized into four groups based on their splicing mechanism: (1) autocatalytic group I, found in protists, bacteria, and bacteriophages; (2) group II, found in fungal and land plant mitochondria and in bacteria and archaea; (3) spliceosomal introns, found in nuclear pre-mRNA genes; and (4) tRNA introns, found in eukaryotic nuclei and in archaea (Haugen *et al.*, 2005; Roy and Gilbert, 2006). Eukaryotic spliceosomal introns vary by two orders of magnitude in their density, from <0.1 to 5.5 introns per gene in fungi to 2.6–9.3 introns per gene in the metazoans (Roy, 2006). Fascinating questions include the mechanisms by which introns are gained and lost, the selective pressures on intron size, and their evolutionary history (Jeffares *et al.*, 2006; Pozzoli *et al.*, 2007). While introns were thought to have arisen late in eukaryotic evolution, a single intron was discovered in the genome of the primitive protozoan *Giardia lamblia* (see Chapter 19) as well as several introns in its close relative *Carpediemonas membranifera* (Nixon *et al.*, 2002; Simpson *et al.*, 2002).

In addition to the issue of introns, eukaryotic genes also occupy a far smaller proportion of the genome than bacterial and archaeal genes. Eukaryotic protein-coding exons occupy just 25% of nematode and insect genomes and less than 3% of human and mouse genomes. Chromosome 13 has the lowest protein-coding gene density (6.5 genes per megabase, with a region of 38 megabases having just 3.1 genes/Mb; Dunham *et al.*, 2004). Chromosome 19 has the highest gene density, with 26 loci per megabase (Grimwood *et al.*, 2004).

(a) Gene with multiple exons



(b) Single exon gene



**FIGURE 8.13** (a) Eukaryotic gene prediction algorithms differentiate several kinds of exons including: those in noncoding regions; initial coding exons that include a start codon; internal exons; and terminal exons that include a stop codon. These exons are built into a model for a predicted gene. (b) In some cases, genes have a single exon and are intronless. The border of exons and introns typically has a GT/AG boundary, but the structure of such genes may still difficult to predict *ab initio*.

Algorithms for finding protein-coding genes in eukaryotes can be divided into two main categories: extrinsic and intrinsic (Stein, 2001; Brent, 2008; Picardi and Pesole, 2010; Alioto, 2012). Extrinsic methods rely on comparisons to external data sources. These include RNA studies that map expressed sequence tags back to genomic loci, or studies of protein sequences to define gene structures. An additional form of extrinsic gene identification is to compare genomic DNA of two related organisms (Novichkov *et al.*, 2001; Morgenstern *et al.*, 2002). By comparing human DNA to pufferfish (*F. rubripes*) DNA, it was possible to discover nearly 1000 putative human genes (Aparicio *et al.*, 2002; Hedges and Kumar, 2002). Intrinsic (also called *ab initio*) methods search for exons and introns based on signals or patterns in the genomic DNA. Extrinsic and intrinsic methods are often used in combination.

The use of RNA data is extremely helpful in annotating eukaryotic genes, and is sometimes considered a gold standard for identifying genomic loci corresponding to exons (and therefore for annotating exon/intron borders). Until the advent of next-generation sequencing, this approach relied on cDNA libraries (introduced in Chapter 10); more recently, RNA-seq has been used. There are notable limitations to these approaches:

- The quality of EST sequence is sometimes low, as clones are often sequenced on only one strand and sequencing errors are common.
- Highly expressed genes are often disproportionately represented, although some cDNA libraries are normalized (Chapter 10).
- Transcripts expressed at low levels may be incompletely characterized by RNA-seq if there is not sufficient depth of coverage.
- ESTs provide no information regarding the genomic location.

Most human genes are alternatively spliced (Buratti *et al.*, 2013). If ESTs are available corresponding to alternatively spliced isoforms, these sequences can be mapped to exons.

Intrinsic programs are also widely used to annotate genomic DNA. A large fraction of predicted genes do not have identifiable orthologs, nor are EST sequences available. It is therefore essential to identify protein-coding genes using *ab initio* (intrinsic) approaches.

**TABLE 8.8 Algorithms for finding genes in eukaryotic DNA. Adapted from Picardi and Pesole (2010), with permission from Springer.**

Program	Description	URL
AAT	Analysis and Automation Tool	<a href="http://aatpackage.sourceforge.net/">http://aatpackage.sourceforge.net/</a>
ASPIC	Extrinsic. Web server	<a href="http://srv00.ibbe.cnr.it/ASPicDB/index.php">http://srv00.ibbe.cnr.it/ASPicDB/index.php</a>
AUGUSTUS	Extrinsic. University of Göttingen	<a href="http://bioinf.uni-greifswald.de/augustus/">http://bioinf.uni-greifswald.de/augustus/</a>
Eugène	Extrinsic	<a href="http://eugene.toulouse.inra.fr/">http://eugene.toulouse.inra.fr/</a>
Exogean	Extrinsic	<a href="http://www.biologie.ens.fr/dyogen/spip.php?rubrique4&amp;lang=fr">http://www.biologie.ens.fr/dyogen/spip.php?rubrique4&amp;lang=fr</a>
FgeneSH	Intrinsic. Ab initio gene finder	<a href="http://www.softberry.com/berry.phtml">http://www.softberry.com/berry.phtml</a>
GAZE	Combiner: extrinsic, intrinsic	<a href="http://www.sanger.ac.uk/resources/software/gaze/">http://www.sanger.ac.uk/resources/software/gaze/</a>
geneid	Intrinsic. Web server from Roderic Guigó	<a href="http://genome.crg.es/geneid.html">http://genome.crg.es/geneid.html</a>
GeneMark	Intrinsic. Georgia Institute of Technology	<a href="http://exon.gatech.edu/GeneMark/">http://exon.gatech.edu/GeneMark/</a>
GenomeScan	Extrinsic	<a href="http://genes.mit.edu/genomescan.html">http://genes.mit.edu/genomescan.html</a>
Genscan	Intrinsic. Based on HMMs	<a href="http://genes.mit.edu/GENSCANinfo.html">http://genes.mit.edu/GENSCANinfo.html</a>
GlimmerHMM	Intrinsic. Generalized HMM-based. From TIGR and the University of Maryland	<a href="http://cbsb.umd.edu/software/glimmerhmm/">http://cbsb.umd.edu/software/glimmerhmm/</a>
GRAILEXP	Extrinsic	<a href="http://compbio.ornl.gov/grailexp/">http://compbio.ornl.gov/grailexp/</a>
JIGSAW	Combiner: extrinsic, intrinsic	<a href="http://www.cbsb.umd.edu/software/jigsaw/">http://www.cbsb.umd.edu/software/jigsaw/</a>
Xpound	Intrinsic. A probabilistic model for detecting coding regions	<a href="http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::xpound">http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::xpound</a>

We discuss the GLIMMER program for annotating genes in bacteria and archaea in Chapter 17.

We obtain 50,000 base pairs of DNA from the beta globin region (chr11:5,245,001–5,295,000) of human genome build GRCh37/hg19 (available as Web Document 8.6). This region includes *HBB*, *HBD*, *HBBP1*, *HBG1*, *HBG2*, and *HBE1* according to RefSeq. (The GENCODE version 17 models for *HGB2* and *HBE1* extend hundreds of kilobases further to positions ~5,530,000 and ~5,670,000, respectively.) The GENSCAN server is available at <http://genes.mit.edu/GENSCAN.html> (WebLink 8.29). Its output is given in Web Document 8.7.

Many eukaryotic gene prediction programs are available; some are listed in Table 8.8. These programs typically produce models of gene structures (exons, introns, alternative splicing) and identify other features such as CpG islands (regions of a higher than expected occurrence of CpG dinucleotides over a particular distance such as 300 base pairs). Often these programs include searches with RepeatMasker to identify classes of repetitive DNA as well as BLAST or BLAST-like searches to identify known genes, proteins, and expressed sequence tags that help to model the gene structure.

As an example of an *ab initio* prediction tool we can run GENSCAN (Burge and Karlin, 1998) via a server, uploading 50 kilobases of DNA spanning the beta globin region on chromosome 11. The resulting annotation partially matches the known globin gene cluster, and includes a prediction of an exon in an intergenic region that has an open reading frame but without support of expressed sequence tag or other RNA-based data.

The difficulty of finding protein-coding genes in genomic DNA is illustrated by the efforts to annotate a typical eukaryotic genome: the *indica* and *japonica* subspecies of the rice genome. Yu *et al.* (2002) obtained 75,659 gene predictions when they submitted their assembled draft version of the rice genome (*indica*) to an FGeneSH web server. Only 53,398 of these predictions were complete (having both initial and terminal exons), about 7500 had only an initial exon, 11,000 had only a terminal exon, and 3400 predicted genes had neither. Additionally, they reported that exon–intron boundaries were often not precisely defined. However, when the finished sequence was obtained from the draft sequence, the estimate of gene content improved dramatically. Sasaki *et al.* (2002) obtained the finished sequence of rice chromosome 1 (subspecies *japonica*) and predicted 6756 genes on this chromosome. In contrast, the draft version of this genome predicted just 4467 genes. The presence of several thousand gaps in the draft sequence precluded the ability to accurately predict complete genes.

As another example of an approach to annotating genes, the *Drosophila* 12 Genomes Consortium (2007) reported the sequencing of ten *Drosophila* species yielding a total of

12 *Drosophila*-related genomes. The genomes were sequenced to varying depths, from over 10 $\times$  coverage to just 2.9 $\times$  coverage. They used four different *de novo* gene prediction algorithms, three homology-based predictors that relied on the well-annotated *Drosophila melanogaster* genome sequence, one predictor (called Gnomon) that combined *de novo* and homology-based evidence, and a gene model combiner (called GLEAN) that reconciled all the predicted genes into a set of consensus models. Quality was assessed in part by measuring RNA transcript levels with microarrays (Chapter 11).

## Finding Genes in Eukaryotic Genomes: EGASP Competition

The ENCODE Genome Annotation Assessment Project (EGASP) was a competition designed to objectively test the performance of a set of gene-finding software. The GENCODE consortium created a “gold standard” by rigorously mapping all the protein-coding genes with the ENCODE regions (Harrow *et al.*, 2006). This was achieved by carefully applying a range of experimental techniques such as 5’ rapid amplification of complementary DNA ends (RACE) and the polymerase chain reaction with reverse transcription (RT-PCR). A total of 434 coding loci were annotated as part of the GENCODE reference set. Only 40% of the GENCODE annotations were within the RefSeq and Ensembl annotation sets, reflecting the discovery of a large number of alternatively spliced isoforms with unique exons.

Given this deep level of annotation of ENCODE regions based on experimental evidence, the EGASP competition consisted of groups that predicted gene structures with the raw sequence data but without prior access to the annotation results (Guigo *et al.*, 2006; Harrow *et al.*, 2006). This allowed false positive and false negative error rates to be assessed. Sensitivity was defined as the proportion of annotated features (nucleotides, exons, or genes) that are predicted correctly, while specificity was defined as the proportion of predicted features that is annotated. The most successful gene prediction methods achieved a maximum sensitivity of 70% at the gene level (for finding at least one correct exon/intron structure), 45% at the transcript level (for correctly predicting all alternatively spliced variants), and 90% at the coding nucleotide level. Only about 3% of the many computationally predicted exons could be experimentally validated, suggesting that over-prediction remains a fundamental problem.

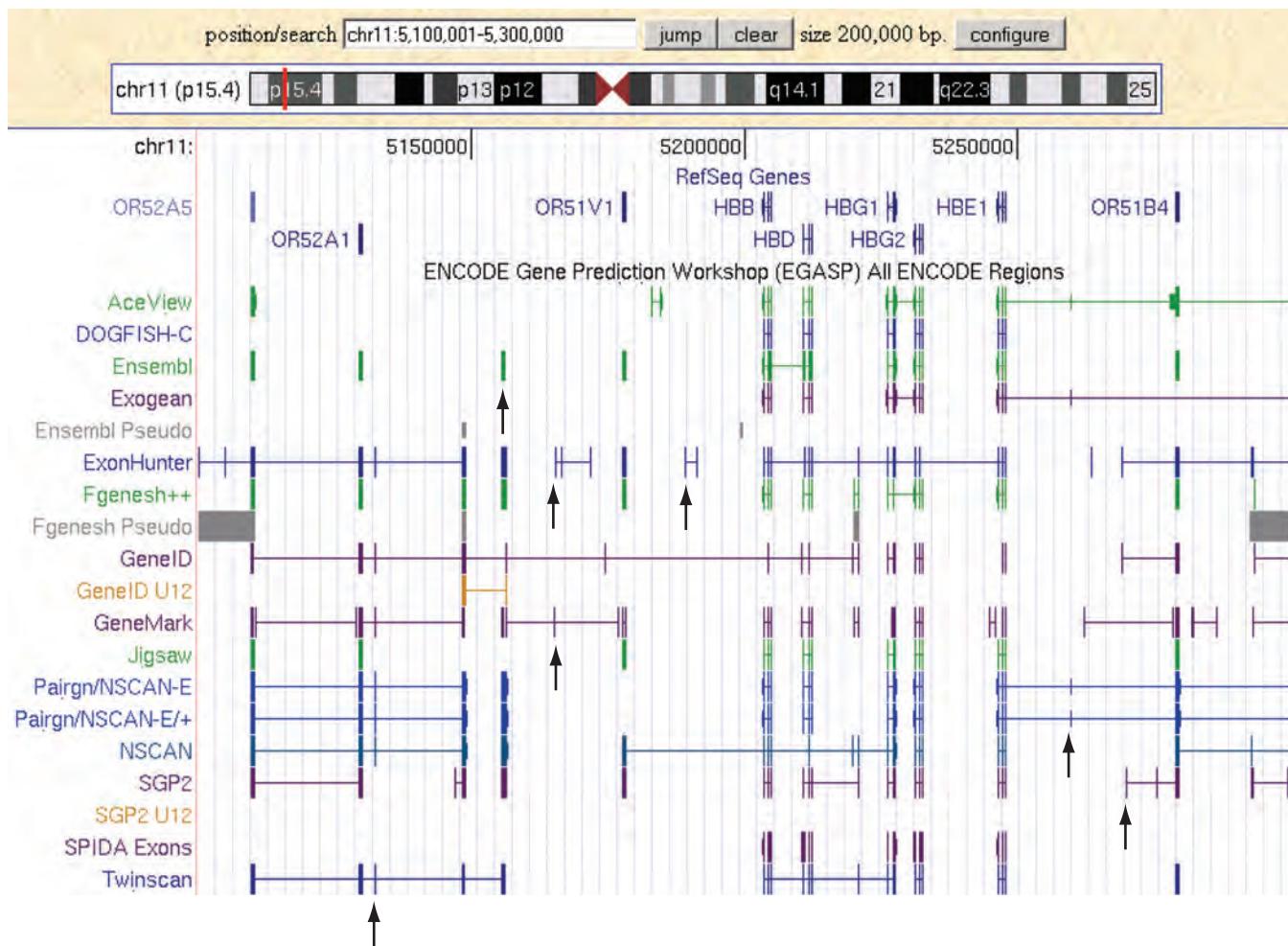
We can view the results of the EGASP competition at the UCSC Genome Browser website (**Fig. 8.14**). There is generally good agreement on the identification of exons, although there is considerable variation in the prediction of complete gene models.

One of the best-performing programs in the GENCODE competition was JIGSAW by Jonathan Allen, Steven Salzberg and colleagues (Allen and Salzberg, 2005; Allen *et al.*, 2006). JIGSAW is an integrative program that combines different sources of evidence into a model of a gene structure. It incorporates models from other gene prediction programs (typically three or more) as well as sequence alignment data and intron splice site prediction programs. It allows separate signal types including start codons, stop codons, and splice junctions (acceptor and donor sites at the 5’ and 3’ ends of introns). In one mode JIGSAW uses a linear combiner to assign a weight to each evidence source, and it maximizes the sum of the evidence (Allen *et al.*, 2004). This can be accomplished without using a training set. In another mode JIGSAW uses a statistical combiner which requires a training set (with examples of known genes) that are used to evaluate the accuracy of various combinations of evidence. Once a model is trained it is applied to a dataset.

For the EGASP competition, JIGSAW predictions were based on training with a variety of inputs including gene finders used by the UCSC annotation database (GENEID, SGP, TWINSCAN, and GENSCAN) as well as the GeneZilla and GlimmerHMM programs. It further incorporated expression evidence from human and nonhuman sources, GC percentage, sequence conservation, and a variety of genomic features such as TATA

We discuss other competitions for proteomics and protein structure (CASP) in Chapters 12 and 13, respectively. The GENCODE Project website is <http://genome.imim.es/gencode/> (WebLink 8.30), including a genome browser. The GENCODE team worked in collaboration with the Human And Vertebrate Analysis aNd Annotation (HAVANA) team at the Sanger Institute (<http://www.sanger.ac.uk/HGP/havana/>, WebLink 8.31).

JIGSAW can be downloaded from <http://cbsc.umd.edu/software/jigsaw/> (WebLink 8.32). Details on the JIGSAW method are presented in Web Document 8.8.



**FIGURE 8.14** In the EGASP competition, protein coding genes were experimentally validated in ENCODE regions. Various gene-finding software tools were used to independently predict gene structures. The beta globin ENCODE region consists of one million base pairs on human chromosome 11p. A portion of 200,000 base pairs is shown (x axis) with tracks for RefSeq genes and EGASP predictions from 19 software programs (y axis tracks). Many of the programs predict exons and/or entire gene structures that are not experimentally confirmed; examples are shown (arrows). Overfitting therefore remains a problem for prediction software. An even greater problem is that a complete, correct gene model is generated for fewer than half of all genes.

Source: ENCODE, courtesy of UCSC.

box and signal peptide sequences, intron phase, and CpG islands. Surprisingly, adding some categories of information (such as training on untranslated regions) diminished rather than improved accuracy (Allen *et al.*, 2006).

### Three Resources for Studying Protein-Coding Genes: RefSeq, UCSC Genes, GENCODE

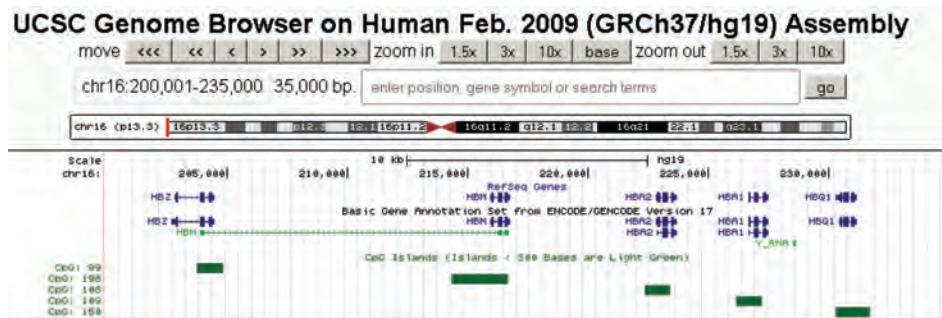
We can compare and contrast three major resources for protein-coding loci available at the UCSC Genome Browser:

1. We introduced the RefSeq project in Chapter 2.
2. The UCSC Genes set includes broader sources (e.g., RefSeq, GenBank, CCDS, Rfam) and includes 10% more genes than the RefSeq track at UCSC, four times as many noncoding genes, and twice as many splice variants.
3. The GENCODE project (Harrow *et al.*, 2012) includes high-quality manual annotation along with automated annotation.

These various tracks were shown in **Figure 8.7a** for the *HBB* locus. There are several notable differences. The 5' region of the *HBB* gene is longer in the Ensembl transcripts and GENCODE tracks than the RefSeq or CCDS tracks. One Ensembl transcript extends about 2 kilobases further to the 5' end than any other transcript. The UCSC Genes entry lists a single exon antisense gene of 23 base pairs (called DL074624) that is not annotated in the RefSeq collection and that lacks homologs in a variety of other species. Additional tracks for human mRNAs and human ESTs offer varying levels of support for these different models of the *HBB* gene. Another example of differences between GENCODE and RefSeq in the alpha globin gene cluster is provided in **Figure 8.15a**.

When gene models differ, a challenge is to know which is correct. Projects that rely on expert manual annotation often yield superior results to automated pipelines. In each particular case, it is however helpful to understand the experimental data that support each

(a) CpG islands in the human alpha globin gene cluster



(b) CpG island associated with *HBA1*

```
>chr16:226174-227254
CGTCCGGGTGCGCGCATTCTCTCCGCCAGGATTGGCGAACGCTCCGGCTCGCACT
CGCTCGCCCGTGTGTTCCCCGATCCCGCTGGAGTCGATGCGCTCCAGCGCTGCCAGGC
CGGGGCGGGGGTGCGGCTGACTTCTCCCTCGTAGGGACGCTCCGGCGCCGAAAGGA
AAGGGTGGCGCTGCGCTCGGGGTGCA CGAGCCGACAGCGCCGACCCCAA CGGGCCGGC
CCCGCCAGCGCCGCTACCGCCCTGCCCGGGCGAGCGGGATGGCGGGAGTGGAGTGGC
GGGTGGAGGGTGGAGACGTCTGGCCCCCGCCCCCGCTGCACTCCCCAGGGGAGGCCAGC
CCGCCGCCGGCCCCCGGCGAGGCCCGCCCGGGACTCCCCCTGCGTCCAGGCCGCC
GGGCTCCGGCCAGCCAATGAGCGCCGCCGGCGTGCCCCCGGCCAGCCAAAGCATA
AACCTCTGGCGCTCGCGACTCTGGTCCCCACAGACTCAGAGAGAACCCA
CCATGGTGTCTCTGGACAAGACCAACGTCAAGGCCGCGCTGGGTAAGGTGGCG
CGCACGCTGGCGAGTATGGTGGAGGCCCTGGAGAGGTGAGGCCCTCCCCTGCTCG
ACCCGGGCTCTCGCCCGCCCGAACGCCACAGGCCACCCCTCAACCGTCTGGCCCCGGACC
CAAACCCCACCCCTCACTCTGCTCTCCCCTGAGGATGTTCTGCTCTCCCCACCA
AGACCTACTTCCCGCACTTCTGAGGCCACGGCTCTGCCAGGGTTAACGGCCAAGGCA
AGAAGGTGGCGACCGCTGACCAACGCCCGTGGCGCACCGTGGAGACATGCCAACGCC
TGTCCGGCCCTGAGCGACCTGCA CGCGACAAGCTTGGGTGGAGGCCCGTCAACTTCAAGG
TGAGCGCGGGCGGGAGCGATCTGGGTGGAGGCCAGATGGCGCCTTCCTCGCAGGGC
AGAGGATCAAGGCCGGTTGCGGGAGGTGACGGCAGGCCCGGCTGGGCCCTGGGCCCTC
G
```

**FIGURE 8.15** CpG islands are associated with the regulation of expression of many eukaryotic genes. (a) The alpha globin gene cluster on human chromosome 16 is shown (in a window of 35,000 base pairs of chr16:200,001–235,000 on the UCSC Genome Browser). Each of the five genes has an associated CpG island, defined as having a GC content of 50% or greater, a length greater than 200 base pairs, and a ratio >0.6 of observed to expected CpG dinucleotides. (b) By clicking on the *HBA2* CpG island, its DNA sequence (chr16:222,370–223,447) is accessed. CpG dinucleotides are highlighted in pink.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

gene model, both in general (sensitivity and specificity of a computational or biochemical approach) and for a specific research question (e.g., you can view the evidence supporting particular gene models).

### Protein-Coding Genes in Eukaryotes: New Paradox

The *C* value paradox is answered based on the variable amounts of noncoding DNA in a variety of eukaryotes. A new paradox is introduced (Claverie, 2001; Betrán and Long, 2002): why are the proteomes of various eukaryotes similar in size, given the enormous phenotypic differences between eukaryotes? As we survey eukaryotic genomes in Chapters 18 and 19, we see that organisms such as worms and flies appear to have about 13,000–20,000 protein-coding genes, while plants, fish, mice, and humans have only slightly more (about 20,000–40,000 genes; Harrison *et al.*, 2002). Why do organisms such as humans, having so much greater biological complexity than insects and nematodes, have not even twice as many genes? The genes of higher eukaryotes employ more complex forms of gene regulation, such as alternative splicing. The architecture of individual genes also tends to be more complex, for example with more domains present in an average human protein relative to insect.

## REGULATORY REGIONS OF EUKARYOTIC CHROMOSOMES

### Databases of Genomic Regulatory Factors

The VISTA Enhancer Browser is available at <http://enhancer.lbl.gov> (WebLink 8.33).

In computer lab exercise (8.11) we use an R package to measure dinucleotide frequencies in genomic DNA.

In addition to predicting the presence of genes, it is also important to predict the presence of genomic DNA features such as promoters, enhancers, silencers, insulators, and locus control regions (Maston *et al.*, 2006; Pennacchio *et al.*, 2013). Such regulatory elements are sometimes called *cis*-regulatory modules (CRMs). Identifying them is difficult compared to finding protein-coding genes because the DNA sequences of interest may be very short (e.g., fewer than a dozen base pairs for transcription factor-binding sites), and conserved between species to variable extents. Algorithms are available for identifying regulatory elements, as well as databases storing compilations of genomic features. **Table 8.9** lists some of these resources, including software tools developed and used in association with ENCODE.

CpG islands represent an example of a regulatory element. The dinucleotide cytosine followed by guanosine (CpG) is approximately five-fold underrepresented in many genomes, partly because the cytosine residue can be exchanged for thymidine by spontaneous deamination. Cytosine residues on CpG dinucleotides are often methylated. This in turn leads to the recruitment of protein complexes that include histone deacetylases capable of removing acetyl groups of histones and therefore inhibiting active transcription. CpG islands are regions of high density of unmethylated CpG dinucleotides and are commonly found in upstream (5') regulatory regions near the transcription start sites of constitutively active “housekeeping” genes. By one criterion, a CpG island is defined as having a GC content  $\geq 50\%$ , a length  $\geq 200$  base pairs, and a ratio of observed to expected number of CpG dinucleotides of  $> 0.6$ . **Figure 8.15a** shows five CpG islands in the human alpha globin locus, visualized using the UCSC Genome Browser, each in the vicinity of an alpha globin gene. The extraordinarily dense number of CpG dinucleotides is evident in one of these islands (**Fig. 8.15b**).

The UCSC Genome Browser offers access to a wealth of additional resources. In the “Regulation” category of annotation tracks, several dozen tracks are available (**Fig. 8.16a**). Some of these elements are shown for a small region (15,000 base pairs) of the beta globin locus (**Fig. 8.16b**). For example, the Open REGulatory ANNOtation database (ORegAnno) compiles regulatory elements from the literature and includes a validation process by expert curators (Griffith *et al.*, 2008). Information in ORegAnno includes promoters,

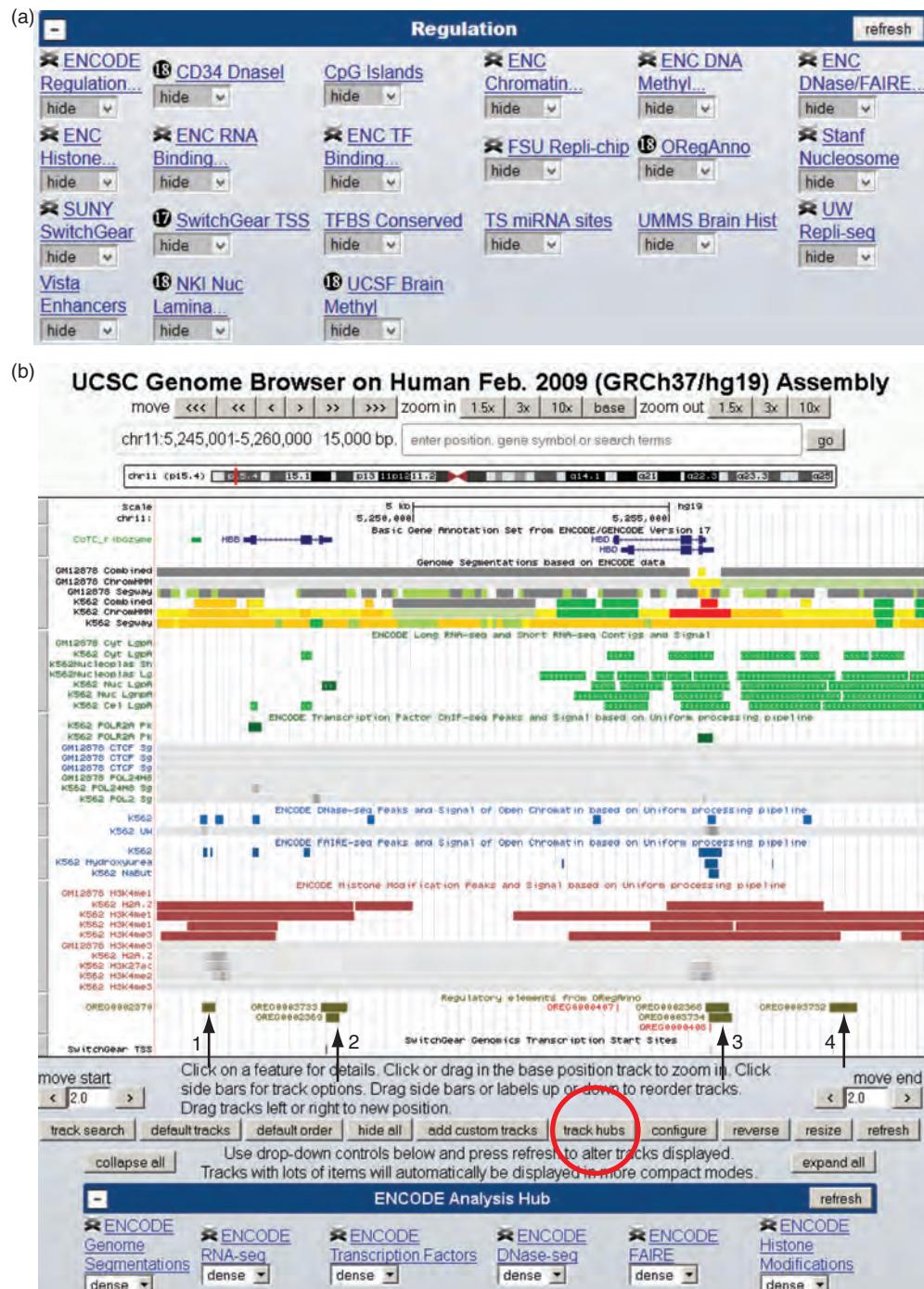
**TABLE 8.9 Software for identifying features of promoter regions in genomic DNA. Additional resources are summarized at <http://www.oreganno.org/oregano/OtherResources.jsp> (WebLink 8.50) and <http://www.gene-regulation.com/pub/programs.html> (WebLink 8.51). ENCODE software tools are described at <https://www.encodeproject.org/software> (WebLink 8.52).**

Program	Description	URL
AliBaba2	Predicts binding sites of transcription factor binding sites in an unknown DNA sequence	<a href="http://www.gene-regulation.com/pub/programs.html">http://www.gene-regulation.com/pub/programs.html</a>
ENCODE software: ENCODE-motifs	Database of transcription factors	<a href="http://www.broadinstitute.org/~pouyak/motif-disc/human/">http://www.broadinstitute.org/~pouyak/motif-disc/human/</a>
ENCODE software: Factorbook	Wiki-style resource for ChIP-Seq data on transcription factors	<a href="http://www.factorbook.org/mediawiki/index.php&gt;Welcome_to_factorbook">http://www.factorbook.org/mediawiki/index.php&gt;Welcome_to_factorbook</a>
ENCODE software: HaploReg	Tool to analyze haplotype blocks	<a href="http://www.broadinstitute.org/mammals/haploreg/haploreg.php">http://www.broadinstitute.org/mammals/haploreg/haploreg.php</a>
ENCODE software: RegulomeDB	Identifies DNA features and regulatory elements in noncoding regions	<a href="http://regulome.stanford.edu/">http://regulome.stanford.edu/</a>
ENCODE software: Spark	For epigenomic data	<a href="http://sparkinsight.org/">http://sparkinsight.org/</a>
Eukaryotic Promoter Database (EPD)	Annotated nonredundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally	<a href="http://epd.vital-it.ch/">http://epd.vital-it.ch/</a>
Open REGulatory ANNOtation database (ORegAnno)	Comprehensive, open access, community-based resource	<a href="http://www.oreganno.org">http://www.oreganno.org</a>
Promoter 2.0 Prediction Server	Technical University of Denmark	<a href="http://www.cbs.dtu.dk/services/promoter/">http://www.cbs.dtu.dk/services/promoter/</a>
Regulatory Sequence Analysis Tools (RSAT)	Université Libre de Bruxelles	<a href="http://rsat.ulb.ac.be/rsat/">http://rsat.ulb.ac.be/rsat/</a>
Transcriptional Regulatory Element Database (TRED)	Cold Spring Harbor Laboratory	<a href="http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home">http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home</a>
TRANSFAC	Database of transcription factors, their genomic binding sites, and DNA-binding profiles	<a href="http://www.gene-regulation.com/index2">http://www.gene-regulation.com/index2</a>

enhancers, transcription factor-bindings sites, and regulatory polymorphisms. As another example, the 7× regulatory potential track is based on regulatory potential scores computed from alignments of seven organisms (human, chimpanzee, rhesus macaque, mouse, rat, dog, and cow; King *et al.*, 2005; Taylor *et al.*, 2006). Scores are based on log ratios of transition probabilities from variable order Markov models, based on the use of a training set. Constrained (conserved) residues in a multiple sequence alignment may have regulatory potential if they are more similar to known regulatory elements than to ancestral repeats (which serve as a model for neutrally evolving DNA). King *et al.* evaluated regulatory regions of the beta globin locus, which includes 23 experimentally determined CRMs; all but three or four of these are conserved in rat and mouse, and just four are conserved in chicken. The regulatory potential method performed better (based on estimates of sensitivity and specificity) than other methods that rely exclusively on conservation of loci among species.

The UCSC Genome Browser options for the “Regulation” category include ENCODE tracks (Fig. 8.16a). Clicking on any of these headers provides access to track display features as well as the methodology and literature citations. You can further select track hubs (see Fig. 8.16b, red circle) to access a wealth of other datasets, including an ENCODE analysis hub. This provides access to data from hundreds of experiments, including genomic segmentation, RNA-seq, transcription factors, and histone modifications (Fig. 8.16b; Gerstein *et al.*, 2012; Wang *et al.*, 2012).

ORegAnno is available online at <http://www.oreganno.org> (WebLink 8.34). Web Document 8.9 lists definitions of several categories of regulatory elements within ORegAnno.



**FIGURE 8.16** Regulatory elements in genomic DNA. (a) The UCSC Genome Browser (GRCh37/hg19 assembly) includes two dozen annotation tracks in the “regulation” category, many of which include analyses of chromatin modifications. (b) The beta globin and delta globin gene loci are shown (15,000 bases at the location chr11:5,245,001–5,260,000) with some of these annotation tracks opened: ORegAnno (arrows 1–4 highlight regulatory elements) and SwitchGear (showing two transcriptional start sites). These tracks highlight regulatory elements surrounding these genes. Additionally, you can open track hubs (red circle) and open large amounts of ENCODE analysis data (see track options at the bottom of the figure). Many of these are opened, showing a variety of regulatory features.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

These options include chromatin immunoprecipitation sequencing (ChIP-seq) experiments, in which antisera directed against specific proteins (such as DNA-binding transcription factors) are used to immunoprecipitate those proteins with their target DNA. This DNA can be amplified and identified by next-generation sequencing. ChIP-seq data using a variety of approaches can be displayed on the UCSC Genome Browser.

Another set of data are from DNase I sensitivity experiments (Sabo *et al.*, 2006). DNase I hypersensitive sites reveal accessible genomic regions that are characteristic of active *cis*-regulatory sequences and transcription start sites in particular (John *et al.*, 2013). ENCODE findings include enriched interactions of exons with promoters and enhancers, and the cataloguing of ~21.9 million DNase I hypersensitive sites across 125 cell and tissue types (Thurman *et al.*, 2012; Mercer *et al.*, 2013).

## Ultraconserved Elements

Comparisons of eukaryotic genome sequences have revealed some highly conserved coding and noncoding DNA sequences. The Ensembl and UCSC Genome Browsers offer comparative genomics annotation tracks, including for conservation. A UCSC track shows the extent of conservation in up to 46 vertebrate species (including mammals, amphibians, birds, and fish) based on phastCons and PhyloP (described in Chapter 5).

Comparison of the human and *Fugu rubripes* genomes that last shared a common ancestor about 450 million years ago revealed many ultraconserved sequences (also called highly conserved elements). Ultraconserved elements are sometimes defined as having a length  $\geq 200$  base pairs that match identically with corresponding regions of the human, mouse, and rat genomes. Bejerano *et al.* (2004) identified 481 such segments, most of which were also highly conserved with the dog and chicken genomes. Many of these elements are distant from any protein-coding gene. These regions are highly constrained evolutionarily (Katzman *et al.*, 2007). Dermitzakis *et al.* (2003) also described ultraconserved sequences on human chromosome 21. In a computer laboratory exercise (8.9), we identify a series of DNA sequences that share 100% nucleotide identity between human and chicken (species that last shared a common ancestor over 300 million years ago).

UCNEbase is a database of ultraconserved elements (Dimitrieva and Bucher, 2013). You can browse these conserved elements and link to view them in the UCSC Genome Browser (including a track with the Bejerano *et al.*, 2004 data).

It might be expected that ultraconserved elements have important functions such that they are so highly conserved under negative selection. Some ultraconserved elements drive tissue-specific expression, while non-exonic ultraconserved elements are depleted in regions of segmental duplications and copy number variants (Chiang *et al.*, 2008). McLean and Bejerano (2008) find that mammalian non-exonic conserved elements are over 300-fold more likely to be conserved during rodent evolution relative to neutral DNA.

These findings on the importance of conserved elements contrast with the results of earlier studies investigating the consequence of deleting them. For example, Nóbrega *et al.* (2004) deleted two large noncoding regions from the mouse genome (consisting of 1511 and 845 kilobases) and created viable homozygous deletion mice. They detected no altered phenotype (and only very minor differences in the expression of neighboring genes). These deletion regions harbored over 1200 noncoding sequences conserved between humans and rodents. It is possible that, under some physiological conditions, the deletions would have large phenotypic consequences; nonetheless, this study suggests that large portions of chromosomal DNA are potentially dispensable.

You can access UCNEbase at  
🌐 <http://ccg.vital-it.ch/UCNEbase/>  
(WebLink 8.35).

## Nonconserved Elements

In analyzing regulatory regions of genomic DNA, a focus has been on identifying conserved noncoding regions as candidates for functionally important loci. Fisher *et al.* (2006) studied regulatory regions near the *RET* gene in zebrafish, and used a transgenic assay to identify a series of teleost sequences that direct ret-specific reporter gene expression. Surprisingly, a series of human noncoding sequences were also able to drive zebrafish gene expression, even though there was no detectable conservation between the human and zebrafish sequences. This highlights how little we understand about transcription factor binding, and suggests that vast amounts of functionally important regulatory sequences are not detectable based on sequence conservation (Elgar, 2006). The ENCODE perspective is that functionally important loci (based on biochemical assays) tend not to be conserved (ENCODE Project Consortium *et al.*, 2007, 2012). Andrew McCallion, Ivan Ovcharenko and colleagues expanded the Fisher *et al.* (2006) findings to show that human/zebrafish regions can retain common regulatory functions in the absence of sequence detectable conservation. Conversely, many highly conserved regions have not been shown to have biochemical function (reviewed in Stamatoyannopoulos, 2012).

Synteny derives from Greek roots meaning “same thread” or “same ribbon.” A common error is to refer to orthologous genes as being syntenic, when instead they share *conserved synteny* (Passarge *et al.*, 1999).

PipMaker and MultiPipMaker are available at <http://pipmaker.bx.psu.edu/pipmaker/> (WebLink 8.36). (“Pip” stands for “percent identity plot.”) VISTA (Visualization Tools for Alignments) is at <http://genome.lbl.gov/vista/index.shtml> (WebLink 8.37). mVISTA (main VISTA) is a program for visualizing genomic alignments, while rVISTA (regulatory VISTA) is used to align transcription factor binding sites. AVID is an alignment algorithm used by the VISTA tools (Bray *et al.*, 2003). A VISTA browser is online at <http://pipeline.lbl.gov/cgi-bin/gateway2> (WebLink 8.38); a typical output is shown in Figure 8.17. This allows human–mouse, mouse–rat, and human–rat genomic DNA comparisons. VISTA also offers a browser for enhancer elements (<http://enhancer.lbl.gov/>; WebLink 8.39).

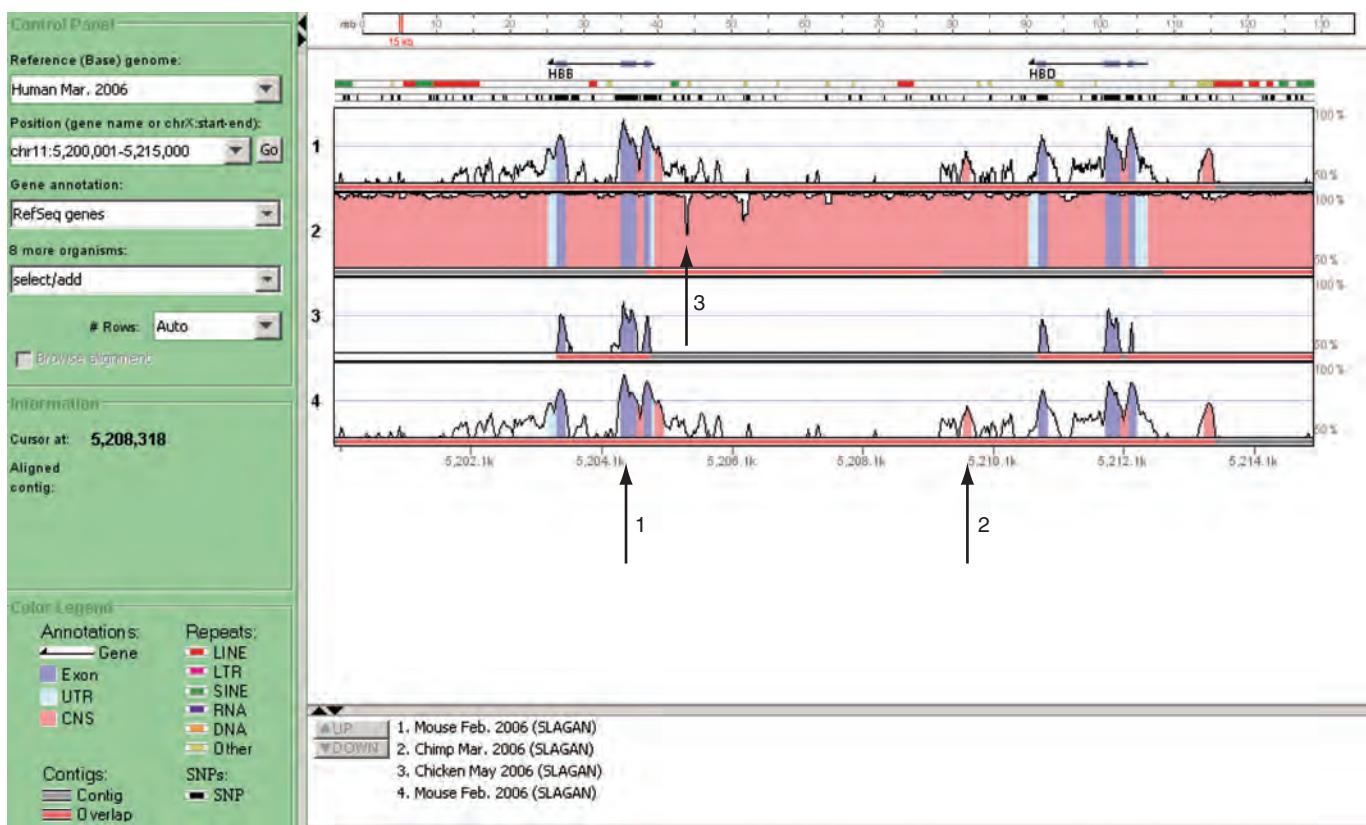
## COMPARISON OF EUKARYOTIC DNA

Comparative genomics is a powerful approach to annotating and interpreting the meaning of genomic DNA from multiple organisms. When we analyze the genomes of organisms that diverged recently (e.g., humans and chimpanzees diverged at 5 MYA) or in the distant past (e.g., mosquitoes and fruit flies diverged at 250 MYA; Zdobnov *et al.*, 2002), it is helpful to align the genomic sequences in order to define conserved regions. Such analyses can provide a wealth of information about the existence and evolution of protein-coding genes and other DNA features, as well as information about chromosomal evolution.

Genes from different organisms that are derived from a common ancestor are orthologs (Chapter 3). In comparing genomic sequences from two (or more) organisms, we may wish to analyze regions in each species having orthologous genes. Such regions are said to have conserved synteny. Synteny denotes the occurrence of two or more gene loci on the same chromosome, regardless of whether or not they are genetically linked. This definition refers to an arrangement of genes along a chromosome within a single species. “Conserved synteny” refers to the occurrence of orthologous genes (i.e., in two species) that are syntenic. As an example, the occurrence of the neighboring genes *RBP4* and *CYP26A1* on human chromosome 10 and mouse chromosome 19 represents conserved synteny.

In order to analyze regions of conserved synteny – or even larger regions of genomic DNA that do not necessarily contain protein-coding genes – it is necessary to perform pairwise alignment and multiple sequence alignment of genomic DNA. We discuss approaches to this for bacteria and archaea in Chapter 17, and in Chapter 5 we discussed algorithms that are useful for the comparison of large DNA queries to databases containing genomic DNA including PatternHunter, BLASTZ, MegaBLAST, BLAT, LAGAN, and EPO.

There are other powerful tools for the comparison of genomic DNA in eukaryotes including PipMaker (Schwartz *et al.*, 2000), VISTA (Mayor *et al.*, 2000; reviewed in Frazer *et al.*, 2003) and MUMmer (Kurtz *et al.*, 2004; see Chapters 16 and 17). The goal of each program is to align long sequences (e.g., thousands to millions of base pairs) while visualizing conserved segments (exons and presumed regulatory regions) as well as large-scale genomic changes (inversions, rearrangements, duplications). It is important to learn both the order and orientation of conserved sequence features. The VISTA browser output for human chromosome 11, including the beta globin and delta globin genes, is



**FIGURE 8.17** The VISTA program for aligning genomic DNA sequences is available through a web browser that can be queried with text or DNA sequence (up to 300,000 bases). The output for a query of the human beta and delta globin gene region is shown here. The x axis shows the nucleotide position along human chromosome 11, and the y axis shows the percent nucleotide identity between human and chimpanzee, mouse, and chicken. A variety of exons (e.g., arrow 1) and conserved noncoding sequences (e.g., arrow 2) are shown. Human and chimpanzee have nearly identical sequences, but divergent regions are easily seen (e.g., arrow 3). By clicking a link (not shown), VISTA data can be output on a version of the UCSC Genome Browser.

Source: VISTA. <http://genome.lbl.gov/vista/index.shtml>. Courtesy of VISTA.

shown in **Figure 8.17**. This includes an alignment to the chimpanzee, mouse, and chicken genomes, highlighting conserved exons and conserved noncoding regions.

## VARIATION IN CHROMOSOMAL DNA

We might think of chromosomes as unchanging entities that define the genome of each species. However, they are dynamic in many ways across large timescales (millions of years), between generations, between individuals in a population, and even with individual lifetimes. A broad variety of cytogenetic changes occur in eukaryotes, allowing an assessment of different types, mechanisms, and consequences of rearrangement (Coghlan *et al.*, 2005).

### Dynamic Nature of Chromosomes: Whole-Genome Duplication

When we compare the genomes of related species, we can observe many types of chromosomal changes. One level is ploidy. In eukaryotes, normal germ cells are haploid while somatic cells are usually diploid. Different cells within an individual can therefore have different ploidy. Ploidy is the number of chromosome sets in a cell, and can vary in many ways. Some single-celled eukaryotes such as *S. cerevisiae* can grow in either the haploid

or diploid state. Triploid *Drosophila* are viable (but with reduced fertility). Although we distinguish the ploidy state in germ cells and somatic cells, ploidy can also vary in somatic cells within an individual. For example, in humans a small fraction of liver cells is typically triploid. In general, an extra germline copy of even one chromosome is lethal in mammals.

One of the dramatic ways that ploidy can change for an entire species is through whole-genome duplication. Mechanistically, a mitotic or meiotic error may cause diploid gametes to form, having two sets of chromosomes. These may fuse with haploid gametes to form triploid zygotes which are unstable but may lead to the formation of stable tetraploid zygotes. When whole-genome duplication occurs within a species, the result is termed autopolyploidy. Such a massive event happened in yeast; in Chapter 18 we review evidence for whole-genome duplication and computational tools to analyze and visualize it. A variety of protozoan, plant, and fish genomes also underwent whole-genome duplication. In the case of the ciliate *Paramecium tetraurelia*, analysis of the genome sequence suggests that there have been at least three whole-genome duplication events (Aury *et al.*, 2006; Chapter 19).

The genomes of two distinct species may merge to generate a novel species (allopolyploidy; Hall *et al.*, 2002). This phenomenon has been described in many plants (Comai, 2000), animals, and fungi. For example, the plant *Arabidopsis suecica* derives from the *A. thaliana* and *Cardaminopsis aerenosa* genomes (Lee and Chen, 2001; Lewis and Pikaard, 2001). Another example of allopolyploidy is the mule, which is the result of a cross between a male donkey (*Equus asinus*,  $2n = 62$ ) and a female horse (*Equus caballus*,  $2n = 64$ ). Mules cannot propagate because they are sterile (they cannot produce functional haploid gametes; see Ohno, 1970).

Ohno (1970) hypothesized that the increased complexity of vertebrates is due to two rounds of whole-genome duplication in early vertebrate evolution. This has been called the 2R hypothesis (reviewed in Dehal and Boore, 2005; Panopoulou and Poustka, 2005). Ohno argued that duplication provided the genetic material to be shaped by mutation and selection to introduce novel functions to organisms (Prince and Pickett, 2002; Taylor and Raes, 2004). There are three advantages of becoming polyploid (Comai, 2005). (1) Hybrids sometimes exhibit an increase in performance relative to their inbred parents, a phenomenon termed heterosis. (2) Gene redundancy occurs, offering the opportunity to mask recessive deleterious alleles by dominant wildtype alleles. Also, one member of a duplicated gene pair may be silenced, up- or down-regulated in its expression level, or regulated in a tissue-specific manner (Adams and Wendel, 2005; Li *et al.*, 2005). The most common fate of duplicated genes is that they become deleted as has been shown in fungi (discussed in Chapter 18), the plants *Arabidopsis thaliana* and *Oryza sativa* (Thomas *et al.*, 2006), and fish (Brunet *et al.*, 2006; Paterson *et al.*, 2006). (3) Self-fertilization may become possible (asexual reproduction).

Another type of chromosomal change that can be fixed in a species is the fusion of two chromosomes. For example, acrocentric chromosomes may be subject to Robertsonian translocation, in which two centromeres fuse (Sljepcevic, 1998). Human chromosome 2, the second largest human chromosome, is derived from two ancestral great ape acrocentric chromosomes (chimpanzee chromosomes 2a and 2b, formerly named 12 and 13; Ijdo *et al.*, 1991; Fan *et al.*, 2002; Martin *et al.*, 2002). The human 2q13 band, near the centromere, contains telomeric repeats in a head-to-head orientation. Over 50 interstitial telomeres have been described (Azzalin *et al.*, 2001; Lin and Yan, 2008).

In addition to fusion, chromosomes can split (fission). As an example, human chromosomes 3 and 21 derive from a larger ancestral chromosome (Muzny *et al.*, 2006). Chromosomal inversions represent another change that can lead to speciation. There are five distinct subtypes of the mosquito *Anopheles gambiae* having varying kinds of paracentric inversions on chromosome 2 (Holt *et al.*, 2002; Ayala and Coluzzi, 2005;

*Paramecium tetraurelia* is exceptional because most of its duplicated genes have not been deleted (Aury *et al.*, 2006; Chapter 19).

In human trisomy 21 (Down syndrome), it is not uncommon for a copy of chromosome 21 to fuse with another acrocentric chromosome.

Nwakanma *et al.*, 2013), and these inversions may lead to speciation by preventing successful chromosomal pairing among members of different subtypes.

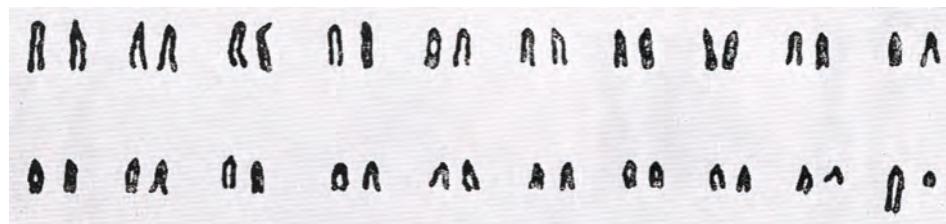
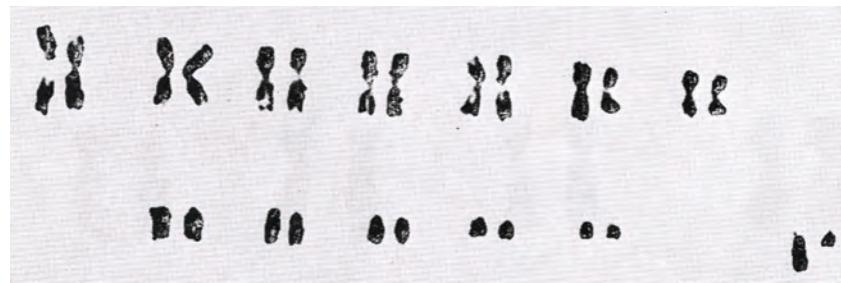
The recent availability of sequences from hundreds of eukaryotic genomes has led to the reconstruction of many ancestral genomes. For example, Kohn *et al.* (2006) described the eutherian karyotype from 100 million years ago, prior to the radiation of mammalian species. Murphy *et al.* (2005) compared the chromosomal architecture of eight species (human, horse, cat, dog, pig, cattle, rat, and mouse) and inferred the structure of their ancestral chromosomes. They characterized the sites of evolutionary breakages, which included subtelomeric and pericentromeric regions in particular.

Large-scale chromosomal changes may lead to the establishment of a new species (speciation). Susumu Ohno (1970) provided an example. The karyotypes of the tobacco mouse *Mus poschiavinus* ( $2n = 26$ ) and the house mouse *Mus musculus* ( $2n = 40$ ) are shown (Fig. 8.18a, b). The ancestral *M. poschiavinus* may have become physically isolated from *M. musculus* and was therefore not able to interbreed. At this time its chromosomes underwent Robertsonian translocations, thus forming a new genome with a reduced number of chromosomes. The F1 progeny form a series of seven trivalents (each from one *poschiavinus* metacentric and two *musculus* acrocentrics; Fig. 8.18c) which are not compatible with survival.

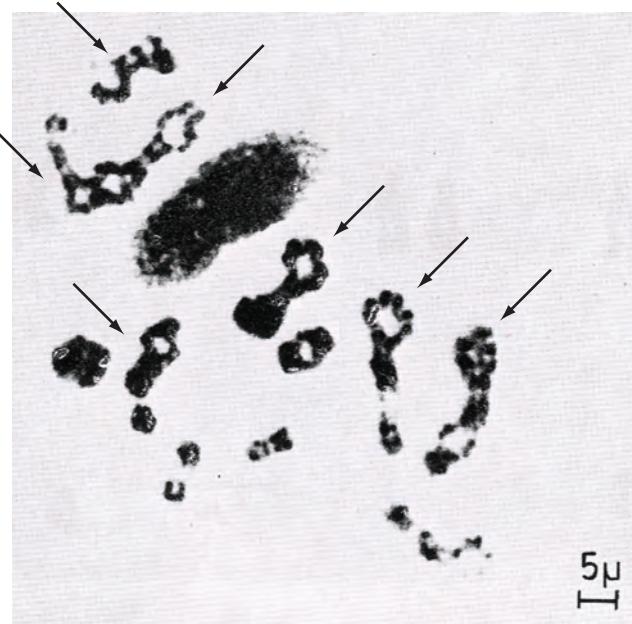
### Chromosomal Variation in Individual Genomes

A comparison of closely related species has revealed many chromosomal changes involving single chromosomes. At the level of the individual organism, many changes to chromosomes occur, sometimes causing disease.

- An individual may acquire an extra copy of an entire chromosome. For example, Down syndrome is caused by a trisomy (triplicated copy) of chromosome 21 (Fig. 8.1b). We discuss this type of disorder in Chapter 21. Aneuploidy (the presence of an abnormal number of chromosomal copies) occurs commonly and is often caused by nondisjunction (Hassold and Hunt, 2001).
- Uniparental disomy may occur, in which both homologous chromosomes are inherited from one parent. We discuss this in more detail in “SNP Microarrays” below. Uniparental disomy is often associated with disease in humans (Kotzot, 2001, 2008).
- A portion of a chromosome may be deleted. Deletions may be terminal or interstitial; an example of a terminal deletion of chromosome 11q is shown in Figure 8.1a (arrow B).
- Segmental duplications commonly occur (introduced above; see also Chapter 20).
- Normal chromosomes from any eukaryotic species can vary between individuals in length, number, and position of heterochromatic segments. For example, the ribosomal DNA repeat segments on the short arms of the five human acrocentric chromosomes vary greatly in length between individuals. A variety of human chromosomes show tremendous polymorphisms in the population, such as portions of chromosome 7 (Chapter 20).
- Fragile sites often occur, sometimes causing chromosomal breaks (Debatisse *et al.*, 2012). These fragile sites can be inherited in a dominant Mendelian fashion.
- Some eukaryotes display chromatin diminution, a form of developmentally programmed DNA rearrangement. Remarkably, chromosomes in somatic cells can fragment, then lose some chromosomal material. Somatic chromosomes can therefore have a different structural organization and a smaller gene number than germline cells. Chromatin diminution could represent an unusual gene-silencing mechanism (Müller and Tobler, 2000). This phenomenon has been observed in at least 10 nematode species, including the horse intestinal parasite *Parascaris univalens* (also called *Ascaris megalocephala*) and the hog parasite *Ascaris suum*.

(a) Ordinary male house mouse (*Mus musculus*,  $2n = 40$ )(b) Male tobacco mouse (*Mus poschiavinus*,  $2n = 26$ )

(c) Male first meiotic metaphase from an interspecific F1-hybrid



**FIGURE 8.18** Robertsonian fusion creates one metacentric chromosome by the fusion of two acrocentrics. (a) Karyotype of the normal mouse, *Mus musculus* ( $2n = 40$ ). (b) Karyotype of the male tobacco mouse (*Mus poschiavinus*,  $2n = 26$ ). Its smaller chromosome number derives from Robertsonian fusion events. (c) Male first meiotic metaphase from an interspecific F1-hybrid. Note seven trivalents (indicated with arrows). Each represents one *poschiavinus* metacentric and two *musculus* acrocentrics.

Source: Ohno (1970). Reproduced with permission from Springer Science + Business Media.

Among the many functional changes that chromosomes undergo, dosage compensation of the X chromosome is a prominent example. In human females, one copy of each X chromosome is functionally inactivated through the action of an X-chromosome inactivation center (XCI; Latham, 2005). Genomic imprinting, the selective silencing of either maternal or paternal copies of genes, is another regulatory mechanism (Morison *et al.*, 2005).

## Structural Variation: Six Types

Structural variation (SV) consists of genomic alterations in DNA copy number, orientation or location (Hall and Quinlan, 2012; Liu *et al.*, 2012). The sizes of structural variants are typically defined as greater than 1 kilobase, or >100 base pairs for insertions/deletions (indels). Six main categories of structural variation are: (1) insertions; (2) deletions; (3) tandem duplications; (4) inversions; (5) translocations; and (6) complex structural variants. A structural variation track from UCSC, based on the Database of Genomic Variants (Iafrate *et al.*, 2004) currently includes structural variation data collected from over 50 publications. These include genomic insertions (e.g., duplications), deletions, inversions, translocations, and other complex rearrangements (Hall and Quinlan, 2012).

Visit the DGV at <http://dgv.tcag.ca/dgv/app/home> (WebLink 8.40).

### *Inversions*

A.H. Sturtevant (1921), a student of Thomas Hunt Morgan, mapped a series of genes and reported that *Drosophila simulans* has an inversion on chromosome III relative to *Drosophila melanogaster*. This example highlights another feature of chromosomal plasticity: while 500 unique inversions are known in *D. melanogaster* (a highly polymorphic species), only 14 unique inversions are known in *D. simulans* (a monomorphic species; Aulard *et al.*, 2004). Different species present varying propensities to undergo chromosomal changes.

In humans and other species, inversions occur commonly. They can be extraordinarily difficult to detect because even DNA sequencing may not reveal changes, and they may be undetectable using conventional cytogenetics. Stefansson *et al.* (2005) described an inversion polymorphism of 900 kilobases that occurs on chromosome 17q21.31 (from 44.1 to 45.0 Mb). This inversion is common in Europeans where it is under positive selective pressure. Surprisingly, the inverted segment occurs in chromosomes having different orientations in two lineages (H1 and H2) which diverged as long as 3 million years ago. As another example, an inversion of a single gene causes a severe form of hemophilia (Antonarakis *et al.*, 1995).

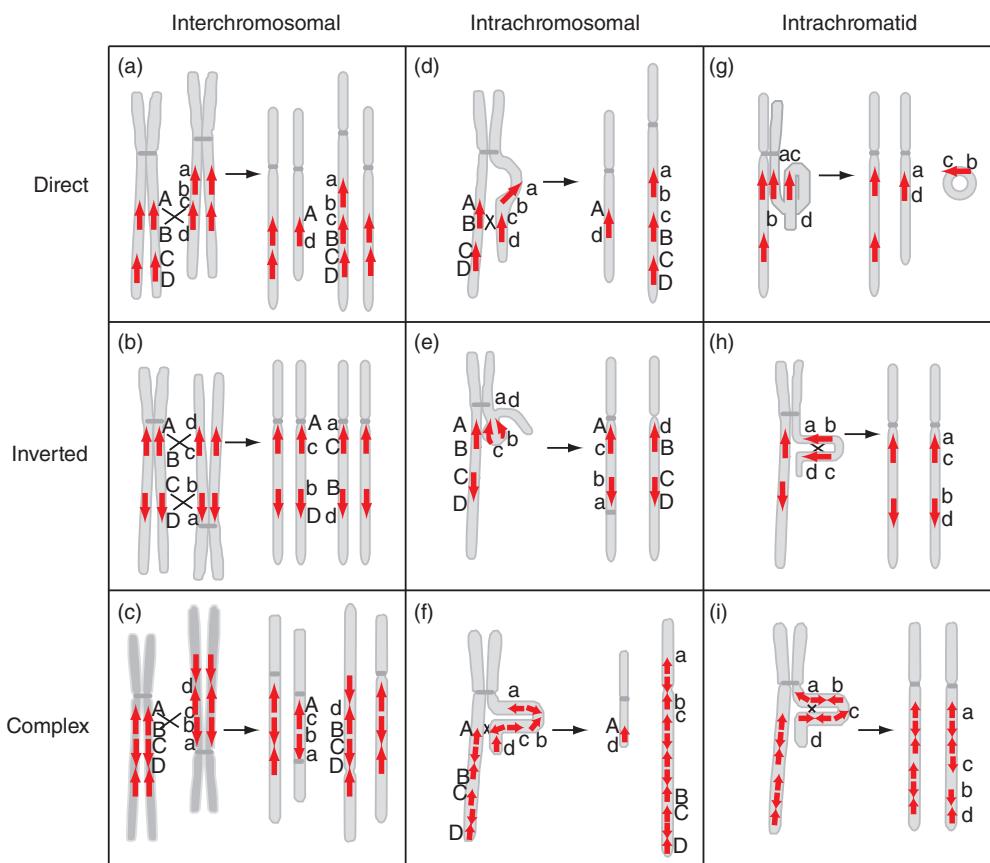
In an innovative approach, Pavel Pevzner and colleagues have used small inversions as evolutionary characters to perform phylogenetic analyses (Chaisson *et al.*, 2006). They estimate that one microinversion occurs per megabase per 66 million years of evolution, and they developed a method to distinguish microinversions (local alignments between orthologous sequences on the reverse strand) from palindromes and inverted repeats. This method is limited to analysis of sequences with sufficient conservation to permit clear assignment of orthology, but its phylogenetic reconstruction matches traditional approaches.

You can read about this hemophilia at the Online Mendelian Inheritance in Man (OMIM) site at NCBI (entry 306700). We describe OMIM in Chapter 21.

### *Mechanisms of Creating Duplications, Deletions, and Inversions*

In the first half of the twentieth century, a variety of detailed models were proposed to explain how genes become duplicated, deleted, or inverted (Darlington, 1932). A major current model is non-allelic homologous recombination mediated by low-copy repeats (i.e., by segmental duplications; Stankiewicz and Lupski, 2002; Bailey and Eichler, 2006). Repetitive DNA of about 10–50 kilobases that occur in two (or more) distinct chromosomal loci can lead to unequal crossing over (Fig. 8.19). These cross-overs can occur intrachromosomally, intrachromosomally, or between sister chromatids (Fig. 8.19, columns). The orientation of the low-copy repeats influences the nature of the rearrangement that occurs; these repeats may occur in a direct orientation, they may be inverted repeats, or they may have a complex structure (Fig. 8.19, rows).

We examine the case of direct repeats in Figure 8.19a. The phrase “non-allelic homologous recombination” refers to meiotic recombination between chromosomes. One



**FIGURE 8.19** Mechanisms of creating genomic rearrangements. Non-allelic homologous recombination (NAHR) based on low-copy repeats (LCRs) or segmental duplications cause these changes. The orientation of the LCRs may be head-to-head (top row), head-to-toe (middle row), or complex (bottom row) involving DNA exchanges that are interchromosomal (left column), intrachromosomal (middle column), or intrachromatid (right column). For each of the nine scenarios the chromosomal configuration is shown as well as the products of unequal crossing over. (a) Unequal cross-overs between directly ordered repeats lead to a duplication and a deletion. (b) Mechanism of forming an inversion. (c) Interchromosomal exchange between inverted repeats causes inversions and can result in duplications and deletions. (d) Mispairing of direct repeats leads to an intrachromosomal deletion/duplication. (e) An inversion results from intrachromosomal unequal exchange between inverted repeats. (f) Complex repeats lead to an intrachromosomal deletion/duplication. (g) A deletion and an acentric fragment result from intrachromatid mispairing due to direct low-copy repeats. (h) An intrachromatid loop of inverted repeats results in an inversion. (i) Complex repeats lead to intrachromatid mispairing and an inversion. Redrawn from Stankiewicz and Lupski (2002) with permission from Elsevier.

chromosome has repeats labeled AB and CD, while the other has ab and cd. The repeats can combine even when they are non-allelic (for example, AB and ab are allelic but AB and CD are non-allelic). Nonetheless, they are homologous and therefore able to pair. Following the cross-over event indicated by the X in **Figure 8.19a**, one copy contains ab cB CD and therefore has a duplication, while the other copy has Ad from the cross-over event and therefore has a deletion.

As indicated in **Figure 8.19**, many other products can result from unequal exchanges. In this way, segmental duplications (low-copy repeats) have been a major force in shaping genome evolution, including the emergence of gene families. In Chapter 21 we present six models by which deletions (or duplications or inversions) may cause disease (Lupski and Stankiewicz, 2005). In other cases the genomic rearrangements, such as altering the

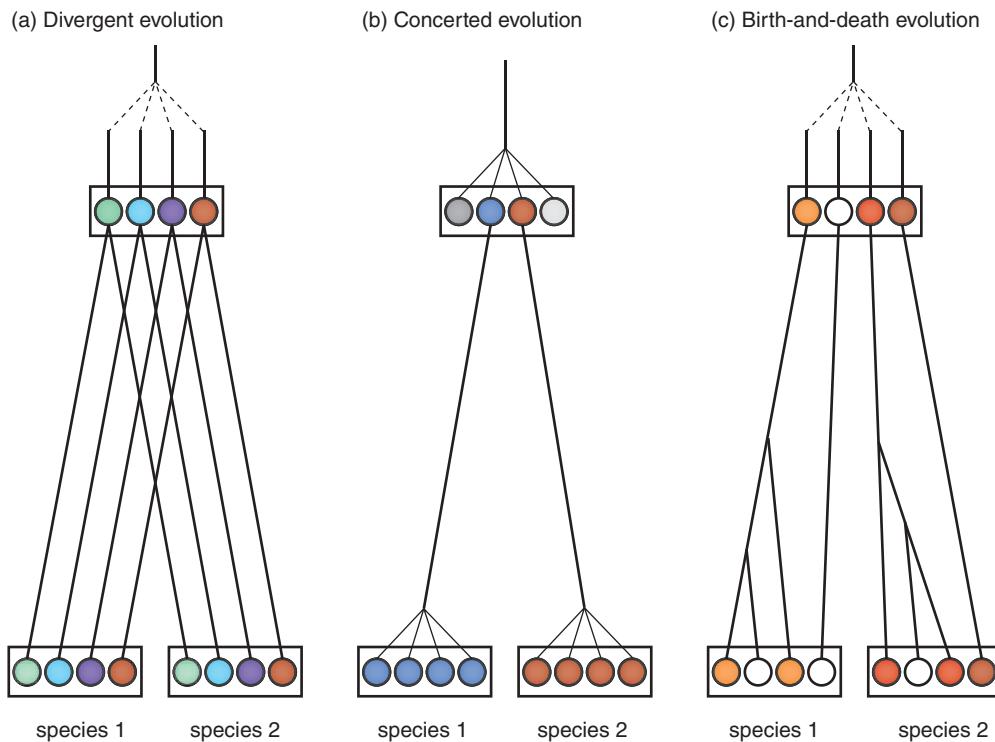
dosage of a gene or fusing two genes together, may present an organism with an innovation that is advantageous and selected for.

The boundaries of segmentally duplicated regions often contain *Alu* repetitive sequences (Bailey and Eichler, 2006). Pericentromeric and subtelomeric regions are also enriched for segmental duplications, with interchromosomal segmental duplications present in 30 out of 42 subtelomeric regions (reviewed in Bailey and Eichler, 2006).

#### **Models for Creating Gene Families**

One prominent aspect of genomes is the occurrence of multigene families. Multigene families (also called superfamilies) consist of a group of paralogs such as the globins. Nei and Rooney (2005) reviewed this topic and described three separate models for their evolution.

1. According to a divergent evolution model, members of a gene family gradually diverge as duplicate genes assume new functions (**Fig. 8.20a**). For example, the alpha and beta globin groups each have multiple members as shown in the phylogenetic tree of **Figure 3.3**. Some of these globins are expressed at specific developmental stages.
2. According to a concerted evolution model, all the members of a gene family evolve in a concerted manner rather than independently (**Fig. 8.20b**). An example of this scenario is the tandemly repeated ribosomal DNA genes. We describe the structure of human rDNA repeats in Chapter 10 (**Fig. 10.7**). Work by Donald Brown and others showed that intergenic regions of ribosomal DNA clusters were more similar within a species than between two related *Xenopus* (frog)



**FIGURE 8.20** Three models for the creation of duplicate genes in multigene families: (a) divergent evolution; (b) concerted evolution; and (c) birth-and-death evolution. Open circles refer to functional genes; closed circles correspond to pseudogenes.

Source: Nei and Rooney (2005). Reproduced with permission from Annual Reviews.

The DNA and protein RefSeq accession numbers for *hsp70Aa* are NM\_169441 and NP\_731651, while for *hsp70Ab* they are NM\_080059 and NP\_524798.

Web Document 8.10 lists some of the gene families cited by Nei and Rooney.

The HapMap website is <http://www.hapmap.org> (WebLink 8.41). Currently (May 2015) dbSNP build 42 includes ~113 million human RefSNPs; see [http://www.ncbi.nlm.nih.gov/ SNP/snp\\_summary.cgi](http://www.ncbi.nlm.nih.gov/ SNP/snp_summary.cgi) (WebLink 8.42).

species. When one member of such a gene cluster acquires a mutation, that change spreads to other members. One mechanism by which this can occur is unequal crossing over. Another proposed mechanism is gene conversion. In gene conversion, one gene (or other DNA element) serves as a donor and, through a form of nonreciprocal recombination, it mediates the conversion of a second gene to form a copy of the first gene. Examples of gene families that have evolved by concerted evolution include the primate U2 snRNA genes, 5S RNA genes in *Xenopus* (which has 9000 to 24,000 members) or humans (which has ~500 members), and heatshock protein genes in *Drosophila*. The *hsp70Aa* and *hsp70Ab* genes are a pair of inverted tandem repeats that are virtually identical in *D. melanogaster* as well as in *D. simulans*. Their within-species identity could provide an example of gene conversion.

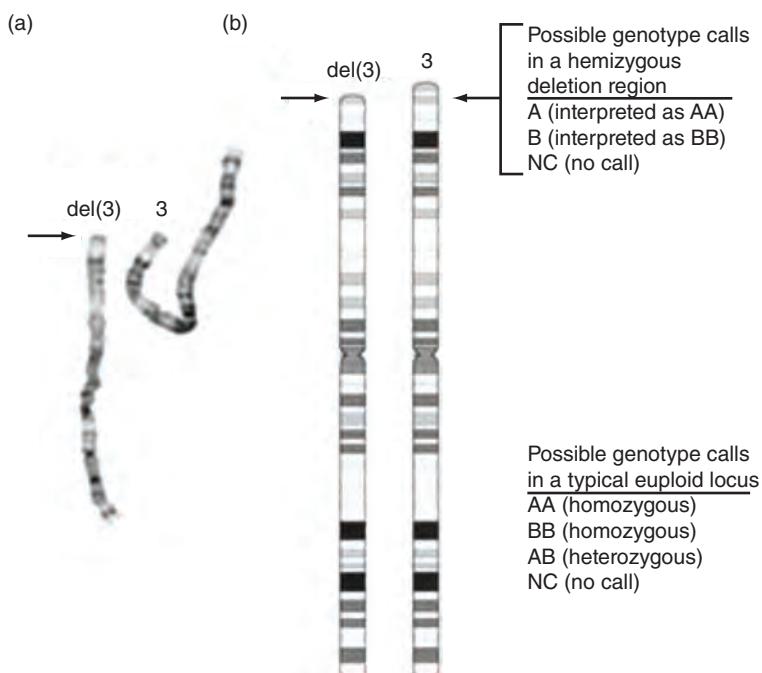
3. A birth-and-death evolution model was proposed by Masatoshi Nei and others (reviewed in Nei and Rooney, 2005; Fig. 8.20c). According to this model, new genes are created by gene duplication. Some duplicates remain in the genome, while others are inactivated (becoming pseudogenes) or deleted. This model was proposed to explain the evolution of the major histocompatibility complex (MHC) genes. MHC proteins bind foreign or self peptides and present them to T-lymphocytes as part of the immune response. MHC class I genes in particular are highly polymorphic due to positive selection on the peptide-binding region (Hughes and Nei, 1989). The birth-and-death model presents a mechanism for the generation of gene diversity that is distinct from concerted evolution or divergent evolution, and explains how new functions can be acquired by duplicate genes.

According to Nei and Rooney, most gene families are subject to birth-and-death evolution. In some cases such as histone genes and the ubiquitins, the birth-and-death process is accompanied by very strong purifying selection that conserves the protein sequences. This selective pressure, rather than the homogenizing properties of gene conversion or unequal cross-over, accounts for the tremendous conservation of these proteins. In other cases a mixed process of concerted evolution and birth-and-death evolution occurs, such as in the alpha globin genes in which *HBA1* and *HBA2* genes encode identical proteins, possibly because of gene conversion.

### Chromosomal Variation in Individual Genomes: SNPs

SNPs represent one of the most commonly occurring forms of variation in all genomes. Figure 8.21 shows an example of two SNPs from the beta globin gene at the Entrez database of SNPs (dbSNP) at NCBI. By convention, each of the variants (C or G in these two cases) is represented as A or B for the major and minor alleles in the population. Most SNPs are biallelic (i.e., there are two rather than three or four variants at a given position) with a range of population frequencies. Possible genotype calls for a diploid sample (such as human) are AA or BB (homozygous) or AB (heterozygous). In regions of hemizygous deletion (where one of two chromosomal copies are deleted), or on the male X chromosome which is by its nature hemizygous, the genotypes are A or B but should never be heterozygous (Fig 8.21.).

The HapMap project was created to identify SNPs in the human genome. It resulted in the determination of over three million SNPs (International HapMap Consortium, 2005, 2007). This resource, available through a HapMap database, initially focused on genotyping of four diverse populations (from northern Europe, Africa, Japan, and China). The SNP data are useful to describe variation between and within populations, including the structure of shared alleles (haplotypes), to characterize recombination rates and the evolution of both nonsynonymous and synonymous SNPs in coding regions.



**FIGURE 8.21** SNP microarray experiments provide information about chromosomal copy number (based on the intensity of hybridization) and genotype (based on alleles detected at each SNP position). (a) Karyotype of chromosome 3 from a patient with a hemizygous deletion (i.e., loss of a portion of one of the two chromosomal copies). The deletion region is indicated with an arrow. (b) Ideogram of chromosome 3. Throughout most of the chromosome there are four possible genotype calls: AA or BB (homozygous calls), AB (heterozygous), or NC (no call). In the deletion region there are three possible calls: an underlying state of A (interpreted by current software packages as a biallelic call, AA), B (interpreted as BB), or no call. There can be no AB calls (unless there is a technical failure). Some software packages detect stretches of homozygous SNPs which, in the presence of a reduced copy number, corresponds to a hemizygous deletion. Note that the human male X chromosome is by its nature hemizygous, and no AB calls are expected other than those that represent genotyping errors.

## TECHNIQUES TO MEASURE CHROMOSOMAL CHANGE

For several decades, karyotyping has been the pre-eminent technique to visualize chromosomes. Today, clinical genetics laboratories routinely use karyotyping to assess the occurrence of aneuploidy as well as smaller changes such as microdeletions and micro-duplications. Typically, deletions that are smaller than about 3 million base pairs are too small to detect. Chromosomal inversions can only be detected if they are large enough to disrupt the banding pattern. Translocations may be balanced (if two chromosomal regions exchange) or unbalanced (if material is gained or lost).

Fluorescence *in situ* hybridization (FISH) offers greatly increased resolution. A bacterial artificial chromosome (BAC) clone, typically consisting of about 200,000 base pairs of genomic DNA inserted into a cloning vector of about 10,000 base pairs, can be labeled with a fluorescent dye then used to probe a spread of metaphase chromosomes on a microscope slide. FISH has been used to refine information about chromosomal anomalies such as microdeletions and translocations.

In 1992 Kallioniemi *et al.* performed comparative genome hybridization (CGH) in which genomic DNA from two samples (such as one diseased and one apparently normal) is isolated, labeled with a green or red fluorescent dye, and hybridized to a normal chromosomal spread. This technique showed regions of gain or loss of DNA sequences, including amplifications seen in tumor cell lines.

## Array Comparative Genomic Hybridization

Array CGH (aCGH) is a high-throughput extension of the CGH technique to microarrays that is useful to detect copy number changes at defined chromosomal loci. It combines the high resolution of FISH with the broad chromosome-wide perspective of karyotyping. An aCGH platform may consist of thousands of BAC clones or oligonucleotides immobilized on the surface of a glass microarray. Genomic DNA is purified from a test sample (e.g., the DNA is isolated from a cell line or blood sample) and a reference sample. If the test sample DNA is labeled with a red dye and the reference is labeled with a green dye, then upon hybridization the signal intensities are comparable. If an amplification or deletion occurs, the log signal intensities deviate from a value of zero. The region of copy number gain or loss may be as small as one single probe (e.g., one base pair for a SNP array, or about 200,000 base pairs for a BAC array). The change may also extend across an entire chromosome arm or entire chromosome. **Figure 8.22** shows an example of a microdeletion on chromosome 2. This resulted in the hemizygous loss of many genes, and intellectual disability in the patient.

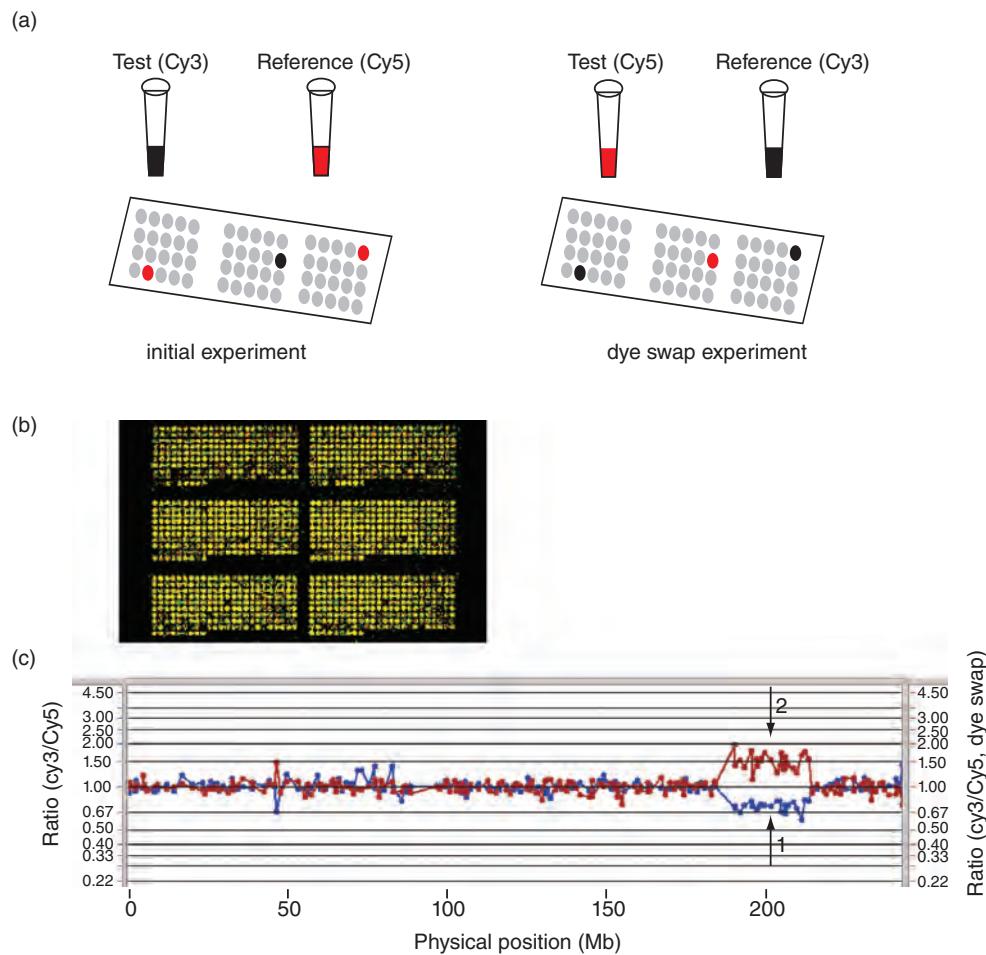
A simple approach has been to apply a ratio threshold to define a region of amplification or deletion. For a gain of one copy, the amount of signal is expected to increase 1.5-fold (from 2 copies in the euploid state to 3 copies), while a hemizygous deletion reduces the copy number two-fold (from 2 copies to 1). On a  $\log_2$  scale, unchanged copy number corresponds to a value of 0 (i.e., a 1:1 ratio), while a gain and loss correspond to +1 and -1  $\log_2$  intensity values, respectively.

Many statistical approaches have been developed to analyze aCGH data. Two estimation problems must be addressed: inferring the number of chromosomal alterations and their statistical significance, and locating the boundaries of such events. Lai *et al.* (2005) tested the accuracy of a group of 11 algorithms. Their comparative study included receiver operating characteristic (ROC) curves, plotting the false positive rate versus the true positive rate. For many test datasets, the 11 algorithms produced dramatically different estimates of copy number changes. The algorithms were all better at detecting large-scale aberrations with a good signal to noise ratio, but faltered with smaller aberrations and noisy data. Some algorithms did not detect particular amplifications or deletions; others either merged a group of alterations or splintered them inappropriately. Overall, one of the best-performing algorithms in the Lai *et al.* (2005) study and another comparative study (Willenbrock and Fridlyand, 2005) was the circular binary segmentation method (CBS; Olshen *et al.*, 2004; Venkatraman and Olshen, 2007). This method divides the genome into regions of equal copy number, assuming that chromosomal gains or losses occur in discrete, contiguous regions. The goal is to identify copy number change-points which partition the chromosome into segments. A likelihood ratio statistic tests the null hypothesis that there is no change against the alternative hypothesis that there is one change at a given location. The null hypothesis is rejected if the test statistic exceeds some threshold; the variance can be estimated from the data by Monte Carlo simulations using a permuted reference distribution.

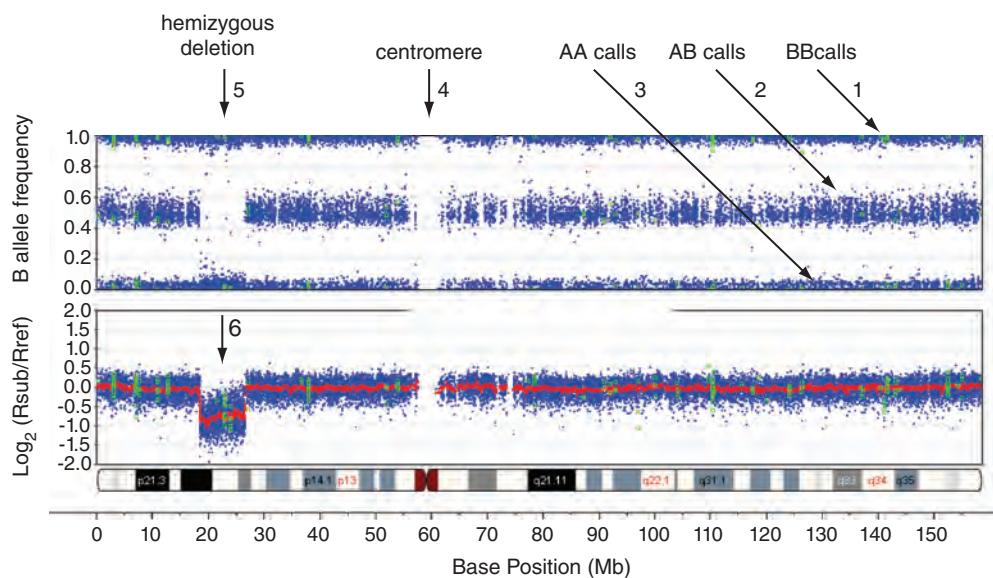
aCGH is one of the techniques that has been used to discover copy number variants (CNVs) in the human genome. There is an astonishing amount of variation between even apparently normal individuals, with large numbers of megabase-sized deletions and duplications. We address this topic in Chapter 20.

## SNP Microarrays

There are many applications of SNPs including mapping polymorphisms in genes and genomes, selecting markers to identify individuals having alleles of interest in large segregating populations, and finding association between genomic regions and segregating



**FIGURE 8.22** Array comparative genome hybridization (aCGH) allows detections of chromosomal gains and losses. (a) Experimental design. Genomic DNA is isolated from a test sample (e.g., from a patient) and a reference (e.g., from a pool of apparently normal controls). The DNA is fragmented then labeled with differently colored fluorescent dyes such as Cy3 and Cy5. In a parallel dye swap experiment, the test and reference samples are labeled with opposite dyes. The samples are coincubated on a microscope slide containing up to tens of thousands of bacterial artificial chromosome (BAC) clones, each of which typically spans 200,000 base pairs and has known chromosomal position. More recently arrays contain densely packed oligonucleotides, although this figure depicts BAC clones. Following hybridization, washing, and image analysis, most BACs on the array have a comparable amount of Cy3 and Cy5 dye (indicated as gray spots on the two slides). A deletion in the test sample is associated with relatively more Cy5 dye in the reference; see the two red spots in the slide at the left. In the dye swap, these two spots appear black, providing an independent validation. An amplification in the test sample results in relatively more Cy3 dye (see the black spot in the slide at left, which appears red in the dye swap experiment to the right). (b) Example of an aCGH image from a scanner. The output includes a spreadsheet that includes quantities of the signal intensities in the Cy3 and Cy5 channels for each BAC clone. (c) Example of the output for chromosome 2. The *x* axis corresponds to chromosome 2 (from the *p* terminus to the *q* terminus). The *y* axis corresponds to the Cy3/Cy5 ratio from the initial experiment and from the dye swap. There are therefore two sets of data points that are superimposed. The test sample is from a patient who has a deletion of about 23 megabases (from 190.5 to 213.8 Mb in chromosome 2q32.2–q34). This deletion is evident as a reduced signal intensity ratio across a group of adjacent BACs (arrow 1). As expected, the dye swap experiment shows a mirror image deviation (arrow 2).



**FIGURE 8.23** SNP profile from chromosome 7 in a patient with a hemizygous deletion. The upper panel shows the B allele frequency from thousands of SNPs across the chromosome, including BB calls (B allele frequency near 1.0; arrow 1), heterozygous AB calls (arrow 2), or homozygous AA calls (arrow 3). In some heterochromatic regions such as the centromere (arrow 4), there are no SNPs and the plot therefore lacks data points. In the region of a hemizygous deletion on 7p (arrow 5), there are essentially no AB calls. The lower panel shows the intensity values corresponding to chromosomal copy number. The y axis is  $\log_2(R_{\text{sub}}/R_{\text{ref}})$ , corresponding to the  $\log_2$  ratio of the intensity value for the subject (i.e., this patient sample) to the intensity values for a reference set (such as mean intensity values for a large set of apparently normal individuals).  $\log_2(R_{\text{sub}}/R_{\text{ref}})$  tends to have a value near 0.0 (the subject and reference data therefore have a one-to-one correspondence), but in the deletion region the  $\log_2$  value is  $-1.0$  (see arrow 6). In regions of homozygous deletion (i.e., two copies deleted; not shown), the  $\log_2$  value tends to be close to  $-5.0$ . In cases of trisomy (not shown), the extra copy causes the B allele frequency to split into four tracks (corresponding to AAA, AAB, ABB, and BBB genotypes) and the intensity values are elevated. Data are from an Illumina microarray with 550,000 SNPs.

traits (Chapter 20). A basic application is to measure chromosomal changes in genomic DNA samples. Several technologies exist to measure vast numbers of SNPs on microarrays, such as a single-base extension strategy from Illumina and an oligonucleotide-based hybridization strategy from Affymetrix. An example of a SNP dataset using the Illumina platform is shown in **Figure 8.23**. The experiment provides information on chromosomal copy number (based on hybridization intensity measurements) and genotype (based on AA, AB, or BB calls). There is a characteristic profile for hemizygous deletions with a lack of heterozygous SNPs.

SNP arrays can provide information on a variety of chromosomal changes beyond those detectable by aCGH or conventional cytogenetics. An example is uniparental disomy in which both homologous chromosomes are inherited from one parent. The term disomy refers to two copies, as opposed to zero (nullsomy), one (monosomy), three (trisomy), or four (tetrasomy). There are two copies of each chromosome, as usual, but the two copies of a single chromosome are derived from just one parent (uniparental disomy). Since each parent has two copies of a given autosome, the result may be uniparental heterodisomy (in which the two copies derived from the mother or the father are different) or uniparental isodisomy (in which the two copies are identical). This is also associated with disease in humans (Kotzot, 2001). SNP arrays can show regions of homozygosity without copy number change. In the absence of copy number change, the cause can be uniparental disomy (Ting *et al.*, 2007).

## Next-Generation Sequencing

It is now possible to sequence an individual's entire genome (whole-genome sequencing) or collection of exons (whole-exome sequencing) at a relatively modest cost. This approach offers the following advantages to karyotyping, aCGH, and SNP arrays:

- WGS (and WES for coding regions) allow us to assay all alleles. In contrast, almost all SNP arrays focus on common variants; WGS and WES are therefore far more comprehensive.
- Unlike SNP arrays and aCGH, WGS and WES provide nucleotide sequences including information about single-nucleotide variants (SNVs, synonymous with SNPs), short insertions/deletion events (indels), structural variants, and heterozygosity.

There are also several notable disadvantages:

- Using WGS (and WES) it is extremely difficult to detect balanced and unbalanced chromosomal translocations that are easily detectable by cytogenetics.
- WGS and WES are less appropriate for the detection of megabase-scale variants such as aneuploidy. By analogy, if you want to understand traffic patterns in a city a traffic helicopter can provide a useful overview of major events (aCGH, SNP arrays) while a street view of every street and alley in the city provides such dense information that it can be hard to see the big picture (WGS, WES).

Some researchers therefore combine next-generation sequencing with other technologies such as SNP arrays.

## PERSPECTIVE

One of the broadest goals of biology is to understand the nature of each species of life: what are the mechanisms of development, metabolism, homeostasis, reproduction, and behavior? Sequencing of a genome does not answer these questions directly. Instead, we must first try to annotate the genome sequence in order to estimate its contents, and then we try to interpret the function of these parts in a variety of physiological and evolutionary processes.

The genome sequences of representative species from all major eukaryotic divisions are now becoming available. This will have dramatic implications for all aspects of eukaryotic biology. For studies of evolution we will further understand mutation and selection, the forces that shape genome evolution.

As complete genomes are sequenced, we are becoming aware of the nature of non-coding and coding DNA. Major portions of the eukaryotic genomic landscape are occupied by repetitive DNA, including transposable elements. The number of protein-coding genes varies from about 2000 in fungi to tens of thousands in plants and mammals. Many of these protein-coding genes are paralogous within each species, such that the “core proteome” size is likely to be on the order of 10,000 genes for many eukaryotes. New proteins are invented in evolution through expansions of gene families or through the use of novel combinations of DNA encoding protein domains.

## PITFALLS

A tremendous need in genomics research is the continued development of algorithms to find protein-coding genes, noncoding RNAs, repetitive sequences, duplicated blocks of sequence within genomes, and conserved syntentic regions shared between genomes. We may then characterize gene function in different developmental stages, body regions, and physiological states. Through these approaches we may generate and test hypotheses about the function, evolution, and biological adaptations of eukaryotes, and hence extract meaning from the genomic data.

We are now in the earliest years of the field of genomics. Many new lessons are emerging:

- Draft versions of genome sequences are extremely useful resources, but gene annotation is an ongoing process that often improves dramatically as a sequence becomes finished.
- It is extraordinarily difficult to predict the presence of protein-coding genes in genomic DNA. This is especially true in the absence of complementary experimental data on gene expression, such as expressed sequence tag information.
- We know little about the nature of noncoding RNA molecules.
- Large portions of eukaryotic genomes consist of repetitive DNA elements. Segmental duplications offer a creative evolutionary opportunity to shuffle DNA within and between chromosomes.
- Comparative genomics is extraordinarily useful in defining the features of each eukaryotic genome.

Understandably, there is great enthusiasm for sequencing of thousands of eukaryotic genomes. However, a concern is that next-generation sequence technology (introduced in Chapter 9) relies on relatively short reads (often using libraries with insert sizes up to 500 base pairs). Alkan *et al.* (2011) compared *de novo* assemblies of several human genomes, based on short read technologies, to deeply characterized references of those genomes. They reported that the *de novo* assemblies were ~16% shorter than the references and were missing >99.1% of validated duplicated sequences (with >2300 coding exons missing). This chapter introduced the repetitive DNA content and other features of eukaryotic chromosomes that must be accounted for in analyzing genome sequences.

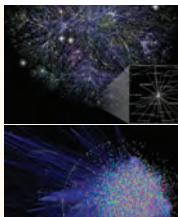
## ADVICE TO STUDENTS

There are so many approaches to the study of chromosomes and DNA. Think about the problem you are trying to solve in order to choose the appropriate analysis route. I have three suggestions for how to approach the material we covered in this chapter. (1) Delve deeper into the Ensembl genome browser, the UCSC Genome and Table Browser, or both. (2) Become familiar with Galaxy. For example, beginning in Galaxy, import data from the UCSC Table Browser or from BioMart, and analyze it to study a set of problems (e.g., “how many microRNAs are annotated within exons?”). (3) We introduced the R packages `biostings` and `biomaRT`; try to become familiar with these to learn how R packages work and explore their functionality. Then browse the Bioconductor website to get a sense of the hundreds of other R packages that are available.

The ENCODE project has enhanced our understanding of chromosomal features including genes and their regulation. One way to begin immersing yourself in this project is to read some of the hundreds of ENCODE studies, starting with the *Nature* portal. It helps if you can begin with a specific question of interest, or begin with beta globin. Try exploring the ENCODE hub and other tracks at the UCSC Genome Browser.

## WEB RESOURCES

We have presented key resources for many eukaryotic organisms and their genome-sequencing websites. An excellent starting point is the Ensembl website (<http://www.ensembl.org/>, WebLink 8.43), which currently includes gateways for the mouse, rat, zebrafish, fugu, mosquito, and other genomes. The UCSC Genome Browser includes an excellent user guide (<http://genome.ucsc.edu/training/index.html>, WebLink 8.44) and many other training resources. A major gateway for ENCODE resources is at the National Human Genome Research Institute (<http://www.genome.gov/10005107>, WebLink 8.45). Another useful gateway is at the journal *Nature* (<http://www.nature.com/encode/>, WebLink 8.46).



# Discussion Questions

**[8-1]** If there were no repetitive DNA of any kind, how would the genomes of various eukaryotes (human, mouse, a plant, a parasite) compare in terms of size, gene content, gene order, nucleotide composition, or other features?

**[8-2]** If someone gave you 1 Mb of genomic DNA sequence from a eukaryote, how could you identify the species? (Assume you cannot use BLAST to directly identify the species.) What features distinguish the genomic DNA sequence of a protozoan parasite from an insect or a fish?

**[8-3]** The computer lab problems below include the use of websites and R. For which problems can you use either approach, and for which do you strongly need to use either web resources or R but not the other?

## **PROBLEMS/COMPUTER LAB**

**[8-1]** This problem encourages you to explore the UCSC Table Browser via Galaxy. How many microsatellites are there in the human genome, and which one in the Table Browser is longest? (1) Visit the UCSC Table Browser. For the group “Variation and Repeats” set the region to “Genome,” the output to BED file, and send the query to Galaxy (with one BED record per whole gene). (2) In Galaxy, use Tools > Text Manipulation > Compute an expression on every row. Subtract c3–c2 (the end position minus the start position with rounding the result). A new column (c5) is generated. (3) Under Tools > Filter and Sort > Sort in descending order based on column c5. The largest satellite element (chr19:43,167,386–43,167,883) extends almost 500 base pairs, all consisting of a repeating pattern of AT residues.

**[8-2]** This problem uses R to search for patterns in DNA. The first 15 nucleotides of the beta globin coding sequence (on human chromosome 11; chr11:5,248,237–5,248,251 of GRCh37/hg19) are 5'-ATGGTGCATCTGACT-3'. (This gene is transcribed on the bottom strand, so the top strand sequence is 5'-AGTCAGATGCACCAT-3'). How often does that pattern occur on chromosome 11, and where? A more detailed version of this exercise is provided as Web Document 8.11. (1) Install R and RStudio (as above) and set your working directory to your favorite folder. Then download the human genome reference sequence and install the biostrings package.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("BSgenome.Hsapiens.UCSC.hg19")
# this may take about 30 minutes to download.
> biocLite("Biostrings")
```

There are therefore two matches. Does the second match correspond to a different globin gene? What happens if you repeat the search, and tolerate up to two mismatches? Try this:

```
> matchPattern(mypattern15rev, Hsapiens$chr11,  
+               max.mismatch=2)
```

**[8-3]** The purpose of this problem is to obtain a typical dataset (in this case, a table of repetitive DNA elements found in a 70,000 base pair region of the globin locus) and plot the results in R. (1) Go to the UCSC Table Browser. Set the genome and build to human GRCh37/hg19; use group “Variation and Repeats,” track “RepeatMasker,” position chr11:5,230,001–5,300,000, output format “select fields from selected table,” output filename ucsc\_chr11\_repeats.txt, and click “get output.” When prompted, select the following fields: swScore (Smith Waterman alignment score), genoStart and genoEnd (genomic positions), strand (+ or – orientation), repName, repClass, and repFamily (name, class, and family of each repeat). Since you specify an

output file name, a plain text file is returned which you can save to a directory. The output has 91 rows plus a header, and while you should obtain this file yourself it is also available as Web Document 8.12. Note that the output columns include Smith–Waterman scores (swScore), repeat class, and repeat family. (2) Open RStudio, and set your working directory to the place you saved your file. Import this text file via the workspace panel (specify that there is a header row). Inspect some basic properties of this dataset.

```
> dim(ucsc_chr11_repeats3)
[1] 91 7
> str(ucsc_chr11_repeats3)
'data.frame': 91 obs. of 7 variables:
 $ X.swScore: int 208 1218 189 1691 12383 1530
   12383 4149 266 797 ...
 $ genoStart: int 5230215 5230647 5231331
5232000
 5232660 5234055 5234278 5235524 5236584
 5236631 ...
 $ genoEnd : int 5230295 5231194 5231407 5232286
 5234055 5234278 5235526 5236191 5236624
 5236773 ...
 $ strand  : Factor w/ 2 levels "-","+": 2 1 1
2 1
 1 1 2 2 1 ...
 $ repName : Factor w/ 51 levels "(A)n","(CA)n",...
 48 29 49 13 38 36 38 38 10 23 ...
 $ repClass : Factor w/ 7 levels "DNA","LINE",
 "Low_complexity",...: 6 2 6 6 2 2 2 5 6 ...
 $ repFamily: Factor w/ 11 levels "Alu","ERVL",
 "ERVL",...: 9 6 9 1 6 6 6 6 10 1 ...
```

### (3) Plot the repeat class as a boxplot.

```
> plot(x = ucsc_chr11_repeats3$repClass,
      y = ucsc_chr11_repeats3$X.swScore,
      main = "Repeat classes in the human beta
              globin locus",
      col = "pink",
      xlab = "repeat class",
      ylab = "SW score")
```

Which repeat class has the highest mean Smith–Waterman scores? You can inspect the plots, or you can invoke the `tapply` command to determine the mean and other summary data.

```
> tapply(ucsc_chr11_repeats3$X.swScore,
      ucsc_chr11_repeats3$repClass, mean)
> tapply(ucsc_chr11_repeats3$X.swScore,
      ucsc_chr11_repeats3$repClass, range)
```

**[8-4]** In this exercise we will use the R package `GenomeGraphs` in RStudio to plot the structure of the beta globin gene, and plot the position of this gene on an ideogram of chromosome 11. We will extract information from Biomart. For more information, browse the `GenomeGraphs` vignette

at the [bioconductor.org](http://bioconductor.org) website, as well as a user's guide written by Steffen Durinck and James Bullard.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("GenomeGraphs")
> options(width=50)
> library(GenomeGraphs)
> mart <- useMart("ensembl", dataset="hsapiens_gene_ensembl")
> gene <- makeGene(id = "ENSG00000244734",
  type="ensembl_gene_id", biomart = mart)
> gdPlot(gene) # save the output as Rplot1
  (a .png file)
> transcript <- makeTranscript
  (id = "ENSG00000244734", type="ensembl_gene_id", biomart = mart)
> gdPlot(list(gene, transcript)) # save the output as Rplot2 (a .png file)
> minusStrand <- makeGeneRegion(chromosome = 11,
  start = 5246696, end = 5248301, strand = "-",
  biomart = mart)
> genomeAxis <- makeGenomeAxis(add53 = TRUE)
  # Add53 shows 5' and 3' ends
> gdPlot(list(genomeAxis, minusStrand))
# This shows a plot with brown boxes for [exons] and genomic coordinates. Save it as Rplot3.
> minStrand <- makeGeneRegion(chromosome = 11,
  start = 5200000, end = 5250000, strand = "-",
  biomart = mart)
> ideogram <- makeIdeogram(chromosome = 11)
> genomeAxis <- makeGenomeAxis(add53=TRUE,
  add35=TRUE)
> gdPlot(list(ideogram, minusStrand, genomeAxis,
  minStrand))
# save as Rplot4.png
```

**[8-5]** The purpose of this exercise is to become familiar with ENCODE resources at UCSC. Visit the ENCODE Experiment Matrix site (<http://encodeproject.org/ENCODE/data-Matrix/encodeDataMatrixHuman.html>) (WebLink 8.47). This includes GM12878 BAM and BAM index (BAI) files; we will learn how to view and manipulate BAM files in Chapter 9. This matrix site has clickable boxes, for example DNase-seq and many other assays for GM12878 (a HapMap individual). Click on such a box and view the data in the UCSC Genome Brower. Select a specific gene (such as *HBB*). What can you learn about its regulation from the ENCODE data matrix?

**[8-6]** Analyze open reading frames in a BAC clone.

- Retrieve a typical *Mus musculus* bacterial artificial chromosome (BAC) from Entrez (e.g., choose BAC T18A20, GenBank accession AC009324).
  - Note the approximate size (in kilobases). Is this a large or a small BAC?
  - Note the approximate number of protein products in it. Bacteria have about one gene per kilobase. How many genes are there per kilobase in this eukaryotic DNA?

(b) Go to the ORF Finder at NCBI.

- From the main page, look at the left sidebar. Choose “Tools for data mining”; then you will see the ORF Finder.
  - Alternatively, from the main page, look at the left sidebar at the top. Choose “Site map” and you will also find a link to the ORF Finder.
  - Paste in the accession number for your BAC. Click OrfFind.
- (c) At the ORF Finder at NCBI, Click on the largest ORF.
- How many amino acids long is it?
  - What is its molecular weight (in kilodaltons)?
  - Is this protein small, average, or large?
  - From which strand of the BAC is this putative gene transcribed? Overall, are there more ORFs on the top or bottom strand or is it about the same?
- (d) Using the ORF Finder at NCBI, BLAST search the ORF of (c) using the default parameters that are given to you.
- This BLAST result reveals many matches to *Mus* proteins. However, note that if you perform a standard BLASTP search using this ORF as a query, you will find matches to many dozens of species. You will also see a match to the Conserved Domain Database.

**[8-7]** Human centromeres typically contain several thousand base pairs of a 171 bp repeat called  $\alpha$ -satellite (accession X07685). First perform a BLASTN search against the nonredundant database. What kinds of database matches do you observe? Second, restrict your BLAST search to nonhumans. Are there matches in primates, rodents, or plants? Why might centromeric repeats have this phylogenetic distribution; would you expect each species to have its own, unique centromeric signature?

**[8-8]** We further explore human centromeric regions. There is an enrichment of segmentally duplicated segments in pericentromeric regions (e.g., within 5 Mb of the centromere; She *et al.*, 2004). How many duplicated segments are present near the centromere of chromosome 11? Is the number in this region greater than that of the chromosome-wide average? (1) Go the UCSC Genome Browser, view the human GRCh37/hg19 build, and view coordinates chr11:48,000,001–58,000,000 (a span of 10 Mb). (2) View the segmental duplication data in the Table Browser; inspect the summary. This region includes 3.4 Mb in gaps (because of the centromere), 328 duplications spanning ~1.8 Mb (182 duplications per Mb). (3)

Repeat this Table Browser analysis across all of chromosome 11. There are 1933 segmental duplications per 135 Mb (14 per Mb).

**[8-9]** Identify ultraconserved elements that share 100% identity between the chicken and human genomes. While there are several approaches, try the following. (1) Go to the UCSC Genome Bioinformatics site (<http://genome.ucsc.edu>). Select the Table Browser. Set the clade to vertebrate clade, the genome to chicken, the group to “Comparative Genomics,” and the track to “Most Conserved.” Under “region” select whole genome. (2) If you obtain the summary statistics at this point, there are over 950,000 items which includes a range of conservation levels. The output format is “all fields from selected table.” Click the filter button, and select scores that are  $\geq 900$  (on a scale from 1 to 1000). There are now only six items (on chicken chromosomes 1, 2, 5, and 7). These are listed in Web Document 8.13 at <http://www.bioinfbook.org/chapter8>. (3) Change the output format to “hyperlinks to Genome Browser.” You can now access the Genome Browser showing these ultraconserved elements and, by clicking the annotation tracks, you can view multiple sequence alignments of the highly conserved DNA.

**[8-10]** Which genes express the greatest extent of copy number gains and losses in the human genome? (1) Use the UCSC Table Browser DGV Structural Variation track (in the Variation and Repeats group), setting the region to genome. The summary/statistics option shows that there are >200,000 items in hg19. (2) Choose the filters option by clicking “create.” Using the pull-down menus select “observedGains” is  $>100$  and “observedLosses” is  $>100$ . Click submit. The summary/statistics tab shows that there are now 18 results. The “Get output” tab allows you to see the 18 entries. These include highly polymorphic HLA genes and *LILRA6* (a leukocyte immunoglobulin-like receptor gene on chromosome 19q). See Web Document 8.14. You can also view one of these genes and its DGV track in the UCSC Genome Browser to see the dramatic copy number variation.

**[8-11]** The dinucleotide CG is often referred to as CpG (p denotes the phosphate linkage between the two residues). In the human genome CpG dinucleotides occur at a very low frequency (about five-fold less than other dinucleotides). What are the frequencies of all dinucleotides at the human beta globin (HBB) locus on chromosome 11? To answer this, obtain this DNA sequence in the FASTA format, import it into the R package SeqinR, and use the count function. This problem is described in

an online book by Avril Coghlan available at <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter1.html> (WebLink 8.48). See also the SeqinR documentation at <http://SeqinR.r-forge.r-project.org/> (WebLink 8.49). (1) Select a region of 60,000 base pairs on chr11:5,240,001–5,300,000, encompassing HBB and other globin genes. You can access the DNA via the Table Browser (select output format > sequence) or from the Genome Browser (view > DNA). You can also view SeqinR documentation for instructions on fetching sequences from NCBI in R. (2) Open an R session. We change our working directory to one containing the FASTA sequence.

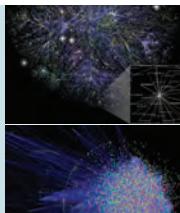
You can view this plot as Web Document 8.15 at the textbook website, <http://bioinfbook.org/chapter8>.

```
> dir() # Looking at the directory we confirm
       the sequence file is present
[1] "chr11_60kb"
# Next we install SeqinR and load its library.
> source("http://bioconductor.org/biocLite.R")
> biocLite("SeqinR")
> library("SeqinR")
> globinDNA <- read.fasta(file = "chr11_60kb")
```

```
# we read the FASTA formatted file into an R
object called globinDNA
> globinseq <- globinDNA[[1]]
> length(globinseq) # We confirm the length of
this sequence is 60 kb
[1] 60000
> count(globinseq,1) # the count function
reports the frequency of each nucleotide
a c g t
18714 12002 11453 17831
> count(globinseq,2) # we specify we want to
know the frequency of all dinucleotides
aa ac ag at ca cc cg ct ga gc gg gt
ta tc tg tt
6470 3103 4271 4870 4443 2932 406 4221 3615 2282
2660 2896 4186 3685 4116 5843
```

Note that the frequency of CpG dinucleotides is indeed substantially lower than that of all other dinucleotides. We can also create a table object with these results (called mydinucleotides) and view a particular result. We can plot the results.

```
> mydinucleotides <- count(globinseq,2)
> mydinucleotides[["cg"]]
[1] 406
> plot(mydinucleotides)
```



## Self-Test Quiz

- [8-1]** The C value paradox is that:
- the nucleotide C is underrepresented in some genomes;
  - the genome size of various eukaryotes correlates poorly with the number of protein-coding genes of the organism;
  - the genome size of various eukaryotes correlates poorly with the biological complexity of the organism; or
  - the genome size of various eukaryotes correlates poorly with the evolutionary age of the organism.
- [8-2]** Hundreds or thousands of sequence repeats, each consisting of a unit of about 4 to 8 nucleotides, are commonly found:
- in interspersed repeats;
  - in processed pseudogenes;
  - in telomeres; or
  - in segmentally duplicated regions.

- [8-3]** You are sequencing the genome of a newly described organism (a slime mold). What is likely to happen if you use RepeatMasker to assess its repetitive DNA content? You set the default setting of RepeatMasker to the settings for human DNA.

- RepeatMasker should successfully identify essentially all of the repetitive DNA. Various repetitive DNA elements are similar enough between organisms to allow this software to work on your slime mold DNA.
- RepeatMasker should identify most of the repetitive DNA. However, because some types of repeats are species-specific, it is likely that there will be many false positive and false negative results.
- RepeatMasker would fail to identify most of the repetitive DNA. Most types of repeats are highly species-specific. It is necessary for you to train the RepeatMasker algorithm on your slime mold DNA in order for the program to work.

- (d) It is not possible to predict, because repetitive DNA may or may not be variable between organisms.

**[8-4]** What is the definition of a gene? Use a recent definition introduced as part of the ENCODE project.

- (a) A gene is a unit of hereditary information localized to a particular chromosome position and encoding one protein.
- (b) A gene is a unit of hereditary information localized to a particular chromosome position and encoding one or more protein products.
- (c) A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.
- (d) A gene is a unit of hereditary information encoding one or more functional products.

**[8-5]** It is extremely difficult for intrinsic (*ab initio*) gene-finding algorithms to predict protein-coding genes in eukaryotic genomic DNA. What is the main problem?

- (a) exon/intron borders are hard to predict;
- (b) introns may be many kilobases in length;
- (c) the GC content of coding regions is not always differentiated from the GC content of noncoding regions; or
- (d) all of the above.

**[8-6]** What are some of the properties of ultraconserved elements?

- (a) They have variable lengths (from 50 to >1000 base pairs) and are nearly perfectly conserved.
- (b) They have variable lengths (from 50 to >1000 base pairs), are nearly perfectly conserved, and typically correspond to protein-coding regions.
- (c) They have lengths  $\geq 200$  base pairs and are perfectly or nearly perfectly conserved between relatively closely related species such as rats and mice.
- (d) They have lengths  $\geq 200$  base pairs and are perfectly or nearly perfectly conserved between rel-

atively distantly related species such as humans and rodents.

**[8-7]** The genomes of two distinct eukaryotic species can sometimes merge to create an entirely new species:

- (a) true; or
- (b) false.

**[8-8]** According to Ohno's 2R hypothesis, whole-genome duplication (polyploidy) offers several advantages. Which of the following is NOT an advantage?

- (a) Hybrids may propagate more successfully than their parents.
- (b) Genes may become redundant, allowing novel functions to emerge.
- (c) Self-fertilization may become possible.
- (d) Self-fertilizing organisms may become able to interbreed.

**[8-9]** Several mechanisms have been proposed by which new gene families are formed. According to the birth-and-death evolution model:

- (a) new genes arise by gene duplication followed by either functional diversification or inactivation;
- (b) genes acquire novel functions as a gradual process that follows gene duplication;
- (c) members of a gene family evolve in a concerted manner; or
- (d) new genes arise and acquire new functions in a coordinated manner dependent on the death of other duplicated genes.

**[8-10]** Single-nucleotide polymorphism (SNP) arrays can reliably detect all of the following phenomena except for:

- (a) deletions;
- (b) duplications;
- (c) inversions; or
- (d) uniparental isodisomy.

## SUGGESTED READING

Our understanding of eukaryotic chromosomes has been transformed by the sequencing and analysis of genomes. The ENCODE Project Consortium papers of 2004 (introducing the project) and 2007 (describing an overview of the results of analyzing 1% of the human genome) provide background, and Stamatoyannopoulos (2012) provides an important overview and perspective on the subsequent phase. The ENCODE Project Consortium

*et al.* (2012) summarize their key findings in a major paper, and ENCODE Project Consortium *et al.* (2011) also provide a useful user's guide to the project.

For a review of repetitive DNA from a comparative genomics perspective, see Richard *et al.* (2008).

## REFERENCES

- Adams, K.L., Wendel, J.F. 2005. Novel patterns of gene expression in polyploid plants. *Trends in Genetics* **21**, 539–543. PMID: 10731132.
- Alioto, T. 2012. Gene prediction. *Methods in Molecular Biology* **855**, 175–201. PMID: 22407709.
- Alkan, C., Sajadian, S., Eichler, E.E. 2011. Limitations of next-generation genome sequence assembly. *Nature Methods* **8**(1), 61–65. PMID: 21102452.
- Allen, J.E., Pertea, M., Salzberg, S.L. 2004. Computational gene prediction using multiple sources of evidence. *Genome Research* **14**, 142–148. PMID: 14707176.
- Allen, J.E., Salzberg, S.L. 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21**, 3596–3603. PMID: 16076884.
- Allen, J.E., Majoros, W.H., Pertea, M., Salzberg, S.L. 2006. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biology* **7** Suppl 1, S9.1–S9.13. PMID: 16925843.
- Ambros, V. 2001. microRNAs: Tiny regulators with great potential. *Cell* **107**, 823–826.
- Amor, D. J., Choo, K. H. 2002. Neocentromeres: Role in human disease, evolution, and centromere study. *American Journal of Human Genetics* **71**, 695–714.
- Antonarakis, S. E., Rossiter, J.P., Young, M. *et al.* 1995. Factor VIII gene inversions in severe hemophilia A: Results of an international consortium study. *Blood* **86**, 2206–2212. PMID: 7662970.
- Aparicio, S., Chapman, J., Stupka, E. *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310. PMID: 12142439.
- Aulard, S., Monti, L., Chaminade, N., Lemeunier, F. 2004. Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**, 137–150.
- Aury, J.M., Jaillon, O., Duret, L. *et al.* 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178. PMID: 17086204.
- Avramova, Z. V. 2002. Heterochromatin in animals and plants. Similarities and differences. *Plant Physiology* **129**, 40–49.
- Ayala F.J., Coluzzi, M. 2005. Chromosome speciation: humans, *Drosophila*, and mosquitoes. *Proceedings of the National Academy of Science, USA* **102** Suppl. 1, 6535–6542. PMID: 15851677.
- Azzalin, C. M., Nergadze, S. G., Giulotto, E. 2001. Human intrachromosomal telomeric-like repeats: Sequence organization and mechanisms of origin. *Chromosoma* **110**, 75–82.
- Bailey, J.A., Eichler, E.E. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics* **7**, 552–564.
- Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., Eichler, E. E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research* **11**, 1005–1017.
- Balakirev, E.S., Ayala, F.J. 2003. Pseudogenes: are they “junk” or functional DNA? *Annual Review of Genetics* **37**, 123–151.
- Bejerano, G., Pheasant, M., Makunin, I. *et al.* 2004. Ultraconserved elements in the human genome. *Science* **304**, 1321–1325.
- Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580.
- Betrán, E., Long, M. 2002. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115**, 65–80.
- Blattner, F. R., Plunkett, G. 3rd, Bloch, C.A. *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474. PMID: 9278503.

- Bray, N., Dubchak, I., Pachter, L. 2003. AVID: A Global Alignment Program. *Genome Research* **13**, 97–102.
- Brent, M.R. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews Genetics* **9**(1), 62–73. PMID: 18087260.
- Britten, R. J., Kohne, D. E. 1968. Repeated sequences in DNA. *Science* **161**, 529–540.
- Brunet, F.G., Roest Crollius, H., Paris, M. et al. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution* **23**, 1808–1816. PMID: 16809621.
- Buratti, E., Baralle, M., Baralle, F.E. 2013. From single splicing events to thousands: the ambiguous step forward in splicing research. *Briefings in Functional Genomics* **12**(1), 3–12. PMID: 23165350.
- Burge, C. B., Karlin, S. 1998. Finding the genes in genomic DNA. *Current Opinion in Structural Biology* **8**, 346–354. PMID: 9666331.
- Cameron, R. A., Mahairas, G., Rast, J.P. et al. 2000. A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proceedings of the National Academy of Science, USA* **97**, 9514–9518. PMID: 10920195.
- Carlton, J. M., Angiuoli, S.V., Suh, B.B. et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512–519. PMID: 12368865.
- Castillo-Davis, C.I. 2005. The evolution of noncoding DNA: how much junk, how much func? *Trends in Genetics* **21**, 533–536.
- Cavalier-Smith, T. 2002. Origins of the machinery of recombination and sex. *Heredity* **88**, 125–141.
- Chaisson, M.J., Raphael, B.J., Pevzner, P.A. 2006. Microinversions in mammalian evolution. *Proceedings of the National Academy of Science, USA* **103**, 19824–19829.
- Chiang, C.W., Derti, A., Schwartz, D. et al. 2008. Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. *Genetics* **180**(4), 2277–2293. PMID: 18957701.
- Choo, K. H. 2001. Domain organization at the centromere and neocentromere. *Developmental Cell* **1**, 165–177.
- Claverie, J. M. 2001. Gene number. What if there are only 30,000 human genes? *Science* **291**, 1255–1257.
- Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H., Stein, L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics* **21**, 673–682. PMID: 16242204.
- Comai, L. 2000. Genetic and epigenetic interactions in allopolyploid plants. *Plant Molecular Biology* **43**, 387–399.
- Comai, L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**, 836–846.
- Cummings, C. J., Zoghbi, H. Y. 2000. Trinucleotide repeats: Mechanisms and pathophysiology. *Annual Review of Genomics and Human Genetics* **1**, 281–328.
- Darlington, C.D. 1932. *Recent Advances in Cytology*. P. Blakiston's Son & Co., Philadelphia.
- Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B., Brison, O. 2012. Common fragile sites: mechanisms of instability revisited. *Trends in Genetics* **28**(1), 22–32. PMID: 22094264.
- Dehal, P., Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* **3**, e314.
- Dermitzakis, E. T., Reymond, A., Lyle, R. et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582. PMID: 12466853.
- Derrien T., Johnson, R., Bussotti, G. et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* **22**(9), 1775–1789. PMID: 22955988.
- Dimitrieva, S., Bucher, P. 2013. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research* **41**(Database issue), D101–109. PMID: 23193254.
- Djebali, S., Davis, C.A., Merkel, A. et al. 2012. Landscape of transcription in human cells. *Nature* **489**(7414), 101–108. PMID: 22955620.

- Dolan, M. 2011. The role of the Giemsa stain in cytogenetics. *Biotechnic and Histochemistry* **86**(2), 94–97. PMID: 21395494.
- Doolittle, W.F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Science, USA* **110**(14), 5294–5300. PMID: 23479647.
- Douglas, S., Zauner, S., Fraunholz, M. *et al.* 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091–1096. PMID: 11323671.
- Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218.
- Dunham, A., Matthews, L.H., Burton, J. *et al.* 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428**, 522–528. PMID: 15057823.
- Echols, N., Harrison, P., Balasubramanian, S. *et al.* 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Research* **30**, 2515–2523. PMID: 12034841.
- Eddy, S. R. 2001. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**, 919–929.
- Eddy, S. R. 2002. Computational genomics of noncoding RNA genes. *Cell* **109**, 137–140.
- Eddy, S.R. 2012. The C-value paradox, junk DNA and ENCODE. *Current Biology* **22**(21), R898–899. PMID: 23137679.
- Eddy, S.R. 2013. The ENCODE project: missteps overshadowing a success. *Current Biology* **23**(7), R259–261. PMID: 23578867.
- Elgar, G. 2006. Different words, same meaning: understanding the languages of the genome. *Trends in Genetics* **22**, 639–641.
- ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology* **9**(4), e1001046. PMID: 21526222.
- ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A. *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816. PMID: 17571346.
- ENCODE Project Consortium, Bernstein, B.E., Birney, E. *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74. PMID: 22955616.
- Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., Trask, B. J. 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Research* **12**, 1651–1662.
- Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., McCallion, A.S. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276–279.
- Flicek, P., Ahmed, I., Amode, M.R. *et al.* 2013. Ensembl 2013. *Nucleic Acids Research* **41**(Database issue), D48–55. PMID: 23203987.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., Hardison, R. C. 2003. Cross-species sequence comparisons: A review of methods and available resources. *Genome Research* **13**, 1–12.
- Frith, M.C., Forrest, A.R., Nourbakhsh, E. *et al.* 2006. The abundance of short proteins in the mammalian proteome. *PLoS Genetics* **2**, e52. PMID: 16683031.
- Gardner, M. J., Hall, N., Fung, E. *et al.* 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511. PMID: 12368864.
- Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., Snyder, M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17**, 669–681.
- Gerstein, M.B., Kundaje, A., Hariharan, M. *et al.* 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414), 91–100. PMID: 22955619.
- Graur, D., Li, W.-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

- Graur, D., Zheng, Y., Price, N. *et al.* 2013. On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* **5**(3), 578–590. PMID: 23431001.
- Green, E.D., Guyer, M.S. 2011. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204–213. PMID: 21307933.
- Gregory, S.G., Barlow, K.F., McLay, K.E. *et al.* 2006. The DNA sequence and biological annotation of human chromosome 1. *Nature* **441**(7091), 315–321. PMID: 16710414.
- Griffith, O.L., Montgomery, S.B., Bernier, B. *et al.* 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research* **36**, D107–D113. PMID: 18006570.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S. R. 2003. Rfam: An RNA family database. *Nucleic Acids Research* **31**, 439–441.
- Grimwood, J., Gordon, L.A., Olsen, A. *et al.* 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535. PMID: 15057824.
- Guigo, R., Flicek, P., Abril, J.F. *et al.* 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biology* **7**, S2.1–31. PMID: 16925836.
- Hall, A. E., Fiebig, A., Preuss, D. 2002. Beyond the *Arabidopsis* genome: Opportunities for comparative genomics. *Plant Physiology* **129**, 1439–1447.
- Hall, I. M., Quinlan, A. R. 2012. Detection and interpretation of genomic structural variation in mammals. *Methods in Molecular Biology* **838**, 225–248. PMID: 22228015.
- Hancock, J. M. 2002. Genome size and the accumulation of simple sequence repeats: Implications of new data from genome sequencing projects. *Genetica* **115**, 93–103.
- Harrison, P. M., Gerstein, M. 2002. Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *Journal of Molecular Biology* **318**, 1155–1174.
- Harrison, P. M., Kumar, A., Lang, N., Snyder, M., Gerstein, M. 2002. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Research* **30**, 1083–1090.
- Harrow, J., Denoeud, F., Frankish, A. *et al.* 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biology* **7** Suppl 1, S4.1–9. PMID: 16925838.
- Harrow, J., Frankish, A., Gonzalez, J.M. *et al.* 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22**(9), 1760–1774. PMID: 22955987.
- Hartl, D. L. 2000. Molecular melodies in high and low C. *Nature Reviews Genetics* **1**, 145–149.
- Hassold, T., Hunt, P. 2001. To err (meiotically) is human: the genesis of human aneuploidy. *Nature Reviews Genetics* **2**, 280–291.
- Hattori, M., Fujiyama, A., Taylor, T.D. *et al.* 2000. The DNA sequence of human chromosome 21. *Nature* **405**(6784), 311–319. PMID: 10830953.
- Haugen, P., Simon, D.M., Bhattacharya, D. 2005. The natural history of group I introns. *Trends in Genetics* **21**, 111–119.
- Hazan, R., Ben-Yehuda, S. 2006. Resolving chromosome segregation in bacteria. *Journal of Molecular Microbiology and Biotechnology* **11**(3–5), 126–139. PMID: 16983190.
- Hedges, S. B., Kumar, S. 2002. Genomics. Vertebrate genomes compared. *Science* **297**, 1283–1285.
- Hillier, L.W., Graves, T.A., Fulton, R.S. *et al.* 2005. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**(7034), 724–731. PMID: 15815621.
- Holt, R. A., Subramanian, G. M., Halpern, A. *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149. PMID: 12364791.
- Hou, G., Le Blancq, S. M., Yaping, E., Zhu, H., Lee, M. G. 1995. Structure of a frequently rearranged rRNA-encoding chromosome in *Giardia lamblia*. *Nucleic Acids Research* **23**, 3310–3317.
- Hughes, A. L., Nei, M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Science, USA* **86**, 958–962.
- Iafrate, A.J., Feuk, L., Rivera, M.N. *et al.* 2004. Detection of large-scale variation in the human genome. *Nature Genetics* **36**(9), 949–955. PMID: 15286789.

- Ijdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., Wells, R. A. 1991. Origin of human chromosome 2: An ancestral telomere–telomere fusion. *Proceedings of the National Academy of Science, USA* **88**, 9051–9055.
- International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**, 1299–1320.
- International HapMap Consortium *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
- Jarstfer, M. B., Cech, T. R. 2002. Effects of nucleotide analogues on *Euplotes aediculatus* telomerase processivity: Evidence for product-assisted translocation. *Biochemistry* **41**, 151–161.
- Jeffares, D.C., Mourier, T., Penny, D. 2006. The biology of intron gain and loss. *Trends in Genetics* **22**, 16–22.
- Jentsch, S., Tobler, H., Muller, F. 2002. New telomere formation during the process of chromatin diminution in *Ascaris suum*. *International Journal of Developmental Biology* **46**, 143–148.
- John, S., Sabo, P.J., Canfield, T.K. *et al.* 2013. Genome-scale mapping of DNase I hypersensitivity. *Current Protocols in Molecular Biology Chapter 27*, Unit 21.27. PMID: 23821440.
- Jones, K. L. 1997. *Smith's Recognizable Patterns of Human Malformation*. W. B. Saunders, New York.
- Jurka, J. 1998. Repeats in genomic DNA: Mining and meaning. *Current Opinion in Structural Biology* **8**, 333–337.
- Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends in Genetics* **16**, 418–420.
- Jurka J., Kapitonov V.V., Kohany O., Jurka M.V. 2007. Repetitive sequences in complex genomes: structure and evolution. *Annual Reviews in Genomics and Human Genetics* **8**, 241–259. PMID: 17506661.
- Kallioniemi, A., Kallioniemi, O.P., Sudar, D. *et al.* 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821.
- Kashi, Y., King, D.G. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* **22**, 253–259.
- Katz, L.A. 2012. Origin and diversification of eukaryotes. *Annual Review of Microbiology* **66**, 411–427. PMID: 22803798.
- Katzman, S., Kern, A.D., Bejerano, G. *et al.* 2007. Human genome ultraconserved elements are ultraselected. *Science* **317**, 915.
- Kidwell, M. G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63.
- King, D.C., Taylor, J., Elnitski, L. *et al.* 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Research* **15**, 1051–1060.
- Knight, J. 2002. All genomes great and small. *Nature* **417**, 374–376.
- Kohn, M., Högel, J., Vogel, W. *et al.* 2006. Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends in Genetics* **22**, 203–210.
- Kotzot, D. 2001. Complex and segmental uniparental disomy (UPD): Review and lessons from rare chromosomal complements. *Journal of Medical Genetics* **38**, 497–507.
- Kotzot, D. 2008. Complex and segmental uniparental disomy updated. *Journal of Medical Genetics* **45**(9), 545–556. PMID: 18524837.
- Kuhn, R.M., Haussler, D., Kent, W.J. 2013. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**(2), 144–161. PMID: 22908213.
- Kurtz, S., Phillippy, A., Delcher, A.L. *et al.* 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5**, R12.
- Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., Monaco, A.P. 2001. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**(6855), 519–523. PMID: 11586359.

- Lai, W.R., Johnson, M.D., Kucherlapati, R., Park, P.J. 2005. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770.
- Latham, K.E. 2005. X chromosome imprinting and inactivation in preimplantation mammalian embryos. *Trends in Genetics* **21**, 120–127.
- Lee, H. S., Chen, Z. J. 2001. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proceedings of the National Academy of Science, USA* **98**, 6753–6758.
- Lewis, M. S., Pikaard, C. S. 2001. Restricted chromosomal silencing in nucleolar dominance. *Proceedings of the National Academy of Science, USA* **98**, 14536–14540.
- Li, W.H., Yang, J., Gu, X. 2005. Expression divergence between duplicate genes. *Trends in Genetics* **21**, 602–607.
- Lima-de-Faria, A. 2003. *One Hundred Years of Chromosome Research and What Remains to be Learned*. Kluwer Academic Publishers, Boston.
- Lin, K.W., Yan, J. 2008. Endings in the middle: current knowledge of interstitial telomeric sequences. *Mutation Research* **658**(1–2), 95–110. PMID: 17921045.
- Liu, P., Carvalho, C.M., Hastings, P.J., Lupski, J.R. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics and Development* **22**(3), 211–220. PMID: 22440479.
- Lorite, P., Carrillo, J. A., Palomeque, T. 2002. Conservation of (TTAGG)(n) telomeric sequences among ants (Hymenoptera, Formicidae). *Journal of Heredity* **93**, 282–285.
- Lowe, T. M., Eddy, S. R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964.
- Lupski, J.R., Stankiewicz, P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics* **1**(6), e49. PMID: 16444292.
- Makalowski, W. 2000. Genomic scrap yard: How genomes utilize all that junk. *Gene* **259**, 61–67.
- Marshall, O.J., Chueh, A.C., Wong, L.H., Choo, K.H. 2008. Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution. *American Journal of Human Genetics* **82**(2), 261–282. PMID: 18252209.
- Martin, C. L., Wong, A., Gross, A. et al. 2002. The evolutionary origin of human subtelomeric homologies: or where the ends begin. *American Journal of Human Genetics* **70**, 972–984. PMID: 11875757.
- Maston, G.A., Evans, S.K., Green, M.R. 2006. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* **7**, 29–59.
- Matsuura, T., Yamagata, T., Burgess, D.L. et al. 2000. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genetics* **26**(2), 191–194. PMID: 11017075.
- Mayor, C., Brudno, M., Schwartz, J.R. et al. 2000. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047. PMID: 11159318.
- McCormick-Graham, M., Romero, D. P. 1996. A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*. *Molecular and Cellular Biology* **16**, 1871–1879.
- McKnight, T. D., Fitzgerald, M. S., Shipp, D. E. 1997. Plant telomeres and telomerases. A review. *Biochemistry (Mosc.)* **62**, 1224–1231.
- McLean, C., Bejerano, G. 2008. Dispensability of mammalian DNA. *Genome Research* **18**(11), 1743–1751. PMID: 18832441.
- Meaburn, K.J., Misteli, T. 2007. Cell biology: chromosome territories. *Nature* **445**, 379–381.
- Melek, M., Davis, B. T., Shipp, D. E. 1994. Oligonucleotides complementary to the *Oxytricha nova* telomerase RNA delineate the template domain and uncover a novel mode of primer utilization. *Molecular and Cellular Biology* **14**, 7827–7838.
- Mercer, T.R., Edwards, S.L., Clark, M.B. et al. 2013. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics* **45**(8), 852–859. PMID: 23793028
- Meyer, L.R., Zweig, A.S., Hinrichs, A.S. et al. 2013. The UCSC Genome Browser database: extensions and updates. *Nucleic Acids Research* **41**(D1), D64–69. PMID: 23155063.

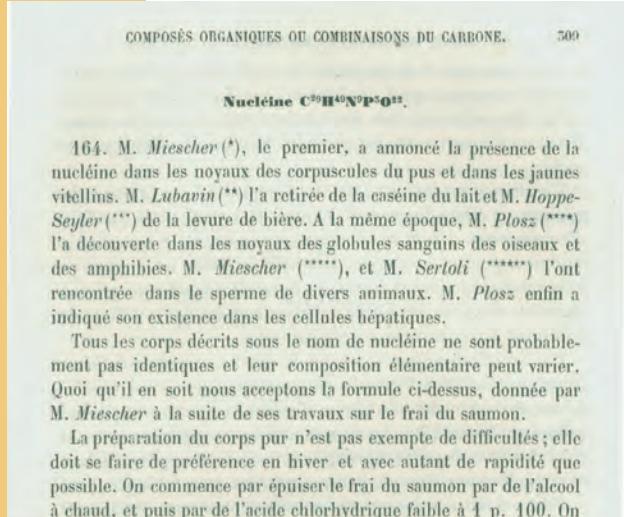
- Morgenstern, B., Rinner, O., Abdeddaïm, S. *et al.* 2002. Exon discovery by genomic sequence alignment. *Bioinformatics* **18**, 777–787. PMID: 12075013.
- Morison, I.M., Ramsay, J.P., Spencer, H.G. 2005. A census of mammalian imprinting. *Trends in Genetics* **21**, 457–465.
- Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M. *et al.* 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology* **13**(8), 418. PMID: 22889292.
- Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915), 520–562. PMID: 12466850.
- Müller, F., Tobler, H. 2000. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parasacaris univalens*. *International Journal of Parasitology* **30**, 391–399.
- Murphy, W.J., Larkin, D.M., Everts-van der Wind, A. *et al.* 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**, 613–617. PMID: 16040707.
- Muzny, D.M., Scherer, S.E., Kaul, R. *et al.* 2006. The DNA sequence, annotation and analysis of human chromosome 3. *Nature* **440**, 1194–1198. PMID: 16641997.
- Nanda, I., Schrama, D., Feichtinger, W. *et al.* 2002. Distribution of telomeric (TTAGGG)(n) sequences in avian chromosomes. *Chromosoma* **111**, 215–227. PMID: 12424522.
- Nei, M., Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**, 121–152.
- Niu, D.K., Jiang, L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and Biophysical Research Communications* **430**(4), 1340–1343. PMID: 23268340.
- Nixon, J. E., Wang, A., Morrison, H.G. *et al.* 2002. A spliceosomal intron in *Giardia lamblia*. *Proceedings of the National Academy of Science, USA* **99**, 3701–3705. PMID: 11854456.
- Nóbrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., Rubin, E.M. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* **431**, 988–993.
- Novichkov, P. S., Gelfand, M. S., Mironov, A. A. 2001. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17**, 1011–1018.
- Nwakanma D.C., Neafsey, D.E., Jawara, M. *et al.* 2013. Breakdown in the process of incipient speciation in *Anopheles gambiae*. *Genetics* **193**(4), 1221–1231. PMID: 23335339.
- Ohno, S. 1970. *Evolution by Gene Duplication*. SpringerVerlag, Berlin.
- Ohno, S. 1972. So much “junk” DNA in our genome. *Brookhaven Symposia in Biology* **23**, 366–370. PMID: 5065367.
- Olshen, A.B., Venkatraman, E.S., Lucito, R., Wigler, M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.
- Ostertag, E. M., Kazazian, H. H., Jr. 2001. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Research* **11**, 2059–2065.
- Panopoulou, G., Pousta, A. J. 2005. Timing and mechanism of ancient vertebrate genome duplications: the adventure of a hypothesis. *Trends in Genetics* **21**, 559–567.
- Passarge, E., Horsthemke, B., Farber, R. A. 1999. Incorrect use of the term synteny. *Nature Genetics* **23**, 387. PMID: 10581019.
- Paterson, A.H., Chapman, B.A., Kissinger, J.C. *et al.* 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends in Genetics* **22**, 597–602.
- Pavlicek, A., Gentles, A.J., Paces, J., Paces, V., Jurka, J. 2006. Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends in Genetics* **22**, 69–73.
- Pei B., Sisu, C., Frankish, A. *et al.* 2012. The GENCODE pseudogene resource. *Genome Biology* **13**(9), R51. PMID: 22951037.
- Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., Bejerano, G. 2013. Enhancers: five essential questions. *Nature Reviews Genetics* **14**(4), 288–295. PMID: 23503198.
- Pevsner J., Reed R.R., Feinstein P.G., Snyder S.H. 1988. Molecular cloning of odorant-binding protein: member of a ligand carrier family. *Science* **241**(4863), 336–339. PMID: 3388043.

- Picardi, E., Pesole, G. 2010. Computational methods for ab initio and comparative gene finding. *Methods in Molecular Biology* **609**, 269–284. PMID: 20221925.
- Pozzoli, U., Menozzi, G., Comi, G.P., Cagliani, R., Bresolin, N., Sironi, M. 2007. Intron size in mammals: complexity comes to terms with economy. *Trends in Genetics* **23**, 20–24. PMID: 17070957.
- Prince, V.E., Pickett, F.B. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics* **3**, 827–837.
- Rebollo R., Romanish M.T., Mager D.L. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**, 21–42. PMID: 22905872.
- Redi, C. A., Garagna, S., Zacharias, H., Zuccotti, M., Capanna, E. 2001. The other chromatin. *Chromosoma* **110**, 136–147.
- Reese, M.G., Guigó, R. 2006. EGASP: Introduction. *Genome Biology* **7** Suppl 1, S1.1–1.3.
- Richard, G.F., Kerrest, A., Dujon, B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Review* **72**(4), 686–727. PMID: 19052325.
- Rosenbloom K.R., Sloan, C.A., Malladi, V.S. et al. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research* **41**(Database issue), D56–63. PMID: 23193274.
- Roy, S.W. 2006. Intron-rich ancestors. *Trends in Genetics* **22**, 468–471.
- Roy, S.W., Gilbert, W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**, 211–221.
- Ruvkun, G. 2001. Molecular biology. Glimpses of a tiny RNA world. *Science* **294**, 797–799.
- Sabo, P.J., Kuehn, M.S., Thurman, R. et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature Methods* **3**, 511–518.
- Sasaki, T. et al. 2002. The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316.
- Schlötterer, C., Harr, B. 2000. *Drosophila virilis* has long and highly polymorphic microsatellites. *Molecular Biology and Evolution* **17**, 1641–1646.
- Schwartz, S., Zhang, Z., Frazer, K.A. et al. 2000. PipMaker: a web server for aligning two genomic DNA sequences. *Genome Research* **10**, 577–586. PMID: 10779500.
- She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpern, A.L., Eichler, E.E. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930.
- Shippen-Lentz, D., Blackburn, E. H. 1989. Telomere terminal transferase activity from *Euplotes crassus* adds large numbers of TTTTGGGG repeats onto telomeric primers. *Molecular and Cellular Biology* **9**, 2761–2764.
- Simpson, A. G., MacQuarrie, E. K., Roger, A. J. 2002. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419**, 270.
- Slijepcevic, P. 1998. Telomeres and mechanisms of Robertsonian fusion. *Chromosoma* **107**, 136–140.
- Smit, A. F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development* **9**, 657–663.
- Smith, N.G., Brandström, M., Ellegren, H. 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84**(5), 806–813. PMID: 15475259.
- South, S.T. 2011. Chromosomal structural rearrangements: detection and elucidation of mechanisms using cytogenomic technologies. *Clinics in Laboratory Medicine* **31**(4), 513–524. PMID: 22118734.
- Speicher, M.R., Carter, N.P. 2005. The new cytogenetics: blurring the boundaries with molecular biology. *Nature Reviews Genetics* **6**(10), 782–792. PMID: 16145555.
- Stamatoyannopoulos, J.A. 2012. What does our genome encode? *Genome Research* **22**(9), 1602–1611. PMID: 22955972.
- Stankiewicz, P., Lupski, J.R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* **18**(2), 74–82. PMID: 11818139.
- Stein, L. 2001. Genome annotation: From sequence to biology. *Nature Reviews Genetics* **2**, 493–503.
- Stefansson, H., Helgason, A., Thorleifsson, G. et al. 2005. A common inversion under selection in Europeans. *Nature Genetics* **37**, 129–137. PMID: 15654335.

- Sturtevant, A.H. 1921. A case of rearrangement of genes in *Drosophila*. *Proceedings of the National Academy of Science, USA* **7**, 235–237.
- Taylor, J.S., Raes, J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics* **38**, 615–643.
- Taylor, J., Tyekucheva, S., King, D.C. *et al.* 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome Research* **16**, 1596–1604.
- Thomas, B.C., Pedersen, B., Freeling, M. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research* **16**, 934–946.
- Thurman, R.E., Rynes, E., Humbert, R. *et al.* 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414), 75–82. PMID: 22955617.
- Tjio, J.H., Levan, A. 1956. The chromosome number of man. *Hereditas* **42**, 1–6.
- Ting, J.C., Roberson, E.D., Miller, N.D. *et al.* 2007. Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. *Human Mutation* **28**, 1225–1235.
- Trask, B.J. 2002. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Reviews Genetics* **3**, 769–778.
- Upcroft, P., Chen, N., Upcroft, J.A. 1997. Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote *Giardia duodenalis*. *Genome Research* **7**, 37–46.
- Vassetzky, N.S., Kramerov, D.A. 2013. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Research* **41**(Database issue), D83–89. PMID: 23203982.
- Vellai, T., Vida, G. 1999. The origin of eukaryotes: The difference between prokaryotic and eukaryotic cells. *Proceedings of the Royal Society of London, B: Biological Science* **266**, 1571–1577.
- Venkatraman, E.S., Olshen, A.B. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663.
- Venter, J. C., Adams, M.D., Myers, E.W. *et al.* 2001. The sequence of the human genome. *Science* **291**, 1304–1351. PMID: 11181995.
- Ventura, M., Antonacci, F., Cardone, M.F. *et al.* 2007. Evolutionary formation of new centromeres in macaque. *Science* **316**, 243–246. PMID: 17431171.
- Waldeyer, W. 1888. Über Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen. *Arch. mikrosk. Anat.* **32**, 1–122.
- Wang, J., Zhuang, J., Iyer, S. *et al.* 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research* **22**(9), 1798–1812. PMID: 22955990.
- Watt, W. B., Dean, A. M. 2000. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annual Review of Genetics* **34**, 593–622.
- Wheeler, T.J., Clements, J., Eddy, S.R. *et al.* 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research* **41**(Database issue), D70–82. PMID: 23203985.
- Willenbrock, H., Fridlyand, J. 2005. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091.
- Wolfsberg, T.G. 2011. Using the NCBI Map Viewer to browse genomic sequence data. *Current Protocols in Human Genetics* Chapter 18, Unit18.5. PMID: 21480181.
- Yasuhara, J.C., Wakimoto, B.T. 2006. Oxymoron no more: the expanding world of heterochromatic genes. *Trends in Genetics* **22**, 330–338.
- Yu, J., Hu, S., Wang, J. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92. PMID: 11935017.
- Zdobnov, E. M., von Mering, C., Letunic, I. *et al.* 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159. PMID: 12364792.



(a) Description of Miescher's discovery of nuclein by his mentor Felix Hoppe-Seyler (1877)



(b) Descriptions of globulin, haemoglobin, and nuclein from Carpenter (1876)

**GLOBULIN OR CRYSTALLIN** is obtained from the crystalline lens. It dissolves in acetic acid, but it is precipitated by exact neutralization with ammonia, and *vice versa*, it is dissolved by ammonia, but is precipitated by exact neutralization with acetic acid. It is completely precipitated from its solutions by carbonic acid.

**HÆMOGLOBIN, HEMATOCRYSTALLIN**, a coloring substance closely allied to the albuminous compounds will be fully described in the chapter on the blood.

**NUCLEIN** has been obtained by Miescher from the nuclei of lymph-corpuscles. It is soluble in alkalies and in alkaline bicarbonates, and is precipitated again by acids.

(c) Adenine and guanine (1891)

CHEMISTRY OF THE LEUCOMAINES. 285

There can be no doubt that adenine and guanine play an important part in the physiological function of the cell nucleus, which, from recent observations, appears to be necessary to the formation and building up of organic matter. It is now known that non-nucleated cells, though capable of living, are not capable of reproduction. We must look, therefore, to the nucleus as the seat of the functional activity of the cell—indeed, of the entire organism. Nuclein, the parent substance of adenine and guanine, is the best known and probably most important constituent of the nucleus, and as such it has been already credited with a direct relation to the reproductive powers of the cell. This chemical view has recently been confirmed by ZACHARIAS, who showed that chromatin of histologists is identical with nuclein. LIEBERMANN has questioned nuclein as being the source of xanthine compounds, but in this he is not supported by the mass of evidence.

(d) Description of adenine (1891)

**ADENINE**, C<sub>5</sub>H<sub>5</sub>N<sub>5</sub>, which was discovered by KOSSEL in 1885, forms the simplest member of the uric acid group of leucomaines, and as such it deserves special attention, inasmuch as it shows most clearly the relation that exists between hydrocyanic acid and the members of this group.

This base was first prepared from pancreatic glands—hence the term adenine, which is derived from the Greek word *αδήνη*, meaning a gland. It has since been shown to occur together with guanine, hypoxanthine, etc., as a decomposition-product of nuclein, and, therefore, it may be obtained from all tissues and organs, animal or vegetable, rich in nucleated cells. Accordingly, it has been found in the kidneys, spleen, pancreatic, thymus and lymphatic glands, in beer-yeast, in spermatic fluids, but not in testicles of the steer; occurs also in tea-leaves. In the latter adenine appears to exist in a preformed condition, since it can be extracted without the use of acid reagents. The thymus

ious sources from which nuclein was purified. Hoppe-Seyler gives the formula C<sub>29</sub>H<sub>49</sub>N<sub>9</sub>P<sub>3</sub>O<sub>22</sub>. (b) After brief entries on globulin and haemoglobin, Carpenter (1876, p. 86) mentions that nuclein is soluble in alkalies. (c) Jules Piccard first identified guanine as a constituent of nuclein, as well as two oxidized purines (xanthine and sarkin, today called hypoxanthine) that are derivatives of adenine. Later, Albrecht Kossel identified pyrimidines in nuclein (he discovered thymine in 1883 and cytosine in 1894). (d) Description of adenine.

Sources: (a) Hoppe-Seyler (1877). (b) Carpenter (1876). (c, d) Vaughan and Novy (1891), pp. 284–285 and p. 283, respectively.

In this chapter we describe how to interpret vast amounts of DNA sequence data. Nucleic acids were first discovered by Johann Friedrich Miescher II (1844–1895), a Swiss chemist. During 1868–1869 he worked in the laboratory of the famous chemist Felix Hoppe-Seyler (1825–1895) where he studied the composition of pus cells. He discovered what he called “nucleins,” a new class of compounds that are rich in phosphorous and that he thought were as important as the proteins. Treatment of pus cell extracts with pepsin led to his identification of an acid fraction (“pure nuclein,” i.e., DNA) and a base fraction (“protamine,” which he thought was an alkaloid but is actually nucleohistones). (a) Description of nuclein from an 1877 French translation of a book by Hoppe-Seyler (p. 309), describing var-

# Analysis of Next-Generation Sequence Data

# CHAPTER 9

*The ability to sequence complete genomes and the free dissemination of sequence data have dramatically changed the nature of biological and biomedical research. Sequence and other genomic data have the potential to lead to remarkable improvement in many facets of human life and society, including the understanding, diagnosis, treatment and prevention of disease; advances in agriculture, environmental science and remediation; and our understanding of evolution and ecological systems.*

—Revolutionary Genome Sequencing Technologies – The \$1000 Genome, RFA-HG-10-012, National Human Genome Research Institute (<http://www.nhgri.nih.gov>)

*Children like puzzles, and they usually assemble them by trying all possible pairs of pieces and putting together pieces that match. Biologists assemble genomes in a surprisingly similar way, the major difference being that the number of pieces is larger. For the last 20 years, fragment assembly in DNA sequencing mainly followed the “overlap–layout–consensus” paradigm (1–6). Trying all possible pairs of pieces corresponds to the overlap step, whereas putting the pieces together corresponds to the layout step of the fragment assembly. Our new EULER algorithm is very different from this natural approach—we never even try to match the pairs of fragments, and we do not have the overlap step at all. Instead, we do a very counterintuitive (some would say childish) thing: we cut the existing pieces of a puzzle into even smaller pieces of regular shape. Although it indeed looks childish and irresponsible, we do it on purpose rather than for the fun of it. This operation transports the puzzle assembly from the world of a difficult Layout Problem into the world of the Eulerian Path Problem, with polynomial algorithms for puzzle assembly (in the context of DNA sequencing).*

—Pavel Pevzner et al. (2001)

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

- explain how sequencing technologies generate NGS data;
- describe the FASTQ, SAM/BAM, and VCF data formats;
- compare methods for aligning NGS data to a reference genome;
- describe types of genomic variants and how they are determined;
- explain types of error associated with alignment, assembly, and variant calling; and
- explain methods for predicting the functional consequence of genomic variants in individual genomes.

## INTRODUCTION

Next-generation sequencing (NGS) technology is revolutionizing biology. Of the many applications of NGS technology the following are particularly significant. The central dogma of biology has been a remarkably useful model since its introduction in the 1950s, and the usefulness of NGS extends to the domains of DNA, RNA, and protein (Shendure and Lieberman Aiden, 2012).

Visit these projects at  
✉ <http://1000genomes.org/> (WebLink 9.1), ✉ <http://www.1000bullgenomes.com/> (WebLink 9.2), and  
✉ <http://www.1001genomes.org/> (WebLink 9.3) at ✉ <http://bioinfbook.org>.

1. NGS technology enables determination of the DNA sequence of genomes across the entire tree of life. For any species of organism, through whole-genome sequencing we can determine reference genomes (i.e., prototypic examples that are representative of an entire species) as a starting point to catalog genomic features and to evaluate genetic variation.
2. Through NGS it is becoming routine to perform re-sequencing of individual genomes. We can compare an individual's sequence to a reference genome (e.g., comparing the genomes of 100 people with a disease to a reference human genome). This allows us to determine individual genetic variation within members of a species at a genome-wide scale. Applications range from the 1000 Genomes Project in humans to the 1000 bull project to the 1001 Genomes Project in plants.
3. We can compare the genetic differences within an individual across different cell types. This allows an assessment of somatic changes – those acquired during development sometime after the zygote forms – in contrast to germline changes, many of which are inherited from an individual's parents. Somatic changes are important in cancer, in which we can sequence the genome from tumors and from nontumorous parts of the body of the same person.
4. NGS applied to RNA (i.e., RNA-seq) allows the measurement of RNA transcript levels. We introduce RNA in Chapter 10 and then address the RNAseq technology and related microarray techniques in Chapter 11.
5. Most life forms on the planet are single-celled organisms, uncultivable as single species in a laboratory. Through metagenomics it is possible to apply NGS technology to environmental samples, in many cases allowing broad, deep surveys of the species in ecological niches from the human gut to soil to seawater.
6. There are many other specialized applications of NGS technology (see the end of this chapter). One example is chromatin immunoprecipitation followed by NGS (called ChIP-Seq) to identify DNA sequences associated with promoter regions (Park, 2009). Another method allows differentially methylated DNA regions to be identified.

In this chapter we first describe NGS technology. We then present a workflow covering the main problems of obtaining raw data, assembly, alignment, variant calling, and interpretation. We introduce the major types of file formats (e.g., FASTQ, SAM/BAM, VCF) and some of the most commonly used software tools for alignment (e.g., BWA), variant calling (SAMtools, GATK), analysis of variants (VCFtools), and functional prediction of variants (SIFT, PolyPhen2, Variant Effect Predictor or VEP).

For DNA sequencing studies, we can think of three main types of experiments. First, a whole genome may be sequenced. While the human genome is ~3.2 gigabases, much of it is repetitive and typically fewer than 3 billion base pairs are amenable to sequencing. Such an experiment will include coverage of the exons that comprise <3% of the genome as well as intronic and vast intergenic regions. Second, an exome may be sequenced. This is a collection of exons corresponding to most of the ~20,300 protein-coding human genes. There are three commonly used platforms, each of which employs biotinylated oligonucleotides complementary to exonic regions. Exons are “captured” or enriched prior to sequencing. Third, a region of interest may be targeted and sequenced. To help explain how NGS data analysis works, we study a relatively small dataset generated by my laboratory. I collaborated with Illumina, Inc. to develop a targeted autism sequencing panel involving the exons corresponding to just 101 genes previously implicated in autism.

I provide sample files (in the FASTQ, BAM, and VCF formats) on the book's website. For those interested in whole-genome and whole-exome sequences, see the resources of the 1000 Genomes Project and the major public repositories at the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI).

Most bioinformatics experts perform analyses using the Linux operating system or the related Mac OS terminal environment. The advantages include: an operating system that is appropriate for importing, storing, and analyzing very large data files; a command-line environment in which the user can execute software programs with control over a program's optional settings; and an architecture in which a cluster can be established to optimize storage, analysis, and scheduling of jobs.

Some NGS analyses are possible through websites or graphical user interface (GUI) programs including Galaxy, UCSC, the Genome Workbench of NCBI, and VEP at Ensembl. We also introduce these web-based tools.

Common products for human exome enrichment are from Agilent (SureSelect at <http://www.genomics.agilent.com/>, WebLink 9.4), NimbleGen (SeqCap at <http://www.nimblegen.com/seqcapez/>, WebLink 9.5), and Illumina (TruSeq at <http://www.illumina.com/truseq.ilmn>, WebLink 9.6).

## DNA SEQUENCING TECHNOLOGIES

Nucleic acids were discovered by Johann Friedrich Miescher (1844–1895) in 1869. He called them “nucleins” because they were present in all cell nuclei. The first complete nucleic acid sequence of a molecule (an alanine tRNA from yeast) was accomplished by Holley *et al.* (1965), who purified tRNAs then treated them with a series of ribonucleases. Another milestone was reached in 1970 when Ray Wu developed a primer extension strategy to sequence nucleotides of DNA; this became the basis of Sanger sequencing. Wu determined the sequence of the two cohesive ends of lambda phage DNA in 1971 (Wu, 1970; Wu and Taylor, 1971).

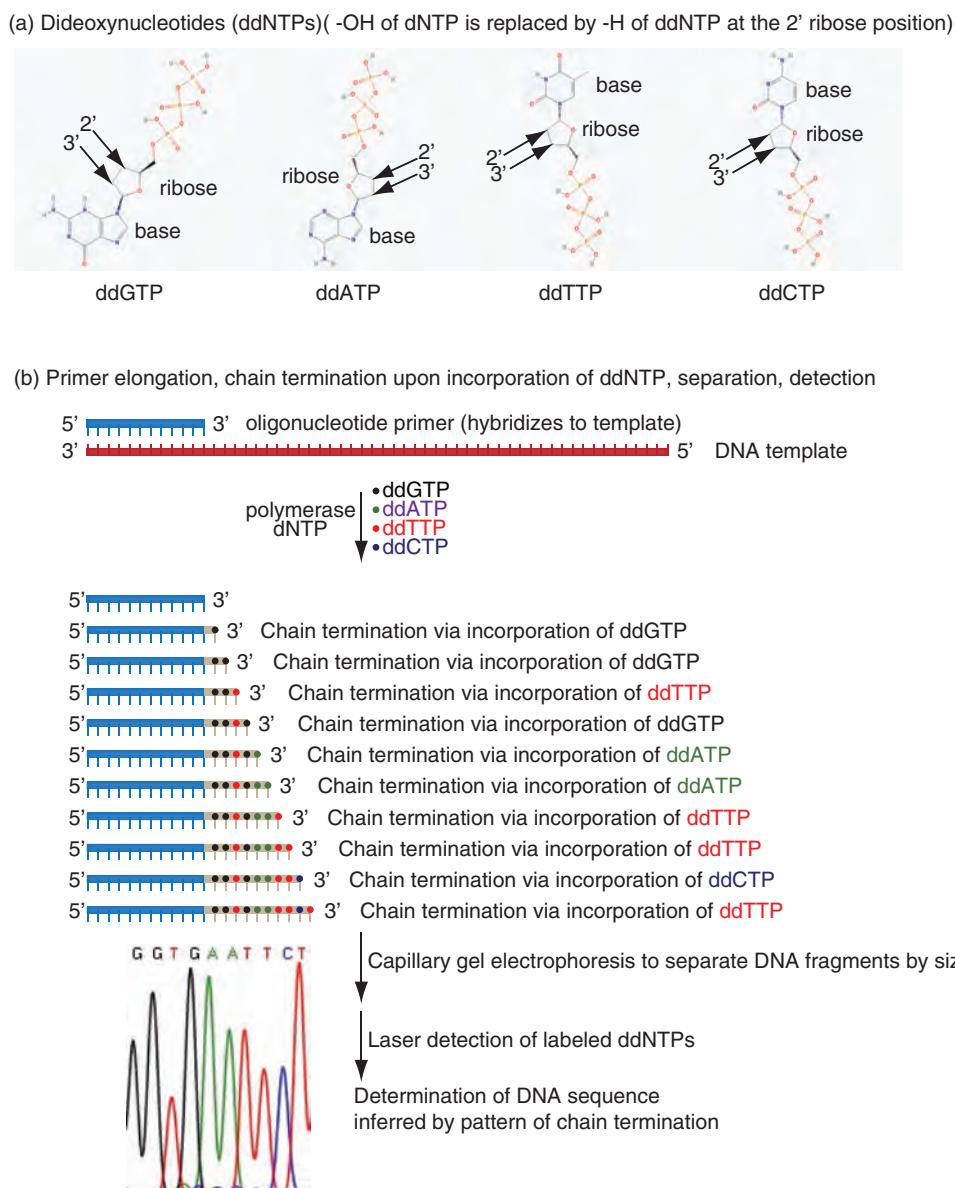
### Sanger Sequencing

Sanger and colleagues (1977) introduced the most commonly used technique for sequencing DNA, now called Sanger sequencing or dideoxynucleotide sequencing (Fig. 9.1). The principle is to obtain a template of interest (such as a fragment of genomic DNA or complementary DNA), denature it to yield single-stranded DNA, and add to it an oligonucleotide primer (typically about 20 nucleotides in length and complementary to the strand being sequenced). In the presence of DNA polymerase I (Klenow fragment) and the four 2'-deoxynucleotides (dNTPs), a second strand is synthesized. This synthesis can be inhibited by the addition of a dideoxynucleotide such as 2',3'-dideoxythymidine triphosphate (ddTTP; see Fig. 9.1a). Separate reactions include ddATP, ddGTP, or ddCTP accompanying the four dNTPs. Each dideoxynucleotide can be incorporated by a polymerase but lacks a 3' hydroxyl group on its ribose moiety; it therefore serves as a chain terminator preventing any further extension. The reaction with ddTTP contains a series of extended fragments, each sharing the same 5' end but terminating at various positions having a T residue. Four reactions are performed, DNA fragments are separated based on size, and the sequence is inferred (Fig. 9.1b). Samples travel by capillary electrophoresis to a detection area within a DNA sequencing machine where a laser excites the fluorophores, producing fluorescence emissions that correspond to the base calls. Improvements include better microfluidic separation devices and superior fluorescence detection (Metzker, 2005). In their 1977 paper, Sanger *et al.* reported that they could read as many as 300 bases in a set of reactions. Current reads can approach 800 or more bases.

From the 1970s through the completion of the human genome sequence in 2003, Sanger sequencing was the dominant method for genome sequencing. A typical sequencing facility can produce very high-quality reads (having an error rate of less than 1% per base; see below). Most large genome sequencing centers still rely on high-throughput Sanger sequencing for a variety of customized applications, such as verifying the sequence of a clone of interest.

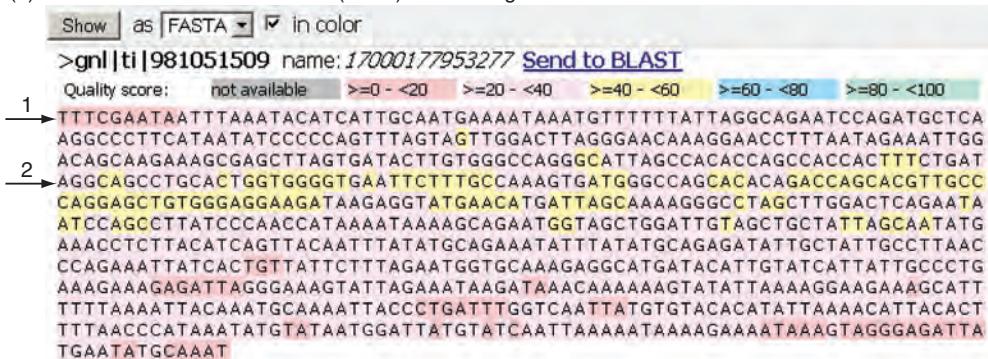
**Figure 9.2** shows a typical Sanger read (corresponding to the human beta globin gene). In addition to a FASTA-format nucleotide sequence, each base is assigned a quality

You can read about a standard Sanger sequencing machine, the Applied Biosystems 3730, at <http://www.3700.com> (WebLink 9.7).



**FIGURE 9.1** DNA sequencing by the Sanger method. (a) Structures of the four modified dideoxynucleotide (ddNTP) bases: 2'-3'-dideoxyguanosine 5'-triphosphate (ddGTP), 2',3'-dideoxyadenosine-5'-triphosphate (ddATP), 2',3'-dideoxythymidine 5'-triphosphate (ddTTP), and 2',3'-dideoxycytidine 5'-triphosphate (ddCTP). The 2' and 3' ribose positions have hydrogen atoms in the ddNTPs, while they have a 3' hydroxyl in DNA. (b) An oligonucleotide primer (in blue) (e.g., a 22-mer or synthetic nucleic acid of length 22 nucleotides) is hybridized to a single-stranded template (red) then extended using a DNA polymerase in the presence of dNTPs and a limited amount of one of the four ddNTPs. Chain termination occurs at one of the sites containing the ddNTP. The resulting synthesized fragments can be separated using a method such as capillary electrophoresis, and the products can be detected to infer the DNA sequence (bottom). The sequence in this example (GGTGAATTCT) corresponds to beta globin (**Fig. 9.2**). Structures are from the NIH PubChem Open Chemistry Database at NCBI (<http://pubchem.ncbi.nlm.nih.gov/>; compounds 446577, 65304, 65051, and 119119).

(a) Genomic DNA in Trace Archive (NCBI) from beta globin locus: FASTA format

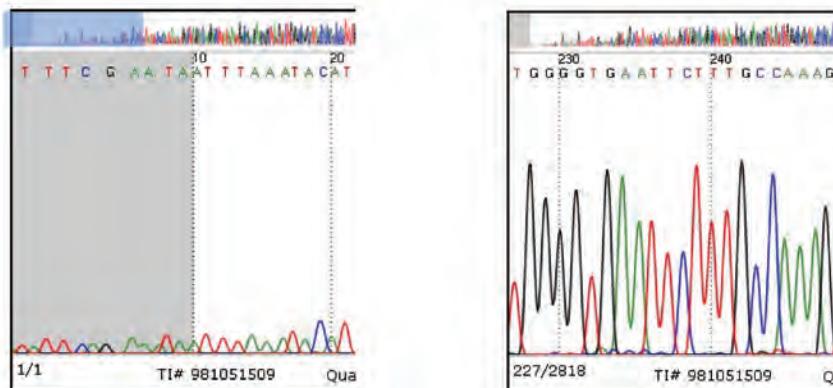


(b) Genomic DNA in Trace Archive (NCBI) from beta globin locus: base quality scores



(c) Sequence traces (region of low quality reads)

(d) Sequence traces (high quality reads)



**FIGURE 9.2** Base quality scores from Sanger sequencing. (a) A sequencing read called a trace (accession `TIJ981051509`) from the Trace Archive repository at NCBI. This archive contains raw data from genome sequencing projects; this trace was obtained by searching the archive by megaBLAST with human beta globin (`NM_000518.4`) as a query. The nucleotides are shown, color-coded according to quality score. (b) A display of the PHRED-scaled quality scores (the bottom rows were deleted). Each nucleotide is assigned a quality score. Sequence traces are shown (c) for the first 21 bases (see arrow 1 of (a)) and (d) bases in the middle portion having very high-quality scores (see arrow 2 of (a)). (c, d) Shown at the same scale. Reads that are ambiguous and difficult to read are associated with low-quality base scores.

*Source:* megaBLAST, NCBI.

**TABLE 9.1** Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, [http://en.wikipedia.org/wiki/DNA\\_sequencer](http://en.wikipedia.org/wiki/DNA_sequencer), and literature cited for each technology.

Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US\$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

Sanger reads and similar data based on gel and capillary platforms are stored at the Trace Archive of NCBI at <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi> (WebLink 9.8). As of October 2014 it contains over 2.1 billion traces. In Chapter 15 we query the Trace Archive in detail using a Perl script.

Next-generation sequencing technologies are sometimes referred to as “second-generation sequencing,” referring to the technology that has followed Sanger sequencing. “Third-generation sequencing” then refers to the currently emerging generation of sequencing tools.

score (see “Topic 2: From Generating Sequence Data to FASTQ” below). For regions having low-quality scores, it can be difficult to call nucleotides accurately and the error rate is high. For example, at a quality score of 20 ( $Q_{20}$ ) there is a  $10^{-2}$  or 1% error rate. Base quality scores below  $Q_{20}$  are often considered suspect, whether obtained by Sanger sequencing or next-generation sequencing. The software tools described in this chapter can filter out bases below a selected threshold.

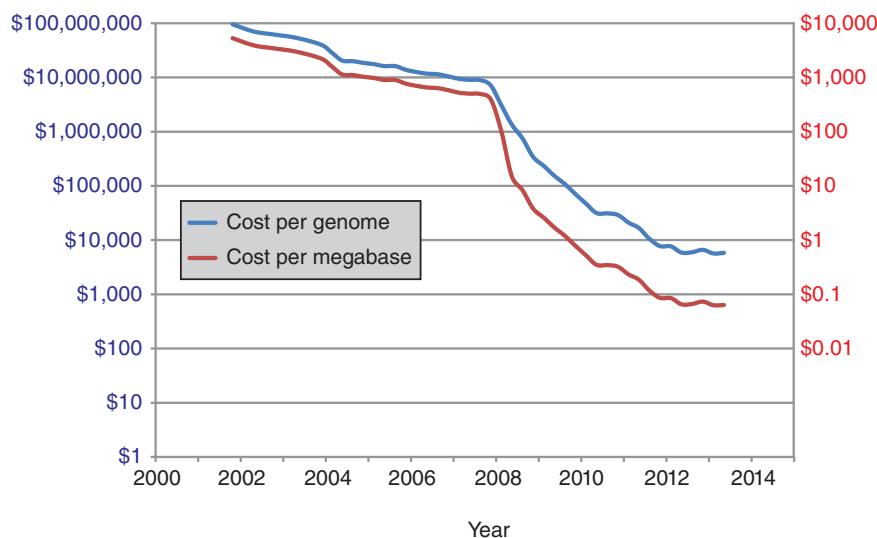
## Next-Generation Sequencing

A group of powerful new sequencing technologies has emerged in recent years. Collectively, these are known as next-generation sequencing. **Table 9.1** compares several properties of five prominent NGS technologies, as well as Sanger sequencing.

- When first introduced many of these platforms read lengths of just 35–50 base pairs, but now it is typical to achieve reads of hundreds of base pairs. The Pacific Biosciences platform is notable for its lengths of thousands of base pairs. This can be extraordinarily important in resolving duplicated regions and in genome assembly (see “Topic 3: Genome Assembly” below).
- The number of sequencing reads produced by each platform ranges from millions to even billions. This massively parallel output is key to the success of NGS.
- The time required for a run has become hours to days. Such reasonable time frames are helping NGS experiments become more routinely used tools.
- The cost per megabase of sequence is an outstanding feature of NGS technologies, in stark contrast to Sanger sequencing. While the human genome project was estimated to have cost US\$ 1–3 billion over a 15 year period, the first genome sequence of an individual (that of Craig Venter, reported by Levy *et al.*, 2007) was estimated to cost US\$ 80 million. We can obtain a whole-genome sequence today for well under US\$ 2000 (Fig. 9.3). This is also reflected in the steep drop in the cost of sequencing one megabase of DNA.
- Each technology introduces different, characteristic types of errors that influence the variants that are called at the end of the data analysis pipeline. Mark DePristo *et al.* (2011) show data from this using the GATK package that we will describe. Michael Snyder and colleagues (Clark *et al.*, 2011) also show this by also analyzing an exome using three different methods for exome enrichment. The capture methods differed in features such as target choice, oligonucleotide bait lengths, and bait density.

## Cyclic Reversible Termination: Illumina

Illumina technology can generate one terabase (Tb) of DNA sequence data (1000 gigabases or Gb) in a single run of its HiSeq machine. The HiSeq X ten instrument generates up



**FIGURE 9.3** The decline of DNA sequencing costs. The cost per megabase (million base pairs) of DNA sequence is shown with respect to a minimum quality score of  $Q_{20}$  (or PHRED<sub>20</sub>). Cost per genome refers to a human-sized genome. The y axis is in logarithmic units. Adapted from Kris Wetterstrand, “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program” (available at <http://www.genome.gov/sequencingcosts/>).

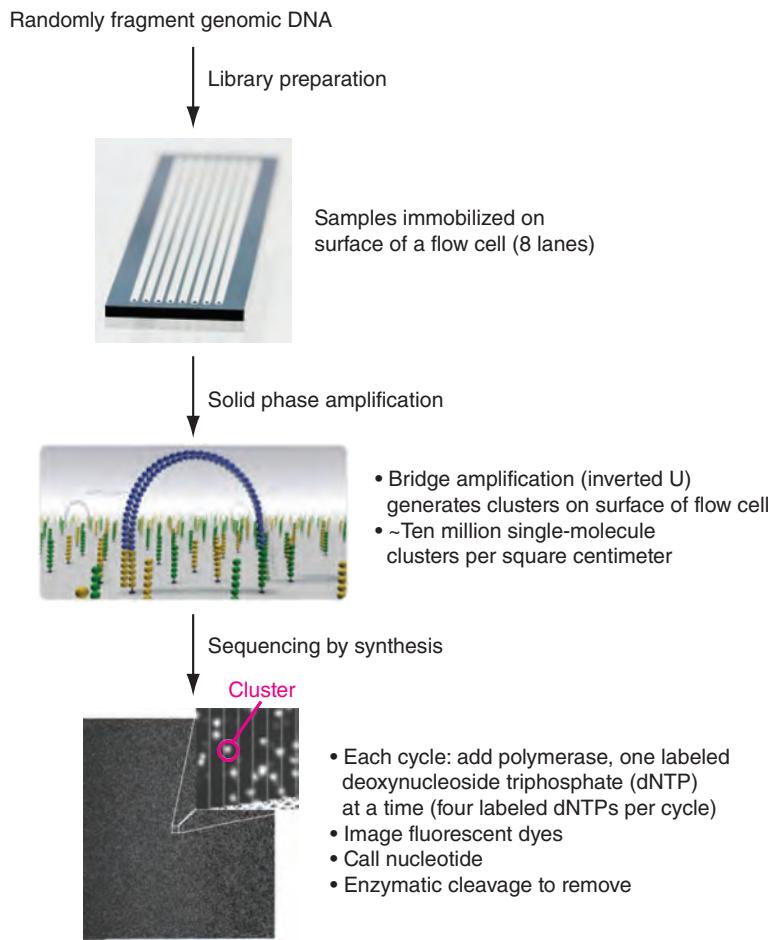
to 1.8 Tb of data in each run. My lab has a small sequencing machine (a MiSeq) that generates 5 Gb of DNA per run in about 24 hours.

Illumina sequencing works on the principle of cycle reversible termination which functions as follows (Fig. 9.4). (a) Genomic DNA is purified and then randomly fragmented. This can be accomplished mechanically by methods such as sonication, shearing, or nebulization, often followed by size selection of the randomly fragmented DNA. Adapters are attached to both ends. (b) Single-stranded DNA fragments are covalently attached to the surface of flow cell channels. (c) The addition of DNA polymerase and unlabeled deoxynucleotides creates solid-phase “bridge amplification” in which the template DNA makes U-shaped loops with both ends attached to the surface of the channel. (d) Double-stranded bridges are formed. The double-stranded molecules are denatured and then amplified to generate dense clusters of template DNA. (e) Four labeled reversible terminators are added (with primer and DNA polymerase). Only a single reversible terminator will be added to each template in a given cycle. As with Sanger sequencing, chain termination will occur at specific bases that cannot elongate. (f) Following laser excitation, the identity of the first base is recorded. (g) For the second cycle, the reversible terminators are removed (by deprotection). All four labeled reversible terminators and the polymerase are again added to the flow cell. The cycles are repeated.

The Illumina system is very fast and generates massive amounts of sequence data. Its read lengths are typically 150 bases or longer, making it particularly appropriate for resequencing projects (Bentley *et al.*, 2008). Paired end reads spanning 600 base pairs are becoming available. The main advantages of this approach relative to Sanger sequencing are its scalability and the elimination of the need for gel electrophoresis. The main advantages relative to pyrosequencing (described in the following section) are that all four bases are present at each cycle and the sequential addition of dNTPs allows homopolymer tracts to be accurately read.

Currently, the Illumina platform accounts for about 80% of all next-generation sequence data that are being generated.

You can learn more about the Illumina system at <http://www.illumina.com/> (WebLink 9.9).



**FIGURE 9.4** Sequencing by Illumina technology. Genomic DNA is randomly fragmented and further processed (e.g., by size selection and addition of adapters). Samples are immobilized on a lane on the surface of a flow cell. Solid phase amplification uses bridge amplification to generate a lawn of clusters. Flow cell image is from [http://res.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf). Bridge amplification image is from <http://systems.illumina.com/systems/miseq/system.html>. Cluster image is from the NIH Open Image database ([http://openi.nlm.nih.gov/detailedresult.php?img=2734321\\_btp383f1&req=4](http://openi.nlm.nih.gov/detailedresult.php?img=2734321_btp383f1&req=4); Whiteford *et al.*, 2009).

## Pyrosequencing

Pyrosequencing is one of the powerful new alternative technologies that has gained prominence (Rothberg and Leamon, 2008). First introduced by Hyman (1988), it forms the core of the 454 Life Sciences Corp. technology that has produced dramatic genome sequencing results (Margulies *et al.*, 2005). That group sequenced and assembled the entire *Mycoplasma genitalium* genome (580,069 bases) with 96% coverage and at 99.96% accuracy with a single run of a sequencing machine. Roche Diagnostics Corporation, the company that owns 454, will soon phase out this technology, but we present it as a major force in sequencing. As of late 2014, over 3000 publications cite the use of this technology.

The 454 Life Sciences Corp. website is <http://www.454.com/> (WebLink 9.10).

A key feature of pyrosequencing is that only one dNTP is added into the reaction at a time. The principle is outlined in **Figure 9.5**. DNA is immobilized on beads that capture (on average) one single-stranded template that is amplified using the polymerase chain reaction (PCR). The template is placed in small (picoliter volume) wells, with 1.6 million wells per plate, and one dNTP is added to the wells per cycle. The reaction mixture

contains the template DNA, a sequencing primer, four enzymes (DNA polymerase I, ATP sulfurylase, luciferase, and apyrase) as well as the substrates adenosine 5' phosphosulfate (APS) and luciferin (**Fig. 9.5a**). In each cycle a single dNTP is added and is incorporated into the nascent strand until a different dNTP is required (**Fig. 9.5b**). Upon incorporation of each dNTP, an equimolar amount of pyrophosphate (PPi) is generated. This PPi is converted to ATP by ATP sulfurylase (**Fig. 9.5c**) and the ATP promotes the luciferase-mediated conversion of luciferin to oxyluciferin with the generation of light (**Fig. 9.5d**). The emitted light is detected with a charge coupled device (CCD) camera. The amount of light is measured over time (**Fig. 9.5e**) to indicate at which position a nucleotide was incorporated; because of the quantitative nature of this process, the incorporation of two nucleotides creates twice the light output. Apyrase degrades both unincorporated dNTPs and excess ATP, clearing the system for repeated cycles with low background noise (**Fig. 9.5f**). In this process dNTPs are systematically added across different cycles, but dATP $\alpha$ S is used in place of the usual dATP because it is efficiently used by DNA polymerase I but is not a substrate for luciferase. A schematic of the output, showing a sequencing read of GACCGTTC, is shown in **Figure 9.5g**.

Pyrosequencing offers many advantages. (1) It is very fast and the cost per base is low relative to Sanger sequencing. (2) One run of an experiment can generate up to 600 megabases of raw nucleotide sequence data, a massive amount. (3) DNA molecules are amplified without the need for bacterial cloning; this is especially helpful for metagenomics and ancient genomics projects. (4) The accuracy of the reads is very high.

There are also several major disadvantages of pyrosequencing technology. The sequencing reads are short (several hundred base pairs), making whole-genome assembly extremely challenging. Another disadvantage is that the machine has difficulty in sequencing homopolymers (e.g., a string of 10 identical nucleotides). Huse *et al.* (2007) compared about 340,000 sequencing reads to reference templates of known sequence and determined the error rates. Errors involved homopolymer effects, insertions, deletions, and mismatches. While these errors were distributed along the length of each read, they found that 82% of all the reads had no errors while only a small percent had a disproportionately large number of errors. By identifying and removing such low-quality reads, they could improve the overall accuracy of the dataset from 99.5% to 99.75% or higher.

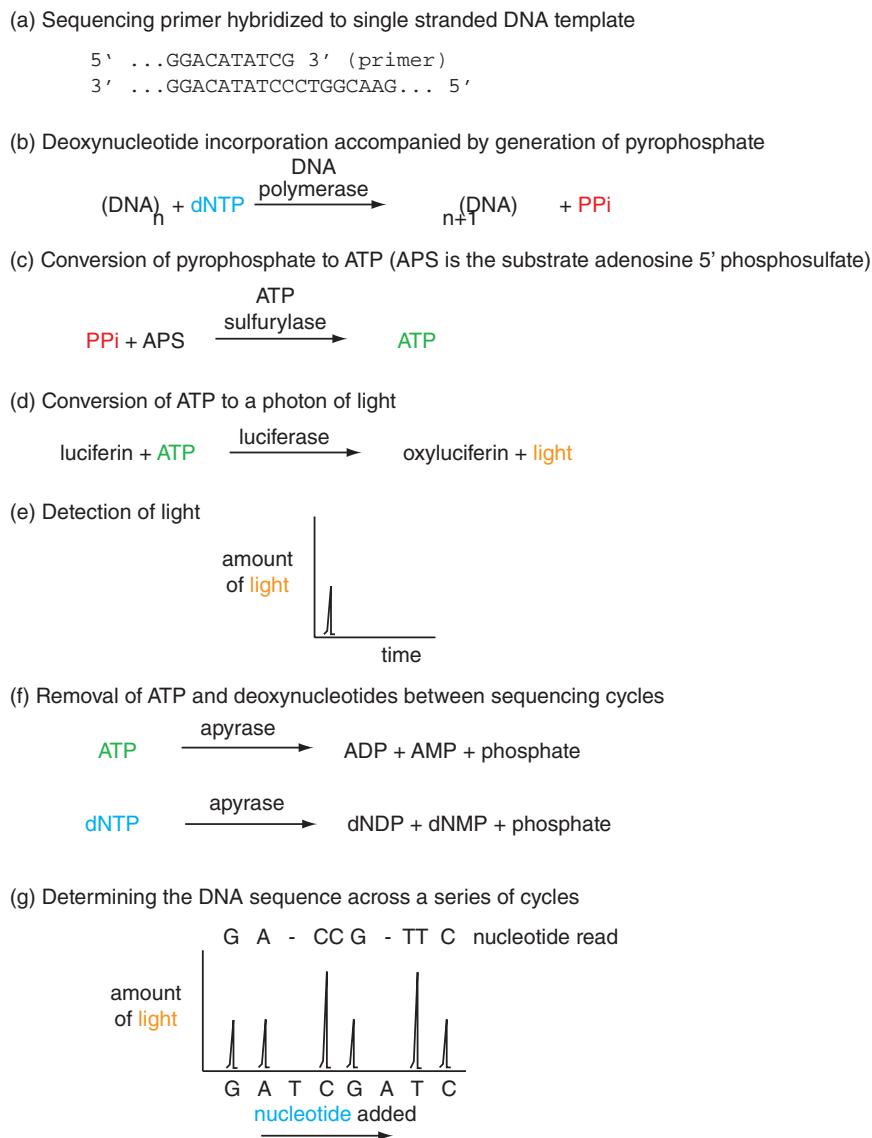
Applications of 454 technology include some of the ancient DNA and metagenomics sequencing projects such as the sequencing of the Neanderthal genome and identifying the microbial community in parts of a mine in Minnesota (Edwards *et al.*, 2006; see also Chapter 15).

### Sequencing by Ligation: Color Space with ABI SOLiD

The Applied Biosystems ABI SOLiD™ offers extremely accurate sequencing with extremely low error rates. Library preparation is similar to that of other sequencing technologies. Genomic DNA is sheared, adapters are ligated, and emulsion PCR is performed to generate bead clones, each of which contains a single insert sequence. Other technologies then use DNA polymerases to incorporate labeled dideoxynucleotide chain terminators. SOLiD is dramatically different, performing sequencing by ligation instead of sequencing by synthesis. A mixture of degenerate oligonucleotides is added to each reaction, having 3–5 Ns (i.e., any residue) followed by one of 16 specific dinucleotides adjacent to the 3' end. Each oligo (with  $n = 16$  possible dinucleotides) is attached to a dye ( $n = 4$ ). Reading a single color does not specify a single base, but rather corresponds to any of four possible dinucleotides. By interrogating each base position twice, the base can be called unambiguously. This approach is referred to as using “color space.” While color space presents a series of data analysis challenges (including problems with assembly; Flicek and Birney, 2009), its strength is its very low base-calling error rate.

The symbol N corresponds to any nucleotide residue. N may refer to a randomly selected residue (e.g., during oligonucleotide synthesis), to a base position at which a nucleotide cannot be determined (e.g., because of limitations of a sequencing technology), or to any base (e.g., the pattern GNNNC refers to a motif in which a G and a C residue are separated by any three nucleotides).

SOLiD technology is described at  
 ↗ <http://www.appliedbiosystems.com> (WebLink 9.11).



**FIGURE 9.5** Pyrosequencing. (a) A single-stranded DNA template is immobilized on a bead and amplified by PCR. After transfer to a small well, primer is added as well as additional enzymes and substrates and one of the four deoxynucleotides (dGTP, dCTP, dTTP or, in place of dATP, the modified nucleotide dATPyS). (b) DNA polymerase I catalyzes the addition of a single deoxynucleotide, releasing pyrophosphate (PPi). If there is a sequence of  $n$  nucleotides in a row in the template DNA, an equimolar amount of PPi will be released. (c) ATP sulfurylase converts a substrate (APS) and PPi to adenosine triphosphate (ATP). (d) Luciferase, in the presence of its substrate luciferin and the ATP, produces a product (oxygenated luciferin) and light. (e) A charge-coupled camera detects the light and provides an intensity measurement over time. The y axis is proportional to the amount of deoxynucleotide that was incorporated, thus specifying whether zero, one, two, or more dNTPs occur in the template DNA in that position. (f) Apyrase cleaves ATP, clearing the system for successive cycles. (g) The light patterns emitted from a series of cycles allow the DNA sequence of the template to be read. The longest reads with current technology approach 1000 bases. Because of the massively parallel nature of this process, tens of millions of base pairs of high-quality sequence can be generated with this technology.

## **Ion Torrent: Genome Sequencing by Measuring pH**

The idea behind Ion Torrent's sequencing technology is remarkably simple. When a DNA polymerase incorporates a nucleotide into a strand of DNA, a hydrogen ion is released. The sequencing machine is essentially a pH meter with 1.2 million wells that can distinguish incorporation of the four bases (Rothberg *et al.*, 2011). It can identify the incorporation of two or more bases as a signal with increased voltage, but a limitation of this technology is its low accuracy with homopolymer runs. This method involves direct detection of the chemical reaction on a semiconductor chip using an ion sensor without scanning, cameras, or light sources.

Ion Torrent is sold by Life Technologies. You can learn more at <http://lifetechnologies.com> (WebLink 9.12). Note that sequence reads generated with this technology should use mappers that are customized for Ion Torrent.

## **Pacific Biosciences: Single-Molecule Sequencing with Long Read Lengths**

Pacific Biosciences enables the real-time sequencing of a DNA molecule using a DNA polymerase and four distinguishable labeled deoxyribonucleoside triphosphates (dNTPs; Eid *et al.*, 2009). The activity of the DNA polymerase was measured in this way, directly observing processive incorporation of nucleotides (at a rate of 4.7 bases/second). This approach relies on a zero-mode waveguide nanophotonic structure: detection of single fluorophores is enabled by attaching a polymerase to the bottom of a well where it can interact with a DNA molecule and receive excitation by a laser light in a volume of just zeptoliters ( $10^{-21}$  L). In this confined space, dNTP concentrations as high as 10  $\mu$ M are achieved.

A great advantage of this technology is that read lengths can average 5000 base pairs, sometimes exceeding 20 kb. Koren *et al.* (2013) reported the advantage of using these long reads to assemble bacterial and archaeal genomes with very high accuracy; we show this in the section “Topic 3: Genome Assembly” below. Another unique strength of Pacific Biosciences technology is that it allows epigenomic analyses concurrently with DNA sequence determination. For example, it can report adenine and cytosine methylation.

The Pacific Biosciences home page is <http://www.pacificbiosciences.com/> (WebLink 9.13).

## **Complete Genomics: Self-Assembling DNA Nanoarrays**

Complete Genomics introduced a platform that produces highly accurate genome sequences (Drmanac *et al.*, 2010). Genomic DNA is fragmented, cloned into circular vectors, and single-stranded vectors containing hundreds of copies of insert sequence are assembled into self-assembling DNA nanoballs. Sequencing is then performed with a ligation chemistry called combinatorial probe anchor ligation sequencing. Complete Genomics further extended their technology to sequencing whole genomes from collections of just 10–20 human cells, permitting accurate haplotyping (Peters *et al.*, 2012).

Visit <http://www.completemicrogenomics.com/> (WebLink 9.14), a BGI company (<http://www.genomics.cn/en/index>, WebLink 9.15). The Complete Genomics website currently offers a series of analysis tools and access to data from 69 whole human genomes.

# **ANALYSIS OF NEXT-GENERATION SEQUENCING OF GENOMIC DNA**

## **Overview of Next-Generation Sequencing Data Analysis**

Overviews of next-generation sequence data analysis have been presented by Stein (2011) and Pabinger *et al.* (2014). We present a broad outline of sequence analysis in **Figure 9.6**, and examine 11 topics as follows. (1) Experimental design and sample preparation, in which it is essential for the biologist to have an intimate role.

The next stages are often performed at a core facility in which experts implement a workflow. However, the biologist can and should understand all the steps that are taken, especially since they are fundamental to the outcome of the experiment. (2) The generation of sequence data and FASTQ formatted files, which also includes quality assessment of FASTQ data. Are the quality scores above an appropriate threshold (such

Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	
Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline	Quality assessment Alignment to reference genome Variant identification Annotation	FASTQ Reference: FASTA Output: SAM/BAM Variant Call Format ( VCF/BCF)
Visualization	Trimming, filtering Software: FastQC	
Prioritization	Software: BWA, Bowtie2	
Storage	Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	VCF
	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	BAM, VCF

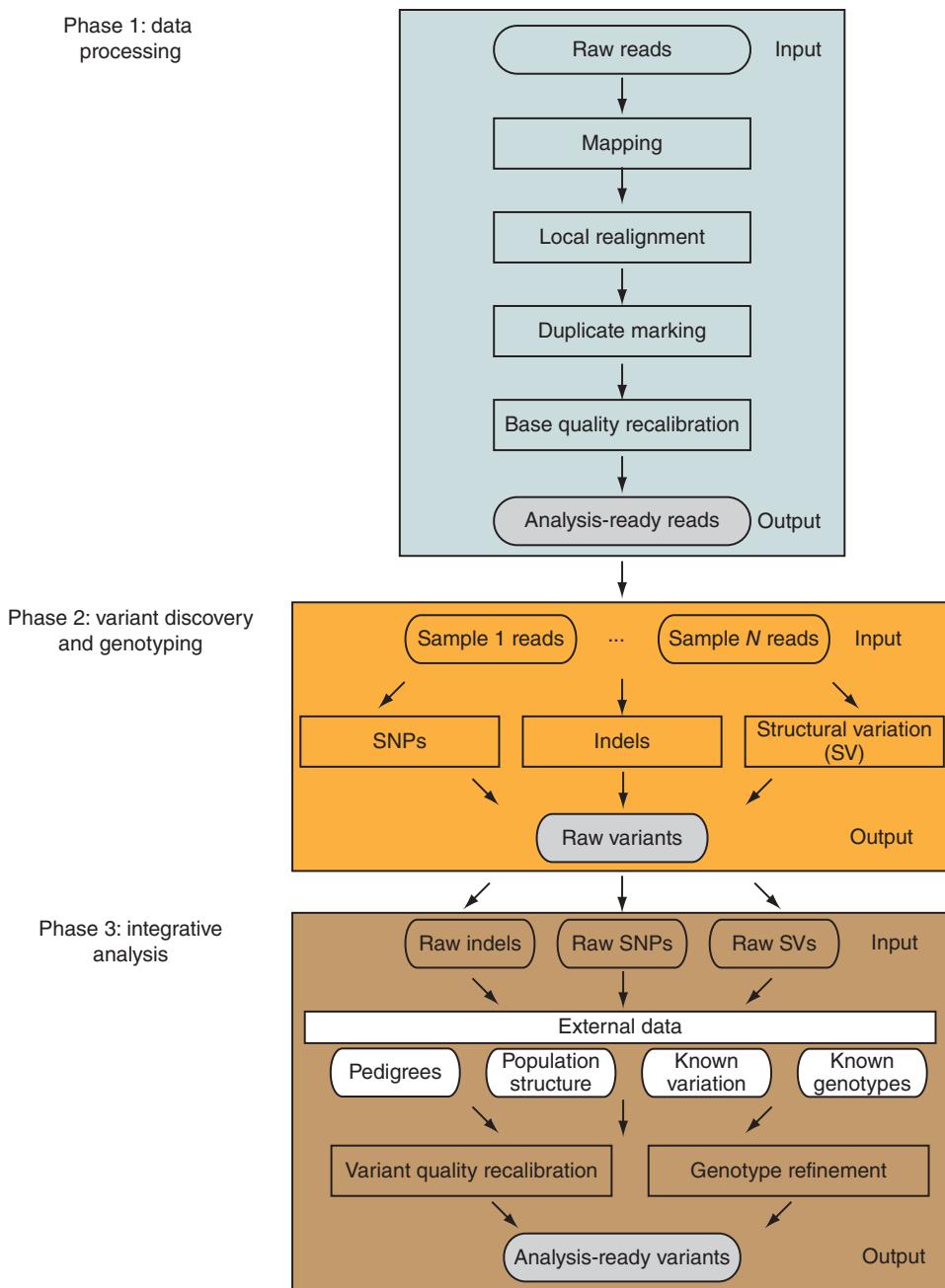
**FIGURE 9.6** Workflow for next-generation sequence experiments: from experimental design to data analysis. We describe software tools and data formats in this chapter.

as an error rate of 1 in 1000)? (3) Genome assembly (if needed). (4) Sequence alignment to a reference genome, including measurement of read depth and alignment of repetitive DNA. (5) The SAM/BAM format and SAMtools software: storing and analyzing aligned reads. (6) Variant calling for single-nucleotide variants. (7) Variant calling for structural variants. Are there insertions/deletions (indels), inversions, or other complex variants? (8) Summarizing variation with the variant call format (VCF). (9) Visualizing next-generation sequence data as well as genomic arithmetic with IGV, BEDTools, and bigBed.

The final steps are often the responsibility of the biologist. (10) Interpreting the biological significance of variants; this may be followed up by validation studies such as targeted sequencing of candidate disease alleles. (11) Depositing (storing and sharing) data in repositories.

There are many variations on these 11 topics. For example, some studies involve pedigrees, studies of somatic variation, or other genetic topics; some involve related technologies such as ChIP-Seq, RNA-seq, and methylation studies. We conclude this chapter by briefly describing some of these alternate topics.

The outline presented in **Figure 9.6** is relatively simplistic. We will use that workflow to gain experience with basic data manipulation. A state-of-the-art workflow, used by many experts, is the Genome Analysis Toolkit (GATK; McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). We show the GATK workflow in **Figure 9.7**, and explain why its approaches are crucial to sensitive, specific analyses of genome sequences.



**FIGURE 9.7** Workflow for variant discovery and genotyping from next-generation DNA sequencing using GATK. In the first phase, raw reads (in the FASTQ format) are mapped to a reference genome, realigned, duplicate reads are removed, and base quality scores are recalibrated. In the second phase variants are identified in the three categories of single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants (SVs). In the third phase, quality scores of variants are recalibrated and genotypes are refined in the context external data sources that inform the analyses. The steps introduced by GATK greatly reduce both false negative and false positive errors. Adapted from DePristo *et al.* (2011), with permission from Macmillan Publishers.

### Topic 1: Experimental Design and Sample Preparation

In order to sequence DNA from a source of interest, genomic DNA must be purified and prepared in the form of a library. For whole-genome sequencing (WGS), genomic DNA is typically fragmented (through nebulization or mechanical shearing) to produce

fragments that are size-selected to some desired range (e.g., 300 base pairs). For targeted sequencing, regions of genomic DNA are enriched (“captured”). One popular approach is whole-exome sequencing (WES) in which DNA corresponding to exons is selectively enriched for sequencing. This is of interest for studies in which coding region variants are studied. Another approach is targeted sequencing of particular loci of interest. This may be a set of genes you are interested in, such as the panel of 101 autism genes we study in this chapter, or 16S ribosomal DNA for metagenomics studies (Chapters 15–17). Another application could be targeted sequencing of a region implicated in disease (e.g., from a genome-wide association study or GWAS, described in Chapter 21).

One main source of excitement about genome sequencing is that it may solve the genetic basis of many diseases. For any research studies involving humans in the United States, it is necessary to obtain Institutional Review Board approval. For clinical studies, IRB approval is not required, but guidelines for protecting patients’ rights are required and laboratories must be accredited for clinical work. There are many essential issues to consider, such as:

- *Informed consent*: do the participants understand the potential risks and benefits of obtaining genetic information?
- *Privacy*: if the sequence data are deposited in a repository such as dbGaP at NCBI (Chapter 21), can privacy be ensured? Privacy involves limiting others’ access to information about a person. It has been suggested that no samples can be adequately deidentified.
- *Confidentiality*: if genetic information is obtained about a person’s genome, is that information protected or misused?
- *Ownership of data*: if a child’s genome is sequenced, the parents’ genomes are also routinely sequenced. This is done to determine which variants are inherited by the child and which variants occur *de novo* (and are therefore potentially clinically relevant). What happens if there are incidental findings, such as a mutation in a cancer gene in a patient with an unrelated condition? What if there are clinically relevant findings in the parents’ genomes, when the focus of the study had been on the child’s genome?

A violation of privacy would be visiting someone’s house and looking in their private clothes drawer. A violation of confidentiality would be telling others what you saw there.

Once genomic DNA is purified and quantified, it is packaged into a library. Paired-end libraries are routinely created; in these, DNA inserts are size-selected and sequenced from both the 5' and 3' ends. (It is possible to make inserts and sequence from only one end.) An advantage of paired-end libraries is that paired reads are generated which can be mapped to a particular genomic location. If that position is unexpectedly far apart this may indicate an insertion has occurred in the sample, while if the reads are closer than expected this can be interpreted as a deletion. We discuss structural variant discovery in “Topic 7: Variant Calling: Structural Variants” below.

## Topic 2: From Generating Sequence Data to FASTQ

When genomic DNA is sequenced, raw image files are typically generated that are used to interpret which nucleotide is called in a given fragment or “read” (Ledergerber and Dessimoz, 2011). For most users, the FASTQ files (rather than the underlying image files) represent the raw reads from which other analyses are performed. Like the FASTA format, the FASTQ format includes a sequence string, consisting of the nucleotide sequence of each read. FASTQ also includes an associated quality score for every base. I have supplied two FASTQ files (forward and reverse reads) from an experiment sequencing autism genes, as well as three other file types which are discussed in “Topic 5: The SAM/BAM Format and SAMtools” and “Topic 8: Summarizing Variation: The VCF Format and VCFtools” below (`.bam`, `.bam.bai`, and `.vcf`). You can copy them from the textbook website to

your Linux machine (or Mac or PC) to study them. We first use `ls` to list the files of our directory, with the `-lh` option to list them fully in human-readable form.

```
$ ls -lh
total 1.5G
-rwxrwxr-x. 1 pevsner pevsner 326M Oct 17 15:52 mysample1.bam
-rwxrwxr-x. 1 pevsner pevsner 3.0M Oct 17 15:52 mysample1.bam.bai
-rwxrwxr-x. 1 pevsner pevsner 574M Oct 17 15:52 mysample1_R1.fastq
-rwxrwxr-x. 1 pevsner pevsner 573M Oct 17 15:52 mysample1_R2.fastq
-rwxrwxr-x. 1 pevsner pevsner 55K Oct 17 15:52 mysample1.vcf
```

The two FASTQ files are available at Web Documents 9.1 and 9.2 at  
<http://bioinfbook.org>. To read about the FASTQ format see  
<http://maq.sourceforge.net/fastq.shtml> (WebLink 9.16).

This shows that the FASTQ files are each about 570 megabytes (if compressed as `mysample1_R1.fastq.gz` files, they each occupy ~230 MB). We can use the word count program `wc -l` (where `-l` specifies lines) to see that each FASTQ file has about 7.2 million rows. Next we can use `head` to view the beginning lines of one of the files.

```
$ head mysample1_R1.fastq
@M01121:5:00000000-A2DTN:1:1101:19726:2176 1:N:0:2
GNCTAACTCTGGCTGAAGGACTAGCTAACGCTGCTGGACAGAGGCCTTGAGGGGCCCTGCCCCACTGTTAT
TCTCAGAGCTGGCATATGGGGAGAGGTGGGTGA
+
A#>>AAAFFBFFGGGG1EGGFHHHHHGHHFGE?EFHGAE0BEEFHHF2AGEGCGCFFGHEFFGHHHG
HHGGHHHHHEFH/GHHGHFEEAE>E/FGEEG</
```

Each FASTQ file has records that are in blocks four lines long. The first line, beginning with the `@` symbol, identifies the record. It may optionally include information about the sequence length or the machine used for sequencing. The second line has the sequence (in upper case), including the nucleotides G, A, T, C, and (as is the case here in the second position) there may be an N for unknown nucleotide. The third line begins with the `+` symbol and typically contains just that character (as in this case), or it can have more information. The fourth line includes the quality scores corresponding to every base. Each quality score is assigned a single character, and the entire quality score string must equal the length of the sequence string.

If you wish to analyze a FASTQ file in Linux, be careful because the `@` and `+` symbols denote the start of lines 1 and 3 of each record but they may also denote a base quality score within the fourth line. A tool such as `grep` should therefore be used with caution in extracting information.

Cock *et al.* (2010) reviewed the three different types of FASTQ file formats: the Sanger standard format (which is currently the most commonly used format); a second format introduced by Solexa, Inc. (now Illumina, Inc.) in 2004; and an Illumina 1.3+ FASTQ format. All three have different meanings because the quality scores are scaled differently. The standard Sanger format relies on quality scores  $Q$  that are also called PHRED scores, defined:

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e) \quad (9.1)$$

where  $P_e$  refers to the estimated probability of a base call error. PHRED scores were introduced in 1988 by Phil Green and colleagues to describe base quality scores from Sanger sequencing. This definition is used in the Sanger FASTQ format. Characters are stored as ASCII printable characters 33–126 (i.e., with an ASCII offset of 33) so the range of possible quality scores is 0 to 93. These are listed in Figure 9.8. A value of 93 corresponds to a probability of  $10^{-9.3}$  that a base call occurred by chance, that is, the read is extremely likely to be correct. At  $Q_{30}$  there is a 1:1000 (i.e.,  $10^{-3}$ ) error rate, a threshold that is often set as a minimum for high-quality reads. For the first four nucleotides GNCG in the example above, the quality scores A#>> correspond to Sanger FASTQ values of 32, 2, 29, and 29. The N residue therefore has an extraordinarily low-quality score.

The 2004 Solexa definition of a quality score was given as:

$$Q_{\text{Solexa}} = -10 \times \log_{10} \left( \frac{P_e}{1 - P_e} \right). \quad (9.2)$$

ASCII refers to the American Standard Code for Information Interchange, a character-encoding scheme that encodes 128 characters. The first 32 codes (numbers 0–31 decimal) are reserved control characters that were not intended to be printed. ASCII 32 is the space character, so Sanger FASTQ files use ASCII 33–126. This corresponds to PHRED qualities 0–93. You can view an ASCII table that defines its one-character symbols at  
<http://www.asciiitable.com/> (WebLink 9.17).

Dec	Char	Dec	Char	Sanger FASTQ	Dec	Char	Sanger FASTQ	Dec	Char	Sanger FASTQ
0	Non-printing	32	Space	64	@	31	96	.	63	
1	Non-printing	33	!	0	65	A	32	97	a	64
2	Non-printing	34	"	1	66	B	33	98	b	65
3	Non-printing	35	#	2	67	C	34	99	c	66
4	Non-printing	36	\$	3	68	D	35	100	d	67
5	Non-printing	37	%	4	69	E	36	101	e	68
6	Non-printing	38	&	5	70	F	37	102	f	69
7	Non-printing	39	'	6	71	G	38	103	g	70
8	Non-printing	40	(	7	72	H	39	104	h	71
9	Non-printing	41	)	8	73	I	40	105	i	72
10	Non-printing	42	*	9	74	J	41	106	j	73
11	Non-printing	43	+	10	75	K	42	107	k	74
12	Non-printing	44	,	11	76	L	43	108	l	75
13	Non-printing	45	-	12	77	M	44	109	m	76
14	Non-printing	46	.	13	78	N	45	110	n	77
15	Non-printing	47	/	14	79	O	46	111	o	78
16	Non-printing	48	0	15	80	P	47	112	p	79
17	Non-printing	49	1	16	81	Q	48	113	q	80
18	Non-printing	50	2	17	82	R	49	114	r	81
19	Non-printing	51	3	18	83	S	50	115	s	82
20	Non-printing	52	4	19	84	T	51	116	t	83
21	Non-printing	53	5	20	85	U	52	117	u	84
22	Non-printing	54	6	21	86	V	53	118	v	85
23	Non-printing	55	7	22	87	W	54	119	w	86
24	Non-printing	56	8	23	88	X	55	120	x	87
25	Non-printing	57	9	24	89	Y	56	121	y	88
26	Non-printing	58	:	25	90	Z	57	122	z	89
27	Non-printing	59	;	26	91	[	58	123	{	90
28	Non-printing	60	<	27	92	\	59	124		91
29	Non-printing	61	=	28	93	]	60	125	}	92
30	Non-printing	62	>	29	94	^	61	126	~	93
31	Non-printing	63	?	30	95	_	62	127	DEL	

**FIGURE 9.8** FASTQ quality scores use ASCII coding symbols. The chart shows ASCII symbols (Char columns indicate characters) corresponding to decimal notation 1–127. The characters for 0–31 (which are not used for printing) are not shown. Decimal 31 is a space, and subsequent characters (decimal 33–126) are used to represent base quality scores in the Sanger FASTQ format. For example, if base calls GATC have quality scores of 28, 30, 25, and 31 then the symbols for their quality scores are =?:@. Adapted from <http://www.lookuptables.com>.

This format uses an ASCII range of 59126 (offset 64) and has a range of values from –5 to 62. It can be interchanged with the Sanger FASTQ format (Cock *et al.*, 2010):

$$Q_{\text{PHRED}} = 10 \times \log_{10}(10^{Q_{\text{Solexa}}/10} + 1). \quad (9.3)$$

Several tools such as Maq (Li *et al.*, 2008) interconvert FASTQ formats. NCBI has converted Solexa to Sanger formatted FASTQ files.

#### *Finding and Viewing FASTQ files*

It is helpful to look at FASTQ files to learn their format and size, and to learn how to manipulate them. You can visit the Sequence Read Archive (SRA) at NCBI to obtain FASTQ files from a vast number of experiments. NCBI provides an SRA ToolKit, a program to download files, or you can browse for files of interest.

To use SRA Toolkit, follow the download instructions from NCBI. On a Linux or OS X computer navigate to the `bin` directory and use the `fastq-dump` utility to download data from a typical SRA file, accession SRR390728. Note that there are six SRA accession

The SRA website at NCBI is at  
<http://www.ncbi.nlm.nih.gov/sra/> (WebLink 9.18). This site includes documentation with detailed instructions for installing and using the SRA ToolKit on a Linux, OS X, or Windows machine.

types. (1) SRA is an SRA submission accession. This is a virtual container holding objects from the other five types. (2) SRP is an SRA study accession containing project metadata, that is, a study summary. (3) SRX is an SRA experiment accession, including metadata, platform, and experimental details. (4) SRS is an SRA sample accession describing the physical sample. (5) SRZ is an historical SRA analysis accession containing a sequence data file and metadata. (6) SRR is an SRA run accession having sequencing data such as SRR390728 that we will use. A given experiment may have multiple runs (SRR files). We will include the argument `-X 3` to specify that we want to print just the first three spots, and `-Z` sends them to standard output.

```
$ fastq-dump -X 3 -Z SRR390728
Read 3 spots for SRR390728
Written 3 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTCACGTTCTCGAGCTTGGTTTCAGCGATGGAGAATGACTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
:::::::::::::::::::9:::665142:::::::::::::::::::96&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGACTGGGT
+SRR390728.2 2 length=72
::::::::::::::::::4:::::3;393.1+4&&5&&:::::::::::::::::::<9;<:::::464262
@SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTACCCGTTTACTTATTATTATTATTGAGACAGAGCATTGGTC
+SRR390728.3 3 length=72
-;;8:::::,*';'-4,44,:&,1,4'./&19:::::669;;99:::::-;3;2;0;+;7442&2/
```

Next, we can add the `-fasta` argument to output three entries of just FASTA formatted data and, with the number 36, we specify that we want 36 bases per line:

```
$ fastq-dump -X 3 -Z SRR390728 -fasta 36
Read 3 spots for SRR390728
Written 3 spots for SRR390728
>SRR390728.1 1 length=72
CATTCTCACGTTCTCGAGCTTGGTTTCAGC
GATGGAGAATGACTTGACAAGCTGAGAGAAGNTNC
>SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTCTACGAGCTTGTGT
TCCAGCTGACCCACTCCCTGGGTGGGGGACTGGGT
>SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTACCCGTTTACTT
ATTATTATTATTATTGAGACAGAGCATTGGTC
```

You can also access FASTQ files via the European Nucleotide Archive (ENA). Enter a text search “1000genomes exome.” (We search for exomes because they are relatively small.) Take the first file (sample accession SRS001696) and click the download option “Fastq files (galaxy).” This directs you to a Galaxy page with the FASTQ file entered into your history. (It has 6.2 million sequences.) Within Galaxy, click the eye icon to view the file, and you can use the tools menu to further analyze the data.

#### *Quality Assessment of FASTQ data*

Several packages are useful to assess the quality of the sequence data including FastQC and ShortRead. There are many types of sequencing errors:

- Errors typically increase as a function of read length. For example, for Illumina technology, with each increasing cycle it becomes increasingly difficult to differentiate the signal to noise ratio for which nucleotide has been incorporated.
- Errors occur as a function of GC content.
- Errors may occur in homopolymer positions. Both pyrosequencing and Ion Torrent technologies are susceptible to this.

These six filetype accessions apply to SRA data. The first letter S refers to NCBI-SRA. For EMBL-SRA data the first letter is E instead of S, for example ERR015959 refers to an ENA run accession with sequence data. For data originating from the DNA Database of Japan DDBJ-SRA, the first letter is D.

The ENA homepage at the European Bioinformatics Institute is <http://www.ebi.ac.uk/ena/> (WebLink 9.19). The Galaxy site you are directed to is <https://usegalaxy.org/> (WebLink 9.20).

FastQC software provides quality control statistics. Data are imported from FASTQ files (or from SAM/BAM; see “Topic 5: The SAM/BAM Format and SAMtools”), and are then analyzed in a stand-alone interactive mode (for small numbers of FASTQ files) or a non-interactive mode for larger pipelines. The output includes the following:

- Basic statistics. This includes information such as the range of sequence lengths and the percent GC content.
- Per base sequence quality. This shows quality scores (y axis) versus base position (x axis).
- Per sequence quality scores shows the number of sequences (y axis) versus mean sequence quality (Phred score; x axis).
- Overrepresented sequences. You can copy these sequences and save them to a text file. You can then try BLAT (particularly if you know the species of the overrepresented sequences), BLASTN (to search for matches across species), batch BLAST, or other tools.
- k-mer content: this shows data for a series of 5-mers (strings of 5 nucleotides) plotted by relative enrichment (y axis) versus position in read (in base pairs; x axis). The expected 5-mer frequency (determined from the base composition of the entire sequence) can be compared to the observed frequency. k-mer counts may be reduced (e.g., when poor quality reads reduce the counts for duplicated sequences) or enriched (e.g., when 5-mers are overrepresented at particular locations along a read, such as in the vicinity of a tag added to the 5' end of sequencing reads).

The FastQC website is <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (WebLink 9.21).

Many dozens of commonly used software tools for trimming are listed at <http://omictools.com> (WebLink 9.22; Henry *et al.*, 2014).

The output of a FASTQC analysis from a 1000 Genomes exome housed at the ENA is saved as Web Document 9.3.

The FASTG Format Specification Working Group issued a document that is available as Web Document 9.4. For a FASTG website see <http://fastg.sourceforge.net/> (WebLink 9.23).

In this FASTG example there are begin and end lines; two scaffolds (chr1 and chr2); a gap of between 4 and 6 bases which is assigned a default value of 5; an ambiguous base that is assigned a C but could be a G (see [1:alt:allele|C,G]), perhaps because there are two alleles present in equal amounts; a stretch of between 8 and 12 AT dinucleotides, given a primary representation of 10 repeats (see [20:tandem:size=(10,8..12)|AT]); and an expression [1:alt|A,T,TT] indicating the occurrence of either A, T, or TT at that location. Note that any FASTG string can be converted to FASTA, although with a loss of information about potential ambiguities.

You can run FastQC on a Linux server. Type the command:

```
$ fastqc mysample1_R1.fastq
```

Within a few seconds, the analysis is complete. You can use a variety of tools to trim (or mask) the sequence reads. It may be necessary to trim if there is evidence that particular base pairs of your reads have low base quality scores, or if there are contaminating primer or adapter sequences that might adversely affect downstream variant calling.

FastQC is also available via Galaxy. On the left sidebar select Tools > NGS: QC > FastQC and execute the FASTQ file. A variety of plots are presented in HTML.

### **FASTG: A Richer Format than FASTQ**

The FASTQ format offers a linear representation of genomic sequence. A working group has proposed an alternative FASTG format (in which G stands for graph). Unlike FASTQ, the FASTG format can represent allelic polymorphism as well as limitations in the assembly in which multiple sequence and/or assembly versions are possible.

An example of a FASTG file format is as follows:

```
#FASTG:begin;
#FASTG:version=1.0:assembly_name="tiny example";
>chr1:chr1;
ACGANNNN[5:gap:size=(5,4..6)]CAGGC[1:alt:allele|C,G]TATACG
>chr2;
4
ACATACGCATATATATATATATATAT[20:tandem:size=(10,8..12)|AT]TCAGG
CA[1:alt|A,T,TT]GGAC
#FASTG:end;
```

## **Topic 3: Genome Assembly**

Genome assemblies offer a consensus representation of a genome, spanning all the chromosomes (and extrachromosomal elements such as organellar genomes and plasmids). When next-generation sequencing is performed on a previously assembled

genome (e.g., when we sequence a person's genome) alignment to the reference genome is performed, but that human reference has already been assembled so further assembly is not required. In contrast, when we sequence the genome of a species that has not previously been characterized, *de novo* ("from new") assembly is required.

Whole-genome assembly involves fragmenting genomic DNA from an organism, then constructing libraries of various sizes (often from 2 kb to 50 kb or even >100 kb). In one approach the ends of cloned inserts are sequenced (producing mate pair reads). As reads are aligned they are organized into contigs such as those found in the Whole-Genome Shotgun (WGS) division of NCBI. Contigs can be ordered and oriented to assemble scaffolds (also called supercontigs). These may contain gaps whose sizes can be estimated. Global statistics for assemblies include: (1) the total number of scaffolds (including those with or without known placement or orientation); (2) the scaffold N50 (the length in base pairs such that scaffolds of this length or longer include 50% of the bases in the assembly); (3) the total number of contigs; and (4) the contig N50 (here the length such that contigs of this length or longer include 50% of the bases in the assembly. N50 is therefore a measure of contiguity, with larger values denoting more complete assemblies.

When we examine genomes across the tree of life in Part III of this book, we will see N50 statistics which give a sense of how completely a genome has been assembled. The Genome Reference Consortium (GRC) which is responsible for human genome assemblies lists the N50 for each human chromosome. For chromosome 11 (harboring the *HBB* gene cluster) the N50 is about 41.5 megabases, while in earlier assemblies (such as NCBI35) it was millions of base pairs shorter.

Many software tools are available for assembly as reviewed by Flicek and Birney (2009), Miller *et al.* (2010), Li *et al.* (2012), Paszkiewicz and Studholme (2010), Henson *et al.* (2012), and Nagarajan and Pop (2013). We provide a detailed example (assembling the *E. coli* genome) in Chapter 15 using the Velvet assembler. Assembly tools vary in speed, ability to process different types of sequence data, scalability, and results. Several are listed in **Table 9.2**. Performing assembly with relatively short reads generated with next-generation sequencing technologies offers particular challenges (Alkan *et al.*, 2011b).

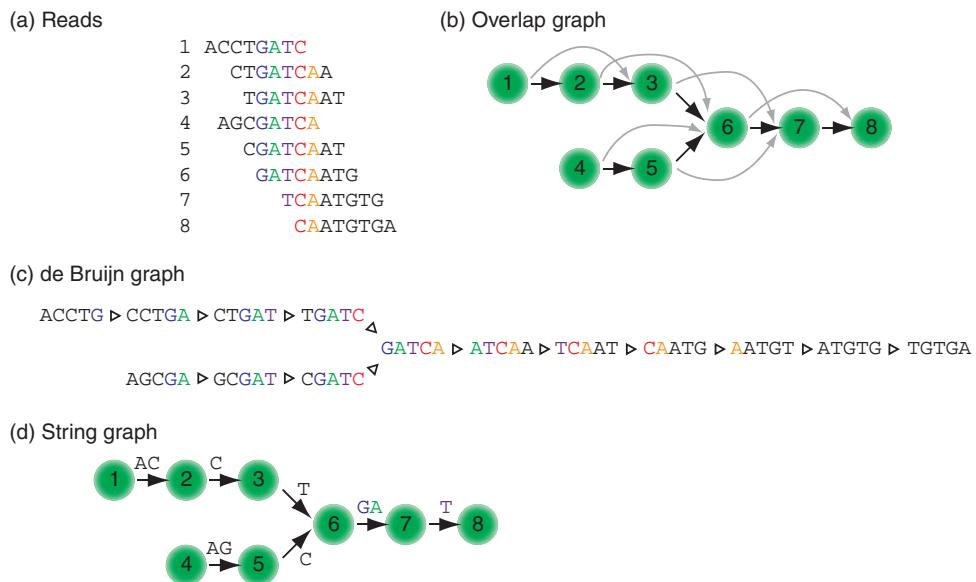
Two main methods are used by assemblers: the overlap/layout/consensus approach and de Bruijn graphs. Consider the eight reads shown in **Figure 9.9a** (from Henson *et al.*, 2012). An overlap graph represents every read as a node (**Fig. 9.9b**). Edges correspond to overlaps (here  $k = 5$  so overlaps are of 5 or more bases). The edges are transitive: larger overlaps can encompass a set of shorter overlaps (see curved arrows). Assemblers using the overlap/layout/consensus approach perform pairwise alignments of all the reads to determine the overlap.

In a de Bruijn graph (**Fig. 9.9c**), sequences are all broken into strings of some fixed length  $k$ . Each node corresponds to a  $k$ -mer such as  $k = 5$  in the figure. An edge separates

The Genome Reference Consortium (GRC) homepage is <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/> (WebLink 9.24). We discuss GRC in Chapter 20 (Fig. 20.5, Table 20.4). NCBI offers assembly resources at <http://www.ncbi.nlm.nih.gov/assembly/> (WebLink 9.25), including an overview of the topic and a glossary of terms. View the GRC N50 statistics at <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data/?build=37> (WebLink 9.26).

**TABLE 9.2. Software for genome assembly.**

Assembler	Reference	URL
ABYSS	Simpson <i>et al.</i> (2009)	<a href="http://www.bcgsc.ca/platform/bioinfo/software">http://www.bcgsc.ca/platform/bioinfo/software</a>
ALLPATHS-LG	Gnerre <i>et al.</i> (2011)	<a href="http://www.broadinstitute.org/software/allpaths-lg/blog/">http://www.broadinstitute.org/software/allpaths-lg/blog/</a>
Bambus2	Koren <i>et al.</i> (2011)	<a href="http://www.cbcb.umd.edu/software">http://www.cbcb.umd.edu/software</a>
CABOG	Miller <i>et al.</i> (2008)	<a href="http://www.jcvi.org/cms/research/projects/cabog/overview/">http://www.jcvi.org/cms/research/projects/cabog/overview/</a>
SGA	Simpson and Durbin (2012)	<a href="https://github.com/jts/sga">https://github.com/jts/sga</a>
SOAPdenovo	Luo <i>et al.</i> (2012)	<a href="http://soap.genomics.org.cn/soapdenovo.html">http://soap.genomics.org.cn/soapdenovo.html</a>
Velvet	Zerbino and Birney (2008)	<a href="http://www.ebi.ac.uk/~zerbino/velvet/">http://www.ebi.ac.uk/~zerbino/velvet/</a>



**FIGURE 9.9** Methods for genome assembly from short reads. (a) Example of 8 aligned reads (note that reads 4 and 5 only partially match reads 1–3). Colored nucleotides are identical for all aligned sequences. (b) Overlap graph represents a solution to the assembly. (c) de Bruijn graph breaks the reads into units of five nucleotide ( $k$ -mers with  $k = 5$  in this example). Colors of nucleotides match (a). Adapted from Henson *et al.* (2012) with permission from Future Medicine.

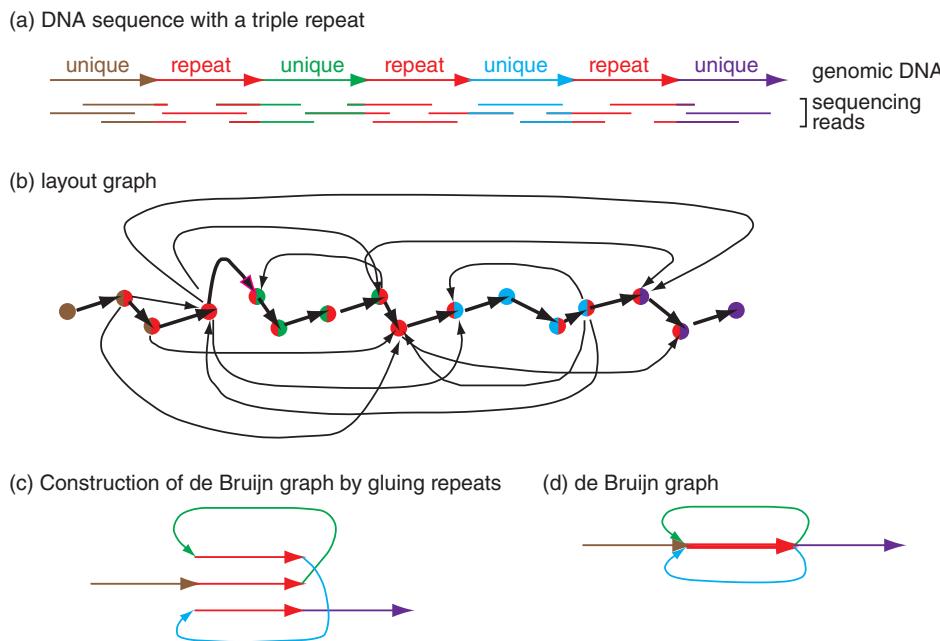
a pair of  $k$ -mers that occur consecutively. Adjacent nodes share  $k-1$  letters (e.g., in our example the first two nodes of length 5 share the bases CCTG). The genome assembly corresponds to some path through the nodes, and assemblers refine the path. For example, the path represented by sequences 4 and 5 might be rejected if it is supported by significantly less read depth than occurs for sequences 1 to 3.

The de Bruijn graph approach, championed by Pavel Pevzner (2001) and colleagues, is especially useful in the assembly of DNA with repetitive regions, as commonly occurs in eukaryotes in particular. Given a region with four unique segments and a repeat that occurs three times (Fig. 9.10a), the overlap/layout/consensus approach introduces a node for every read (Fig. 9.10b). The repetitive DNA introduces many ways in which two nodes can be connected. Pevzner *et al.* suggest viewing the DNA as a “thread” with repeat regions covered in “glue” that binds them (Fig. 9.10c). The corresponding de Bruijn graph (Fig. 9.10d) therefore represents each repeat as an edge rather than as a collection of nodes, and leads to efficient solutions for identifying the optimal paths.

Genome assembly is facilitated by having longer sequencing reads. Figure 9.11 shows three de Bruijn graphs for the assembly of the 4.2 megabase *E. coli* genome. This is a circular, bacterial genome. At  $k=50$  the graph is severely tangled; at  $k=1000$  it is greatly simplified; and at  $k=5000$  the graph is completely resolved into a single contig encompassing the entire genome. Such a large  $k$  value is made possible through long-read sequencing with the Pacific Biosciences technology. When used in combination with the Illumina platform (that offers a lower error rate) it has been possible to achieve an extraordinarily high accuracy of bacterial genome assembly (Koren *et al.*, 2013).

#### Competitions and Critical Evaluations of the Performance of Genome Assemblers

The Assemblathon competition was introduced to compare the performance of assemblers. In Assemblathon 2, 21 teams submitted 43 assemblies for three nonmammalian vertebrate genomes (the bird *Melopsittacus undulatus*, the fish *Maylandia zebra*, and the snake



**FIGURE 9.10** Efficient assembly of repetitive DNA regions using a de Bruijn graph. (a) A genomic DNA segment is shown having four unique segments and three repeats. (b) The layout graph represents these repeats with a complex set of possible paths. (c) The de Bruijn graph is constructed by “gluing” repeats. (d) The de Bruijn graph represents repeat regions as edges rather than as a set of vertices in the layout graph.

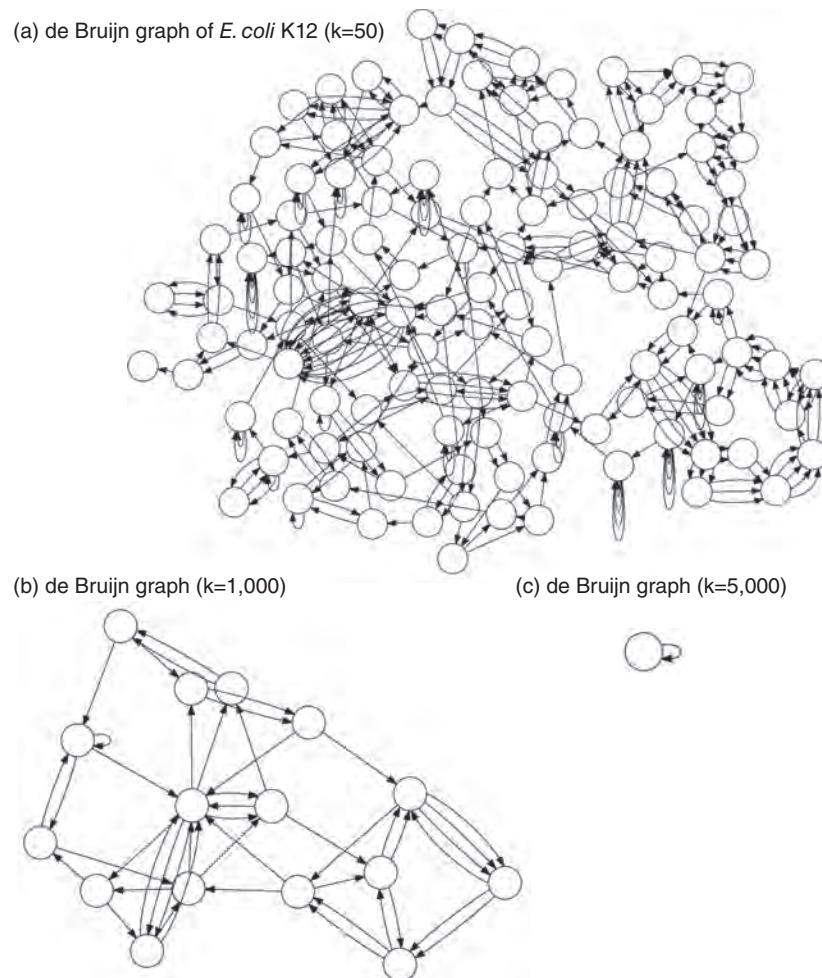
Source: Pevzner *et al.* (2001). Reproduced with permission from National Academy of Sciences.

*Boa constrictor constrictor*; Bradnam *et al.*, 2013). Since the genomes had not been previously sequenced this study served to compare assembly software methods. The main conclusion was that the various assemblers produced vastly different results, lacking consistency within and between species assemblies. The authors proposed 10 metrics for the performance of assemblers, including coverage (i.e., determining what portion of a reference genome was assembled), multiplicity (whether repeats were collapsed), measuring how many of a core set of 458 eukaryotic genes were mapped, and relating the assembly to optical map data as an approach to defining accuracy.

In an effort to critically evaluate assembly methods, Salzberg *et al.* (2012) conducted a Genome Assembly Gold-standard Evaluation (GAGE). They chose eight leading software tools and applied them to four short-read datasets (all involving the Illumina platform): two previously finished bacterial genomes, the bumble bee *Bombus impatiens* (for which the true assembly was not previously reported), and human chromosome 14. Their main metrics were the contig and scaffold N50 sizes. Their results and conclusions included the following:

- Many published genome sequences, including the human (Lander *et al.*, 2001), mouse (Mouse Genome Sequencing Consortium *et al.*, 2002), and panda (Li *et al.*, 2010) genomes, and even the Assemblathon project, do not include an assembly workflow and are therefore not reproducible. GAGE (Salzberg *et al.*, 2012) includes detailed instructions for using each of the eight assemblers they tested.
- Error correction of the sequence data is a critical step in assembly. Examples of errors are  $k$ -mers occurring just once or twice in a dataset (these are likely base-calling errors) and untrimmed adapter sequences. Following data cleaning, the N50 contig size increased 30-fold in one assembly.

The Assemblathon website is  
<http://assemblathon.org/>  
 (WebLink 9.27). Bradnam  
*et al.* (2013) lists 91 authors. We  
 describe the 458 genes that are  
 part of Core Eukaryotic Genes  
 Mapping Approach (CEGMA) in  
 Chapter 15.



**FIGURE 9.11** Improvements in assembly with increasing sequence length. *Escherichia coli* K12 MG1655 (having a circular genome of 4.64 megabases) was assembled. Each node is a contig with edges indicating relationships that cannot be resolved unambiguously due to the occurrence of repeats. de Bruijn graphs are shown for (a)  $k = 50$  (hundreds of contigs are evident in a complex pattern); (b)  $k = 1,000$  (the graph is greatly simplified); and (c)  $k = 5,000$  (the graph is fully resolved).

Source: Koren et al. (2013). Licensed under the Creative Commons Attribution License 2.0.

- If two contigs are erroneously joined, the N50 contig size will appear to be greatly improved when in fact the assembly is worse. It is therefore essential to identify and correct such errors.
- Both the degree of contiguity and the correctness of the assembly using the eight software packages varied widely (and the correctness was not well correlated with the contiguity).

The NHGRI human genome sequence quality standards are available online at <http://www.genome.gov/10000923> (WebLink 9.28). We survey all the finished human chromosomes in Chapter 20 on the human genome.

#### *The End of Assembly: Standards for Completion*

How can we decide when a genome has been successfully assembled? The National Human Genome Research Institute (NHGRI) has established standards for the human genome. “Finished sequence” refers to a DNA region of 99.99% or greater accuracy ( $\geq Q_{40}$ ), ideally with no gaps. Finished sequence applies particularly to bacterial artificial chromosomes (BACs). “Finished chromosomes” should have sequence contiguity across euchromatic regions, spanning  $\geq 95\%$  of the chromosome. (It is acknowledged that sequencing across heterochromatic regions is technically more difficult.) Any gaps must be characterized by size and orientation, and annotated.

## Topic 4: Sequence Alignment

Our goal is to align sequences to a reference genome. In the case of whole-genome sequencing of a human sample, this entails aligning FASTQ-formatted sequence reads to a human genome reference that is given in the FASTA format. You can obtain this reference from sources such as Ensembl, NCBI, and UCSC. In the case of whole-exome sequencing (WES), this may involve alignment to the entire genome reference or to a FASTA file corresponding to exons. For our example of a targeted autism sequencing panel, our reference consists of genomic DNA corresponding to the exons from just 101 genes. We can look at this FASTA file in Linux; the `#` symbol marks the start of a comment (given in green). The suffix `.fa` refers to a file in the FASTA format.

```
$ head targeted101genes.fa # displays the beginning of the file
$ tail targeted101genes.fa # displays the end of the file
$ grep ">" targeted101genes.fa | less
$ grep ">" targeted101genes.fa | wc
```

The `grep` command-line utility grabs the rows having a character (or characters) of interest (in this case, the `>` symbol which appears at the start of each gene entry), and by piping this (with the `|` symbol) to `less` we can view the results. By piping the result to `wc` we invoke the word count program which indicates how many lines (i.e., rows) have the `>` symbol (this should be 101 for this example).

There are many popular aligners, including BWA (Li and Durbin, 2009, 2010), Bowtie2, SOAP, MAQ, and Novoalign (Table 9.3), which vary in speed and accuracy. Two main alignment methods involve hash tables and Burrows–Wheeler compression (Chapter 5). A challenge is that sequencing reads are often short (<100–400 base pairs, depending on the technology) and they may align to multiple genomic locations. There are often repetitive regions, including segmental duplications that often encompass genes of interest, and each aligner must adopt a strategy to assign genomic positions. Each technology has some error rate that complicates unambiguous alignment.

We will use Bowtie2, a command-line program (Langmead and Salzberg, 2012). We must first build an indexed database, specifying the file having sequences to be indexed, and the name of the output file. For a small set of sequences this takes several seconds, while for an entire genome the indexing may require hours. Typically this requires a high-performance computing environment (e.g., a Linux machine with at least 8 GB of memory and many terabytes of storage). First we obtain a reference genome in the FASTA format

You can view the list of 101 human genes in Web Document 9.5, and the reference sequences are available in Web Document 9.6.

If you want to download a human genome reference, visit the UCSC bioinformatics site

> Downloads > Genome Data  
> Human > chromosomes  
(<http://hgdownload.cse.ucsc.edu/downloads.html>,

WebLink 9.29). These files are also available concatenated into one file (chromosomes 1-22, X, Y, mitochondrial NC\_012920) at [ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human\\_g1k\\_v37.fasta.gz](ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz) (WebLink 9.30). That compressed file is ~850 MB.

`grep` stands for g/re/p (globally search a regular expression and print).

**TABLE 9.3** Alignment software programs that natively generate SAM files.

Program	Description	URL
BFAST	Blat-like Fast Accurate Search Tool for Illumina and SOLiD reads.	<a href="https://secure.genome.ucla.edu/index.php/BFAST">https://secure.genome.ucla.edu/index.php/BFAST</a>
Bowtie	Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in SAMtools-C.	<a href="http://bowtie-bio.sourceforge.net/bowtie2/index.shtml">http://bowtie-bio.sourceforge.net/bowtie2/index.shtml</a>
BWA	Burrows–Wheeler Aligner for short and long reads.	<a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a>
LASTZ	Aligner for both short and long reads.	<a href="http://www.bx.psu.edu/miller_lab/">http://www.bx.psu.edu/miller_lab/</a>
Novoalign	An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.	<a href="http://novocraft.com/">http://novocraft.com/</a>
SNP-o-matic		<a href="http://snpolomatic.sourceforge.net/">http://snpolomatic.sourceforge.net/</a>
SSAHA2	Classical aligner for both short and long reads.	<a href="http://www.sanger.ac.uk/Software/analysis/SSAHA2/">http://www.sanger.ac.uk/Software/analysis/SSAHA2/</a>

Source: SAMtools.

You can access Bowtie2 from  
 ☎ <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>  
 (WebLink 9.31).

(ours is a small file called `targeted101genes.fa`) and create an indexed database (here called `targeted101genes.fa.fai`).

```
$ bowtie2-build targeted101genes.fa targeted101genes.fa.fai
```

The index facilitates subsequent processing of the reference, which is especially important for whole-genome alignments. As an alternative, we can download a reference human genome from NCBI (using `wget`) and then build an index to the entire human genome (chromosomes 1–22 and X):

```
$ wget ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/
human_g1k_v37.fasta.gz
$ bowtie2-build human_g1k_v37.fasta human_g1k_v37indexed
```

Note that you can obtain copies of the human and mouse genomes that are pre-indexed for use in Bowtie2 at NCBI. Pre-indexed genomes from dozens of species are available at igenomes (☞ [http://support.illumina.com/sequencing/sequencing\\_software/igenome.html](http://support.illumina.com/sequencing/sequencing_software/igenome.html), WebLink 9.32). You should always check how recently such resources have been updated.

Depending on the processor, this step can take an hour.

Next we are ready to align the FASTQ files to the indexed database.

```
$ bowtie2 -x indexed_autism101 -1 mysample1_R1.fastq -2 mysample1_R2.fastq
-S sample1.sam
```

Here the command `-x indexed database` provides the prefix of the indexed files. `-1 sample1/B1_S1_L001_R1_001.fastq` refers to one set of paired end reads, while `-2 sample1/B1_S1_L001_R2_001.fastq` refers to the other matching set of paired end reads. `sample2.sam` specifies the name for the output file. This output consists of a SAM file (our next topic) as well as information regarding the percent of reads that aligned to the reference. To align these same FASTQ files to a human genome reference, we invoke:

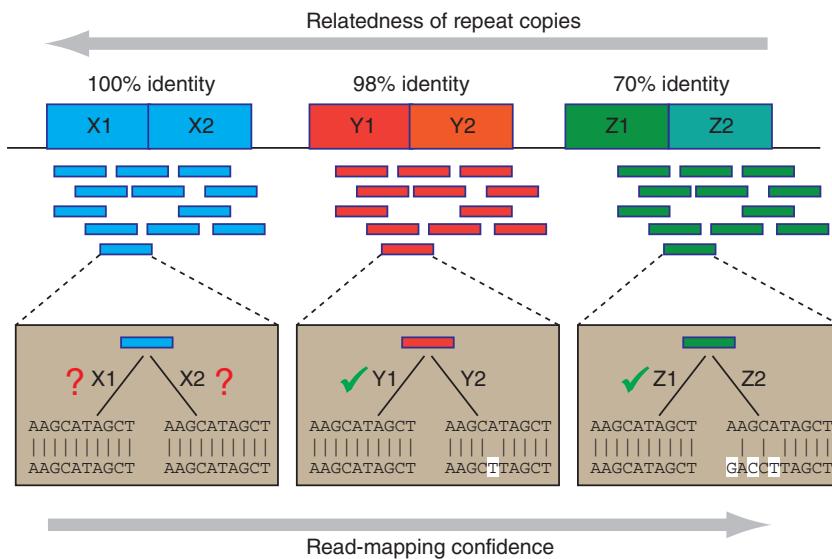
```
$ bowtie2 -x human_g1k_v37indexed -1 mysample1_R1.fastq -2 mysample1_R2.fastq
-S mysample1g1k.sam
```

Here the output SAM file is ~1.4 GB (with 3.6 million lines), whether the alignment is to the indexed library from the small set of exons or the entire human genome. By using a complete set of human exomes, you can assess whether the reads map to paralogs of the 101 genes or other repetitive elements in the genome. In this particular example the alignment is returned with statistics including a 99.04% overall alignment rate: about 1.8 million reads, all of which were paired, and of which 1.6 million (88%) aligned concordantly exactly one time. As we proceed with our analyses we will see why some reads do not pair concordantly.

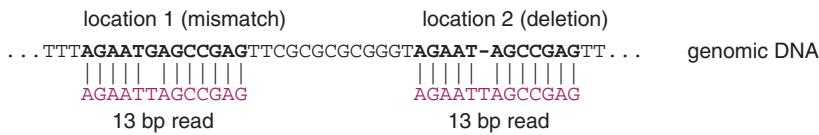
### *Alignment of Repetitive DNA*

We have seen that repetitive DNA is a challenge for assembly. Consider that half the human genome consists of repetitive DNA and other genomes have even more; transposable elements span over 80% of the maize genome. Beyond assembly, this also leads to a tremendous technical challenge for alignment to a reference genome: how should reads that match repetitive elements be aligned? Todd Treangen and Steven Salzberg (2011) discuss how repeats introduce ambiguous assemblies and alignments, sometimes producing biases and errors. They show how two nearly identical repeats cannot be mapped with confidence, while tandem repeats (or other repetitive elements) can be mapped with more confidence as their similarity decreases (**Fig. 9.12a**). In another scenario, a given short read may map to one genomic locus with a mismatch, while it maps equally well to another locus harboring a deletion (**Fig. 9.12b**). The choice of how a mismatch versus a deletion is weighted may determine where the read aligns, potentially leading to a misalignment.

(a) Read mapping confidence increases as relatedness of repeats decreases



(b) Ambiguity mapping a mismatch versus a deletion



**FIGURE 9.12** Ambiguities in mapping repetitive reads. (a) As the relatedness of two copies of a DNA repeat decreases, the confidence in the reads increases. Three tandem repeats are shown: one pair sharing 100% nucleotide identity (X1, X2 in blue), a pair sharing 98% identity (Y1, Y2 in red), and a pair with 70% identity (Z1, Z2 in green). Left: zooming in on a single read (dashed lines leading to box), the read maps equally well to X1 and X2 and thus the mapping confidence is low. Center: an occasional mismatch helps increase confidence that the read aligns to Y1 rather than Y2. Right: multiple mismatches in Z2 indicate that the correct placement of the read is at repeat copy Z1. (b) A 13 base pair read maps to two locations. The position on the left has a single mismatch, while that on the right aligns to a position having a deletion. Various alignment and assembly algorithms require decisions as to how to weigh mismatches versus indels, potentially leading to errors. Adapted from Treangen and Salzberg (2011) with permission from Macmillan Publishers Ltd.

### Genome Analysis Toolkit (GATK) Workflow: Alignment with BWA

One of the most widely used workflows for next-generation sequence analysis involves the Genome Analysis Toolkit (GATK; McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013). GATK uses BWA for alignment instead of Bowtie2 or other leading aligners. The BWA workflow is similar, requiring FASTQ files and a reference genome in the FASTA format. The reference is indexed, and a sequence dictionary is also created using the Picard package. Before performing the alignment, GATK further requires read group information which consists of meta-data about your experiment: the name of each DNA sample; the platform; the library from which the DNA was sequenced; and the particular lane of a flow cell that was used. Read group information is provided at the SAM specification. Inclusion of such meta-data is crucial to the ability of GATK to call variants with high sensitivity and specificity as we describe below. Whenever you work with a SAM/BAM file, you should preserve the header information.

BWA is available from <http://bio-bwa.sourceforge.net/> (WebLink 9.33). Picard is a program written in Java, available from <http://broadinstitute.github.io/picard/> (WebLink 9.34), used to manipulate SAM files. SAM specification is at <http://samtools.sourceforge.net/SAMv1.pdf> (WebLink 9.35).

## Topic 5: The SAM/BAM Format and SAMtools

In our workflow we have aligned FASTQ reads to a reference sequence, using Bowtie2 or BWA, to create a SAM file. The SAM (sequence alignment/map) format is commonly used to store next-generation sequence alignments. SAM files can be easily converted to the BAM (binary alignment/map) format. BAM is a binary representation of SAM, compressed by the BGZF library, and contains the same information as the SAM file. Because they are easily interchangeable, we may refer to the SAM/BAM format. This format is very popular, with many datasets available in repositories (such as the Sequence Read Archive at NCBI, the 1000 Genomes Project, and the Cancer Genome Atlas).

The SAM format includes a header section (having lines beginning with the character @) and an alignment section. The file is tab-delimited, and there are 11 mandatory fields (**Table 9.4**) which we examine in our autism panel (**Fig. 9.13**). We use the `samtools view` command to display the first of over a million rows (each row corresponding to a read that has been aligned to a reference genome). Twelve fields are shown, including the sequence (beginning AATCT...) followed by the corresponding quality scores. The CIGAR string refers to a notation system for variants. Here string “148M2S” shows 148 matches and 2 soft-clipped (unaligned) bases. Standard CIGAR operations are M (match), I (insertion), and D (deletion). Extended CIGAR options are N (skipped bases on reference), S (soft clipping), H (hard clipping), and P (padding).

SAMtools is a library and a software package (Li *et al.*, 2009). We can use it to analyze alignments in the SAM/BAM input, accomplishing the following tasks:

- convert from other alignment formats, or between SAM and BAM formats;
- sort and merge alignments;
- index alignments (once sorted, BAM file can be indexed generating a BAI file used in downstream analyses);
- view alignments in the pileup format (as shown below with the `samtools view` command);
- remove PCR duplicates (this procedure, called duplicate marking or “dedupping,” removes reads that are redundant); and
- call two classes of variants: single-nucleotide polymorphisms (SNPs) and small indels.

**TABLE 9.4. SAM format mandatory fields. There may be additional optional fields.**

Number	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based left-most POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('-' if same as RNAME)
8	MPOS	1-based left-most Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Source: Li *et al.* (2009). Reproduced with permission from Oxford University Press.

Visit the SAMtools website  
 <http://samtools.sourceforge.net/> (WebLink 9.36) for more information about all these fields including the CIGAR format.  
 SAMtools was developed by Heng Li and colleagues.

- (1) The query name of the read is given (`M01121...`)
- (2) The flag value is 163 (this equals 1+2+32+128)
- (3) The reference sequence name, `chrM`, refers to the mitochondrial genome
- (4) Position 480 is the left-most coordinate position of this read
- (5) The Phred-scaled mapping quality is 60 (an error rate of 1 in  $10^6$ )
- (6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:00000000-A2DTN:1:2111:20172:15571      163      chrM
480      60      148M2S =      524      195      AATCTCATCAAT
ACAACCCCTGCCCATCCTACCCAGCACACACACACCCTGCTAACCCCATACCCGAAC
AACCAAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG BBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FEEEEEHBGEFFFGGGG@</0
1BG21222>F21>F11FGFG1@1?GC<G1?1?FGDGGF=GHFFFH-.-
RG:Z:Sample7  XC:i:148 XT:A:U NM:i:3 SM:i:37
AM:i:37 X0:i:1 X1:i:0 XM:i:3 X0:i:0 XG:i:0 MD:Z:19C109C0A17
```

- (7) An = sign shows that the mate reference matches the reference name
- (8) The 1-based left position is 524
- (9) The insert size is 195 bases
- (10) The sequence begins `AATCT` and ends `ACGGG` (its length is 150 bases)
- (11) Each base is assigned a quality score (from BBBBB ending `FHC.-`)
- (12) This read has additional, optional fields that accompany the MiSeq analysis

**FIGURE 9.13** Anatomy of a SAM file. The `samtools view` command was used in Linux to view a BAM file, and the `| less` command sends the output to a single screen of data at a time. (A typical SAM/BAM file has millions of rows.) A single record of the output is shown, with 12 features as indicated.

Source: SAMtools.

Next let's convert a SAM file to a BAM file.

```
$ samtools view -bS sample1_bowtie2.sam > sample1_bowtie2.bam
$ samtools sort sample1_bowtie2.bam sample1_bowtie2_sorted
$ samtools faidx targeted101genes.fa
$ samtools index sample1_bowtie2_sorted.bam
```

SAMtools view takes an input SAM file, and here the `>` symbol in Linux specifies that the output should be sent to a file called `sample1_bowtie2.bam`. We then sort and index the BAM file. To view it, invoke `samtools tview` as follows.

```
$ samtools tview mysample1.bam
```

In that program, type `?` to get a help menu; type `g` to go to a chromosomal location. We select `chrX:153,296,000` to go to an exon within *MECP2* (an X-linked gene that when mutated causes Rett syndrome; see Chapter 21). We can view the aligned reads from the BAM file, showing base quality scores (Fig. 9.14a) or mapping quality scores (Fig. 9.14b). Later when we call a variant this is a convenient way to quickly see the read depth, base quality scores, mapping quality scores, and other features at that locus.

In the GATK workflow, before variants are called the BAM file undergoes duplicate marking with Picard rather than SAMtools. GATK also performs local realignment around indels. This is an important step because indels are often flanked by mismatches that are mapping artifacts but might appear to be authentic single-nucleotide variants.

(a) SAMTools `tview` visualization of reads from a BAM file (base quality view)

(b) SAMTools `tview` (mapping quality view)

**FIGURE 9.14** Using SAMtools to view sequence reads from a BAM file at genomic coordinates of interest. By using the `samtools tview` command you enable a genome-wide view of the reads. Using commands accessed via a help menu, you can view the same reads colored by (a) base quality or (b) mapping quality. The read depth is relatively low to the left side. The quality scores are colored blue (for 0–9), green (10–19), yellow (20–29), or white ( $\geq 30$ ). Underlining represents secondary or orphan reads. This viewer is useful to quickly assess the quality at genomic loci of interest, such as positions having single-nucleotide variants.

*Source:* SAMtools.

Once FASTQ files from your sample have been aligned to a reference genome, it may be assumed that variants (both single-nucleotide variants and indels) can be called by inspecting the alignment and tabulating the differences. The problem with this approach is that many sources of errors occur in the process of sequencing and performing alignment: bias may occur in how libraries are prepared and amplified; sequencing technologies all have associated error rates; and mapping has error rates (as shown in Fig. 9.12). We mentioned that GATK performs local alignment around indels. GATK further performs base quality score recalibration: even the quality scores associated with each base call in a FASTQ file have different types of error. GATK applies an empirically derived error model and adjusts the base quality scores. You can compare base quality scores at a given genomic position before and after this adjustment and see changes assigned to particular bases. DePristo *et al.* (2011) provide a dramatic example of the effects of the GATK pipeline. They sequenced sample NA12878 (a well-characterized DNA from a participant in the 1000 Genomes Project) and performed alignment with BWA. They found that 15% of the reads spanning homozygous deletions were misaligned. Realignment by GATK corrected many of these reads (6.6 million reads in 950,000 regions spanning 21 Mb).

As we evaluate software it is critical to have a dataset that represents a “gold standard” from which we know that the information consists of true positive results. The Genome in a Bottle Consortium was launched to develop standards for DNA sequencing. It provides datasets such as FASTQ files (from ~300 $\times$  sequence coverage) and high-quality variant calls from NA12878 and several mother/father/child trios.

### **Calculating Read Depth**

Read depth (or depth of coverage) is a basic design consideration. If a library is sequenced more often (e.g., analyzing it on multiple lanes of a flow cell) this will offer greater depth and more statistical power to detect variants. At the same time, obtaining deeper coverage is relatively expensive. For a typical whole-genome sequence using Illumina technology that generates 150 nucleotide paired end reads, coverage of 30 $\times$  to 50 $\times$  is obtained, meaning that on average any given base in the genome is covered by 30–50 independent sequencing reads. For whole-exome sequencing (which often spans ~50 megabases or 1–2% of the genome), depth of coverage is often 100 $\times$  or more. For targeted sequencing, such as the autism panel, depth of sequencing ranges from 30 $\times$  to 300 $\times$  depending on factors such as the number of samples that are run simultaneously. When we performed targeted sequencing of a disease-associated mutation that occurs at a low mutant allele frequency (1–18%), we used 13,000-fold median depth of coverage (Shirley *et al.*, 2013).

Lander and Waterman (1988) considered the assembly of reads into contigs (contiguous sequences). The redundancy of coverage  $c$  is a function of the number of reads  $N$ , the average length of each read  $L$ , and the length of the region (e.g., genome  $G$ ) being sequenced (Lander and Waterman, 1988; reviewed by Li *et al.*, 2012):

$$c = \frac{LN}{G}. \quad (9.4)$$

30 $\times$  coverage of a genome implies that there is an average of 30 reads covering any single base in the genome. Of course, there is variability in the distribution of read coverage across the genome, and some bases will be covered with a far higher or lower number of reads. Higher coverage enables improved statistical power in calling heterozygotes and other variants.

The number of contigs that need to be sequenced to achieve a particular read depth depends on the parameters read length  $L$ , sequencing depth  $c$ , genome size  $G$ , and also the minimum length of an overlap between reads  $T$ . The probability a base is not sequenced was derived by Lander and Waterman (1988) and is given by

$$P_0 = e^{-c} \quad (9.5)$$

from which it is possible to estimate the depth of coverage needed to sequence DNA (**Table 9.5**). Next-generation sequencing technologies use relatively short reads. Li *et al.* (2012) note that 30-fold depth of coverage with 50 base pair reads gives an assembly equivalent to 10-fold depth of coverage for 500 base pair reads.

We can use SAMtools to calculate read depth from our sorted BAM file. The output of the `samtools depth` command is a line-by-line readout of the depth at each position. Instead of viewing that large output, we can pipe the result (with `|`) to the `awk` program and specify that we want to calculate the average read depth.

```
$ samtools depth sample1_bowtie2_sorted.bam | awk '{sum+=$3} END
{ print "Average = ",sum/NR}'
Average = 105.838
```

### **Finding and Viewing BAM/SAM files**

There are two prominent places to obtain BAM/SAM files. First, SRA Toolkit at NCBI provides access to BAM files, and the SRA Toolkit software development kit (SDK)

You can visit the Genome in a Bottle website at <http://genomeinabottle.org> (WebLink 9.37). The project was initiated by NIST. NA12878 refers to DNA of a woman from Utah of northern European descent as part of the Centre de l'Étude du Polymorphisme Humain (CEPH) project. This DNA is available from the Coriell Institute for Medical Research in Camden, NJ (<https://catalog.coriell.org>). GM12878 refers to the lymphoblastoid cell line (LCL) from which the DNA was purified; thousands of DNA samples and LCLs are available from Coriell.

**TABLE 9.5.** Probability that a base is sequenced, according to Equation (9.5).

Fold coverage	$P_0$	Percent not sequenced	Percent sequenced
0.25	$e^{-0.25} = 0.78$	78	22
0.5	$e^{-0.5} = 0.61$	61	39
0.75	$e^{-0.75} = 0.47$	47	53
1	$e^{-1} = 0.37$	37	63
2	$e^{-2} = 0.135$	13.5	87.5
3	$e^{-3} = 0.05$	5	95
4	$e^{-4} = 0.018$	1.8	98.2
5	$e^{-5} = 0.0067$	0.6	99.4
6	$e^{-6} = 0.0025$	0.25	99.75
7	$e^{-7} = 0.0009$	0.09	99.91
8	$e^{-8} = 0.0003$	0.03	99.97
9	$e^{-9} = 0.0001$	0.01	99.99
10	$e^{-10} = 0.000045$	0.005	99.995

Source: Lander and Waterman (1988). Reproduced with permission from Elsevier.

The SRA website is <http://www.ncbi.nlm.nih.gov/sra/> (WebLink 9.38). 1000 Genomes alignments as BAM files are available at <http://www.1000genomes.org/data> (WebLink 9.39).

To access autism SAM and BAM files, see Web Document 9.7 at <http://bioinfbook.org> (WebLink 9.40).

IGV software is available (upon registration) from <https://www.broadinstitute.org/igv/> (WebLink 9.41).

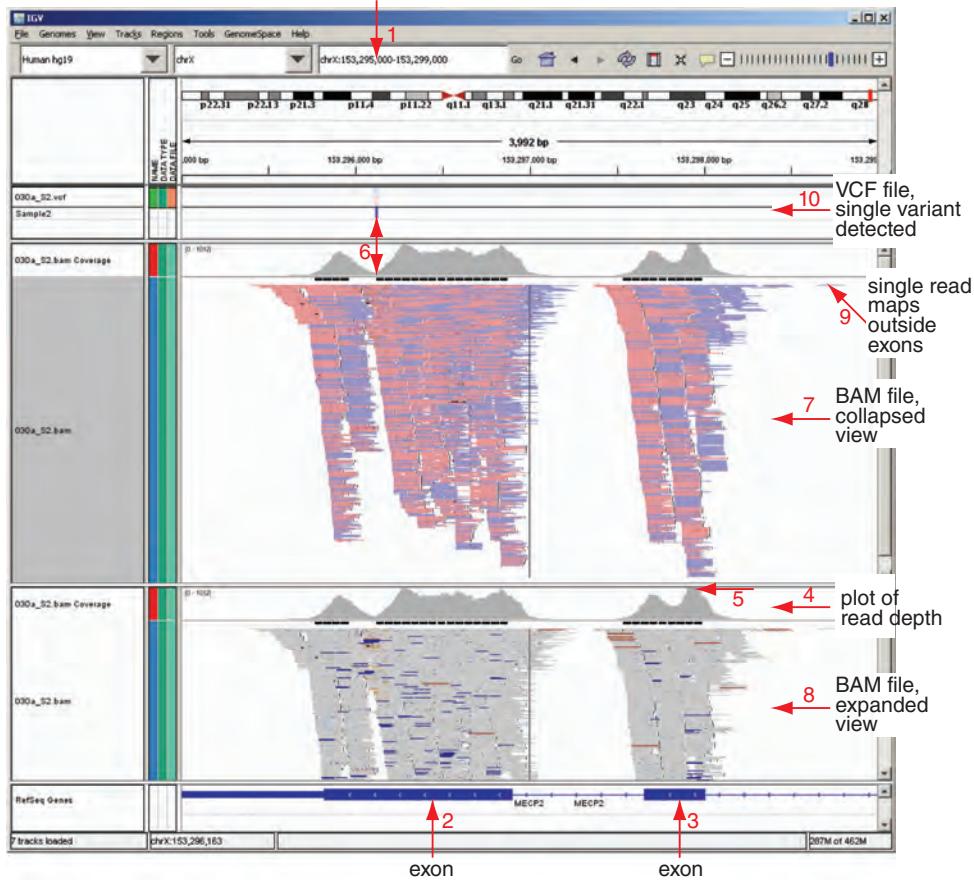
offers programmatic access to BAM files. Second, the 1000 Genomes Project stores BAM files corresponding to whole-genome and whole-exome sequences with a current goal of providing data for >2000 individuals.

I have placed a small BAM file (based on sequencing the exons corresponding to 101 autism-related genes) on the textbook website, as well as the corresponding SAM file that can be viewed in a text editor.

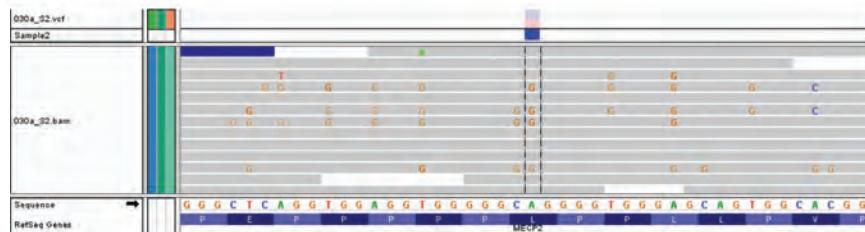
Next we can view BAM file data with Integrative Genomics Viewer (IGV) software (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2012). Once you install IGV, upload a BAM file and view a genomic region of interest. For targeted autism example we can again search for *MECP2* (Fig. 9.15a). We then refine our query to chrX:153,295,000–153,299,000 to view a 4000 base pair window. The *MECP2* gene structure is shown at the bottom, with exons displayed as thick blue rectangles; two exons are visible. We show the BAM file alignments at two levels of resolution. For each, a gray wiggle plot shows peaks of coverage centered over the exons, reaching a maximum of about 1000-fold depth of coverage. (This is excellent coverage for alignment and variant calling.) Figure 9.15a shows the reads (shaded according to forward or reverse strand). Figure 9.15b shows the single-nucleotide variant at base pair resolution. IGV software is highly flexible, using “data tiling” in which data are pre-computed at multiple resolution scales. This facilitates zooming from whole-genome to single base pair views. IGV allows multiple BAM files (or other file types such as VCF; see below) to be viewed simultaneously. It can be customized. For example, a text file with the gene symbols for all 101 autism-related genes can be uploaded and used to display data from just those loci of interest.

#### Compressed Alignments: CRAM File Format

It is important to compress raw sequence files because of their enormous size. CRAM files, developed at the European Nucleotide Archive (EMBL-EBI), represent a format of compressed BAM-like files offering better lossless compression than BAM and full compatibility with BAM (Hsi-Yang Fritz *et al.*, 2011). The JAVA-based `cramtools`

(a) IGV display of a BAM file (at two resolutions) and a VCF in the *MECP2* gene region

(b) IGV display of the variation at base pair resolution



**FIGURE 9.15** The Integrative Genomics Viewer (IGV). (a) Once a data file is loaded, a genomic locus can be queried or a gene symbol can be entered (arrow 1). Here two exons of *MECP2* on Xq28 are shown (arrows 2, 3). A BAM file has been uploaded twice. The coverage is shown (arrow 4), including a peak at which the read depth is ~1000 (exact values are obtained by mousing over the position). Some areas have very low depth of coverage (e.g., arrow 6). The BAM file was uploaded twice to display the collapsed view (arrow 7) showing all the reads at once, as well as an expanded view (arrow 8) which requires scrolling to see all the reads. IGV facilitates exploration of reads such as the single read that maps outside an exon (arrow 9). A variant call format (VCF) file is also uploaded (arrow 10), indicating a single variant called in this region (arrow 6). That variant is rejected because of an artifact (strand bias) and appears to occur at a position of very low read depth. (b) It is possible to zoom in to base pair resolution, in this case to assess the called variant in more detail. Courtesy of Integrative Genomics Viewer (IGV).

package (available from github or from the ENA website at EMBL-EBI) interconverts BAM and CRAM files. CRAM files are also read using Picard (see “Genome Analysis Toolkit (GATK) Workflow” above).

## Topic 6: Variant Calling: Single-Nucleotide Variants and Indels

We have preprocessed our BAM file(s) and are now ready to call variants (reviewed in Nielsen *et al.*, 2011). These consist of single-nucleotide variants (SNVs; also referred to as single-nucleotide polymorphisms or SNPs) and indels. Note that indels are left-shifted by convention by various aligners and in software such as SAMtools and GATK, for example for a called two base pair deletion:

```
GGATATATCC (reference)
GG--ATATCC (read with two base deletion)
```

The indel position is therefore shifted to the left as far as possible. However, an indel can be represented at different positions, even while representing the same haplotype. An alternative solution is to right-shift an indel, in this case with the same nucleotides (AT) deleted, as follows. Note that this choice can have important functional consequences, affecting the accuracy of alignment and the nature of the variant call format files (introduced below) that represents variation.

The GATK LeftAlignIndels tool left-aligns indels within a BAM file.

```
GGATATATCC (reference)
GGATAT--CC (read with same two base deletion)
```

The SAMtools package can call variants as follows:

```
$ samtools mpileup -S -f targeted101genes.fa -g
sample1_bowtie2_sorted.bam > sample1_bowtie2.bcf
$ samtools mpileup -S -f targeted101genes.fa -g
sample2_bowtie2_sorted.bam > sample2_bowtie2.bcf
```

We therefore create a .bcf file that summarizes single-nucleotide variants and indels in our sample. The .bcf is in binary format, and can be processed and then converted to nonbinary variant call format (VCF).

```
$ bcftools view -bvcg sample1_bowtie2.bcf >
sample1_bowtie2raw.bcf
$ bcftools view sample1_bowtie2raw.bcf > sample1variants.vcf
```

These variants can then be annotated to evaluate their biological significance.

GATK also calls variants using its HaplotypeCaller (Van der Auwera *et al.*, 2013), also resulting in a VCF file. The HaplotypeCaller calls SNPs and indels simultaneously, and does so by discarding existing mapping information at regions of variation and performing local *de novo* assembly of haplotypes. Thresholds are used to distinguish high- and low-confidence variant calls.

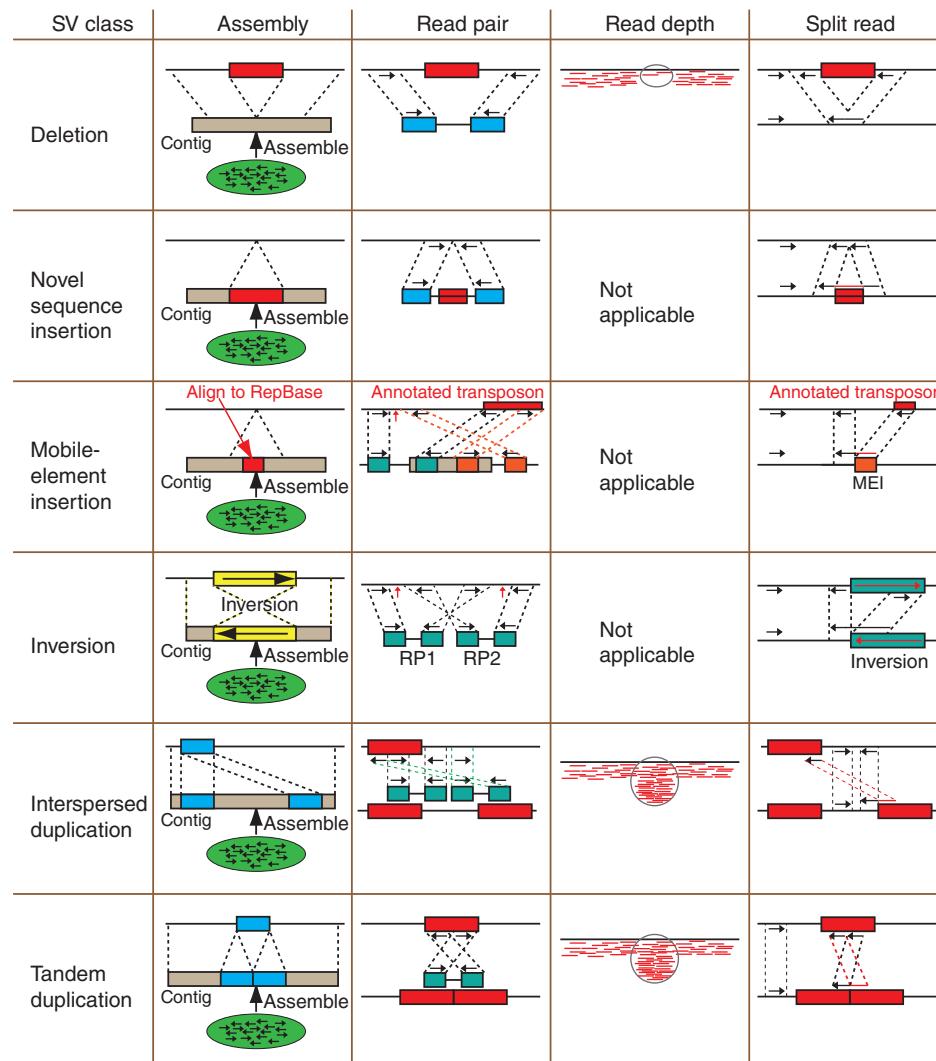
Several groups have assessed variant callers including Liu *et al.* (2013) and O’Rawe *et al.* (2013). Nielsen *et al.* (2011) conclude that base calling and quality score calculation should use well-benchmarked methods (such as GATK or SOAPsnp). The aligner is critical; they recommend sensitive tools such as Novoalign and Stampy. SNP calling should use data from all individuals in a sample simultaneously, and methods incorporating information about linkage disequilibrium (the relationships of neighboring variants on a single haplotype block) to improve accuracy. Ghoshal Lyon and colleagues (O’Rawe *et al.*, 2013) sequenced 15 human exomes, compared five workflows for variant calling, and observed just 57% concordance with up to 5% of variants called as unique to each pipeline. This underscores the complexity of variant analysis from genome and exome sequencing experiments.

## Topic 7: Variant Calling: Structural Variants

In Chapter 8 we introduced types of structural variation. Next-generation sequence data can be analyzed to identify assorted changes (reviewed in Medvedev *et al.*, 2009; Alkan *et al.*, 2011a; Koboldt *et al.*, 2012). **Figure 9.16** shows four approaches to structural variation detection at the levels of assembly, paired read analysis, read depth (i.e., depth of coverage), and split reads (in which only one read of a mate pair maps to a reference genome). With these approaches it is possible to assess six classes of structural variation (Alkan *et al.*, 2011a; although additional complex variants may occur; see Medvedev *et al.*, 2009):

- Deletions.** Paired end reads are useful to identify deletions (as well as insertions) because such reads have an expected distance (depending on the size of the library inserts) and orientation. A discordant pair can indicate a deletion if the read pairs align closer than expected. While the paired end read approach is powerful, it is

Alkan *et al.* (2011a) define an indel as up to 50 base pairs and a copy number variant as >50 base pairs. In previous years insertions, deletions, and inversions were usually defined as greater than 1 kilobase, but the higher resolution of next-generation sequencing inspires this revised definition.



**FIGURE 9.16** Four approaches to identifying structural variants (columns) used to identify six types of structural variation (rows). See text for details. In each panel the upper line corresponds to the reference genome sequence, and the lower line is the contig or scaffold (for assembly) or the aligned reads. Red arrows indicate breakpoints. MEI: mobile-element insertion.

Source: Alkan *et al.* (2011). Reproduced with permission from Macmillan Publishers Ltd.

challenging to apply it to regions of repetitive DNA. Read depth is also a useful approach for identifying deletions because the number of reads mapping to a genomic locus is expected to follow a Poisson distribution and will be proportional to the copy number. Some analysis software accounts for differences in read depth across the genome (such as reduced depth in areas of very high or very low GC content). A limitation of read depth is that its breakpoint resolution is not as exact as for read pair and split read approaches. Split end reads occur when the two reads of a pair align to different genomic loci; Pindel is an example of software that identifies such structural anomalies (Ye *et al.*, 2009).

2. *Novel sequence insertions.* Split read analysis is particularly useful for finding novel sequences (when one pair of a read aligns). A limitation of the read pair strategy is that insert sizes of a library follow a distribution and are not all exactly the same length.
3. *Mobile element insertion.* Read pair approaches can be used to detect mobile element insertions, particularly if the read lengths are sufficiently long (e.g., greater than the size of a typical *Alu* element, 300–400 base pairs). Alternatively, split reads are useful to characterize such insertions.
4. *Inversions.* Inversions can be identified when paired reads unexpectedly map to the same strand. The inversion breakpoints may have complex changes, not usually detectable by read depth analysis (although many inversions involve small, complex rearrangements at the breakpoints that are detectable by copy number change).
5. *Interspersed duplications.* An increased read depth at a genomic region indicates that an insertion has occurred. It indicates absolute copy number, but provides no information about where that extra material maps. Read depth approaches can therefore identify interspersed duplications but cannot distinguish between them and tandem duplications.
6. *Tandem duplications.* Both paired end reads and split reads can identify tandem duplications, in some cases resolving breakpoints to single base pair resolution. As sequencing technology continues to evolve and offer longer read lengths, such variation will become easier to identify.

The Pindel homepage is <http://gmt.genome.wustl.edu/packages/pindel/> (WebLink 9.42).

BreakDancer is available at <http://gmt.genome.wustl.edu/breakdancer/current/> (WebLink 9.43). We mentioned Tandem Repeats Finder in Chapter 8.

Documentation and downloads are available at <http://vcftools.sourceforge.net/> (WebLink 9.44).

A variety of software tools have been developed to detect structural variation (some are reviewed in Koboldt *et al.*, 2012). These include BreakDancer (Chen *et al.*, 2009) which includes a read pair approach and identifies indels, inversions, and translocations from 10 base pairs to 1 megabase or more.

## Topic 8: Summarizing Variation: The VCF Format and VCFtools

In our workflow we have now obtained reads (in the FASTQ format), aligned them to a reference genome (with files in the SAM/BAM format), and called variants. The variant call format (VCF) is a file format for storing DNA variation data such as single-nucleotide variants (SNVs; also called single-nucleotide polymorphisms or SNPs), insertions/deletions (indels), structural variants, and annotations. The VCF format and VCFtools have been described by Danecek *et al.* (2011).

We can look at a VCF file from our autism targeted sequence panel (**Fig. 9.17**). The VCF file includes a header (marked on each row with two hash characters, ##) then a field definition line (starting with a single # character) that begins the data section. To begin, we look at that field definition line and a single row from a data section (type `less mydata.vcf`). A VCF file may include data from multiple samples (e.g., individuals), but in our case it corresponds to a single sample. Each line (row) of the data section corresponds to a variant at one genomic position (or region). There are eight mandatory tab-delimited fields, listed in the field definition line and given in **Table 9.6**. The VCF format allows variants such as SNPs, insertions, deletions, replacements, and large structural variants to be represented. Examples (adapted from Danecek *et al.*, 2011) are given in **Figure 9.17**.

**TABLE 9.6 Columns of a VCF file.**

Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GXQ in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the called variant depth to the total depth (VF), and minimum of {genotype quality assuming variant position, genotype quality assuming nonvariant position} (GXQ).
Sample	No	Sample identifiers define the samples included in the VCF file

VCFtools is a command-line tool (for Unix-based systems). For basic operations, you can include `--vcf <filename>` or `--gzvcf <filename>` to specify whether to analyze an uncompressed or gzipped file. Some of the commands in VCFtools require that you operate on a compressed VCF file. Given file `mydata.vcf`, we can compress with `gzip` (and uncompress with `gunzip`):

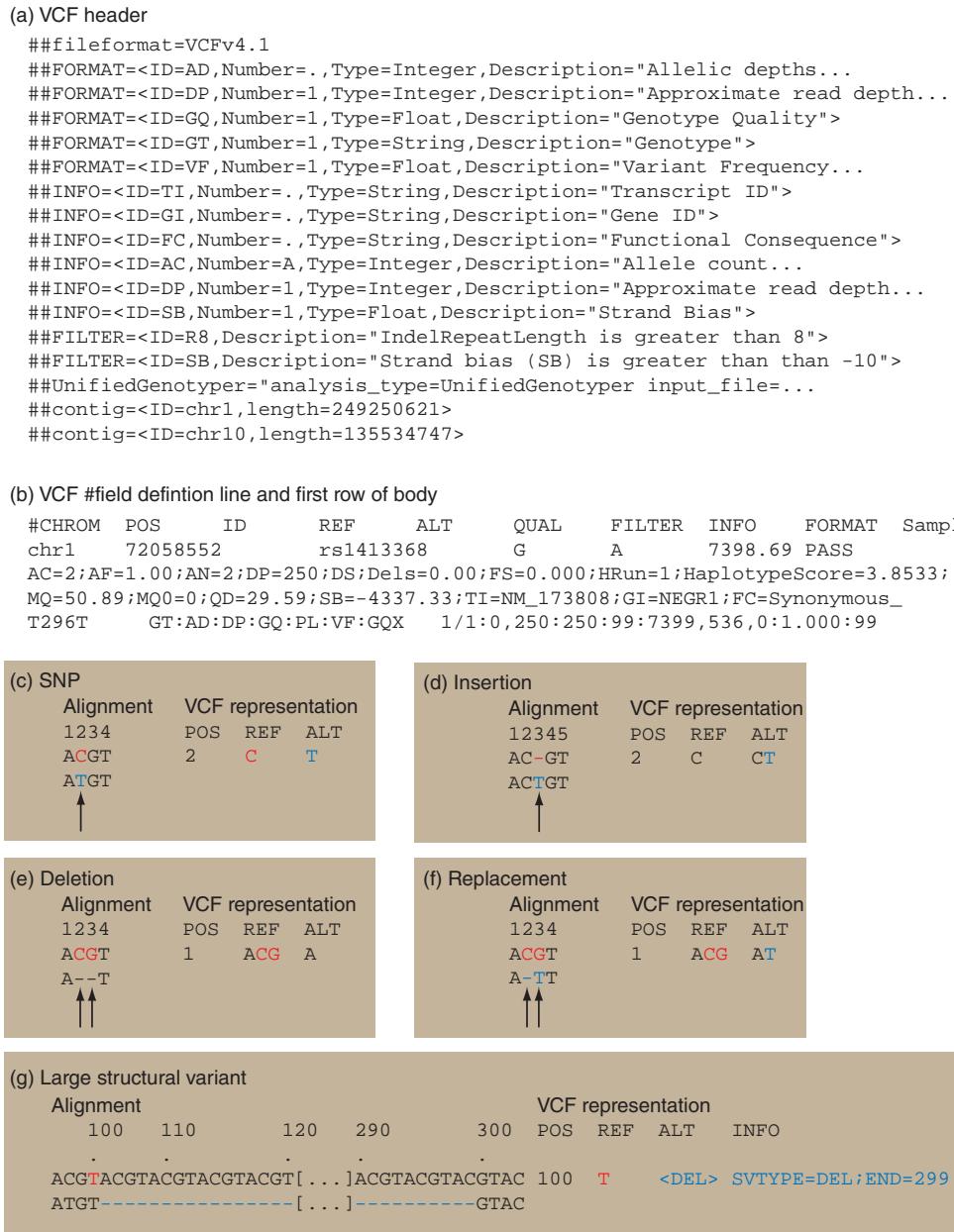
```
$ gzip test.vcf # this creates test.vcf.gz
```

By using the `vcf-stats` command we can summarize some statistics such as the number of each type of SNP (A>C, A>G, A>T, etc.), the number of indels, and the number of heterozygous and homozygous variants. By invoking

```
$ vcftools --gzvcf mydata.vcf.gz --depth
```

we can obtain the mean depth of coverage for each individual in the VCF. Commands can show the read depth at each variant position, genotype data, and transition/transversion statistics. Additional VCFtools commands allow you to merge, query, reorder, annotate, and compare VCF files, as described in the online VCFtools manual. What variants occur on chromosome 11 in the beta globin region? We can see that there are seven sites (these can be listed separately):

```
$ vcftools --vcf ~/data/sample1.vcf --chr 11 --from-bp 5200000 --to-bp 5300000
VCFtools - v0.1.12
(C) Adam Auton and Anthony Marcketta 2009
Parameters as interpreted:
--vcf /Users/pevsner/data/sample1.vcf
--chr 11
--to-bp 5300000
--from-bp 5200000
After filtering, kept 1 out of 1 Individuals
After filtering, kept 7 out of a possible 79824 Sites
Run Time = 0.00 seconds
```



**FIGURE 9.17** Description of a variant call format (VCF) file. Such files contain rows that define the position and nature of variants. In addition to mandatory fields, they may include rich functional annotation. (a) Header section (a portion of the rows are shown). (b) Field definition line and example of a row from the body of the file. Examples of particular variants represented in the VCF including: (c) a single-nucleotide polymorphism; (d) an insertion; (e) a deletion; (f) a replacement; and (g) a large structural variant. Adapted from Danecek *et al.* (2011) with permission from Oxford University Press and P. Danecek.

What are the differences in variants between the VCF files of two individuals? We can specify two VCF files of interest and send the output to a file called `diffs`.

```
$ vcftools --vcf ~/data/sample1.vcf --diff ~/data/sample2.vcf --out diffs
```

Given a BAM file, the choice of variant calling strategy can produce very different VCF files. The GATK philosophy includes calling variants with HaplotypeCaller with

very lenient thresholds in order to attain very high sensitivity (i.e., not missing variants at a cost of potentially calling false positives). GATK next uses a variant quality score recalibration step to assign a probability to each variant call (involving a training model in which the log-odds ratio of a variant being a true positive relative to a false positive is assessed). It then filters the raw call set, aiming to achieve good specificity and sensitivity and yielding high-quality variant calls.

For the training set, variants that occur in the HapMap and 1000 Genomes projects are considered likely to be true (see Chapter 20 for a description of these projects). Variants in dbSNP, many of which have not been validated, are not included in the training set. Further annotations used in refining the variant calls include the depth of coverage, the variant quality as a function of depth of coverage, the presence of strand bias (which typically suggests false positive results), and the distance from the end of the read (where false positives occur more often at read ends). Recalibration is performed for both SNPs and indels, producing new VCF files. The GATK website and Van der Auwera *et al.* (2013) provide additional details on these methods.

#### **Finding and Viewing VCF files**

We can find VCF files at the ENA website. For example, a text search for “1000 genomes” includes analysis results (e.g., analysis accession ERZ015345) with links to VCF files that can be downloaded or sent to Galaxy. NCBI offers VCF files at its FTP site, as does the 1000 Genomes Project. Both the 1000 Genomes Project and Ensembl offer a Data Slicer tool allowing you to output a VCF from a particular HapMap individual or population (Chapter 20); this can also be restricted to a chromosomal location. The VCF is downloadable for further study.

Earlier we viewed a BAM file in IGV (Fig. 9.15). That figure also includes a corresponding VCF. You can also view VCF data in the UCSC Genome Browser as described above for BAM files.

### **Topic 9: Visualizing and Tabulating Next-Generation Sequence Data**

We have shown that SAMtools and IGV are both excellent tools to visualize genomic data. Many other resources are available. For example, Jim Kent and colleagues (2010) introduced the BigWig and BigBed formats to enable visualization and analysis of large datasets on the UCSC Genome Browser. BigWig and BigBed are compressed binary indexed files (as are BAM files), and they are viewable at multiple resolutions (as with IGV).

You can also view BAM and VCF files at the UCSC site by posting the data on an http, https, or ftp location and then pointing to it. I have placed indexed BAM and VCF files on the textbook’s website. To view them, visit the UCSC Genome Browser and create a custom track. Enter the text:

```
track type=bam name="My BAM"
bigDataUrl=http://bioinfbook.org/chapter9/WebDoc9-1/mysample1.bam
```

You can then view the data on the genome browser.

BEDtools is described as a Swiss army knife of tools to enable “genome arithmetic.” It allows you to compare, intersect, and summarize genomic features in a variety of common formats (BED, BAM, GTF, GFF, VCF) (Quinlan and Hall, 2010).

BEDtools takes a BED, BAM, or other file and allows you to determine information or perform tasks such as the following:

- Use `bedtools intersect` to find the base pair overlap between a set of sequence alignments and a feature you are interested in such as genes, repeats, microRNAs, etc.

As an example of a large set of human VCF files at NCBI, visit [ftp://ftp.ncbi.nih.gov/snp/organisms/human\\_9606/VCF/](ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/) (WebLink 9.45). For a set of human VCFs, see [http://www.ncbi.nlm.nih.gov/variation/docs/human\\_variation\\_vcf/](http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/) (WebLink 9.46). That page lists VCFs for dbSNP, ClinVar (a clinical variation resource described in Chapter 21), common variation, and other categories. For 1000 Genomes Project VCFs visit <http://www.1000genomes.org/data> (WebLink 9.47).

Instructions for creating BigBed files are available at <http://genome.ucsc.edu/goldenPath/help/bigBed.html> (WebLink 9.48). BigWig files are described at <http://genome.ucsc.edu/goldenPath/help/bigWig.html> (WebLink 9.49).

Many file formats are defined at the UCSC website <http://genome.ucsc.edu/FAQ/FAQformat.html> (WebLink 9.50). We described the BED format in Chapter 2.

- Use `bedtools bamtobed` to convert BAM alignments to the BED format or vice versa.
- Use `bedtools window` to find all genes that are within some distance upstream or downstream of genes, CNVs, or other features of interest.
- Use `bedtools closest` to find the closest *Alu* sequence (or any other feature of interest) to each gene in your BED file.
- Use `bedtools subtract` to remove features such as a BED file specifying introns.
- Use `bedtools coverage` to calculate the read depth in windows of various sizes that span the genome; this tool can also create a BEDGraph for viewing at UCSC.
- Use `bedtools shuffle` to randomly place all discovered variants in the genome (with options to avoid placing them in locations such as gaps or repeats).
- Use `bedtools slop` to mask all regions in the genome except those of interest, such as the exons corresponding to the 101 genes related to autism.

The BEDtools homepage is <https://github.com/arq5x/bedtools2> (WebLink 9.51) and documentation is at <http://bedtools.readthedocs.org/en/latest/> (WebLink 9.52). It was created by Aaron Quinlan.

To learn about BEDtools, complete the following steps: (1) download and install; (2) obtain BED files to study; (3) intersect; (4) closest features; (5) merge; (6) calculate genome coverage; and (7) window. We'll use BED files for these examples, but these approaches are relevant to analyses with BAM, GTF/GFF, VCF and other files.

### 1. Download and install BEDtools.

```
$ mkdir bedtools # Working on a Mac laptop, let's start by making a
# directory called bedtools
$ mv ~/Downloads/bedtools2-2.19.1/ ~/bedtools/ # we'll move the
# downloaded directory from Downloads
$ cd bedtools/ # navigate into the directory called bedtools
$ ls # Look inside our directory; it has the bedtools directory we just
# downloaded and copied
bedtools2-2.19.1
$ cd bedtools2-2.19.1/
$ ls # Here are the files
LICENSE README.md
bin docs genomes scripts test
Makefile RELEASE_HISTORY data genome obj src
$ make # this command compiles the software
```

We can use `sudo` to obtain administrator privileges and copy the binaries from `bin/` to `/usr/local/bin` directory. This will allow us to invoke BEDtools commands without needing to specify a path to the binaries directory.

```
$ sudo cp bin/* /usr/local/bin/
```

### 2. BEDtools can operate on several file types including BED, GFF, BAM, and VCF. For these examples we will use only BED files obtained from the UCSC Table Browser. For each, download a BED file. Then copy those BED files to your `bedtools/data` directory.

```
$ pwd # "Print working directory" shows current location
/Users/pevnsner/bedtools/bedtools2-2.19.1/data
$ cp ~/Downloads/chr11*.bed . # We copy into the current directory
$ ls # We list files in the current directory
chr11_hg19_UCSC_codingexons.bed
chr11_hg19_RefSeqCodingExons.bed
chr11_hg19_hg38diff.bed
chr11_hg19_RepeatMasker.bed
chr11_hg19_SegmentalDups.bed
```

3. Next, we use the BEDtools intersect function. The general format is:

```
$ bedtools intersect -a reads.bed -b genes.bed
```

For our example we will look for the overlap between all RefSeq coding exons on chromosome 11 and the file listing differences between a GRCh37 (a popular human genome assembly sometimes called hg19) and GRCh38 (sometimes called hg38, it is a newer assembly that was released in December 2013).

```
$ bedtools intersect -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_hg38diff.bed | head -5
chr11 369803 369954 NM_178537_cds_0_0_chr11_369804_f 0 +
chr11 372108 372212 NM_178537_cds_1_0_chr11_372109_f 0 +
chr11 372661 372754 NM_178537_cds_2_0_chr11_372662_f 0 +
chr11 372851 372947 NM_178537_cds_3_0_chr11_372852_f 0 +
chr11 373025 373116 NM_178537_cds_4_0_chr11_373026_f 0 +
$ bedtools intersect -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_hg38diff.bed | wc -l # This shows the number of exons
# having differences
9586
$ wc -l chr11_hg19_* # We can list the number of entries in various BED
# files
21352 chr11_hg19_RefSeqCodingExons.bed
239924 chr11_hg19_RepeatMasker.bed
1933 chr11_hg19_SegmentalDups.bed
31523 chr11_hg19_UCSC_codingexons.bed
366 chr11_hg19_hg38diff.bed
```

UCSC Genes encompass more gene models than the more conservative RefSeq genes. We have downloaded BED files having all coding exons from each source. We now report those entries that are in UCSC but have no overlap with RefSeq coding exons.

```
$ bedtools intersect -a chr11_hg19_UCSC_codingexons.bed -b
chr11_hg19_RefSeqCodingExons.bed -v | head
chr11 130206 131373 uc009ybr.3_cds_0_0_chr11_130207_r 0 -
chr11 131466 131469 uc009ybr.3_cds_1_0_chr11_131467_r 0 -
chr11 130206 131373 uc001lnw.3_cds_0_0_chr11_130207_r 0 -
chr11 131466 131469 uc001lnw.3_cds_1_0_chr11_131467_r 0 -
chr11 130206 131087 uc001lnx.4_cds_0_0_chr11_130207_r 0 -
$ bedtools intersect -a chr11_hg19_UCSC_codingexons.bed -b
chr11_hg19_RefSeqCodingExons.bed -v | wc -l
421
```

There are therefore 421 such UCSC coding exons along chromosome 21 that are not in RefSeq coding exons.

4. Use the `closest` program. For every RefSeq coding exon we find the closest gap on the chromosome. The entire BED file of gaps looks like:

```
chr11 0 10000
chr11 10000 60000
chr11 1162759 1212759
chr11 50783853 50833853
chr11 50833853 51040853
chr11 51040853 51090853
chr11 51594205 51644205
chr11 51644205 54644205
chr11 54644205 54694205
chr11 69089801 69139801
chr11 69724695 69774695
chr11 87688378 87738378
chr11 96287584 96437584
chr11 134946516 134996516
chr11 134996516 135006516
```

Here are the first entries showing which gap each RefSeq coding exon is closest to.

```
$ bedtools closest -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_gaps.bed
chr11 193099 193154 NM_001097610_cds_0_0_chr11_193100_f 0 +
chr11 10000 60000 # this ends the first record
chr11 193711 193911 NM_001097610_cds_1_0_chr11_193712_f 0 +
chr11 10000 60000 # end of second record
chr11 194417 194450 NM_001097610_cds_2_0_chr11_194418_f 0 +
chr11 10000 60000
chr11 193099 193154 NM_145651_cds_0_0_chr11_193100_f 0 +
chr11 10000 60000
chr11 193711 193911 NM_145651_cds_1_0_chr11_193712_f 0 +
chr11 10000 60000
chr11 194417 194450 NM_145651_cds_2_0_chr11_194418_f 0 +
chr11 10000 60000
```

- The BEDtools `merge` command is widely useful. Our RepeatMasker BED file has many overlapping entries that we can merge as follows. We return the number of entries that are merged.

```
$ bedtools merge -i chr11_hg19_RepeatMasker.bed -n | head
chr11 60904 61254 1
chr11 61314 61346 1
chr11 61405 61671 1
chr11 61674 61908 1
chr11 62074 62151 1
chr11 62157 62320 1
chr11 62346 62931 1
chr11 62966 64003 2
chr11 64053 64794 1
chr11 64828 67807 4
```

- Genome Coverage lets us ask questions such as “How much of chromosome 11 is spanned by gaps?” We use the `-g` argument to specify the human genome build we were using (several are included in the `genomes` directory of the bedtools download).

```
$ bedtools genomecov -i chr11_hg19_gaps.bed -g ../genomes/human.hg19.
genome
chr11 0 131129516 135006516 0.971283
chr11 1 3877000 135006516 0.0287171
genome 0 3133284264 3137161264 0.998764
genome 1 3877000 3137161264 0.00123583
```

The answer is 2.87% of the chromosome, and 0.1% of the genome is spanned by gaps. The Genome Coverage output includes five columns: (1) the chromosome or entire genome; (2) the depth of coverage from the features in the input file, that is, 0 or 1 in this example; (3) the number of bases on the chromosome (or across genome) with a depth equal to column 2; (4) the size of chromosome (or entire genome) in base pairs, that is, ~135 Mb for chromosome 11 or 3137 Mb for the entire genome; and (5) the fraction of bases on chromosome (or entire genome) with depth equal to column 2.

How much of chromosome 11 does *not* include RefSeq coding exons? We now use a BED file of exons. The answer is 98.5% as we see from the first line of output of this command:

```
$ bedtools genomecov -i chr11_hg19_RefSeqCodingExons.bed -g ../genomes/
human.hg19.genome
chr11 0 133031219 135006516 0.985369
```

7. With BEDtools window we can determine how many RefSeq coding exons are positioned within 40,000 base pairs of a gap on chromosome 11.

```
$ bedtools window -a chr11_hg19_RefSeqCodingExons.bed -b
chr11_hg19_gaps.bed -w 40000 | wc -l
16
```

The answer is 16. We can see the first three of them:

```
$ bedtools window -a chr11_hg19_RefSeqCodingExons.bed -b chr11_hg19_
gaps.bed -w 40000 | head -3
chr11 1244353 1244423 NM_002458_cds_0_0_chr11_1244354_f 0 +
chr11 1162759 1212759
chr11 1246910 1246967 NM_002458_cds_1_0_chr11_1246911_f 0 +
chr11 1162759 1212759
chr11 1247434 1247506 NM_002458_cds_2_0_chr11_1247435_f 0 +
chr11 1162759 1212759
```

The Genome Workbench from NCBI offers another way to visualize next-generation sequence data. We introduced it in Chapter 2. You can obtain a BAM file and its associated BAM index file then load them into Genome Workbench using the File > Open pull-down (**Fig. 9.18a**). (Alternatively, under the Project Tree View click the BAM option to load a BAM file.) This generates a coverage graph that you can view graphically. There are hundreds of track options. **Figure 9.18b** shows a region including two exons of the *HBB* gene, along with tracks for associated variants, RefSeq accessions for the gene, mRNA, and protein, a bar graph showing the depth of coverage, and aligned reads. The bottom portion includes tracks showing the regions that are captured by the Agilent, NimbleGen, Illumina, and 1000 Genomes Project workflows (note their differences). This example highlights how versatile and accessible the Genome Workbench is, and how you can work with any indexed BAM file to explore and analyze any genomic region of interest.

You can access Genome Workbench from <http://www.ncbi.nlm.nih.gov/tools/gbanch/> (WebLink 9.53). For this example we downloaded the file NA19240.chrom11.SLX.maq .SRP000032.2009\_07.bam (i.e., a BAM file containing aligned reads from chromosome 11 of HapMap individual NA19240 sequenced by Illumina technology). This BAM file is 7.6 GB, and its associated BAI (BAM index) file is ~400 kB. Both files are available at Web Document 9.8 at <http://bioinfbook.org>. They can also be downloaded from [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot\\_data/data/NA19240/alignment/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/data/NA19240/alignment/) (WebLink 9.54).

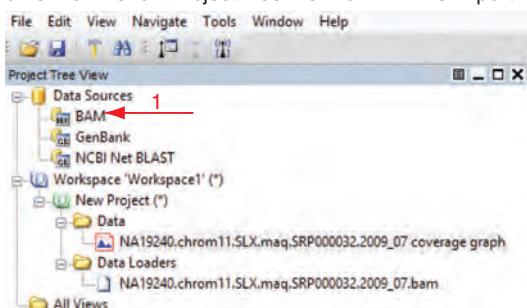
## Topic 10: Interpreting the Biological Significance of Variants

A typical human genome harbors about 3.5 million single-nucleotide variants, 600,000 indels, and a variety of other variants. Which of these are neutral (not affecting phenotype), and which are deleterious (possible disease-causing)? Several main strategies have been developed and are closely related to the topics we covered in Part I of this book: scoring matrices, pairwise and multiple sequence alignment, and sequence conservation.

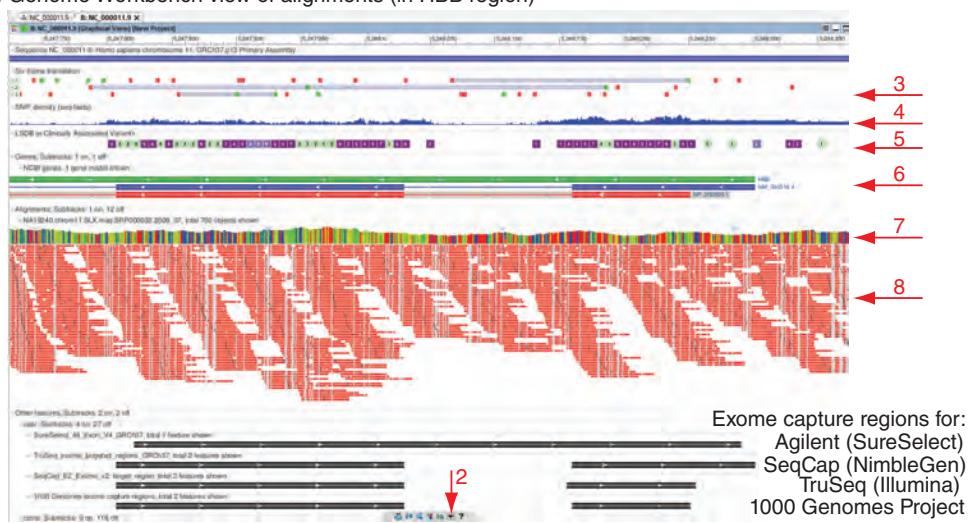
One approach to finding disease-associated mutations is to focus on nonsynonymous variants (those which alter the specified amino acids) as opposed to synonymous variants (those in coding regions that do not specify changes in the amino acids). A major premise is that synonymous variants are neutral, although it is possible that such changes are deleterious (e.g., they could affect splicing accuracy, mRNA structure, translation, and protein folding; Sauna and Kimchi-Sarfaty, 2011). Approximately 25–30% of all nonsynonymous SNPs are predicted to disrupt protein function; consequently, those SNPs tend to be eliminated by purifying selection and are rare in human populations (Ng and Henikoff, 2006). The major human disease databases are Online Mendelian Inheritance in Man (OMIM) and the proprietary Human Gene Mutation Database (HGMD) (see Chapter 21); in these databases about half the disease-associated variants are nonsynonymous. Furthermore, disease-associated amino acid substitutions occur preferentially at evolutionarily conserved amino acid positions (Miller and Kumar, 2001).

We begin by describing two prominent software tools which were developed before the era of next-generation sequencing: Sorting Tolerant from Intolerant (SIFT) and Polymorphism Phenotyping-2 (PolyPhen) (Flanagan *et al.*, 2010). Although they

(a) Genome WorkBench: Project Tree View for BAM file import



(b) Genome Workbench view of alignments (in HBB region)



**FIGURE 9.18** Genome Workbench from NCBI is useful to view and analyze BAM files. (a) BAM files can be uploaded using File > Open or by clicking the BAM link (arrow 1). (b) Large numbers of user-selected tracks are available using menus at the bottom (arrow 2) or at the upper right of each track (not shown). The BAM file includes alignments spanning human chromosome 11 in individual NA19240. This view includes a six-frame translation of the region (arrow 3), SNPs (arrow 4), clinically associated variants (arrow 5), two exons of the *HBB* gene with associated RefSeq identifiers (arrow 6), a histogram of read depth (arrow 7), packed reads (arrow 8), and annotations of the exome capture regions for several technologies. Hundreds of other annotation tracks are available.

Source: Genome Workbench, NCBI.

continue to be popular and have prominent roles in bioinformatics analyses, we see that newer tools offer dramatically better sensitivity and specificity.

SIFT, first introduced in 2001, offers a web server (Kumar *et al.*, 2009; Sim *et al.*, 2012). Given a protein query it performs a PSI-BLAST search (Chapter 5), builds a multiple sequence alignment, and calculates normalized probabilities of occurrence of each amino acid at each position. Positions with normalized probabilities below a threshold (typically 0.05) are predicted to be deleterious; values  $\geq 0.05$  are called tolerated. SIFT also calculates a conservation value ranging from 0 (all 20 amino acids are observed at that position) to  $\log_2 20$  ( $=4.32$ ) when just one amino acid is observed at a given position without substitutions.

PolyPhen (Ramensky *et al.*, 2002) has a similar approach and also incorporates structural information, using empirically derived rules to predict whether nonsynonymous variants are possibly or probably damaging (two separate prediction categories). PolyPhen-2 extends this method using eight sequence-based predictive features and three structure-based features (Adzhubei *et al.*, 2010, 2013). It reports the naïve Bayes posterior probability that a mutation is damaging and reports true and false positive rates.

ID	Chr: bp	Alleles	Source	AA	AA coord	*	SIFT	PolyPhen
rs121909815	11:5248247	A/G	dbSNP	V/A	2		0	0.119
rs121909830	11:5248247	A/C	dbSNP	V/G	2		0.01	0.007
rs33958358	11:5248248	C/T/A	dbSNP	V/L	2		0.01	0.001
rs33958358	11:5248248	C/T/A	dbSNP	V/M	2		0	0.271
rs35906307	11:5248245	G/A	dbSNP	H/Y	3		0.02	0.135
rs35906307	11:5248245	G/A	dbSNP	H/Y	3		0.02	0.135
rs63750720	11:5248241	A/T/G	dbSNP	L/Q	4		0	0.802
rs63750720	11:5248241	A/T/G	dbSNP	L/P	4		0	0.931
HbVar_2753	11:5248241	A/G	PhenCode	L/P	4		0	0.931
rs34126315	11:5248242	G/C/T	dbSNP	L/M	4		0.04	0.127
rs34126315	11:5248242	G/C/T	dbSNP	L/V	4		0.03	0.007
HbVar_2683	11:5248242	G/T	PhenCode	L/M	4		0.04	0.127
rs63750605	11:5248238	G/T	dbSNP	T/N	5		0	0.064
rs281864509	11:5248239	T/G	dbSNP_ClinVar	T/P	5		0	0.185
HbVar_2682	11:5248239	T/G	PhenCode	T/P	5		0	0.185
rs63750605	11:5248238	G/T	dbSNP	T/N	5		0	0.064
rs281864509	11:5248239	T/G	dbSNP_ClinVar	T/P	5		0	0.185
HbVar_2682	11:5248239	T/G	PhenCode	T/P	5		0	0.185
rs34769005	11:5248235	G/C/A	dbSNP	P/L	6		0.29	0.069
rs34769005	11:5248235	G/C/A	dbSNP	P/R	6		0.5	0.108
rs33912272	11:5248236	G/A/C	dbSNP	P/S	6		0.72	0.001
rs33912272	11:5248236	G/A/C	dbSNP	P/A	6		0.94	0
rs77121243	11:5248232	T/A/C	dbSNP	E/V	7		0.06	0.213
rs77121243	11:5248232	T/A/C	dbSNP	E/G	7		0.1	0.076

**FIGURE 9.19** SIFT and PolyPhen scores are provided in the Variation Table of the Ensembl genome browser. A portion of the entry for human beta globin (*HBB*) is shown. Each row represents a variant; columns indicate dbSNP (or other) identifier, chromosome and position of the variant, reference and alternate alleles, database source, reference and alternate alleles, amino acid position in the beta globin protein, and SIFT and PolyPhen predictions. Several rows have been removed for clarity, and additional columns of information may be added at the website. Note the lack of agreement with SIFT and PolyPhen predictions. A well-known mutation, known since the 1960s as E6V (glutamate at position 6 substituted by valine), is correctly listed as E7V (blue rectangle) but note that it is listed as neutral by both SIFT and PolyPhen even though it is a cause of sickle-cell anemia (see Chapter 21).

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

We can view both SIFT and PolyPhen results at the Ensembl genome browser. For the human *HBB* gene there are currently ~6800 annotated variants (viewable using the Variation Table) including over 700 missense variants (Fig. 9.19). This highlights the accessibility of variant annotation results at Ensembl, and also the great discrepancy that often occurs between SIFT and PolyPhen scores. A study introducing a likelihood ratio test found just 5% of predictions were shared by the three tools, while 76% of predictions were unique to one of the tools (Chun and Fay, 2009).

Many software tools accept genomic variants (typically from a VCF file) and provide functional annotation. ANNOVAR is a prominent package (Chang and Wang, 2012) that includes gene-based and region-based annotations. Ensembl offers a Variant Effect Predictor (VEP) that reports the location of variants (e.g., coding exons, introns) and their predicted effects. NCBI's Variation Reporter accepts VCF files. Exomiser is a Java program that uses ANNOVAR code and UCSC KnownGene transcript definitions. Its output conveniently includes information on relevant mouse models. Each of these packages is available as a web-based or command-line tool. As a specific example, I analyzed a VCF file from a whole-exome sequencing experiment using the NCBI Variation Reporter. There were ~80,000 variant alleles, including ~1100 novel alleles at known locations and ~4700 novel alleles at novel locations. A total of ~1900 variant alleles have clinical information (such as an Online Mendelian Inheritance in Man or OMIM allelic variant; see Chapter 21). The entire data file (of ~355,000 rows and 30 columns) is available as Web Document 9.9. This type of file is most easily studied in a UNIX-like operating system where tools such as `grep` can be used to extract information of interest.

There are at least 40 software packages that call variants neutral or deleterious (Tchernitchko *et al.*, 2004; Hicks *et al.*, 2011; Jaffe *et al.*, 2011; Thusberg *et al.*, 2011;

ANNOVAR, developed by Kai Wang, can be accessed from <http://www.openbioinformatics.org/annovar/> (WebLink 9.55). Ensembl's VEP is available at [http://www.ensembl.org/Homo\\_sapiens/Tools/VEP](http://www.ensembl.org/Homo_sapiens/Tools/VEP) (WebLink 9.56), the NCBI Variation Reporter is at <http://www.ncbi.nlm.nih.gov/variation/tools/reporter> (WebLink 9.57), and Exomiser from the Wellcome Trust Sanger Institute is at <http://www.sanger.ac.uk/resources/databases/exomiser> (WebLink 9.58). In addition to VCF, these tools sometimes accept BED, GVF, HGVS (Human Genome Variation Society), or other formats.

Visit the VAAST homepage from the group of Mark Yandell at  
 ↗ <http://www.yandell-lab.org/software/vaast.html> (WebLink 9.59).

Lopes *et al.*, 2012; Liu and Kumar, 2013; Shihab *et al.*, 2013). How can we decide which is best? As with any bioinformatics software, the key is to assess error rates. We can ask whether a software tool calls a variant deleterious if that variant is derived from OMIM or HGMD (presumably constituting a true positive finding). We can ask if it calls a variant neutral if it derives from the 1000 Genomes Project, dbSNP, or other sources of apparently normal individuals; false positive results occur when a software tool calls such neutral variants deleterious. (It has been noted that dbSNP contains an unknown mixture of neutral and deleterious variants, and even participants in the 1000 Genomes Project who are defined as apparently normal also harbor some number of deleterious variants.) For PolyPhen-2, at a false positive rate of 20% (e.g., 2 of every 10 variants that are called deleterious are actually neutral), the true positive prediction rate is 92% (Adzhubei *et al.*, 2010). This involves analysis of a dataset including >3100 variants from UniProt (Chapter 2) that are annotated as causing Mendelian disease. PolyPhen-2 has a 73% true positive prediction rate in another dataset that includes all ~13,000 UniProt variants annotated as disease-causing, as well as ~9000 variants not annotated as disease-causing.

We turn to VAAST, a software package that offers greatly improved sensitivity and specificity (Yandell *et al.*, 2011; Hu *et al.*, 2013). Four files are required to begin: a VCF (or the related GVF format) for both the target (case) variants and the background (control) variants; a list of genes or other features to be scored (e.g., a file called `genes.gff3`) in the GFF format; and a reference genome in the FASTA format (e.g., `mygenome.fasta`). There are three steps to using VAAST.

1. A variant annotation tool (VAT) annotates variants based on functional effects such as introducing missense mutations or splice site mutations. These annotations are introduced in a new column in the output file (e.g., `patientvariants.vat.gvf`). A typical VAT command is as follows:

```
$ VAT -f genes.gff3 -a mygenome.fasta patientvariants.gvf >
patientvariants.vat.gvf
```

2. Next a variant selection tool (VST) generates “condenser” files (.cdr extension) for the target and background sets. The .cdr files are analogous to the query of a BLAST search. VST can perform operations such as finding the union of all variants, or the intersection or complement of genomic variants in a series of .vat.gvf files. For example, we can produce an output containing the union of variant loci present in three files:

```
$ VST -ops 'U(0..2)' patientvariants.vat.gvf file2.vat.gvf
file3.vat.gvf > my_vst_output.cdr
```

Different .vat.gvf files are used for the target and background sets. A Perl script allows a quality check to confirm that the allele frequencies are not significantly different between these two. If they do differ, the analysis may be flawed because many differences will be called due to the underlying genetic differences in these two groups.

3. Variant analysis is then performed with VAAST. For example, we can run:

```
$ VAAST --mode lrt --outfile myoutput genes.gff3 background.cdr
my_vst_output.cdr
```

Here the option `--mode lrt` specifies a composite-likelihood ratio test. This scores features according to the difference in frequencies of variants in the target and background genomes. According to a null model the frequency of a variant is the same in a control population and a case population of interest (e.g., the genome

of a patient or set of patients). According to an alternative model these frequencies differ. VAAST further considers the likelihood that a substitution does not contribute to disease (by using amino acid substitution data drawn from neutral sources such as BLOSUM62), while it assesses the likelihood of deleterious amino acid substitutions by incorporating a model of disease-associated changes from OMIM.

A receiver operator curve (ROC) is shown for six methods in Web Document 9.10, indicating the better performance by VAAST. At a given false positive rate of 5%, the true positive rate for VAAST 2.0 (and VAAST 1.0; Hu *et al.*, 2013) is far better than for MutationTaster (Schwarz *et al.*, 2010), SIFT, and PolyPhen-2.

VAAST offers improved accuracy because of its use of a composite likelihood ratio test. VAAST is also able to score variants in both coding and noncoding regions and, in contrast to SIFT and PolyPhen-2 (which are restricted to regions of aligned phylogenetically conserved amino acids), it can score variants at any coding or noncoding position. (SIFT scores 60% of the protein-coding portion of the genome, PolyPhen-2 scores 81%, while VAAST scores essentially all of it.) The current version of VAAST includes phylogenetic conservation in the form of PhastCons scores (Chapter 6) to weight estimates of amino acid changes being deleterious or neutral. The output of VAAST includes a list of variants ranked by lowest probability value.

MutationTaster offers a web server at <http://www.mutationtaster.org/> (WebLink 9.60). It uses a Bayes classifier to assign a probability as to whether an alteration is a disease mutation or a neutral polymorphism. Its training set includes >390,000 disease mutations from HGMD and >6.8 million neutral SNPs and indels from the 1000 Genomes Project.

## Topic 11: Storing Data in Repositories

Next-generation sequencing experiments can generate many hundreds of gigabases of DNA sequence in a single day. This may correspond to terabytes of image data. For many sequencing centers, it has become routine to need storage on the scale of petabytes.

There are four main options for the storage of large datasets.

1. An investigator who receives data from a core facility (often in the form of an external drive) can maintain the data on a local server.
2. Raw data can be deposited in a repository, similar to the way gene expression datasets are routinely stored at ArrayExpress (at EBI) or the Gene Expression Omnibus (GEO at NCBI). The National Institutes of Health introduced the Sequence Read Archive (SRA) for this function. In many cases the data are submitted by investigators as sorted BAM files (which can be converted back to FASTQ files if needed). In some cases FASTQ files are made available. Other large-scale resources that summarize human variation are The Cancer Genome Atlas (TCGA), the 1000 Genomes Project, the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), and the Exome Aggregation Consortium (ExAC).
3. Cloud computing can be used to provide access to data. Cloud computing refers to the fee-based use of servers that are owned by a company (such as Amazon or Google).
4. Raw data can be discarded. Some have suggested that the expense of storing large amounts of data on a computer server has become more expensive than maintaining the DNA and re-running the sequencing experiment if needed. According to this model, processed data (e.g., BAM files) are maintained.

The ROC curve in Web Document 9.10 shows the results for a set of common and rare variants from HGMD and the 1000 Genomes Project; for rare variants (having minor allele frequency far less than 1%), the ROC curve shows even better performance by VAAST.

The TCGA website is <http://cancergenome.nih.gov/> (WebLink 9.61). The NHLBI ESP website is <http://evs.gs.washington.edu/EVS/> (WebLink 9.62). The ExAC browser at <http://exac.broadinstitute.org/> (WebLink 9.63) currently provides variation data from >63,000 exomes.

## SPECIALIZED APPLICATIONS OF NEXT-GENERATION SEQUENCING

Next-generation sequencing data in repositories offer a treasure trove of information. As one example, many exome and targeted sequencing experiments incidentally capture highly abundant mitochondrial DNA. MitoSeek software (Guo *et al.*, 2013) allows you to easily extract mitochondrial sequences. It also reports mitochondrial copy number,

We discuss the human mitochondrial genome in Chapter 20. MitoSeek is available from <https://github.com/riverlee/MitoSeek> (WebLink 9.64). We describe it in Chapter 21 in the context of mitochondrial disease.

heteroplasmy (presence of varying mitochondrial genomes within an individual), somatic mutations, and structural variants.

There are many other ways useful information can be extracted from archived DNA sequence data (reviewed in Samuels *et al.*, 2013). Some pathogenic eukaryotes have endosymbionts such as the obligate endobacterium *Wolbachia* (an alpha-proteobacterium that inhabits parasitic filarial nematodes as well as spiders, insects, and mites). These bacterial endosymbionts are often uncultivable and are therefore difficult to study. Salzberg *et al.* (2005) searched an NCBI database housing *Drosophila* genomic DNA and discovered three *Wolbachia* strains, assembling one genome to 95% completion.

In addition to DNA sequencing, dozens of new applications of next-generation sequencing have emerged (Shendure and Lieberman Aiden, 2012). These include:

- RNA sequencing (RNA-seq) allows measurement of steady-state RNA levels, as described in Chapters 10 and 11.
- Chromatin immunoprecipitation sequencing (ChIP-Seq) is used to measure protein–DNA interactions (Park, 2009). Protein bound to genomic DNA is cross-linked with formaldehyde. The DNA is sheared; protein–DNA complexes are immunoprecipitated with antisera directed specifically against protein targets of interest (such as DNA-binding transcription factors). The genomic DNA fragments are then isolated, sequenced, and mapped to a reference genome.
- MicroRNAs (introduced in Chapter 10) are small noncoding RNAs that are essential regulators in a variety of pathways. (Over half the human transcriptome is thought to be regulated by miRNAs.) To identify the endogenous messenger RNAs that are targeted by miRNAs, a recent approach is ultraviolet cross-linking and immunoprecipitation coupled to next-generation sequencing (CLIP-seq; Chi *et al.*, 2009).
- Many cytosine residues in the eukaryotic genome are methylated, particularly at CpG dinucleotides. Methylation sequencing (methyl-seq) has been applied to characterize such changes genome-wide (Huss, 2010; Ku *et al.*, 2011). When a sample is treated with bisulfite, cytosine residues are deaminated to uracil. By comparing the sequences of samples with and without bisulfite treatment, it is possible to infer methylation status. There are many related approaches such as treatment of genomic DNA with methylation-sensitive restriction enzymes, or the study of other epigenetic markers such as 5-hydroxymethylcytosine (Branco *et al.*, 2011).
- DNase-seq and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) allow the sequencing of nucleosome-depleted genomic regions, which is then used to map regulatory regions of chromatin (Song *et al.*, 2011).

## PERSPECTIVE

Next-generation sequencing (NGS) technology is revolutionizing biology. We are now able to catalog genetic variation at unprecedented depth. Studies of human diseases now routinely include hundreds of pedigrees, and in some cases thousands. Studies such as the 1000 Genomes Project catalog variation across geographic populations worldwide. NGS has been applied to sequencing genomes across the tree of life, enabling a deeper understanding of myriad biological principles.

In this chapter we outlined 11 topics in sequence analysis. For those who have never performed these analyses it should be straightforward to obtain files containing data in the main formats (FASTQ, BAM/SAM, VCF) and explore them. For this it is especially helpful to work in the Linux operating system, essential for detailed analyses. For those who are less familiar with Linux it is still possible to work with web-based tools (such as Galaxy, UCSC) as well as resources at Ensembl and NCBI (e.g., Genome Workbench can manipulate and display BAM files).

In the future it is likely that the pace of sequencing will continue to increase, as there is no end in sight to the kinds of questions about biological principles that can be addressed with sequenced genomes, exomes, or targeted regions. Technological breakthroughs are on the immediate horizon that will likely enable much longer read lengths at reduced costs. This will facilitate sequencing across regions of repetitive DNA, improve the ability to detect structural variation, and continue to expand the number of species and individuals whose genomes are analyzed. It is commonly said that the greatest bottleneck in applying this technology is bioinformatics analysis. While hundreds of software tools are introduced each year and there is no single “best practice” for data analysis, there is still tremendous opportunity for students to master this fascinating field.

## PITFALLS

There are many applications of NGS technology. One interest is in identifying rare variants in individuals, and further determining which of these cause disease. The steps we outlined in this chapter (including base call quality assessment, alignment of sequence reads to a reference genome, calling variants, and interpreting their significance) are extraordinarily complex and the methods are in flux. Various software packages (such as aligners) are based on dramatically different assumptions about issues such as thresholds for quality scores and how repetitive sequences are mapped to the genome. There is not one simple reference genome, and when you sequence a genome (or exome or targeted region of interest) there are many workflows you can apply to your data. It is likely that if you begin with the same raw FASTQ data and apply two reasonably popular workflows, the final list of variants you obtain will differ substantially. To summarize, it is appropriate to develop your workflow carefully and analyze your results critically.

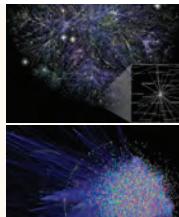
## ADVICE FOR STUDENTS

In my experience, many graduate students (and others) who have backgrounds in biology design next-generation-sequencing-based experiments, whether this involves DNA sequencing, ChiP-seq, RNA-seq, or related high-throughput techniques. In my opinion it is essential to consult a biostatistician regarding experimental design. Do you have adequate sample size? Have you balanced and randomized the design to help minimize batch effects that can destroy your experiment? Data are typically generated in a core facility or at a company, and in some cases data analysis is also performed on your behalf. Once the data arrive, many biologists who are new to bioinformatics prefer to receive a list of results of interest. I feel that it is critical for you to understand the entire data analysis workflow from beginning to end. Even if you are not yet an expert in performing these analyses, you should become educated enough to understand how the data were processed, what assumptions were made (assuming someone else has done the main analyses for you), and how to interpret the results. Take ownership of the project! This includes reading the primary literature which almost always includes benchmarks that explain the performance of one particular set of tools relative to existing software. If they have been generated in a core, or even by someone else in your lab, obtain and study the raw FASTQ files, the BAM files, and the VCF files. Galaxy provides an excellent web-based environment to learn about next-generation sequence analysis tools, and it can be a stepping-stone to working in a command-line environment in Linux.

If you are not currently acquiring this type of dataset, go to the 1000 Genomes website where you can download all these various files as well as extensive documentation. Define a question or a set of questions (e.g., “what is the extent of variation at the beta globin locus?”), and then obtain practical experience working with the data. Keep in mind the many places you can go for help (see Web Resources for this chapter).

## WEB RESOURCES

Several forums are dedicated to discussing issues related to next-generation sequencing. These include Biostars (<https://www.biostars.org/>, WebLink 9.65) and Seqanswers (<http://www.seqanswers.com>, WebLink 9.66). At Biostars be sure to explore the various sections such as ChIP-seq and Assembly as well as the tutorials on a variety of next-generation sequencing topics. For popular resources such as UCSC, Ensembl, NCBI, and Galaxy, as well as a variety of software tools, you can join user's groups to share information, questions, and answers. This can help you to learn and to keep pace with ongoing developments. Over 4000 software tools, including many dedicated to next-generation sequence analysis, are listed at <http://omictools.com> (WebLink 9.22; Henry *et al.*, 2014).



## Discussion Questions

**[9-1]** It was suggested in 2013 that 90% of the world's information has been accumulated in the past two years. Assuming that is correct, and considering NGS in particular, what is the likely change in that rate in the coming years?

**[9-2]** What categories of error are associated with a sequencing experiment? What approaches does GATK take to account for error? What tradeoffs does GATK make between sensitivity and specificity? Suppose you perform a study sequencing the exomes of affected individuals and their family members from 10 pedigrees. If you conduct a simplified sequence analysis workflow without the kinds of adjustments made by GATK, what do you think the consequence would be?

**[9-3]** In the future, public repositories will have 100,000 and then 1 million human genome sequences. What impact will such resources have on the study of variation and disease? Will there be a "human knockout collection" describing the phenotypes of many individuals who have homozygous deletions of each particular gene?

### PROBLEMS/COMPUTER LAB

**[9-1]** This exercise focuses on FASTQ files on the Mac, Windows, or Linux platforms. Obtain a set of FASTQ files from a repository. (1) Visit the Sequence Read Archive (SRA) at NCBI (<http://www.ncbi.nlm.nih.gov/sra/>, WebLink 9.18). (2) Click "Browse samples." Enter the search query "1000genomes"; currently there are over 400 samples. To focus on one, select NA19240 (<http://www.ncbi.nlm.nih.gov/biosample/SRS000214>, WebLink 9.67). (3) Examine the quality statistics of these FASTQ files using FASTQC either on the command line or with FASTQC at Galaxy. Note that the Tools panel of Galaxy includes an option to upload FASTQ files from Get Data > "Upload file from your computer." You can also download FASTQ files from the European Nucleotide Archive (ENA)

as mentioned in this chapter. (4) Align the FASTQ files to a reference using Bowtie or BWA.

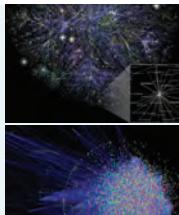
**[9-2]** Analyze BAM files using IGV. Explore viewing a chromosome (e.g., the beta globin region) at broad resolution then zoom into base pair resolution. Explore the options for coloring the reads. Select a list of any five gene symbols, upload them as a custom list, and simultaneously view the reads for those five regions.

**[9-3]** What is the frequency of variants in the *HBB* gene relative to the *HLA* locus? Consider using the UCSC Genome/Table Browser (build GRCh37) at positions chr6:29,570,005–33,377,699 (from *GABBR1* to *KIFC1*) and chr11:5,240,001–5,300,000 (including *HBB*). For more information on the MHC haplotype project visit <http://www.ucl.ac.uk/cancer/medical-genomics/mhc> (WebLink 9.68).

**[9-4]** We described a series of BEDtools examples. Obtain your own BED files (e.g., from the UCSC Table Browser) and further explore its tools. The BEDtools website offers many other suggestions for creatively exploring the genome.

**[9-5]** Perform variant annotation using SIFT and PolyPhen-2. (You may also use VAAST; because of its licensing restrictions a public tool is not available, although a license is freely available for academic use.) Select variants that are known to occur in *HBB*, or select another gene of interest. Of the publicly available individual human genome sequences, which have *HBB* variants that are predicted to be deleterious? As one approach, visit a region of interest at the UCSC Genome Browser and select the Variant Annotation Integrator. This provides data from SIFT, PolyPhen-2, Mutation Taster, GERP, and other resources.

**[9-6]** Explore the Ensembl annotation resources. Use the Variant Effect Predictor (in Linux) to predict the consequences of variants. Use the Data Slicer to create VCF files from various individuals and/or ethnic populations.



## Self-Test Quiz

- [9-1]** A notable limitation of pyrosequencing is that:
- its read lengths tend to be short;
  - its error rate can be high for homopolymers;
  - its error rate can be high for pyrimidines; or
  - it takes an extremely long time to complete a run.
- [9-2]** Most sequencing technologies produce raw data in what format?
- FASTA;
  - FASTG;
  - FASTQ; or
  - FASTX.
- [9-3]** FASTQ files include information on the quality of:
- each run;
  - each read group;
  - each base; or
  - each alignment.
- [9-4]** As genome assemblies improve,
- the scaffolds tend to invert;
  - the FASTQ base quality scores increase;
  - the contig N50 size decreases; or
  - the contig N50 size increases.
- [9-5]** Two repeats can be aligned to a genome with more accuracy if they are:
- inverted;
  - somewhat similar;

- extremely similar; or
- GC rich.

- [9-6]** SAM/BAM files store:
- FASTA records;
  - sequence alignments;
  - assemblies; or
  - VCF data.

- [9-7]** BEDtools is designed for:
- “genome arithmetic,” for example to compare two genomic regions;
  - variant calling;
  - variant prioritization; or
  - alignment to a reference genome.

- [9-8]** IGV
- calls single-nucleotide variants and indels;
  - views SAM files but not VCF;
  - views BAM files and VCF files; or
  - views FASTQ files and BCF files.

- [9-9]** A VCF file
- stores single-nucleotide variants (SNVs) only;
  - stores SNVs and insertions/deletions (indels);
  - stores SNVs, indels, and structural variants (SVs); or
  - stores SNVs, indels, SVs, and inversion strand monomorphisms.

## SUGGESTED READING

Lincoln Stein (2011) has written a succinct overview of next-generation sequencing workflows, while Paul Flicek and Ewan Birney (2009) provide an excellent review of alignment and assembly. Koboldt *et al.* (2010) describe next-generation sequencing of human genomes, including quality assessment and production issues. Pabinger *et al.* (2014) surveyed 205 tools in the categories of quality assessment, alignment, variant identification, variant annotation, and visualization.

## REFERENCES

- Adzhubei, I.A., Schmidt, S., Peshkin, L. *et al.* 2010. A method and server for predicting damaging missense mutations. *Nature Methods* 7(4), 248–249. PMID: 20354512.
- Adzhubei, I., Jordan, D.M., Sunyaev, S.R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics Chapter 7*, Unit 7.20. PMID: 23315928.

- Alkan, C., Coe, B.P., Eichler, E.E. 2011a. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12**(5), 363–376. PMID: 21358748.
- Alkan, C., Sajadian, S., Eichler, E.E. 2011b. Limitations of next-generation genome sequence assembly. *Nature Methods* **8**(1), 61–65. PMID: 21102452.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P. *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218), 53–59. PMID: 18987734.
- Bradnam, K.R., Fass, J.N., Alexandrov, A. *et al.* 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**(1), 10. PMID: 23870653.
- Branco, M.R., Ficz, G., Reik, W. 2011. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics* **13**(1), 7–13. PMID: 22083101.
- Carpenter, W.B. 1876. *Principles of Human Physiology*. Henry C. Lea, Philadelphia.
- Chang, X., Wang, K. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. *Journal of Medical Genetics* **49**(7), 433–436. PMID: 22717648.
- Chen, K., Wallis, J.W., McLellan, M.D. *et al.* 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**(9), 677–681. PMID: 19668202.
- Chi, S.W., Zang, J.B., Mele, A., Darnell, R.B. 2009. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**(7254), 479–486. PMID: 19536157.
- Chun, S., Fay, J.C. 2009. Identification of deleterious mutations within three human genomes. *Genome Research* **19**(9), 1553–1561. PMID: 19602639.
- Clark, M.J., Chen, R., Lam, H.Y. *et al.* 2011. Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology* **29**(10), 908–914. PMID: 21947028.
- Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**(6), 1767–1771. PMID: 20015970.
- Danecek, P., Auton, A., Abecasis, G. *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15), 2156–2158. PMID: 21653522.
- DePristo, M.A., Banks, E., Poplin, R. *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**(5), 491–498. PMID: 21478889.
- Drmanac, R., Sparks, A.B., Callow, M.J. *et al.* 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**(5961), 78–81. PMID: 19892942.
- Edwards, R.A., Rodriguez-Brito, B., Wegley, L. *et al.* 2006. Using pyrosequencing to shed light on deep marine microbial ecology. *BMC Genomics* **7**, 57. PMID: 16549033.
- Eid, J., Fehr, A., Gray, J. *et al.* 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910), 133–138. PMID: 19023044.
- Flanagan, S.E., Patch, A.M., Ellard, S. 2010. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers* **14**(4), 533–537. PMID: 20642364.
- Flicek, P., Birney, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**(11 Suppl), S6–S12. PMID: 19844229.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Gnerre, S., Maccallum, I., Przybylski, D. *et al.* 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Science, USA* **108**(4), 1513–1518. PMID: 21187386.
- Guo, Y., Li, J., Li, C.I., Shyr, Y., Samuels, D.C. 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**(9), 1210–1211. PMID: 23471301.
- Henry, V.J., Bandrowski, A.E., Pepin, A.S., Gonzalez, B.J., Desfeux, A. 2014. OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)* **2014**, pii: bau069. PMID: 25024350.
- Henson, J., Tischler, G., Ning, Z. 2012. Next-generation sequencing and large genome assemblies. *Pharmacogenomics* **13**(8), 901–915. PMID: 22676195.

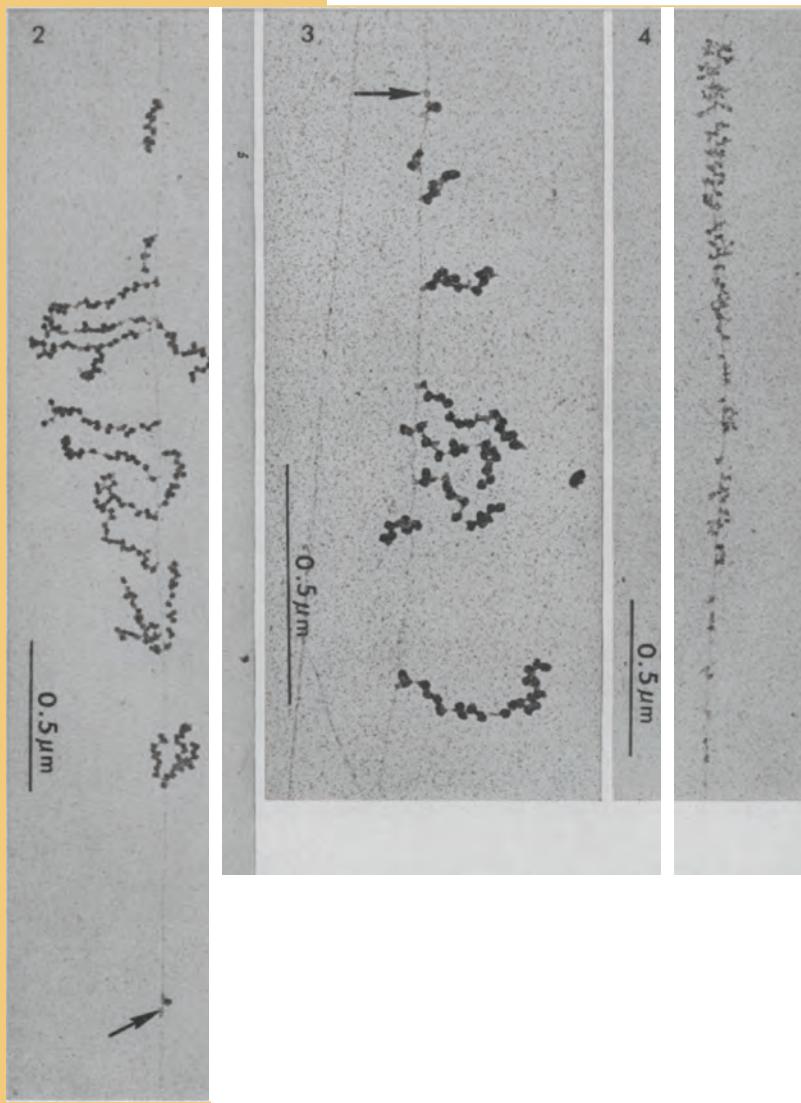
- Hicks, S., Wheeler, D.A., Plon, S.E., Kimmel, M. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Human Mutations* **32**(6), 661–668. PMID: 21480434.
- Holley, R.W., Apgar, J., Everett, G.A. *et al.* 1965. Structure of a ribonucleic acid. *Science* **147**(3664), 1462–1465. PMID: 14263761.
- Hoppe-Seyler, F. 1877. *Traité d'Analyse Chimique Appliquée à la Physiologie et à la Pathologie*. Librairie F. Savy, Paris.
- Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., Birney, E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research* **21**(5), 734–740. PMID: 21245279.
- Hu, H., Huff, C.D., Moore, B. *et al.* 2013. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology* **37**(6), 622–634. PMID: 23836555.
- Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**(7), R143. PMID: 17659080.
- Huss, M. 2010. Introduction into the analysis of high-throughput-sequencing based epigenome data. *Briefings in Bioinformatics* **11**(5), 512–523. PMID: 20457755.
- Hyman, E.D. 1988. A new method of sequencing DNA. *Analytical Biochemistry* **174**(2), 423–436. PMID: 2853582.
- Jaffe, A., Wojcik, G., Chu, A. *et al.* 2011. Identification of functional genetic variation in exome sequence analysis. *BMC Proceedings* **5**, Supplement 9, S13. PMID: 22373437.
- Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., Karolchik, D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**(17), 2204–2207. PMID: 20639541.
- Koboldt, D.C., Ding, L., Mardis, E.R., Wilson, R.K. 2010. Challenges of sequencing human genomes. *Briefings in Bioinformatics* **11**(5), 484–498. PMID: 20519329.
- Koboldt, D.C., Larson, D.E., Chen, K., Ding, L., Wilson, R.K. 2012. Massively parallel sequencing approaches for characterization of structural variation. *Methods in Molecular Biology* **838**, 369–384. PMID: 22228022.
- Koren, S., Treangen, T.J., Pop, M. 2011. Bambus 2: scaffolding metagenomes. *Bioinformatics* **27**(21), 2964–2971. PMID: 21926123.
- Koren, S., Harhay, G.P., Smith, T.P. *et al.* 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* **14**(9), R101. PMID: 24034426.
- Ku, C.S., Naidoo, N., Wu, M., Soong, R. 2011. Studying the epigenome using next generation sequencing. *Journal of Medical Genetics* **48**(11), 721–730. PMID: 21825079.
- Kumar, P., Henikoff, S., Ng, P.C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4**(7), 1073–1081. PMID: 19561590.
- Lander, E.S., Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**(3), 231–239. PMID: 3294162.
- Lander, E.S., Linton, L.M., Birren, B. *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921. PMID: 11237011.
- Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359. PMID: 22388286.
- Ledergerber, C., Dessimoz, C. 2011. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics* **12**(5), 489–497. PMID: 21245079.
- Levy, S., Sutton, G., Ng, P.C. *et al.* 2007. The diploid genome sequence of an individual human. *PLoS Biology* **5**, e254. PMID: 17803354.
- Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760. PMID: 19451168.
- Li, H., Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5), 589–595. PMID: 20080505.
- Li, H., Ruan, J., Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**(11), 1851–1858. PMID: 18714091.

- Li, H., Handsaker, B., Wysoker, A. *et al.* 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079. PMID: 19505943.
- Li, R., Fan, W., Tian, G. *et al.* 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463**(7279), 311–317. PMID: 20010809.
- Li, Z., Chen, Y., Mu, D. *et al.* 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in Functional Genomics* **11**(1), 25–37. PMID: 22184334.
- Liu, L., Kumar, S. 2013. Evolutionary balancing is critical for correctly forecasting disease-associated amino acid variants. *Molecular Biology and Evolution* **30**(6), 1252–1257. PMID: 23462317.
- Liu, X., Han, S., Wang, Z., Gelernter, J., Yang, B.Z. 2013. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS One* **8**(9), e75619. PMID: 24086590.
- Lopes, M.C., Joyce, C., Ritchie, G.R. *et al.* 2012. A combined functional annotation score for non-synonymous variants. *Human Heredity* **73**(1), 47–51. PMID: 22261837.
- Luo, R., Liu, B., Xie, Y. *et al.* 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**(1), 18. PMID: 23587118.
- Margulies, M., Egholm, M., Altman, W.E. *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057), 376–380. PMID: 16056220.
- McKenna, A., Hanna, M., Banks, E. *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**(9), 1297–1303. PMID: 20644199.
- Medvedev, P., Stanciu, M., Brudno, M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**(11 Suppl), S13–20. PMID: 19844226.
- Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Research* **15**, 1767–1776. PMID: 16339375.
- Miller, J.R., Delcher, A.L., Koren, S. *et al.* 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**(24), 2818–2824. PMID: 18952627.
- Miller, J.R., Koren, S., Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**(6), 315–327. PMID: 20211242.
- Miller, M.P., Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics* **10**(21), 2319–2328. PMID: 11689479.
- Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K. *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915), 520–562. PMID: 12466850.
- Nagarajan, N., Pop, M. 2013. Sequence assembly demystified. *Nature Reviews Genetics* **14**(3), 157–167. PMID: 23358380.
- Ng, P.C., Henikoff, S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics* **7**, 61–80. PMID: 16824020.
- Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **12**(6), 443–451. PMID: 21587300.
- O’Rawe, J., Jiang, T., Sun, G. *et al.* 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* **5**(3), 28. PMID: 23537139.
- Pabinger, S., Dander, A., Fischer, M. *et al.* 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* **15**(2), 256–278. PMID: 23341494.
- Park, P.J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669–680. PMID: 19736561.
- Paszkiewicz, K., Studholme, D.J. 2010. De novo assembly of short sequence reads. *Briefings in Bioinformatics* **11**(5), 457–472. PMID: 20724458.
- Peters, B.A., Kermani, B.G., Sparks, A.B. *et al.* 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**(7406), 190–195. PMID: 22785314.
- Pevzner, P.A., Tang, H., Waterman, M.S. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of National Academy of Science, USA* **98**(17), 9748–9753. PMID: 11504945.

- Quinlan, A.R., Hall, I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842. PMID: 20110278.
- Ramensky, V., Bork, P., Sunyaev, S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research* **30**(17), 3894–3900. PMID: 12202775.
- Robinson, J.T., Thorvaldsdóttir, H., Winckler, W. *et al.* 2011. Integrative genomics viewer. *Nature Biotechnology* **29**(1), 24–26. PMID: 21221095.
- Rothberg, J.M., Leamon, J.H. 2008. The development and impact of 454 sequencing. *Nature Biotechnology* **26**(10), 1117–1124. PMID: 18846085.
- Rothberg, J.M., Hinz, W., Rearick, T.M. *et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**(7356), 348–352. PMID: 21776081.
- Salzberg, S.L., Dunning Hotopp, J.C., Delcher, A.L. *et al.* 2005. Serendipitous discovery of Wolbachia genomes in multiple *Drosophila* species. *Genome Biology* **6**(3), R23. PMID: 15774024.
- Salzberg, S.L., Phillippy, A.M., Zimin, A. *et al.* 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22**(3), 557–567. PMID: 22147368.
- Samuels, D.C., Han, L., Li, J. *et al.* 2013. Finding the lost treasures in exome sequencing data. *Trends in Genetics* **29**(10), 593–599. PMID: 23972387.
- Sanger, F., Nicklen, S., Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science, USA* **74**, 5463–5467. PMID: 271968.
- Sauna, Z.E., Kimchi-Sarfaty, C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics* **12**(10), 683–691. PMID: 21878961.
- Schwarz, J.M., Rödelsperger, C., Schuelke, M., Seelow, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* **7**(8), 575–576. PMID: 20676075.
- Shendure, J., Lieberman Aiden, E. 2012. The expanding scope of DNA sequencing. *Nature Biotechnology* **30**(11), 1084–1094. PMID: 23138308.
- Shihab, H.A., Gough, J., Cooper, D.N. *et al.* 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human Mutations* **34**(1), 57–65. PMID: 23033316.
- Shirley, M.D., Tang, H., Gallione, C.J. *et al.* 2013. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in *GNAQ*. *New England Journal of Medicine* **368**(21), 1971–1979. PMID: 23656586.
- Sim, N.L., Kumar, P., Hu, J. *et al.* 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research* **40**(Web Server issue), W452–457. PMID: 22689647.
- Simpson, J.T., Durbin, R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**(3), 549–556. PMID: 22156294.
- Simpson, J.T., Wong, K., Jackman, S.D. *et al.* 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**(6), 1117–1123. PMID: 19251739.
- Song, L., Zhang, Z., Grasfeder, L.L. *et al.* 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research* **21**(10), 1757–1767. PMID: 21750106.
- Stein, L.D. 2011. An introduction to the informatics of “next-generation” sequencing. *Current Protocols in Bioinformatics* **Chapter 11**, Unit 11.1. PMID: 22161566.
- Tchernitchko, D., Goossens, M., Wajeman, H. 2004. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. *Clinical Chemistry* **50**(11), 1974–1978. PMID: 15502081.
- Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**(2), 178–192. PMID: 2251747.
- Thusberg, J., Olatubosun, A., Vihinen, M. 2011. Performance of mutation pathogenicity prediction methods on missense variants. *Human Mutations* **32**(4), 358–368. PMID: 21412949.
- Treangen, T.J., Salzberg, S.L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**(1), 36–46. PMID: 22124482.

- Van der Auwera, G.A., Carneiro, M.O., Hartl, C. *et al.* 2013. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **11**, 11.10.1–11.10.33.
- Vaughan, V.C., Novy, F.G. 1891. *Ptomaines, Leucomaines, and Bacterial Proteids; Or, The Chemical Factors in the Causation of Disease*. Lea Brothers & Co., Philadelphia.
- Whiteford, N., Skelly, T., Curtis, C. *et al.* 2009. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics* **25**(17), 2194–2199. PMID: 19549630.
- Wu, R. 1970. Nucleotide sequence analysis of DNA. I. Partial sequence of the cohesive ends of bacteriophage lambda and 186 DNA. *Journal of Molecular Biology* **51**, 501–521. PMID: 4321727.
- Wu, R., Taylor, E. 1971. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology* **57**, 491–511. PMID: 4931680.
- Yandell, M., Huff, C., Hu, H. *et al.* 2011. A probabilistic disease-gene finder for personal genomes. *Genome Research* **21**(9), 1529–1542. PMID: 21700766.
- Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**(21), 2865–2871. PMID: 19561018.
- Zerbino, D.R., Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**(5), 821–829. PMID: 18349386.





Miller and colleagues (1970, p. 394) visualized gene expression, depicting *Escherichia coli* chromosomal DNA (oriented vertically as a thin strand in each figure) in the process of transcription and translation. As mRNA is transcribed from the genomic DNA and extends off to the side, polyribosomes (dark objects) appear like beads on a string, translating the mRNA to protein.

Source: Miller *et al.* (1970). Reproduced with permission from AAAS.

# Bioinformatic Approaches to Ribonucleic Acid (RNA)

# CHAPTER 10

*When dealing with molecular sequences an evolutionist feels a sense of liberation; he is no longer confined to the world of “higher forms.” From the vantage point provided by molecular data, he now gazes over the Cambrian “wall” that had obstructed his temporal perspective. He can now scan the full panorama of Earth’s four-billion-year evolutionary history. Color has been added to his monochromatic, morphocentric view of the evolutionary process, as physiological and molecular characters take on phylogenetic significance. The static and relatively superficial paleontological link between biology and geology becomes engulfed by the far more interesting long-term interplay between the evolution of the physical planet and that of the organisms inhabiting it. This great burgeoning of our evolutionary perspective has been brought about largely through sequence characterizations of one molecular species, ribosomal RNA (rRNA).*

—Gary Olsen and Carl Woese (1993), p. 113

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

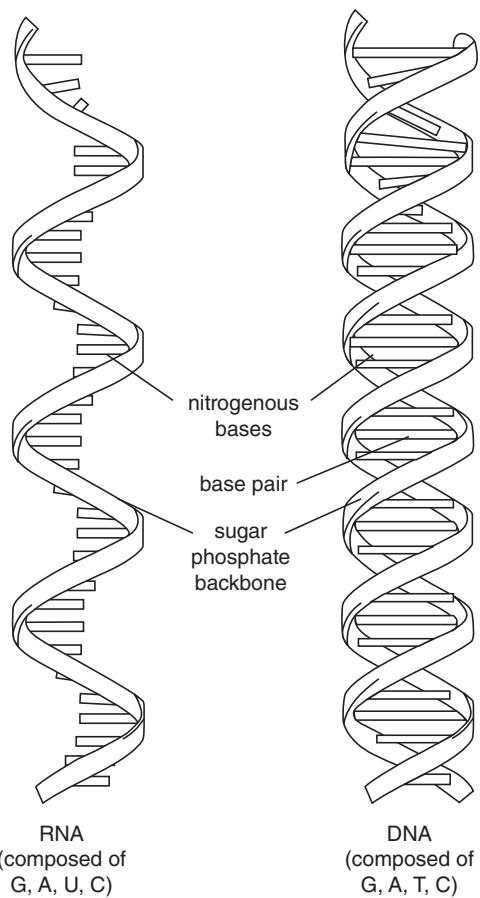
- describe the major categories of coding and noncoding RNA;
- compare and contrast techniques for measuring steady-state RNA levels; and
- compare and contrast the use of microarrays and RNA-seq for measuring mRNA levels.

## INTRODUCTION TO RNA

The word “gene” was introduced by Johannsen in 1909 to describe the entity that determines how characteristics of an organism are inherited. Classic studies by Beadle and Tatum (1941) in the fungus *Neurospora* showed that genes direct the synthesis of enzymes in a 1:1 ratio. As early as 1944, Oswald T. Avery demonstrated that deoxyribonucleic acid (DNA) is the genetic material. Avery *et al.* (1944) showed that DNA from bacterial strains with high pathogenicity could transform strains with low to high pathogenicity. Further experiments involving bacterial transformation, performed by Frederick Griffith, Avery, McLeod, McCarthy, Hotchkiss, and Hershey confirmed that DNA is the genetic material.

James Watson and Francis Crick (1953) proposed the double helical nature of DNA in 1953 (Fig. 10.1). Soon after, Crick (1958) could formulate the central dogma of molecular

You can learn about some of the original discoveries concerning nucleic acids by reading about their Nobel Prize awards. Albrecht Kossel was awarded the Nobel Prize in 1910 for characterizing nucleic acids ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1910/](http://nobelprize.org/nobel_prizes/medicine/laureates/1910/), WebLink 10.1). Beadle and Tatum were awarded Nobel Prizes in 1958 for their one gene-one enzyme hypothesis (see [http://nobelprize.org/nobel\\_prizes/medicine/laureates/1958/](http://nobelprize.org/nobel_prizes/medicine/laureates/1958/), WebLink 10.2). Severo Ochoa and Arthur Kornberg shared a 1959 Nobel Prize “for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid” ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1959/](http://nobelprize.org/nobel_prizes/medicine/laureates/1959/), WebLink 10.3). Although Oswald Avery was the first to show that DNA is the genetic material, he did not receive a Nobel Prize.



**FIGURE 10.1** Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). While DNA usually adopts a double helical conformation, RNA tends to be single stranded. A notable exception is the double-stranded base pairing of many noncoding RNAs to form stem-loop structures, described in this chapter. Adapted from National Human Genome Research Institute (<http://www.genome.gov/glossary/>).

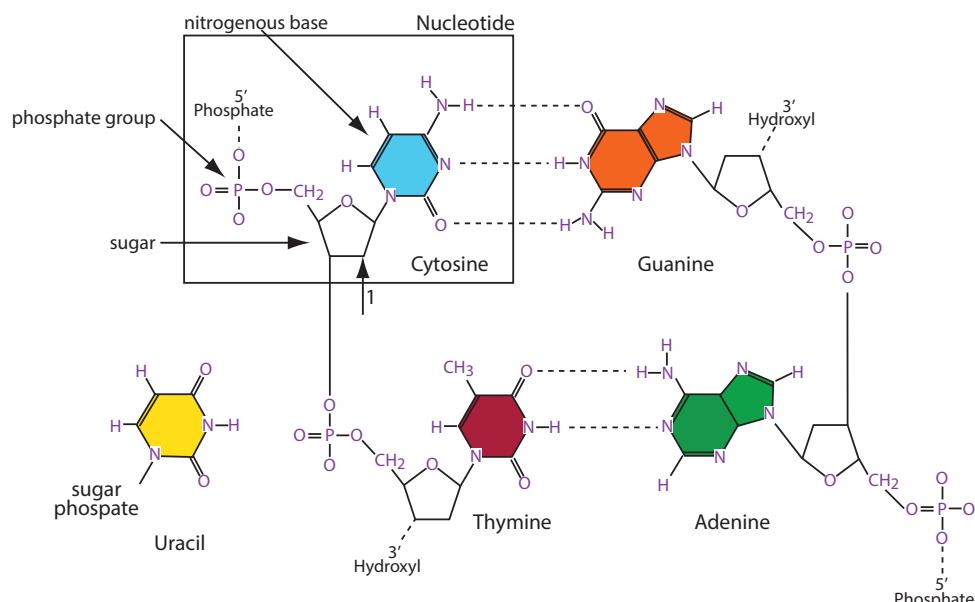
Francis Crick, James Watson, and Maurice Wilkins shared the 1962 Nobel Prize in Physiology or Medicine “for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material.” See [http://nobelprize.org/nobel\\_prizes/medicine/laureates/1962/](http://nobelprize.org/nobel_prizes/medicine/laureates/1962/) (WebLink 10.4).

biology that DNA is transcribed into RNA then translated into protein. Crick wrote (1958, p. 153) that the central dogma:

states that once ‘information’ has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the *precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

In this article Crick further postulated the existence of an adaptor molecule to convert the information from codons in RNA to amino acids in proteins; transfer RNA (tRNA) was indeed later identified.

During the 1960s the genetic code was solved (e.g., Nirenberg, 1965), showing the relationship between messenger RNA codons and the amino acids that are specified. This completed a detailed model for the flow of genetic information from DNA to RNA to protein. However, even by the 1950s, this model was called into question by the nature of RNA. Why did hybridization experiments suggest that only a minute fraction of RNA was complementary to DNA of genes? RNA could be purified away from DNA and proteins, and then shown to separate into discrete bands on density gradients having sedimentation coefficients of 23S, 16S, and 4S. The 23S and 16S species were found to localize to ribosomes and constituted about 85% of all RNA in bacteria. tRNA was found to constitute



**FIGURE 10.2** The nucleotide bases include the purines guanine and adenine and the pyrimidines thymine, uracil (which substitutes for thymine in RNA), and cytosine. These nitrogenous bases are attached to ribose sugars and triphosphate groups. In the case of DNA, the ribose lacks an oxygen side group (arrow 1) that is present in RNA. Redrawn and modified from the National Human Genome Research Institute talking glossary of genetic terms (<http://www.genome.gov/glossary/?id=143>).

about 15% of all RNA. Surprisingly, mRNA was found to represent only a small percentage of total RNA (about 1–4%).

DNA consists of the four nucleotides adenine, guanine, cytosine, and thymidine (A, G, C, T). It can be transcribed into ribonucleic acid (RNA), consisting of the nucleotides A, G, C, and U (uracil; Fig. 10.2). RNA has a backbone consisting of the five-carbon sugar ribose with a purine or pyrimidine base attached to each sugar residue. A phosphate group links the nucleoside (i.e., the sugar with base) to form a nucleotide.

The process of transcription of DNA results in the formation of RNA molecules in two broad classes. The first class is coding RNA, formed when DNA is transcribed into messenger RNA (mRNA). This mRNA is subsequently translated into protein on the surface of a ribosome in a process mediated by transfer RNA (tRNA) and ribosomal RNA (rRNA) as well as by proteins. A second class is noncoding RNA, in which RNA products are transcribed from DNA function without being further translated into protein. We next discuss noncoding and coding RNA from a bioinformatics perspective. There is considerable excitement about many recent advances in our understanding of all classes of RNAs, as we begin to recognize their diverse functional properties. By the 1980s the extraordinary versatility of RNA began to be appreciated when, in addition to the three major RNAs (rRNA, tRNA, mRNA), RNAs with catalytic properties were discovered. Previously, nucleic acids had been considered molecules underlying heredity while proteins functioned as enzymes or other modulators of cellular processes (see Chapter 12). The discovery of ribozymes is consistent with a model of the early evolution of life on Earth in which RNA was the first genetic material, prior to the emergence of DNA. Another implication is that RNA has many potential functional roles in the cell beyond serving as an intermediary between DNA and protein. For example, rRNA catalyzes peptide bond formation during translation.

The 1968 Nobel Prize in Physiology or Medicine was awarded to Robert Holley, Har Khorana, and Marshall Nirenberg “for their interpretation of the genetic code and its function in protein synthesis.” Visit [http://nobelprize.org/nobel\\_prizes/medicine/laureates/1968/](http://nobelprize.org/nobel_prizes/medicine/laureates/1968/) (WebLink 10.5).

The purines include adenine and guanine, and the pyrimidines include cytosine, thymine, and uracil. To view their structures, visit the NCBI website, enter their names into a search of Entrez, and view the results in the PubChem database.

Sidney Altman and Thomas Cech shared the 1989 Nobel Prize in Chemistry "for their discovery of the catalytic properties of RNA." See [http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1989/](http://nobelprize.org/nobel_prizes/chemistry/laureates/1989/) (WebLink 10.6). Altman characterized RNA enzymes (ribozymes) in the bacterium *Escherichia coli*, while Cech studied ribozymes in *Tetrahymena thermophila*. An example of a human gene encoding a noncoding RNA with enzymatic activity is RNA component of mitochondrial RNA processing endoribonuclease (RMRP, accession NR\_003051.3, assigned to chromosome 9p21-p12).

The RefSeq accession for the human *Xist* is NR\_001564.2, spanning 19,296 base pairs. The accession for murine "antisense Igf2r RNA" (*Air*) on chromosome 17 is NR\_002853.2 (1,176 base pairs).

In addition to Rfam and MirBase there are many other excellent noncoding RNA databases such as RNADb (Pang *et al.*, 2007) at <http://research.imb.uq.edu.au/rnadb/> (WebLink 10.7). See Washietl and Hofacker (2010) for a review of databases.

You can access the Rfam database at <http://www.sanger.ac.uk/resources/databases/> (WebLink 10.8) or <http://rfam.janelia.org/> (WebLink 10.9). Release 12.0 (July 2014) has >2400 families of noncoding RNA genes and over 19 million regions.

Throughout this chapter we use human chromosome 21 to demonstrate the nature of various RNAs. This is among the smallest human chromosomes (about 48 million base pairs) and one of the five human chromosomes having ribosomal DNA clusters that produce rRNA. We also use the globins as examples.

## NONCODING RNA

The major classes of noncoding RNAs are tRNA and rRNA, which together account for approximately 95% of all RNAs in a given eukaryotic cell. Other noncoding RNAs, discussed in the following sections, include small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), microRNA, and short interfering RNA (siRNA). Beyond tRNA and rRNA, relatively few noncoding RNAs have had their functions defined. A prominent example of a functionally characterized noncoding RNA is *Xist* (X (inactive)-specific transcript) encoded by the *XIST* gene. This RNA is located in the X inactivation center of the X chromosome and functions in X chromosome inactivation. While males have one copy of the X chromosome (with XY sex chromosomes), females have two copies of which one is inactivated in every diploid cell of mammalian and some other species. *Xist* is expressed from the inactive X and binds to its chromatin, facilitating chromosome inactivation (Borsani *et al.*, 1991). Another functional noncoding RNA is *Air* which functions at the *Igf2R* locus (Sleutels *et al.*, 2002). Some genes that are present in two copies are imprinted, that is, expressed selectively from an allele of one parent. In mouse, noncoding *Air* RNA is required to suppress expression of three genes (*Igf2r*, *Slc22a2*, *Slc22a3*) from the paternal chromosome. It is notable that many noncoding RNAs are very poorly conserved between species, and we explore this for *XIST* and *Air* in problem (10.1) at the end of this chapter.

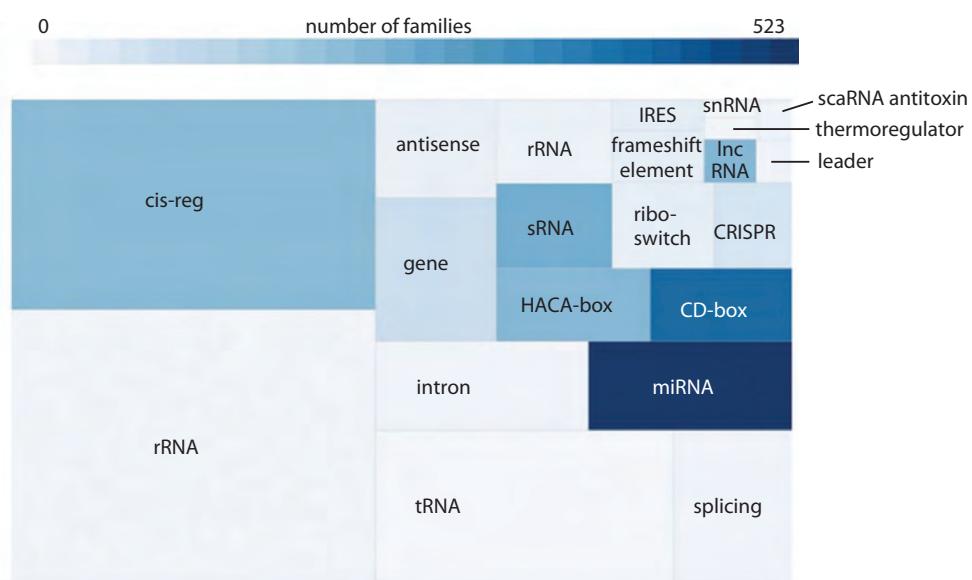
The abundant and well-characterized noncoding RNAs (tRNA, rRNA, and mRNA) have central roles in translation. The smaller and relatively more poorly characterized noncoding RNAs have been proposed to have a broad variety of functions in the regulation of gene expression, development, and assorted physiological and pathophysiological processes. In the following sections we introduce several prominent databases that collect information about noncoding RNAs. These include Rfam (discussed in the following section), MirBase (Kozomara and Griffiths-Jones, 2011), and RNACentral (Bateman *et al.*, 2011). We also introduce two main methods for predicting noncoding RNA structures: a comparative method that is based on multiple sequence alignments of RNAs, and the thermodynamic approach that seeks the minimum free energy of a structure (Hofacker and Lorenz, 2014). Analysis of noncoding RNAs is reviewed by Washietl (2010), Washietl *et al.* (2012), and Nawrocki and Eddy (2013).

## Noncoding RNAs in the Rfam Database

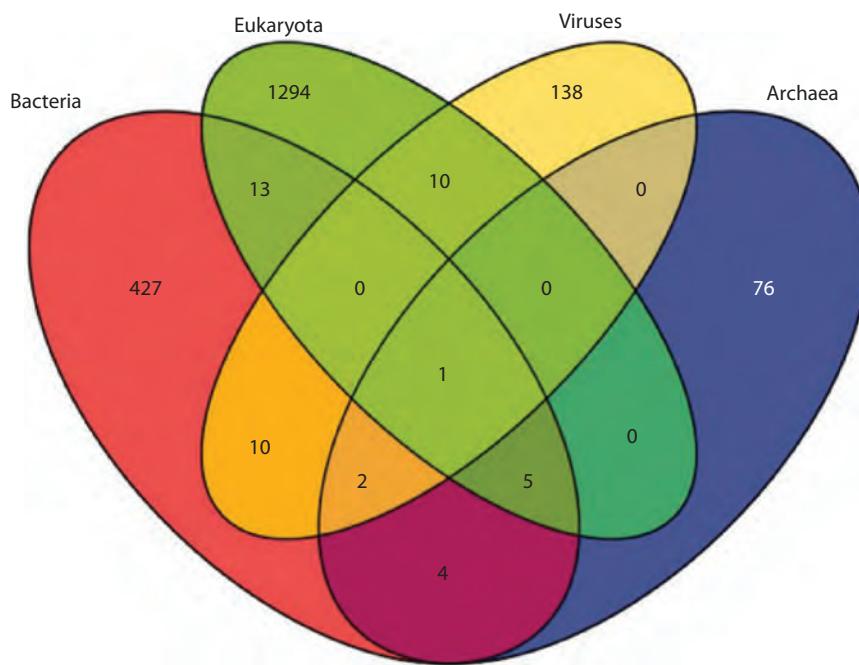
We introduced the Pfam database for protein families in Chapter 6 as an important bioinformatics resource. The Rfam database serves a comparable role in characterizing RNA families (Nawrocki *et al.*, 2015). Rfam includes RNA alignments, consensus secondary structures, and covariance models (discussed in the following). Each Rfam family has a covariance model that is a statistical model of that family's sequence and structure.

The contents of Rfam permit a survey of all currently known noncoding RNAs (Fig. 10.3). These include several well-characterized families that span all three domains of life: tRNAs, rRNAs, SRP RNA (responsible for protein export) and RNaseP (necessary for tRNA maturation). Table 10.1 lists the most abundant RNA families in Rfam for all species.

(a) Rfam sequence space and numbers of families



(b) Rfam taxonomic groupings



**FIGURE 10.3** The Rfam database noncoding RNAs. (a) Types of noncoding RNAs. The number of families is proportional to color. The number of annotated regions is proportional to the size of the rectangles. (b) Taxonomic coverage of Rfam families across the three domains of life (bacteria, archaea, and eukaryotes) and viruses. Families are categorized according to the taxa covered by the seed sequences. Adapted from Burge *et al.* (2013), with permission from Oxford University Press.

**TABLE 10.1 A list of the 13 Rfam entries with the largest number of members. No. full: number of members of the Rfam family (for the full dataset rather than the seed alignment of representative members), rounded to the nearest thousand; Id: the average percent identity of the full alignments.**

Name	Accession	No. full	Ave. len. (full)	Id	Type	Description
5_8S_rRNA	RF00002	376,000	152	69	Gene; rRNA	5.8S ribosomal RNA
tRNA	RF00005	298,000	73	46	Gene; tRNA	tRNA
5S_rRNA	RF00001	229,000	116	60	Gene; rRNA	5S ribosomal RNA
UnaL2	RF00436	101,000	54	78	Cis-reg	UnaL2 LINE 3' element
HIV_POL-1_SL	RF01418	83,000	113	77	Cis-reg	HIV pol-1 stem loop
U6	RF00026	72,000	105	77	Gene; snRNA; splicing	U6 spliceosomal RNA
mtDNA_ssA	RF01853	62,000	104	67	Gene; antisense	Mitochondrial DNA control region secondary structure A
Intron_gpl	RF00028	60,000	365	36	Intron	Group I catalytic intron
Intron_gpll	RF00029	51,000	87	54	Intron	Group II catalytic intron
Hammerhead_1	RF00163	49,000	59	70	Gene; ribozyme	Hammerhead ribozyme (type I)
RRE	RF00036	44,000	337	97	Cis-reg	HIV Rev response element
HIV_GSL3	RF00376	39,000	84	82	Cis-reg	HIV gag stem loop 3 (GSL3)
SNORA7	RF00409	26,000	140	79	Gene; snRNA; snoRNA; HACA-box	Small nucleolar RNA SNORA7

Source: Rfam 11.0. Reproduced under the Creative Commons Zero licence, CC0.

When you search a sequence against the Rfam database, the results include a bit score in the familiar log-odds ratio format:

Chromosome 21p (the short arm of chromosome 21) is about 12 million base pairs in length and contains rDNA clusters (described below) and a total of three RefSeq coding genes (*TEKT4P2*, *TPTE*, and *BAGE*) in assembly GRCh37. Chromosome 21q (the long arm) extends for about 35 million base pairs and has 553 RefSeq genes. Web Document 10.1 at <http://bioinfbook.org> shows the results of a search for Rfam entries from chromosome 21, performed using BioMart at Ensembl. Web Document 10.2 shows a more inclusive GFF file downloaded from the Rfam site at Janelia Farm. GFF files are also suitable for analysis with BEDTools and related software (Chapter 9).

$$\text{bit score} = \log_2 \left( \frac{P_{\text{CM}}}{P_{\text{null}}} \right). \quad (10.1)$$

That is, a positive bit score indicates a significant match in which the query sequence given the covariance model is more likely than the query sequence given the null model.

We can survey typical noncoding RNAs by viewing an Rfam summary of those present on human chromosome 21 (Fig. 10.4). The Rfam site offers a general feature format (GFF) file of queries such as chromosome 21 entries. This has 22 distinct families in 65 regions. These include a tRNA gene, an rRNA gene, small nuclear genes involved in splicing, small nucleolar genes, and microRNAs. We will next examine these various noncoding RNA types.

## Transfer RNA

Transfer RNA molecules carry a specific amino acid and match it to its corresponding codon on an mRNA during protein synthesis. tRNAs occur in 20 amino acid acceptor groups corresponding to the 20 amino acids specified in the genetic code. tRNA forms a structure consisting of about 70–90 nucleotides folded into a characteristic cloverleaf. Key features of this structure include a D loop, an anti-codon loop which is responsible for recognizing messenger RNA codons, a T loop, and a 3' end to which aminoacyl tRNA synthetases attach the appropriate amino acid specific for each tRNA.

Family	Start	End	Bits score				
<u>tRNA</u>	9,734,391	9,734,325	31.22	<u>SNORD7A</u>	17,657,089	17,657,017	59.88
<u>RSV RNA</u>	9,990,192	9,989,909	36.74	<u>mir-10</u>	17,911,414	17,911,485	69.08
<u>RSV RNA</u>	10,142,311	10,142,595	36.83	<u>let-7</u>	17,912,152	17,912,227	62.65
<u>Metazoa SRP</u>	10,380,661	10,380,378	122.56	<u>lin-4</u>	17,962,567	17,962,636	76.22
<u>SNORA7O</u>	10,385,953	10,386,047	42.25	<u>U1</u>	18,091,317	18,091,476	91.09
<u>tRNA</u>	10,492,972	10,492,907	26.05	<u>U6</u>	18,803,865	18,803,965	62.92
<u>tRNA</u>	10,493,037	10,492,973	37.46	<u>tRNA</u>	18,827,177	18,827,107	63.87
<u>mir-548</u>	11,052,015	11,051,932	82.85	<u>Metazoa SRP</u>	18,878,771	18,879,046	64.51
<u>U6</u>	14,419,904	14,420,010	66.41	<u>Y RNA</u>	18,899,565	18,899,458	41.00
<u>U6</u>	14,993,898	14,994,004	76.41	<u>Y RNA</u>	18,949,116	18,949,224	40.22
<u>U6</u>	15,340,916	15,340,810	63.69	<u>RSV RNA</u>	19,938,102	19,937,818	72.79
<u>5S rRNA</u>	15,443,192	15,443,307	42.69	<u>U1</u>	20,717,465	20,717,629	93.53
<u>DRNA</u>	15,448,359	15,448,271	68.52	<u>U6</u>	21,728,164	21,728,060	49.35
<u>U6</u>	16,986,602	16,986,708	75.19	<u>7SK</u>	21,728,965	21,729,208	75.15
<u>U6</u>	17,407,829	17,407,733	41.70	<u>mir-492</u>	21,798,181	21,798,066	40.04
				<u>U4</u>	23,577,511	23,577,651	73.90
				<u>U2</u>	24,654,231	24,654,058	62.66

**FIGURE 10.4** Noncoding RNA families in the Rfam database that are assigned to human chromosome 21. Only a portion of the entries is shown.

Source: Rfam (<http://www.sanger.ac.uk/Software/Rfam/index.shtml>). Reproduced with permission from Genome Research Ltd.

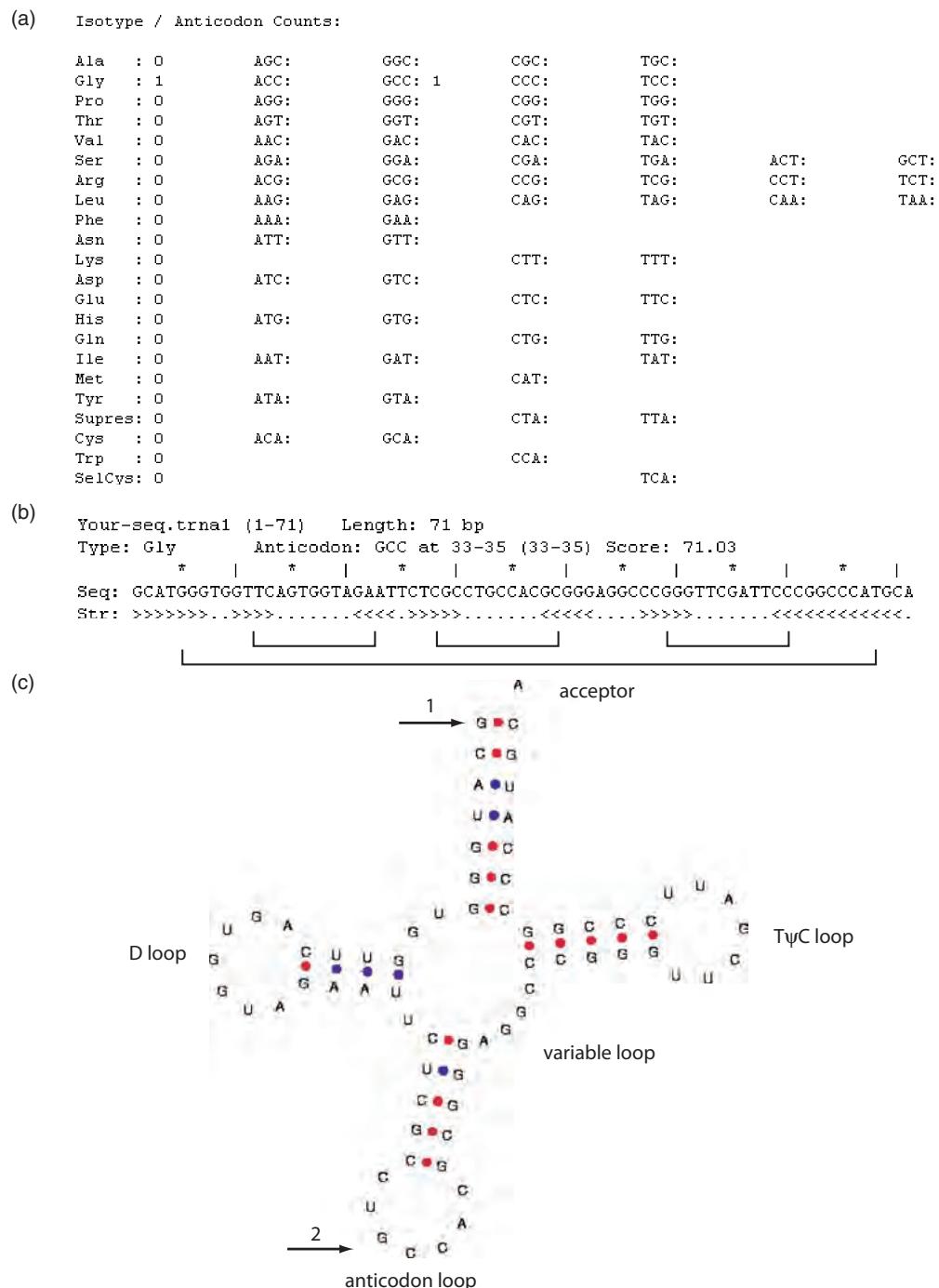
We demonstrate a computational approach to identifying tRNAs using the tRNA-Scan-SE program (Lowe and Eddy, 1997) via its webserver (Schattner *et al.*, 2005). As an input, we use a tRNA known to be assigned to human chromosome 21. The output includes the anticodon counts (Fig. 10.5a), listing the anticodons that have been identified corresponding to the 20 amino acids as well as stop codons and the modified amino acid selenocysteine. In this example, the isotype is GCC indicating that this is a glycine tRNA (in the genetic code glycine is encoded by GGG, GGA, GGT, or GGC; the GCC anticodon matches the GGC codon). Other information in the output shows the predicted tRNA secondary structure in a bracket notation (Fig. 10.5b) as well as a model of its structure (Fig. 10.5c).

tRNA-Scan-SE produces just one false positive per 15 billion nucleotides of random DNA sequence. It achieves high sensitivity and specificity by combining the output of three separate methods of tRNA identification (Lowe and Eddy, 1997). There are three stages. First, it runs two programs that find tRNAs in DNA (or RNA) sequences. One program identifies conserved intragenic promoter sequences found in prototypic tRNAs, and also requires base pairings that occur in tRNA stem-loop “cloverleaf” structures (Fichant and Burks, 1991). The other program searches for signals that occur in eukaryotic RNA polymerase III promoters and terminators (Pavesi *et al.*, 2004). The results of these two programs are merged. In the second stage, tRNA-Scan-SE analyzes the sequences using a covariance model or stochastic context-free grammar (SCFG; Eddy and Durbin, 1994). A covariance model or SCFG is a probabilistic model of RNA secondary structure and sequence consensus, allowing insertions, deletions, and mismatches (Box 10.1). The covariance model includes a training step based on over 1000 previously characterized tRNAs. In the third stage tRNA-Scan-SE performs a secondary structure prediction and identifies the anticodon of the tRNA. tRNAs with introns and tRNA pseudogenes are further identified.

The approach adopted by tRNA-Scan-SE involves the alignment of multiple RNA sequences in order to infer a common structure of each family based on the two interrelated

The tRNA-Scan-SE server is available at <http://lowelab.ucsc.edu/tRNA-Scan-SE/> (WebLink 10.11), <http://selab.janelia.org/tRNA-Scan-SE/> (WebLink 10.12) or <http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms:tRNA-Scan> (WebLink 10.13). You can also visit Todd Lowe's site to download tRNA-Scan-SE and run it locally. The human chromosome 21 tRNA is given in Web Document 10.3 at <http://www.bioinfbook.org/chapter10>. (This 71-base-pair sequence also matches nucleotides 84511 to 84581 of clone AP001670.1.) In Chapter 3 we introduced Dotlet for pairwise alignments. Try using it (<http://myhits.isb-sib.ch/cgi-bin/dotlet>, WebLink 10.14) with the human tRNA as a query against itself, employing a small window size to find the internally matching stem-loop structures.

You can explore the properties of covariance models from Rfam using the CMCompare webserver (Eggenhofer *et al.*, 2013).



**FIGURE 10.5** Identification of tRNAs using the tRNAscan-SE server. 71 base pairs of DNA were input corresponding to a known human chromosome 21 tRNA. (a) Anticodon counts. These indicate that the input sequence includes a single tRNA having an anticodon that pairs with glycine codons GGC. (b) The predicted secondary structure of the tRNA. (c) Graphic of the predicted secondary structure showing the characteristic cloverleaf pattern of tRNAs. Note that the RNA nucleotides (A, G, C, U) are used, while in panel (b) the DNA nucleotides (A, G, C, T) are used. The first nucleotide is indicated (arrow 1), as is the anticodon GCC (arrow 2).

Source: tRNAscan-SE server (<http://lowelab.ucsc.edu/tRNAscan-SE/>). Courtesy of T. Lowe, Lowe Lab.

## BOX 10.1 STOCHASTIC CONTEXT-FREE GRAMMARS, OR COVARIANCE MODELS

Hidden Markov models (HMMs) are probabilistic models that are useful in many areas of bioinformatics to identify features in sequences such as conserved residues that define a particular protein family (Chapter 5), or nucleotide residues that constitute the structure of a gene. Stochastic context-free grammars (SCFG; Sakakibara *et al.*, 1994) or covariance models (Eddy and Durbin, 1994) constitute another class of probabilistic models that account for long-range correlations along a sequence that occur because of base pairing of noncoding RNA sequences that is required to form appropriate secondary structure such as a stem. Eddy and Durbin (1994) introduced a covariance model in which an RNA sequence is described as an ordered tree in which there are states M (including match states, insert states, and delete states), symbol emission probabilities (these are assigned to specific bases according to the 16 possible pairwise nucleotide combinations or the four unpaired nucleotides), and state transition probabilities (scores assigned to changing states such as entering an insert state). They found that the information content in the secondary structure of tRNA molecules is comparable to that of the primary sequences.

SCFGs are comparable to covariance models. The input of a SCFG is a multiple sequence alignment of noncoding RNAs (such as tRNAs; Sakakibara *et al.*, 1994). The SCFG models how to derive the observed sequences based on a set of “production rules.” Production rules and their associated probabilities define a grammar. The advantages of a SCFG are that its parameters are derived from known RNA sequences and structures, and its probabilistic framework yields confidence estimates on its predictions. SCFGs (like HMMs) originate from the field of language processing (speech recognition).

The Rfam covariance models generated by software called Infernal do not provide expect (*E*) values, but they do offer bit scores. These are derived from log-odds ratios of the probability that a sequence matches a covariance model divided by the probability that the sequence was generated by a random model.

properties of primary sequence and secondary structure. Such an approach is motivated by the fact that noncoding RNAs may diverge over time in a way that preserves each molecule’s base-paired structure while conserving only a limited amount of sequence similarity between homologous RNAs.

A distinct approach to determining RNA structures is to estimate the minimum free energy of folding. This thermodynamic approach was pioneered by Zuker and Stiegler (1981). It is implemented in a variety of programs including the Vienna RNA package (Hofacker, 2003; Lorenz *et al.*, 2011) which incorporates several folding algorithms. A sample output using the Vienna RNA webserver, using a chromosome 21 tRNA sequence as input, is shown in **Figure 10.6**.

In sequencing complete genomes it is of interest to identify all the tRNA genes which are often among the largest gene families. In the human genome, there are over 600 tRNA genes. The reason for so many genes is the necessity for large amounts of tRNAs to enable protein synthesis to occur in all cells throughout life. Two major database resources are the Genomic tRNA Database (GtRNAdb; Chan and Lowe, 2009) and TFAM (Tåquist *et al.*, 2007). A summary of the number of tRNA genes in selected organisms is presented in **Table 10.2**.

### Ribosomal RNA

Ribosomal RNA molecules form structural and functional components of ribosomes, the subcellular units responsible for protein synthesis. rRNA constitutes approximately 80–85% of the total RNA in a cell. In eukaryotes, synthesis of rRNA occurs in the nucleolus, a specialized structure within the nucleus. Purified ribosomes include particles that migrate at characteristic sedimentation coefficients upon centrifugation through a gradient (**Table 10.3**). In bacteria these include the 70S ribonucleoprotein particle that is composed of 30S and 50S subunits, further containing three major rRNA forms (16S, 23S, and 5S). In eukaryotes the 80S ribonucleoprotein particle consists of a 40S and 60S ribosomal RNA subunits that are further processed to generate 18S, 28S, and 5.8S subunits.

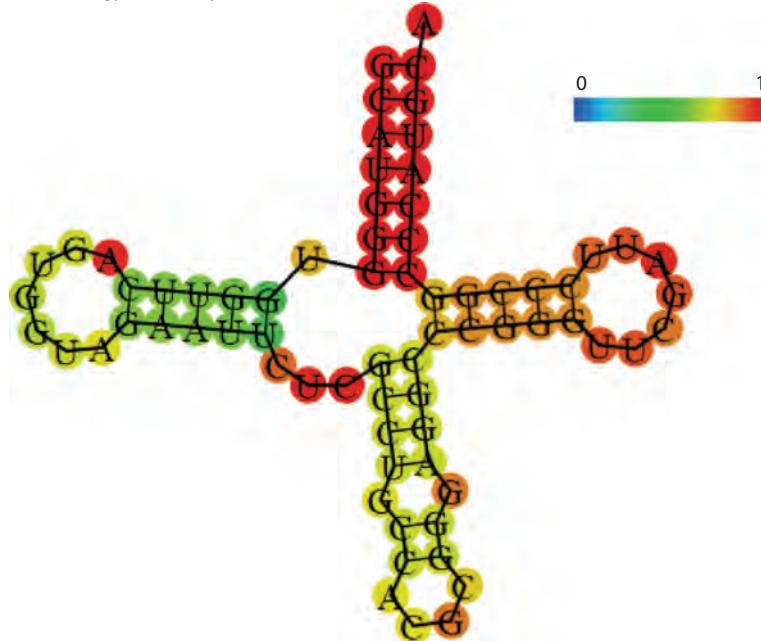
The Vienna RNA package is available at <http://www.tbi.univie.ac.at/RNA/> (WebLink 10.15).

The Genomic tRNA Database (G-tRNA-db) from the laboratory of Todd Lowe is available at <http://lowelab.ucsc.edu/GtRNAdb/> (WebLink 10.16) and contains tRNA identifications of many genomes made using tRNAscan-SE. TFAM is available at <http://tfam.ucmerced.edu/> (WebLink 10.17) and is especially useful for classifying tRNAs having unusual modifications. Another very useful resource is the tRNA database of Mathias Sprinzl and Konstantin Vassilenko available at <http://trnadb.bioinf.uni-leipzig.de/> (WebLink 10.18).

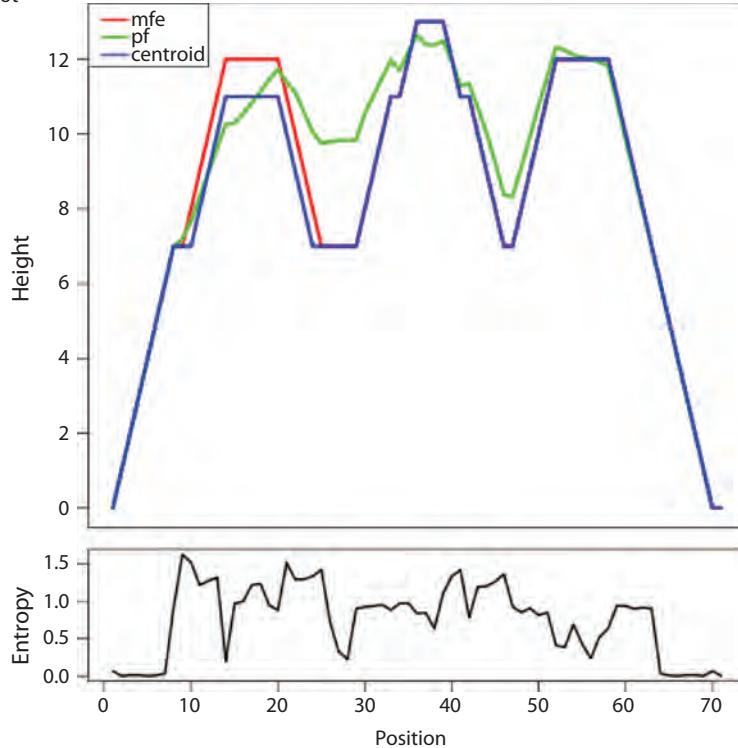
(a) Minimum free energy prediction (colored by base pairing probability): analyzing a tRNA at with Vienna 2.0



(b) Minimum free energy secondary structure



(c) Mountain plot



**FIGURE 10.6** RNA structure prediction based on the minimum free energy of folding. A chromosome 21 sequence known to encode a tRNA (see Fig. 10.5 and Web Document 10.3) was analyzed using the Vienna RNA web server. (a) Optimal predicted structure of RNA using bracket notation. Unpaired nucleotides are represented as dots, while base-paired nucleotides are represented by a pair of matching parentheses. The minimum free energy was  $-35.96$  kcal/mol. (b) Predicted structure of the RNA including stems (double-stranded regions with base pairing) and loops (single-stranded regions). (c) Plot of the minimum free energy (mfe) and a positional entropy measure (pf; y axis) versus the nucleotide position of the input DNA sequence (x axis).

Source: Vienna RNA web server, 2014. Reproduced with permission from I. Hofacker.

**TABLE 10.2 Summary of the number of tRNA genes in selected organisms. The “other” category refers to selenocysteine tRNAs (TCA), suppressor tRNAs (CTA, TTA), or tRNAs with undetermined or unknown isotypes. Additionally, some organisms have tRNAs with introns (e.g., human, 32; *P. falciparum*, 1; *Arabidopsis*, 83).**

Organism	Common name	No. tRNAs decoding the 20 amino acids	No. predicted pseudogenes	Other	Total
<i>Homo sapiens</i>	Human	506	110	9	625
<i>Pan troglodytes</i>	Chimpanzee	456	0	3	459
<i>Mus musculus</i>	Mouse	432	0	3	435
<i>Canis familiaris</i>	Dog (Canfam1)	898	0	8	906
<i>Drosophila melanogaster</i>	Fruit fly	298	4	2	304
<i>Saccharomyces cerevisiae</i>	Baker’s yeast	286	6	3	295
<i>Arabidopsis thaliana</i>	Plant	630	8	1	639
<i>Plasmodium falciparum</i>	Malaria parasite	35	0	0	35
<i>Methanococcus jannaschii</i>	Archaeon	36	0	1	37
<i>Escherichia coli K12</i>	Bacterium	86	1	1	88
<i>Mycobacterium leprae</i>	Bacterium	45	0	0	45

Source:  <http://genome.ucsc.edu>, courtesy of UCSC.

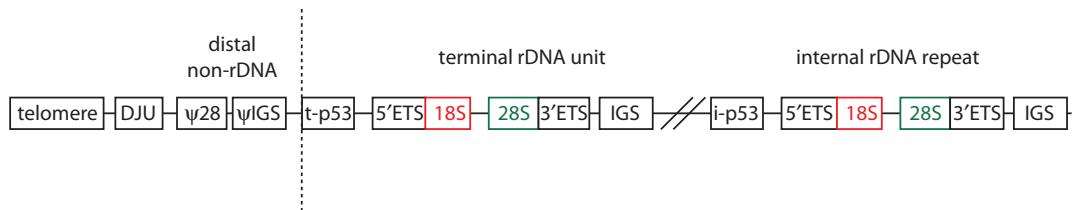
rRNA derives from a multicopy ribosomal DNA (rDNA) gene family. In humans these families are localized to the p arms (i.e., short arms) of the five acrocentric chromosomes (13, 14, 15, 21, and 22; Henderson *et al.*, 1972). The rDNA loci consist of a repeat unit, about 43 kilobases in length, of which 13 kilobases are transcribed and the remainder are nontranscribed spacers (Fig. 10.7). The rDNA genes are identified as *RNR1* (mitochondrially encoded 12S RNA), *RNR2* (mitochondrially encoded 16S RNA), *RNR3*, *RNR4*, and *RNR5*. In the human genome, there are typically ~400 copies of the rDNA repeat. These loci share a high degree of sequence conservation in a process of homogenization that involves both concerted evolution through recombination and gene conversion.

Ribosomal RNA genes have a complex repetitive structure, tremendous conservation across loci on different chromosomes, and enormous variability in the size of

We discussed the structure of the chromosome in Chapter 8, including explanations of mechanisms for conserving sequence identity across chromosomal loci such as concerted evolution and gene conversion. The five acrocentric chromosomes have a centromere positioned near an end of the chromosome rather than in the center.

**TABLE 10.3 Major forms of rRNA in bacteria and eukaryotes. S: sedimentation coefficient; MW: molecular weight. Accession numbers are provided for *E. coli* and human rRNAs. Adapted from NCBI and Dayhoff *et al.* (1972, p. D352).**

Domain	RN	MW	Ribosomal subunits	rRNA species	Function	Accession number	No. base pairs	RFAM accession
Bacteria	70S	$2.6 \times 10^6$	30S (small)	16S	Binding mRNA	M25588.1	1504	RF00177
			50S (large)	23S	Peptide bond formation	M25458.1	542	RF02541
				5S		M24300.1	120	RF00001
Eukaryotes	80S	$4.3 \times 10^6$	40S (small)	18S	Binding mRNA	NR_003286.2	1869	RF01960
			60S (large)	28S	Peptide bond formation	NR_003287.2	5070	RF02543
				5.8S		NR_003285.2	156	RF00002
				5S		NR_023363.1	121	RF00001



**FIGURE 10.7** Structure of a eukaryotic ribosomal DNA repeat unit. A region of an acrocentric chromosome is depicted from the telomere (left side, denoted by an end of the chromosome) to a distal non-rDNA region (containing sequences DJU and two pseudogene regions), then a distal junction (vertical dotted line). To the right (3' end) of this distal junction a terminal rDNA unit is shown; this unit is repeated internally many times, with each unit sharing identical or nearly identical DNA sequence. This region is found in GenBank accession U67616 (8353 base pairs including a variety of repetitive DNA elements and 28S rDNA pseudogenes) and U13369 (42,999 base pairs including transcribed spacers, DNA encoding 18S, 5.8S, and 28S rRNA, and various repetitive DNA elements). IGS: intergenic spacer (also called nontranscribed spacer); ITS: internal transcribed spacer. Adapted from Gonzalez and Sylvester (2001), with permission from Elsevier.

RefSeq accession numbers having the format NR\_123456 consist of noncoding transcripts including structural RNAs and transcribed pseudogenes. Four human RefSeq accessions for rRNA are given in Web Document 10.4 at <http://www.bioinfbook.org/chapter10> along with a 43-kilobase sequence from which they are derived.

RDP is online at <http://rdp.cme.msu.edu/index.jsp> (WebLink 10.19). Release 11 (September 2014) contains over 3 million 16S rRNA sequences. For an example of human genomic DNA sequence that you can use as an input to search Rfam or RDP for rRNA families, see Web Document 10.5 at <http://www.bioinfbook.org/chapter10>.

You can access the ARB project at <http://www.arb-home.de/> (WebLink 10.20). It was developed by Wolfgang Ludwig and colleagues at the Technical University, Munich. ARB refers to arbor (Latin for tree) while silva is Latin for forest. The SILVA website (including a browser) is <http://silva.mpi-bremen.de/> (WebLink 10.21).

the loci between individuals. They are therefore not currently incorporated into the reference human genome at NCBI, UCSC, or Ensembl. To identify human rRNA RefSeq sequences from GenBank, follow the following steps. (1) From the home page of NCBI, navigate to NCBI Nucleotide and restrict the search to human using the search builder. (2) Currently (May 2015) there are nearly 11 million entries. Click “rRNA” under the “molecule types” filter. (3) There are now 30 RefSeq entries corresponding to 5.8S rRNA (e.g., NR\_003285; 156 base pairs), 28S rRNA (NR\_003287; 5070 base pairs), 18S rRNA (NR\_003286; 1869 base pairs), and 45S rRNA (NR\_046235; 13,357 base pairs). For each, the chromosomal assignment is to the acrocentric p-arms.

rDNA sequences are particularly important for phylogenetic analyses across life forms (including the three domains of bacteria, archaea, and eukaryotes). They are uniquely useful because they are closely conserved enough to permit trusted multiple sequence alignments, while they are specific enough to each species that they permit accurate classification. Furthermore, rDNA can be sequenced from environmental samples such as soil or water from which vast numbers of species exist but cannot be cultured (see Chapter 15). Further, rDNA genes are generally not subject to lateral gene transfer (discussed in Chapter 17); that is a form of inheritance in which genes are transmitted horizontally across species rather than being inherited through generations within a species, and it can confound phylogenetic analyses.

Currently there are over 150,000 16S rRNA sequences in GenBank (see Schloss and Handelsman, 2004). There are several major databases of rRNA sequences including the Ribosomal Database Project (RDP; Cole *et al.*, 2007). RDP includes millions of aligned and annotated rRNA sequences, one-third from cultivated bacterial strains and two-thirds from environmental samples. Alignment is performed against a bacterial rRNA alignment model using a stochastic context-free grammar (Box 10.1) as described by Sakakibara *et al.* (1994).

The ARB project is another major resource for RNA studies (Ludwig *et al.*, 2004). It includes a UNIX-based program with a graphical interface that provides software tools to analyze large rRNA databases (such as those imported from the RDP). The related SILVA database includes small subunit (16S, 18S) and large subunit (23S, 28S) rRNA from bacteria, archaea, and eukaryotes. Sequences are downloadable from a browser in the FASTA or other formats.

RNAmer is a hidden Markov model approach to identifying rRNA genes, particularly in newly sequenced genomes (Lagesen *et al.*, 2007). It is useful for searching with large amounts of DNA (e.g., up to 20 million nucleotides) to identify the genomic loci of rRNA genes.

**TABLE 10.4 Examples of human noncoding spliceosomal RNAs.**

Name	Accession	Chromosome	Length (base pairs)
RNU2-1	NR_002716.3	17 q12-q21	188
RNU4-1	NR_003925.1	12q24.31	144
RNU5F-1	NR_002753.5	1p34.1	116
RNU6-2	NR_002752.2	10p13	107

## Small Nuclear RNA

Small nuclear RNA (snRNA) is localized to the nucleus and consists of a family of RNAs that are responsible for functions such as RNA splicing (in which introns are removed from genomic DNA to generate mature mRNA transcripts) and the maintenance of telomeres (chromosome ends). snRNAs associate with proteins to form small nuclear ribonucleoproteins (snRNPs).

You can access RNAmmer at

↗ <http://www.cbs.dtu.dk/services/RNAmmer/> (WebLink 10.22).

The spliceosome is a nuclear complex that includes hundreds of proteins and the five snRNAs U1, U2, U4, U5, and U6 (Valadkhan, 2005). Properties of several of these snRNAs are given in **Table 10.4**. In humans there are about 100 copies of the U4 gene (Bark *et al.*, 1986; Rfam family RF00015) and there are about 1200 copies of the U6 snRNA, including many pseudogenes (nonfunctional genes; Rfam family RF00026). Pseudogenes of protein-coding genes are relatively straightforward to detect because the interruption of an open reading frame can be recognized (see Chapter 8). The identification of nonfunctional, noncoding RNAs presents a far greater challenge because there are no landmarks such as open reading frames, and functional noncoding RNAs are routinely found to have divergent sequences.

## Small Nucleolar RNA

In eukaryotes, ribosome biogenesis occurs in the nucleolus. This process is facilitated by small nucleolar RNAs (snoRNAs), a group of noncoding RNAs that process and modify rRNA and small nuclear spliceosomal RNAs. The two main classes of snoRNAs are C/D box RNAs, which methylate rRNA on a 2'-O-ribose position, and H/ACA box RNAs, which convert uridine to pseudouridine in rRNA. **Table 10.5** presents several online databases that list snoRNAs.

Computational approaches have facilitated the discovery of snoRNAs. For example, after the genome of the yeast *Saccharomyces cerevisiae* was completely sequenced (see Chapter 18), snoRNAs remained challenging to identify. Lowe and Eddy (1999) used a covariance model to identify 22 snoRNAs whose function in methylating rRNA they subsequently confirmed.

77 snoRNAs have currently been annotated in *S. cerevisiae* (*Saccharomyces* Genome Database, ↗ <http://www.yeastgenome.org>, WebLink 10.23).

## MicroRNA

MicroRNAs (miRNAs) are noncoding RNA molecules of approximately 22 nucleotides that have been identified in animals and plants. Since their discovery in the 1990s there

**TABLE 10.5 Small nucleolar RNA (snoRNA) resources.**

Database	Focus	URL
Plant snoRNA database	Arabidopsis snoRNAs	<a href="http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home">http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home</a>
Yeast snoRNA database	H/ACA and C/D box snoRNAs	<a href="http://people.biochem.umass.edu/fournierlab/snornadb/main.php">http://people.biochem.umass.edu/fournierlab/snornadb/main.php</a>
SnoRNABase	human H/ACA and C/D box snoRNAs	<a href="https://www-snorna.biotoul.fr/">https://www-snorna.biotoul.fr/</a>

has been tremendous interest because of their potential functional roles in regulating gene expression (Pasquinelli, 2012). The earliest members of this family to be identified were the *lin-4* and *let-7* gene products of the worm *Caenorhabditis elegans* (Pasquinelli and Ruvkun, 2002). Those genes were identified through positional cloning in a forward genetics strategy: a worm mutant having a defective cell lineage was identified, and a mutation in the *lin-4* RNA was shown to account for the phenotype (Lee *et al.*, 1993). Subsequently, many other miRNA candidates have been identified by complementary DNA (cDNA) cloning of size-selected RNA samples. More recently, next-generation sequencing has been applied in the form of miRNA-seq (for examples of analysis tools see Cho *et al.*, 2013; Humphreys and Suter, 2013). The major function of microRNAs appears to be the downregulation of protein function by inhibiting the translation of protein from mRNA or by promoting the degradation of mRNA.

We can examine a typical microRNA by visiting miRBase, a repository of miRNA data (Kozomara and Griffiths-Jones, 2011). It is possible to browse by organism and find a group of microRNAs assigned to human chromosome 21. Currently, these include hsa-let-7c, hsa-mir-99a, hsa-mir-125b-2, hsa-mir-155, and hsa-mir-802. The entry for let-7c includes the predicted stem-loop structure, the results of next-generation sequencing (including the number of reads per million, in this case across >70 experiments), the genomic coordinates on chromosome 21, a description of neighboring microRNAs (e.g., hsa-mir-99a is less than 10 kilobases away), and database links (e.g., to the European Molecular Biology Laboratory, Rfam, and the Human Genome Organization official nomenclature).

MiRBase also provides links to predicted targets of each microRNA. These targets are RNA transcripts that are potentially regulated by a given microRNA (Rajewsky, 2006; Ritchie *et al.*, 2013). The main approaches to predicting targets include:

- Sequence-based approaches matching complementarity between the ~8 nucleotide seed region of an miRNA and the 3' untranslated region of potential targets. Since this potentially includes gaps, mismatches, and G/U base pairing, the number of possible targets may be extremely large.
- Sequence conservation of 3' UTR target sites across species.
- Analysis of miRNA and messenger RNA (mRNA) expression data. In most cases, increased expression of an miRNA is associated with decreased expression of its target.
- Analysis of thermodynamic stability of the microRNA:mRNA duplex.

Predictions in miRBase are linked from six databases: Diana (predicting ~1000 target genes for the case of has-let-7c; Paraskevopoulou *et al.*, 2013); microRNA (predicting ~5400 target genes; Betel *et al.*, 2008); MiRanda (210 predicted targets; Wang, 2008); RNA22 (>15,000 predictions; Miranda *et al.*, 2006); TargetScan (>1000 predictions; Friedman *et al.*, 2009); and Pictar (~600 predicted targets; Krek *et al.*, 2005). These predictions serve as useful guides to potential targets, although most have not been experimentally validated.

It can be challenging to distinguish an authentic microRNA from other classes of noncoding (or coding) RNA. Ambros *et al.* (2003) proposed a series of definitions of microRNAs based on two criteria regarding their expression:

1. microRNAs consist of an RNA transcript of about 22 nucleotides based on hybridization of the transcript to size-fractionated RNA. Typically, this is accomplished by a Northern blot in which total RNA is purified from a sample such as a cell line, electrophoresed on an agarose gel, transferred to a membrane, and probed with a radioactively labeled form of the candidate miRNA. This experiment shows the size of the RNA, its abundance, and whether the probe hybridizes to multiple RNA species in a sample.

miRBase is available at <http://www.mirbase.org/> (WebLink 10.24). Release 21 (June 2014) includes 29,000 entries for precursors including from >200 species. Chromosome 21 microRNAs are available in Web Document 10.6 at <http://www.bioinfbook.org/chapter10>. We can download a GFF3 format file of all human miRNAs (or miRNAs from over 200 other species) from <ftp://mirbase.org/pub/mirbase/CURRENT/genomes/> (WebLink 10.25). This currently includes >24,000 hairpin precursor miRNAs expressing >30,000 mature miRNA products. For target predictions, databases include MiRanda (part of MiRBase); TargetScan at <http://www.targetscan.org> (WebLink 10.26); Pictar at <http://pictar.mdc-berlin.de/> (WebLink 10.27); DIANA at [http://diana.pcbi.upenn.edu/cgi-bin/micro\\_t.cgi](http://diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi) (WebLink 10.28); and RNAHybrid at <http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/> (WebLink 10.29).

2. The ~22 nucleotide candidate should be present in a library of cDNAs that is prepared from size-fractionated RNA.
- Ambros *et al.* (2003) proposed three additional criteria concerning miRNA biogenesis:
3. The miRNA should have a precursor structure (typically 60–80 nucleotides in animals) that potentially folds into a stem (or hairpin) with the ~22 nucleotide mature miRNA located in one arm of the hairpin. Such a structure is predicted by RNA-folding programs such as mfold (Mathews *et al.*, 1999).
4. Both the ~22 nucleotide miRNA sequence and its predicted fold-back precursor secondary structure must be phylogenetically conserved.
5. Dicer is a protein that functions as a ribonuclease and is involved in processing small noncoding RNAs. There should be increased precursor accumulation in organisms having reduced Dicer function.

Ideally, a putative miRNA meets all five of these criteria, although in practice a subset (such as 1 and 4) may be sufficient.

We describe cDNA libraries in “Analysis of Gene Expression in cDNA Libraries”.

A RefSeq accession for the human Dicer protein is NP\_085124.2.

## Short Interfering RNA

In 1998 Andrew Fire, Craig Mello and colleagues reported that double-stranded RNA introduced into the nematode *Caenorhabditis elegans* can suppress the activity of a gene (Fire *et al.*, 1988). This process is called RNA interference (RNAi). They found that gene silencing occurred when they injected annealed, double-stranded RNA, but not either sense or antisense RNA alone. The silencing was specific to each target gene they studied (such as *unc-22*), and depended upon the injection of double-stranded RNA corresponding to exons rather than introns or promoter sequences. Messenger RNA that is targeted by RNAi is degraded prior to translation, with double-stranded RNA targeting homologous mRNAs in a catalytic manner. This process depends on an RNA-inducing silencing complex (RISC) that includes an endonuclease (to cleave mRNA) and the nuclease Dicer that converts large double-stranded RNA precursors to short interfering RNA.

It is now recognized that RNA interference has many functional implications for eukaryotic cells. RNAi can protect plant and animal cells against infection by single-stranded RNA viruses. RNAi further protects cells from the harmful action of endogenous transposons. These are mobile genetic elements that comprise portions of the human and other genomes. The RNAi mechanism also offers an experimental approach to systematically inhibit the function of genes in mammalian systems; we consider this approach in Chapter 14 (functional genomics).

Andrew Fire and Craig Mello were awarded the 2006 Nobel Prize in Physiology or Medicine “for their discovery of RNA interference - gene silencing by double-stranded RNA.” See [http://nobelprize.org/nobel\\_prizes/medicine/laureates/2006/](http://nobelprize.org/nobel_prizes/medicine/laureates/2006/) (WebLink 10.30).

## Long Noncoding RNA (lncRNA)

In the course of the past decade long noncoding RNAs (lncRNAs) have emerged as the targets class of transcripts in mammalian genomes, with proposed roles in silencing or activating target genes (Lee, 2012; Kornienko *et al.*, 2013). *Xist* and *Air*, described above, are prominent examples. The ENCODE project consortium characterized lncRNAs as part of the GENCODE effort (Djebali *et al.*, 2012), and Derrien *et al.* (2012) defined them according to their location relative to protein-coding genes:

- Antisense RNAs include transcripts that overlap an exon on its opposite strand.
- Large intergenic noncoding RNAs (lncRNAs) have lengths >200 base pairs.
- Sense overlapping transcripts consist of a coding gene within an intron on the same strand.
- Sense intronic transcripts are localized within introns (but do not intersect exons).
- Processed transcripts do not contain an open reading frame (ORF).

A human lncRNA catalog described by Cabili *et al.* (2011) is available at [http://www.broadinstitute.org/genome\\_bio/human\\_lincrnas/](http://www.broadinstitute.org/genome_bio/human_lincrnas/) (WebLink 10.31). As of February 2015 it includes over 14,000 lncRNAs. Derrien *et al.* (2012) note that their ENCODE project catalog has just 39% overlap with that of Cabili *et al.*, leading to the question of how to evaluate error rates in noncoding RNA annotation.

The ENCODE project hosts a human gene annotation catalog at <http://www.gencodegenes.org/> (WebLink 10.32). Version 18 (2013) includes >13,000 lncRNAs. Another resource is IncRNome (Bhartiya *et al.*, 2013), a database of human lncRNAs (<http://genome.igib.res.in/IncRNome/>, (WebLink 10.33)).

The ENCODE findings from Derrien *et al.* (2012) included the following:

- The majority of human lncRNA transcripts are intergenic (and therefore correspond to lncRNAs).
- Most lncRNAs lack coding potential (as expected).
- 98% of lncRNAs are spliced, and 42% of them have two exons. Introns tend to be longer than those of protein-coding genes (median 2.3 versus 1.6 kilobases).
- lncRNAs are under modest purifying selection (more than for neutrally evolving ancestral repeats, but less than for protein-coding genes).
- lncRNA genes have transcription start site histone profiles that are similar to those of protein-coding genes. Their expression tends to be highly cell-type specific.

## Other Noncoding RNA

Noncoding RNAs have been named according to size (e.g., lncRNA, microRNA), cellular localization (rRNA, snRNA, snoRNA), function (mRNA, tRNA), or position and orientation (antisense RNA) (Guenzl and Barlow, 2012). PIWI-interacting RNAs (piRNAs) represent another class of noncoding RNAs, in this case named by their interacting partners. These noncoding RNAs mediate silencing of genes encoding PIWI proteins (Luteijn and Ketting, 2013). Studies in *Drosophila* remarkably show that a protein-coding gene can be converted into a piRNA-producing locus which transcribes piRNAs which silence that gene. This effect may be maintained over generations, and could involve altered chromatin structure and/or recruitment of protein machinery needed for piRNA processing.

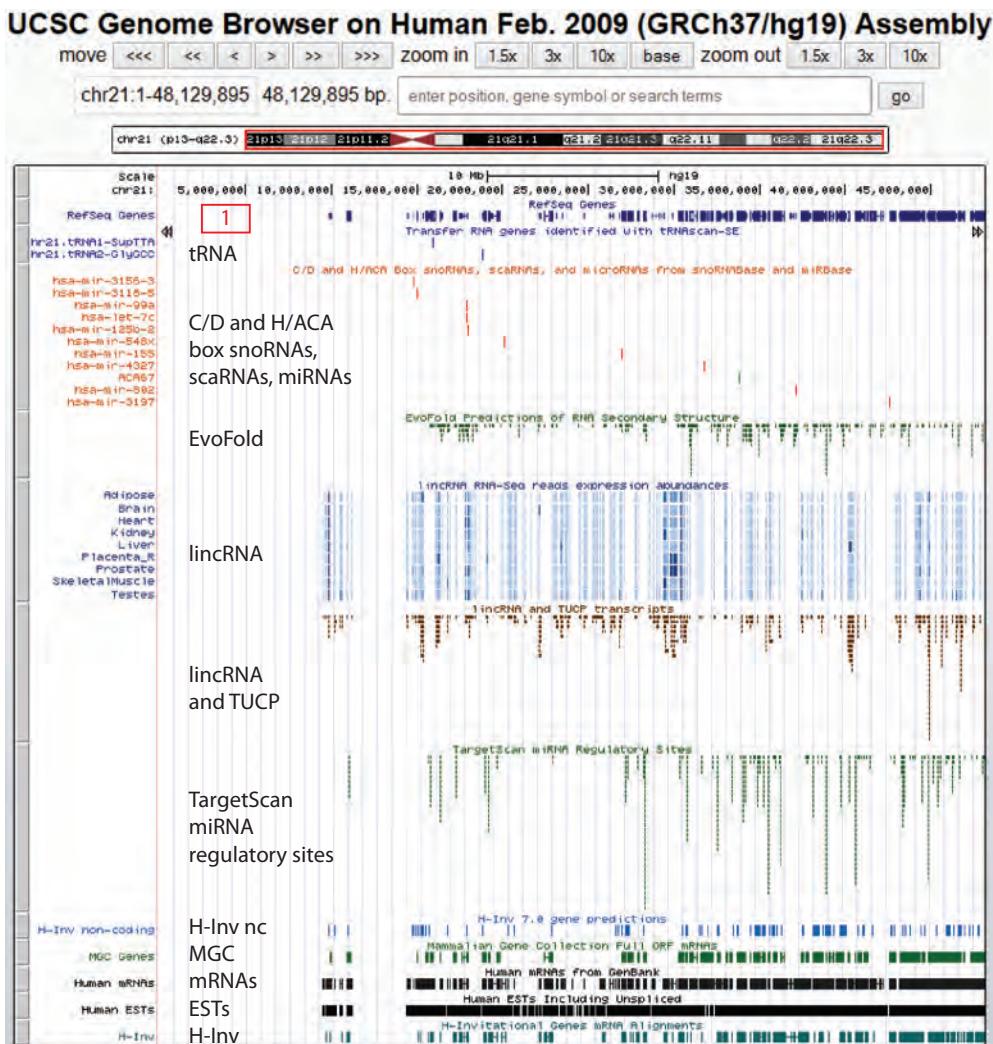
## Noncoding RNAs in the UCSC Genome and Table Browser

As the human genome and other vertebrate genomes continue to be sequenced and analyzed in increasing depth, the Ensembl and UCSC Genome Browsers have emerged as essential tools for visualizing genomic data (introduced in Chapter 2). For noncoding RNAs we can view human chromosome 21 and display a series of user-selected annotation tracks. The following tracks are visible at the resolution of the entire chromosome 21 (48 million base pairs; **Fig. 10.8**):

- A tRNAscan-SE track shows the two chromosome 21 tRNAs we discussed in “Transfer RNA” above.
- A track from miRBase and snoRNABase shows: (1) precursor forms of microRNAs (pre-miRNAs); (2) C/D box small nucleolar RNAs (C/D box snoRNAs); (3) H/ACA box snoRNAs; and (4) small Cajal body-specific RNAs (scaRNAs) (Lestrade and Weber, 2006). There are 11 such noncoding RNA genes on this chromosome, including 10 pre-miRNAs and one snoRNA (ACA67).
- Evofold (Pedersen *et al.*, 2006) shows RNA secondary structure predictions based on phylogenetic stochastic context-free grammars.
- A lncRNA track shows RNA-seq data from 22 tissues encompassing over 450 annotated lncRNAs (Cabili *et al.*, 2011). The expression abundances can be displayed, or the presence of lncRNA and transcripts of uncertain coding potential (TUCP).
- The TargetScanS miRNA Regulatory Sites track shows putative miRNA binding sites in the 3' untranslated region of RefSeq genes. These sites are predicted by the TargetScanS program (Lewis *et al.*, 2005).
- Noncoding gene predictions from the H-Invitational Gene Database are shown.
- We also display data on expressed coding genes (discussed below).

As we have seen, the UCSC Table Browser is complementary to the Genome Browser. Suppose we want to know the exact number of EvoFold entries that occur on

The miRNA data are from  
 ↗ <http://www.mirbase.org/> (WebLink 10.34). The  
 snoRNABase is available online  
 at ↗ <http://www-snorna.biotoul.fr/> (WebLink 10.35).



**FIGURE 10.8** Viewing the genomic landscape of noncoding RNAs on human chromosome 21. To recreate this display, visit <http://genome.ucsc.edu> and select Genome Browser. Set the clade to vertebrate, the genome to human, the assembly to GRCh37/hg19 (different assemblies have varying annotation tracks available), the position to chr21, and click submit. All of chromosome 21 is displayed (about 48.1 million base pairs). You can specify which annotation tracks to select using a series of pull-down menus; under the Genes and Gene Prediction Tracks category select RefSeq genes, tRNA Genes, EvoFold, sno/miRNA, lincRNA, H-Inv, and Mammalian Gene Collection (MGC) Genes. Additional tracks from the mRNA and EST Tracks group are human mRNAs, human expressed sequence tags (ESTs), and H-Inv. The TargetScan miRNA sites track is from the Regulation group.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

chromosome 21. From the full view of chromosome 21 click the “Table Browser” link on the top bar. Choose the table of interest (e.g., EvoFold) and click “summary/statistics” to see that there are 306 EvoFold items. For sno/miRNAs there are 11 items; by clicking “get output” you can obtain their genomic coordinate positions. How many RefSeq genes overlap these EvoFold regions on chromosome 21? To answer this, simply click the “intersection” button and, from the Genes and Gene Prediction Tracks group, select RefSeq genes; at the time of writing (February 2015), using the GRCh37/hg19 genome build the answer is 133.

## INTRODUCTION TO MESSENGER RNA

Gene expression occurs when DNA is transcribed into RNA. Each eukaryotic cell contains a nucleus with some 2000–60,000 protein-coding genes, depending on the organism. However, at any given time the cell expresses only a subset of those genes as mRNA transcripts. The set of genes expressed by a genome is sometimes called the transcriptome. A conventional view that emerged since the “one gene, one enzyme” hypothesis of Beadle and Tatum, which continued through the establishment of the central dogma of molecular biology, is that genes correspond to discrete loci and are transcribed to mRNA in order to make a protein product. We now appreciate that the situation is vastly more complex because of the existence of noncoding RNAs, the interruption of genes by introns, the existence of alternative splicing to generate different mRNA transcripts that often produce distinct protein products, and the pervasive transcription of most nucleotide bases in the genome. We discuss these topics below. Furthermore, while humans, chimpanzees, and mice all have an extremely closely related set of about 20,000 protein-coding genes per genome, what distinguishes the phenotypic expression of each species may depend on the intricacies of the regulation of gene expression. Gene expression is typically regulated in several basic ways:

- by region (e.g., brain versus kidney);
- in development (e.g., fetal versus adult tissue);
- in dynamic response to environmental signals (e.g., immediate–early response genes that are activated by a drug);
- in disease states; or
- by gene activity (e.g., mutant versus wildtype bacterium).

The comparison of gene expression profiles has been used to address a variety of biological questions in an assortment of organisms. For viruses and bacteria, studies have focused both on viral and bacterial gene expression and also on the host response to pathogenic invasion. Among eukaryotes, gene expression studies, and in particular microarrays, have been employed to address fundamental questions such as the identification of genes activated during the cell cycle or throughout development. In multicellular animals cell-specific gene expression has been investigated, and the effect of disease on gene expression has been studied in rodents and primates (including humans). In recent years, gene expression profiling has become especially important in the annotation of genomic DNA sequences. When the genome of an organism is sequenced, one of the most fundamental issues is to determine which genes it encodes (Chapters 15–19). Large-scale sequencing of expressed genes, such as those isolated from cDNA libraries (described in “Analysis of Gene Expression in cDNA Libraries” below), is invaluable in helping to identify gene sequences in genomic DNA.

### mRNA: Subject of Gene Expression Studies

Consider what is measured in gene expression studies. In most cases, total RNA is isolated from cells of interest. (Sometimes, polyadenylated RNA is isolated.) This RNA is readily purified using chaotropic agents that separate RNA from DNA, protein, lipids, and other cellular components. In this way steady-state RNA transcript levels can be measured, reflecting the activity of a gene. Gene expression is regulated in a set of complex steps that can be divided into four categories: transcription, RNA processing, mRNA export, and RNA surveillance (Maniatis and Reed, 2002) (**Fig. 10.9**).

1. *Transcription.* Genomic DNA is transcribed into RNA in a set of highly regulated steps.

In the 1970s, sequence analysis of genomic DNA revealed that portions of the DNA (called exons) match the contiguous open reading frame of the corresponding mRNA,

For the range of gene content in eukaryotic genomes see Chapters 18–20.

In addition to viewing gene expression as a dynamically regulated process, we can also view proteins and metabolites as regulated dynamically in every cell. See Chapter 12.

while other regions of genomic DNA (introns) represent intervening sequences that are not present in mature mRNA.

2. **RNA Processing.** Introns are excised from pre-mRNA by the spliceosome, a complex of five stable small nuclear RNAs (snRNAs) and over 70 proteins. Alternative splicing occurs when the spliceosome selectively includes or excludes particular exons (Modrek and Lee, 2002). Pre-mRNA also is capped at the 5' end. (Eukaryotic mRNAs contain an inverted guanosine called a cap.) Mature mRNA has the unique property among nucleic acids of having a long string of adenine residues attached to its 3' end. This tract is typically preceded by the polyadenylation signal AAUAAA or AUUAAA, located 10–35 nucleotides upstream. Polyadenylation of mRNA is extremely convenient from an experimental point of view, because an oligonucleotide (consisting of a string of thymidine residues attached to a solid support, oligo(dT) resin) can be used to rapidly isolate mRNA to a high degree of purity. In some cases gene expression studies employ total RNA, while many others employ mRNA.
3. **RNA Export.** After splicing occurs, RNA is exported from the nucleus to the cytoplasm where translation occurs. Note that the phrase “gene expression profiling” is commonly used to describe the measurement of steady-state cytoplasmic RNA transcript levels, but may not be precisely correct. “RNA transcript level profiling” is what is performed, and the actual expression of genes is an activity that is not directly measured.
4. **RNA Surveillance.** An extensive RNA surveillance process allows eukaryotic cells to scan pre-mRNA and mRNA molecules for nonsense mutations (inappropriate stop codons) or frame-shift mutations (Maquat, 2002). This nonsense-mediated decay mechanism is important in the maintenance of functional mRNA molecules. Additional mechanisms control the half-life of mRNAs, targeting them for degradation and therefore regulating their availability.

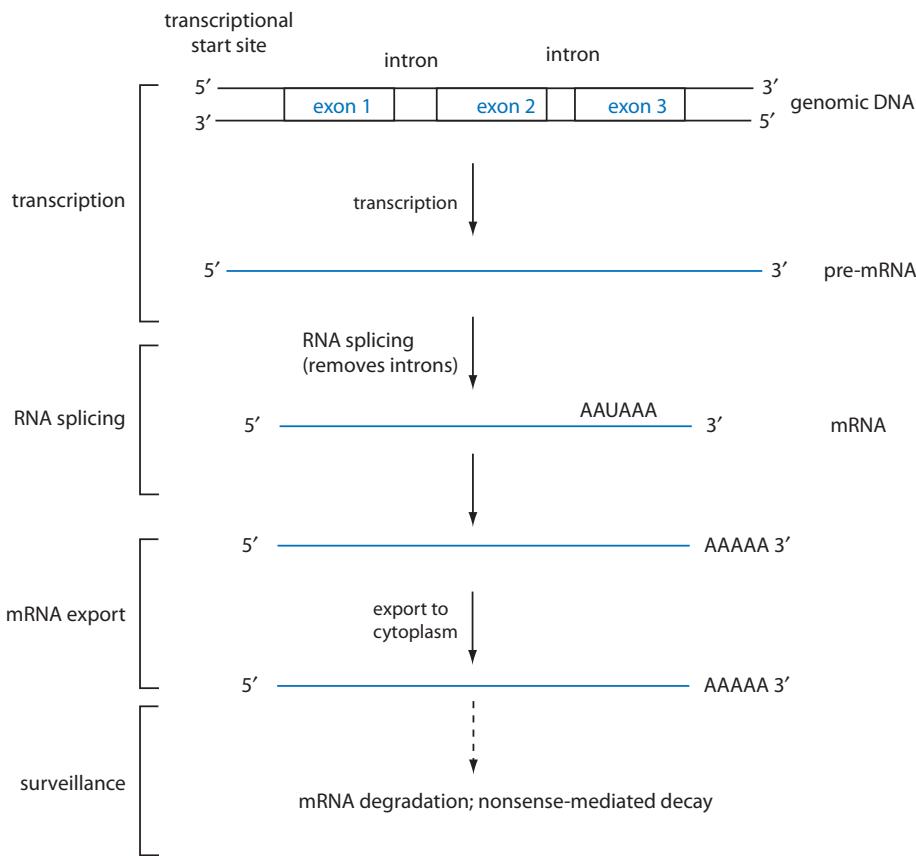
Let us consider human  $\alpha 2$  globin mRNA as an example of a transcript. The function of the globin genes has been characterized in detail. The two alpha globin genes, *HBA1* and *HBA2*, encode proteins sharing 100% identical amino acid sequence. However, the *HBA2* mRNA transcript and protein are expressed at levels about three-fold higher than the mRNA and protein products of the *HBA1* gene (Liebhaber *et al.*, 1986). We can view the *HBA2* gene using the UCSC Genome Browser. There are three exons as shown in **Figure 10.10a**. The exons are interrupted by introns; to view this, try performing a BLASTN search of the RefSeq DNA sequence for *HBA2* against the corresponding region of genomic DNA (**Fig. 10.10b**). (We also see this by using HBA2 protein as a BLAT query.) Matches to the exons are evident as pairwise alignments, but the introns (absent from the mature mRNA and therefore not part of the NM\_000517 entry) do not match the genomic reference. By zooming in on the first exon of *HBA2*, we can see that it is transcribed along the top strand (from left to right beginning at the short arm of chromosome 16; **Fig. 10.10c**). The RefSeq track shows the portion of the first exon that is at the 5' untranslated end (left side), then the coding portion of the exon is displayed with a thickened bar (**Fig. 10.10c**). Here the third or bottom reading frame begins with a methionine and continues to correspond to the protein sequence encoded by *HBA2*.

The *HBA2* gene locus includes portions corresponding to the coding region as well as 5' and 3' untranslated regions (UTRs). These UTRs typically contain regulatory signals such as a ribosome binding site near the start methionine and a polyadenylation signal (often AATAAA) in the 3' UTR. In the case of alpha 2 globin, the 3' UTR contains three cytosine-rich (C-rich) segments that are critical for maintaining the stability of the mRNA (Waggoner and Liebhaber, 2003). Specific RNA-binding proteins interact with the 3' UTR which adopts a stem-loop structure. Mutations that disrupt this region can lead to destabilization of  $\alpha$ -globin mRNA, causing a form of the disease  $\alpha$ -thalassemia (Chapter 21).

Richard J. Roberts and Phillip A. Sharp received the 1993 Nobel Prize in Physiology or Medicine for their discovery of “split genes.” See <http://www.nobel.se/medicine/laureates/1993/> (WebLink 10.36).

A molecule in *Drosophila* provides an extraordinary example of alternative splicing. The Down syndrome cell adhesion molecule (DSCAM) gene product potentially exists in more than 38,000 distinct isoforms (Schmucker *et al.*, 2000; Celotto and Graveley, 2001). The gene contains 95 exons (e.g., NM\_001273835.1). Functionally, multiple DSCAM proteins may confer specificity to neuronal connections in *Drosophila*.

Some alignments of RNA-derived sequences and the corresponding genomic DNA have mismatches. These discrepancies may reflect polymorphisms or errors associated with either the sequencing of genomic DNA or cDNA. One way to decide which sequence has an error is to look for consistency. If multiple, independently derived genomic DNA clones or expressed sequence tags have the identical nucleotide sequence in a region of interest, you can be more confident that sequence is correct. See Chapter 15 for a further discussion.



**FIGURE 10.9** RNA processing of eukaryotic genes. Genomic DNA contains exons (corresponding to the mature mRNA) and introns (intervening sequences). After DNA is transcribed, pre-mRNA is capped at the 5' end and splicing removes the introns. A polyadenylation signal (most commonly AAUAAA) is recognized, the RNA is cleaved by an endonuclease about 10–35 nucleotides downstream, and a polyA polymerase adds a polyA tail (typically 100–300 residues in length). Polyadenylated mRNA is exported to the cytoplasm where it is translated on ribosomes into protein. An RNA surveillance system involving nonsense-mediated decay degrades aberrant mRNAs; a dashed line indicates that RNA surveillance machinery can also degrade pre-mRNAs.

### Low- and High-Throughput Technologies to Study mRNAs

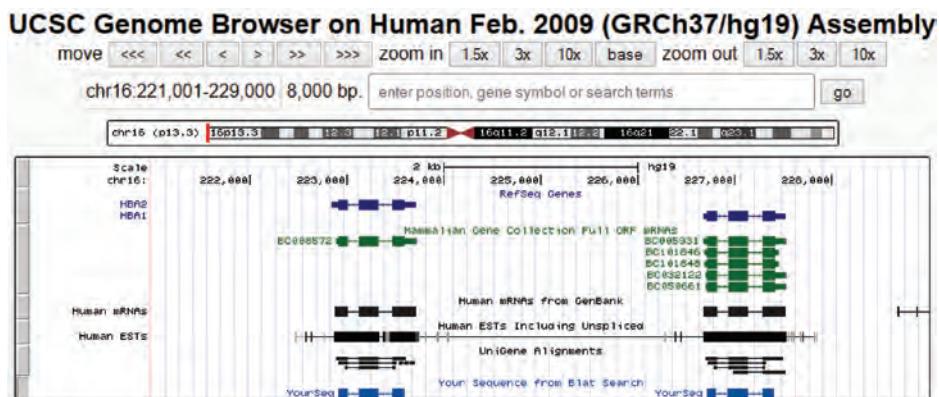
We focus on three techniques for the study of mRNAs: complementary DNA (cDNA) libraries, microarrays using the Affymetrix platform, and RNA-seq.

In recent decades, gene expression has been studied using a variety of techniques such as Northern blotting, the polymerase chain reaction with reverse transcription (RT-PCR), and the RNase protection assay. Each of these approaches is used to study one transcript at a time. In Northern blotting, RNA is isolated, electrophoresed on an agarose gel, and probed with a radioactive or fluorescently labeled cDNA derived from an individual gene. Quantitative RT-PCR (qRT-PCR) employs specific oligonucleotide primers to exponentially amplify specific transcripts as cDNA products. RNase protection is used to quantitate the amount of an RNA transcript in a sample based upon the ability of a specific *in vitro* transcribed cDNA to protect an endogenous transcript from degradation with a ribonuclease. Gene expression may be compared in several experimental conditions (such as normal versus diseased tissue, cell lines with or without drug treatment). The signals may be quantitated. Signals are also normalized to a number of housekeeping genes or other controls that are expected to remain unchanged in their expression levels.

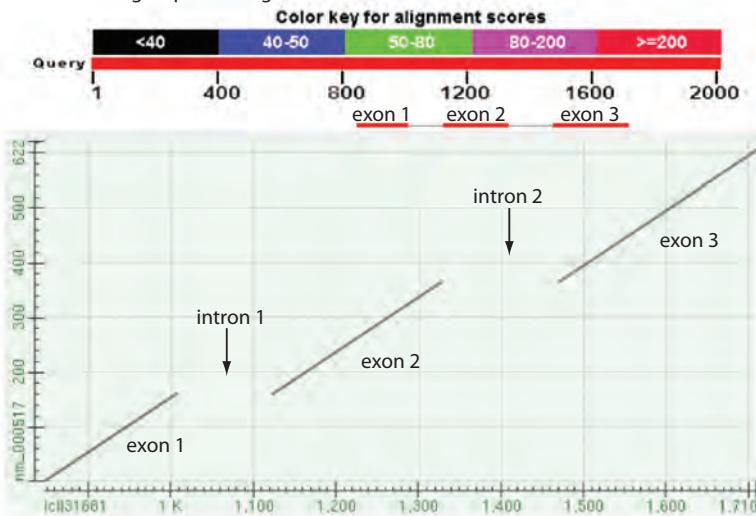
From one point of view, low-throughput techniques such as Northern blotting and quantitative RT-PCR are laborious and do not give as much information as high-throughput

The enzyme reverse transcriptase, present in retroviruses, is an RNA-dependent DNA polymerase (i.e., it converts RNA to DNA).

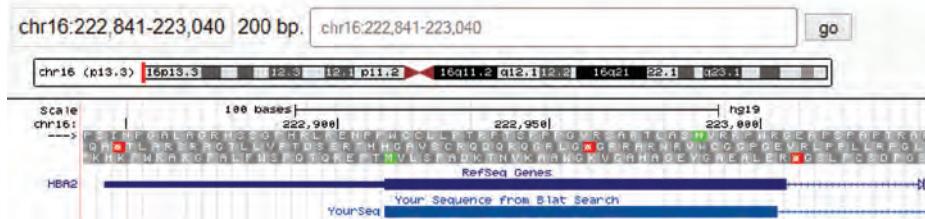
(a) *HBA1* and *HBA2* gene region on human chromosome 11



(b) MegaBLAST of *HBA2* coding sequence to genomic DNA reveals introns



(c) Exon 1 of *HBA2* (including nucleotides encoding protein amino terminus)



**FIGURE 10.10** The HBA2 mRNA in the context of the corresponding genomic DNA. (a) The adjacent *HBA2* and *HBA1* genes of human chromosome 16 are displayed using the UCSC Genome Browser. The ideogram (chromosomal diagram) shows that the region zoomed in on is at the telomeric region of the p arm of chromosome 16. A window size of 8000 base pairs is displayed. The RefSeq Genes track shows the three exons of *HBA2*. Additional tracks show human mRNAs and expressed sequence tags (ESTs), and the result of a BLAT search using alpha globin protein (NP\_000508.1) as a query. (b) To compare the mRNA sequence of HBA2 to its corresponding genomic DNA sequence, megaBLAST was performed at NCBI (Chapter 5). The sequences were NM\_000517.4 and a chromosome 16 genomic contig (RefSeq accession NT\_010393.16, nucleotides 162,000–164,000) that spans the *HBA2* gene locus. Note that the graphic shows three red bars (exons) separated by gaps corresponding to introns. Similarly, the dotplot shows 100% identity for the three exons and gaps at the site of introns. (c) A detailed view of the first exon of *HBA2*, including the beginning of the protein-coding sequence (the start methionine is highlighted in green at the bottom of the three reading frames and matches the start of the protein sequence from BLAT).

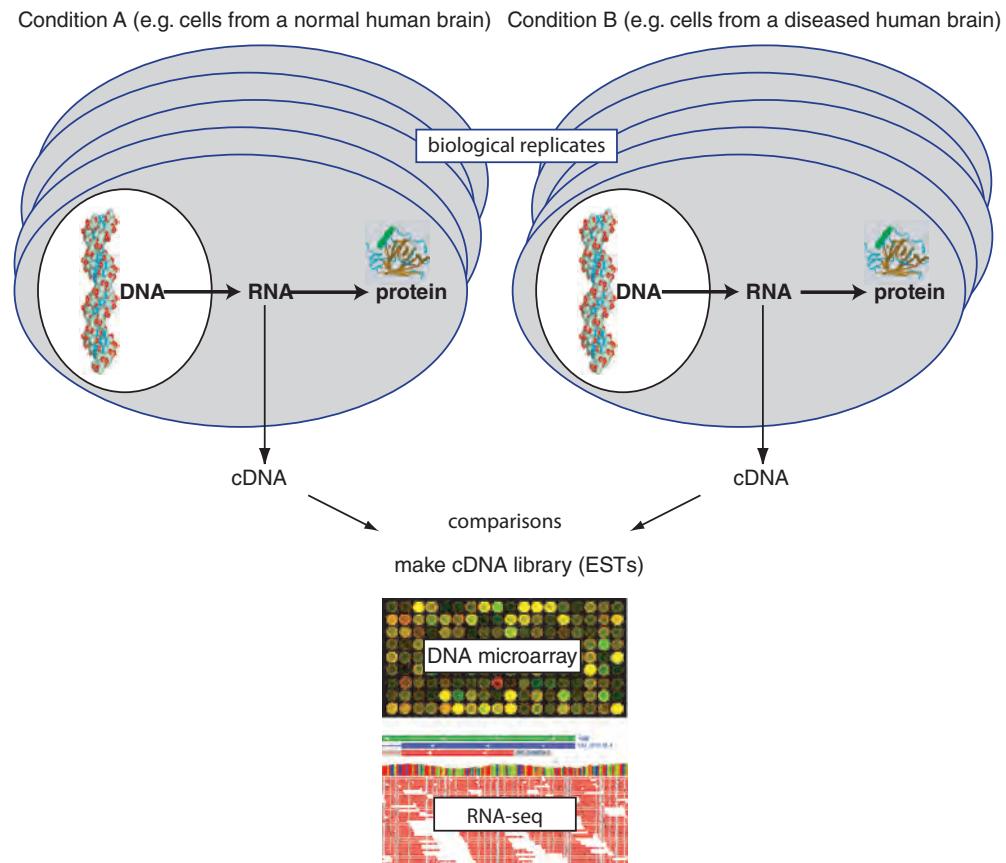
Sources: (a), (c) <http://genome.ucsc.edu>, courtesy of UCSC and Regents of the University of California. (b) MegaBLAST, NCBI.

technologies. From another point of view they remain “gold standards” and provide trusted confirmation of results from high-throughput experiments.

In contrast to these approaches, several high-throughput techniques have emerged that allow a broad survey of gene expression. A global approach to gene expression offers two important advantages over the study of the expression of individual genes:

- A broad survey may identify individual genes that are expressed in a dramatic fashion in some biological state. For example, global comparisons of gene expression in assorted human tissues can reveal which individual transcripts are expressed in a region-specific manner.
- High-throughput analyses of gene expression can reveal patterns or signatures of gene expression that occur in biological samples. This may include the coordinate expression of genes whose protein products are functionally related. We examine tools for the analysis of gene expression data (such as clustering trees) in Chapter 11.

Several high-throughput approaches to gene expression are displayed in **Figure 10.11**. In each case, total RNA or mRNA is isolated from biological samples obtained from two



**FIGURE 10.11** Gene expression can be measured with a variety of high-throughput technologies. In most cases, two biological samples are compared such as a cell line with or without drug treatment, cells with or without viral infection, or aged versus neonatal rat brain. RNA can be converted to cDNA, allowing broader surveys of transcription in a cell. In this chapter and the next we examine several approaches to gene expression. cDNA libraries can be constructed, generating expressed sequence tags (ESTs). These can be electronically compared in UniGene. Complex cDNA mixtures can be labeled with a fluorescent molecule and hybridized on DNA microarrays, which contain cDNA or oligonucleotide fragments corresponding to thousands of genes. High-throughput sequencing of cDNA libraries (RNA-seq) represents a powerful approach to comparing transcripts in two samples.

(or more) conditions that are compared. The RNA is typically converted to cDNA using reverse transcriptase. Complementary DNA is inherently less susceptible to proteolytic or chemical degradation than RNA, and cDNA can readily be cloned, propagated, and sequenced. cDNA may be packaged into libraries and studied (see the following section). RNA may also be labeled to measure transcripts on microarrays, and cDNA libraries may be sequenced by next-generation sequencing (RNA-seq).

There are many other techniques to study gene expression, such as serial analysis of gene expression (see Web Document 10.7).

### Analysis of Gene Expression in cDNA Libraries

The sequencing of cDNA libraries allows the location and quantity of RNA transcripts to be measured. RNA is expressed from some region at some location. One purifies RNA from sources such as the roots, stem, and leaves of a plant at various developmental stages, or human brain at autopsy from those diagnosed with a disease or controls. RNA is converted to cDNA and packaged into a library. The cDNA inserts, called expressed sequence tags (ESTs), may then be sequenced. dbEST is a database of ESTs at NCBI; currently it contains tens of millions of ESTs from a variety of organisms. The UniGene database further partitions these ESTs into nonredundant clusters that generally correspond to expressed genes (Sayers *et al.*, 2012).

Web Document 10.8 provides a diagram showing how a cDNA library is constructed. A summary of the number of ESTs in GenBank is available at [http://www.ncbi.nlm.nih.gov/genbank/dbest/dbest\\_summary/](http://www.ncbi.nlm.nih.gov/genbank/dbest/dbest_summary/) (WebLink 10.37). UniGene is accessed via <http://www.ncbi.nlm.nih.gov/unigene/> (WebLink 10.38).

Each cluster has some number of sequences associated with it, from one (*singletons*) to almost 50,000 (Table 10.6). Of the 130,000 clusters in Table 10.6, half are singletons suggesting that these may be genes expressed so rarely that they have only been observed once. These singletons may represent portions of the genome that are transcribed without representing functional genes (see “The Pervasive Nature of Transcription” below).

**TABLE 10.6 Cluster sizes for human entries in UniGene (Build 236, *Homo sapiens*).  
GAPDH: glyceraldehyde-3-phosphate dehydrogenase.**

Cluster size	Number of clusters	Example(s) of genes in cluster
1	64,371	
2	12,760	
3–4	10,859	Transcribed locus, strongly similar to NP_032247.1 hemoglobin subunit epsilon-Y2 [ <i>Mus musculus</i> ]
5–8	10,637	Transcribed locus, strongly similar to NP_001077424.1 hemoglobin alpha, adult chain 2 [ <i>Mus musculus</i> ]
9–16	7,177	Hemoglobin, theta 1; hemoglobin, beta pseudogene 1
17–32	4,815	Hemoglobin, mu; neuroglobin
33–64	4,557	Hemoglobin, zeta
65–128	4,117	Hemoglobin, delta
129–256	3,889	Hemoglobin, epsilon 1; cytoglobin
257–512	3,858	
513–1024	1,982	
1,025–2,048	729	Hemoglobin, alpha 1; myoglobin; hemoglobin, gamma A
2,049–4,096	224	Hemoglobin, beta; hemoglobin, gamma G
4,097–8,192	56	Hemoglobin, alpha 2
8,193–16,384	20	Albumin, GAPDH; ubiquitin C; tubulin, alpha 1b; ferritin, light polypeptide
16,385–32,768	4	Actin, beta; myelin basic protein; Eukaryotic translation elongation factor 1 alpha 1; Uncharacterized LOC100507412
32,769–65,536	1	EEF1A1

Source: UniGene, NCBI.

**TABLE 10.7 Ten largest cluster sizes in UniGene for human entries. Values are rounded to the nearest 1000.**

UniGene Identifier	Cluster size	Gene symbol	Gene name
Hs.586423	48,000	EEF1A1	Eukaryotic translation elongation factor 1 alpha 1
Hs.535192	27,000	EEF1A1	Eukaryotic translation elongation factor 1 alpha 1
Hs.520640	26,000	ACTB	Actin, beta
Hs.551713	21,000	MBP	Myelin basic protein
Hs.426704	20,000	LOC100507412	Uncharacterized LOC100507412
Hs.520348	16,000	UBC	Ubiquitin C
Hs.418167	16,000	ALB	Albumin
Hs.524390	16,000	TUBA1B	Tubulin, alpha 1b
Hs.510635	16,000	IGHG1	Immunoglobulin heavy constant gamma 1 (G1m marker)
Hs.544577	15,000	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
Hs.180414	15,000	HSPA8	Heat shock 70kDa protein 8
Hs.370247	15,000	APLP2	Amyloid beta (A4) precursor-like protein 2

Source: UniGene, NCBI.

Other genes (such as actin and tubulin) are expressed at very high levels. Even some EST clusters that do not correspond to known, annotated genes are highly represented. The largest cluster sizes represented in UniGene are described in **Table 10.7** for humans and in **Table 10.8** for nonhuman organisms.

Web Document 10.9 provides more detailed background on cDNA libraries and UniGene, including tools to extract libraries in UniGene. As an example, we can use the Digital Differential Display (DDD) tool at UniGene to compare cDNA libraries from different regions of the body or different conditions. **Figure 10.12** shows an example in

**TABLE 10.8 Ten largest cluster sizes in UniGene for nonhuman entries. Cluster size is the number of sequences rounded to the nearest thousand.**

UniGene Identifier	Species	Cluster size	Gene Name
Cin.19067	<i>Ciona intestinalis</i> (vase tunicate; yellow sea squirt)	48,000	Clone:citb001e24, full insert sequence
Bfl.2106	<i>Branchiostoma floridae</i> (Florida lancelet)	31,000	Transcribed locus, strongly similar to NP_007768.1 NADH dehydrogenase subunit 1
Bt.107724	<i>Bos taurus</i> (cow)	22,000	Chymotrypsinogen B1-like
At.46639	<i>Arabidopsis thaliana</i> (thale cress)	16,000	Ribulose bisphosphate carboxylase small chain 1A
Cin.30513	<i>Ciona intestinalis</i>	15,000	ATP-binding cassette sub-family D member 2-like
Dr.31797	<i>Danio rerio</i> (zebrafish)	13,000	Eukaryotic translation elongation factor 1 alpha 1, like 1
Dr.75552	<i>Danio rerio</i>	13,000	Actin, alpha, cardiac muscle 1b
Rn.202968	<i>Rattus norvegicus</i> (Norway rat)	13,000	Albumin
Ta.11048	<i>Triticum aestivum</i> (bread wheat)	13,000	Small subunit
Ssc.6512	<i>Sus scrofa</i> (pig)	12,000	Mitochondrial ATPase 6 mRNA, L transcript, partial sequence

Source: UniGene, NCBI (using the search query 11700:65536[sequence count] NOT txid9606[organism]).

**TABLE 10.9 Fisher's 2×2 exact test used to test null hypothesis that a given gene (gene 1) is not differentially regulated in two pools. Adapted from Claverie (1999).**

	Gene 1	All other genes	Total
Pool A (e.g., brain)	Number of sequences assigned to gene 1 ( $g_{1A}$ )	Number of sequences in this pool NOT gene 1 ( $N_A - g_{1A}$ )	$N_A$
Pool B (e.g., pancreas)	Number of sequences assigned to gene 1 ( $g_{1B}$ )	Number of sequences in this pool NOT gene 1 ( $N_B - g_{1B}$ )	$N_B$
Total	$c = g_{1A} + g_{1B}$	$C = (N_A - g_{1A}) + (N_B - g_{1B})$	

### Digital Differential Display (DDD)

DDD is a tool for comparing EST profiles in order to identify genes with significantly different expression levels ([More about DDD](#)).

Species: *Homo sapiens* (human)      [Start Over](#)  
 Pool A: brain      13 libraries, 70610 ESTs      [Edit Pool](#)  
 Pool B: heart      8 libraries, 29064 ESTs      [Edit Pool](#)  
[New Pool](#)

### Differential Display Results

The following genes (UniGene entries) display statistically significant differences in EST counts by the Fisher Exact Test.

A brain	B heart		UniGene Entry
0.0000	0.0360	<a href="#">Hs.418167</a>	Albumin (ALB)
0.0195	0.0000	<a href="#">Hs.551713</a>	Myelin basic protein (MBP)
0.0024	0.0213	<a href="#">Hs.298280</a>	ATP synthase, H <sup>+</sup> transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle (ATP5A1)
0.0001	0.0182	<a href="#">Hs.435369</a>	Four and a half LIM domains 1 (FHL1)
0.0125	0.0002	<a href="#">Hs.654422</a>	Tubulin, alpha 1a (TUBA1A)
0.0100	0.0210	<a href="#">Hs.586423</a>	Eukaryotic translation elongation factor 1 alpha 1 (EEF1A1)
0.0000	0.0087	<a href="#">Hs.657271</a>	LIM domain binding 3 (LDB3)
0.0080	0.0001	<a href="#">Hs.1787</a>	Proteolipid protein 1 (PLP1)
0.0078	0.0000	<a href="#">Hs.514227</a>	Glial fibrillary acidic protein (GFAP)

**FIGURE 10.12** Digital differential display (DDD) is used to compare the content of expressed sequence tags (ESTs) in cDNA libraries from UniGene. Thousands of libraries have been generated by isolating RNA from a tissue source (such as pancreas, heart, or brain), synthesizing cDNA, packaging the cDNA in a cDNA library, and sequencing up to thousands of cDNA clones (ESTs) from each library. The clones in each library (or in pools of libraries) may be compared using DDD. This site is accessed from the NCBI UniGene site; choose *Homo sapiens*, then select “Library digital differential display.” At this site, click boxes corresponding to any library (or set of libraries) then select a second library (or second set of libraries) for comparison. Result of an electronic comparison of cDNA libraries using the DDD tool of UniGene. The results are displayed as a list of genes (with UniGene accession numbers) for genes that are preferentially expressed in one or the other pool of libraries. Here, transcripts expressed preferentially in heart are shown (e.g., albumin and a cardiac muscle ATP synthase). Other transcripts (e.g., those encoding the glial proteins myelin basic protein and glial fibrillary acidic protein) are more highly represented in brain-derived libraries.

Source: UniGene, NCBI.

In UniGene, click statistics, *Homo sapiens*, then "library browser" to see the range of clones that are sequenced in typical libraries. Currently (February 2015), there are ~8700 human cDNA libraries in UniGene having at least 1000 sequences and ~8000 smaller libraries.

which 13 libraries from human brain are shown to selectively expressed transcripts such as myelin basic protein and glial fibrillary acidic protein (glial proteins not expected to be expressed in heart), while heart preferentially expresses albumin. A probability value is associated with each transcript using a Fisher's exact test (Box 10.2). This test does not require the number of clones being compared to be identical.

Some pitfalls involved in analyzing expression data in cDNA libraries include the following:

- Investigators choose which libraries to construct, and there is likely to be bias toward familiar tissues (such as brain and liver) and bias away from more unusual tissues. The rat nose contains over two dozen secretory glands, almost all of which are of unknown function, but for most of these glands cDNA libraries have never been constructed.
- The depth to which a library is sequenced affects its ability to represent the contents of the original cell or tissue. A cDNA library is expected to contain a frequency of clones that faithfully reflects the abundance of transcripts in the source material. By sequencing only 500 clones, it is unlikely that many low-abundance transcripts will be represented when the contents of the entire library are analyzed. In practice, cDNA libraries are sequenced to varying depths. An advantage of RNA-seq is its potentially extreme depth of coverage.
- Another source of bias is in library construction. Many libraries are normalized, a process in which abundant transcripts become relatively underrepresented while rare transcripts are represented more frequently. The goal in normalizing a library is to minimize the redundant sequencing of highly expressed genes and to therefore discover rare transcripts (Bonaldo *et al.*, 1996). It would be inappropriate to compare normalized and nonnormalized libraries directly using a tool such as UniGene's differential display. For RNA-seq and microarrays, it is also essential to prepare RNA under identical conditions between the sources being compared.
- ESTs are often sequenced on one strand only, rather than thoroughly sequencing both top and bottom strands. There is therefore a substantially higher error rate than is found in finished sequence. (We discussed sequencing error rates in Chapter 9.)
- Chimeric sequences can contaminate cDNA libraries. For example, two unrelated inserts are occasionally cloned into a vector during library construction.

## BOX 10.2 FISHER'S EXACT TEST

Fisher's exact test is used to test the null hypothesis that the number of sequences for any given gene in the two pools (e.g., insulin in pancreas versus brain) is the same in either pool (**Table 10.9**).

The *p* value for a Fisher's exact test is given by

$$P = \frac{N_A! N_B! c! C!}{(N_A + N_B)! g1_B! (N_A - g1_A)! (N_B - g1_B)!}. \quad (10.1)$$

The null hypothesis (that gene 1 is not differentially regulated between brain and muscle) is rejected when the probability value *p* is less than 0.05/*G*, where 0.05 is the nominal threshold for declaring significance and *G* is the number of UniGene clusters analyzed (*G* is therefore a conservative Bonferroni correction; see Chapter 11).

While the NCBI website employs Fisher's exact test, other statistical approaches to cDNA library comparison have been described. In particular, Stekel *et al.* (2000) developed a log-likelihood procedure to assess the probability that gene expression differences observed in a comparison of two or even multiple cDNA libraries are due to genuine transcriptional differences and not sampling errors.

We perform a Fisher's exact test in R using the `fisher.test` function in the `stats` R package in Chapter 11.

## Full-Length cDNA Projects

While UniGene is an example of a database that incorporates information on ESTs and protein-coding genes, it is also of interest to catalog, characterize, and make available collections of cDNAs. There are two main forms of cDNAs: those having full-length protein-coding sequences (typically including some portions of the 5' and 3' untranslated regions), and expression clones in which the protein-coding portion of the cDNA is cloned into a vector that permits protein expression in the appropriate cell type (Temple *et al.* 2006). There are many important resources for obtaining cloned, high-quality, full-length cDNAs. We will next introduce four of the many available cDNA resources.

1. The Functional Annotation of the Mouse (FANTOM) project provides functional annotation of the mammalian transcriptome (Maeda *et al.*, 2006). Currently, over 100,000 full-length mouse cDNAs have been annotated including both coding and nonprotein-coding transcripts. These have been mapped to genomic loci using BLAT, BLASTN, and other search tools. The annotation categories included artifacts (such as contaminants from other species or chimeric mRNAs) and coding sequences (complete, 5'- or 3' truncated, 5' or 3' untranslated regions only, immature, with or without insertion/deletion errors, stop codons, coding for selenoproteins, or mitochondrial transcripts). Upon analyzing transcription start and stop sites, the 5' and 3' boundaries of over 180,000 transcripts were identified (Carninci *et al.*, 2005). This study lead to the identification of over 5000 previously unidentified mouse proteins. Another astonishing conclusion of the FANTOM project is that antisense transcription, in which clustered cDNA sequences on one strand at least partially match to the opposite strand, occurs for 72% of all genome-mapped transcriptional units (Katayama *et al.*, 2005).
2. The H-Invitational Database provides an integrative annotation of human genes including gene structures, alternative splicing isoforms, coding as well as noncoding RNAs, single-nucleotide polymorphisms (Chapter 8), and comparative results with the mouse (Takeda *et al.*, 2013). A total of 21,037 human gene candidates were analyzed, corresponding to 41,118 full-length cDNAs. We saw H-Invitational UCSC tracks in **Figure 10.8**.
3. The Mammalian Gene Collection (MGC) is an NIH project that originally aimed to gather full-length cDNA clones for all human and mouse genes, but has subsequently expanded to include rat, cow, frog, and zebrafish (MGC Project Team *et al.*, 2009). Its site can be searched by BLAST, and its database contents can be viewed at UCSC (**Fig. 10.8**). MGC clones are distributed through the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) consortium.
4. Another important cDNA resource is the Kazusa mammalian cDNA set, called “KIAA” genes (Nagase *et al.*, 2006). This project focuses on characterizing full-length cDNAs that encode particularly large genes. Clones are described and distributed through the HUGE database (Kikuno *et al.*, 2004).

## BodyMap 2 and GTEx: Measuring Gene Expression Across the Body

Two prominent projects have emerged for the study of tissue-specific gene expression across the human body. The Genotype-Tissue Expression (GTEx) project focuses on gene expression and regulation including information on genetic variation (allowing the measurement of expression quantitative trait loci or e-QTLs; see “e-QTLs” below). Human Body Map 2.0, a project lead by Illumina, Inc., measures gene expression across 16 tissues using RNA-seq. You can view the expression of any human gene of interest starting from its NCBI Gene entry. For example, for the *HBB* Gene entry navigate to the genome browser on that page. Select the option to “Configure Tracks,” select the Expression category, and choose from dozens of BodyMap 2 display options.

You can access the FANTOM project at <http://fantom.gsc.riken.go.jp/> (WebLink 10.39).

The H-Invitational database is available at <http://www.h-invitational.jp/> (WebLink 10.40). Hosted by the Japan Biological Information Research Center (JBIRC), this site features a highly informative genome browser.

The Mammalian Gene Collection (MGC) website is <http://mgc.nci.nih.gov/> (WebLink 10.41). It includes ~30,000 human clones (corresponding to ~17,500 nonredundant genes) as of February 2015. The IMAGE consortium website (<http://www.imageconsortium.org/>, WebLink 10.42) can be queried for clones from a number of species.

The HUGE database is at <http://www.kazusa.or.jp/huge/> (WebLink 10.43). You can see examples of Northern blots at the HUGE database (e.g., <http://www.kazusa.or.jp/huge/gfimage/northern/html/KIAA0012.html>, WebLink 10.44).

You can access the GTEx portal at <http://www.gtexportal.org/> (WebLink 10.45). Its data are released through dbGaP at NCBI (e.g., accession phs000424.v4.p1). Human Body Map 2.0 data can be accessed via <http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/> (WebLink 10.46) or as Series GSE30611 from the Gene Expression Omnibus at NCBI (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611>, WebLink 10.47).

## MICROARRAYS AND RNA-SEQ: GENOME-WIDE MEASUREMENT OF GENE EXPRESSION

By 2000 DNA microarrays had emerged as a powerful technique to measure mRNA transcripts (gene expression). They have been used more than any other technique to assess differences in mRNA abundance in different biological samples. The use of microarrays has increased rapidly since the pioneering work of Patrick Brown and colleagues at Stanford University, Jeffrey Trent and colleagues at the NIH, and others (De Risi *et al.*, 1996).

By 2010 RNA-seq had arrived as an even more powerful technique (McGettigan, 2013; Mutz *et al.*, 2013). Many consider it likely that RNA-seq may soon supplant microarrays as the method of choice for gene expression profiling. Since they have the common main purpose of identifying steady-state mRNA levels in biological samples, we introduce them together here and explain how to analyze array and RNA-seq data in Chapter 11.

A microarray is a solid support (such as a glass microscope slide) on which DNA of known sequence is deposited in a regular grid-like array. The DNA may take the form of cDNA or oligonucleotides, although other materials (such as genomic DNA clones; Chapter 8) may also be deposited. Typically, several nanograms of DNA are immobilized on the surface of an array. RNA is extracted from biological sources of interest, such as cell lines with or without drug treatment, tissues from wildtype or mutant organisms, or samples studied across a time course. The RNA (or mRNA) is often converted to cDNA (or cRNA in the case of the popular Affymetrix platform), labeled with fluorescence, and hybridized to the array. During this hybridization, cDNAs or cRNAs derived from RNA molecules in the biological starting material can hybridize selectively to their corresponding nucleic acids on the microarray surface. Following washing of the microarray, image analysis and data analysis are performed to quantitate the signals that are detected. Through this process, microarray technology allows the simultaneous measurement of the expression levels of thousands of RNA transcripts (genes) represented on the array.

RNA-seq applies the same next-generation sequencing technology we described in Chapter 9 to the sequencing of cDNAs derived from RNA sources of interest. The resulting reads are mapped to the transcriptome (e.g., to a set of all exons). The two main measurements are the amount of each transcript present in a sample and the quantitation of exons to infer alternative splicing events. Compared to microarrays, RNA-seq offers additional features (Ozsolak and Milos, 2011; Costa *et al.*, 2013):

- While microarrays depend on prior selection of transcripts for which RNA levels are measured, RNA-seq makes no prior assumptions about which RNA species are present in the samples being assayed. New transcripts may therefore be identified.
- RNA-seq offers a far broader dynamic range, spanning six orders of magnitude for polyadenylated mRNA (and four orders of magnitude for nonpolyadenylated RNAs; Djebali *et al.*, 2012).
- RNA-seq experiments are scalable: deeper sequencing coverage yields improved power to detect variants (such as mutations) and to detect transcript expressed at low levels.
- RNA-seq is used to map transcription start sites (TSSs) at base pair resolution.
- RNA-seq is useful to define patterns of alternative splicing, including previously unannotated fusions between expressed transcripts, and quantitative assessment of the differences in expression of particular exons.
- RNA-seq may be adapted to characterizing small noncoding RNAs (by size-selecting small RNA for analysis).
- It may be possible to isolate RNA from hosts and pathogens together to simultaneously identify RNA changes in both (called “dual RNA-seq” by Westermann *et al.*, 2012).

Both microarrays and RNA-seq also come with particular disadvantages. Both are expensive enough that many biologists analyze too few replicates (see Stage 1 below). Both are subject to artifacts that destroy the usefulness of the experiment. For example, if RNA is extracted from a set of control samples on a Monday and a set of experimental samples on a Tuesday, then any observed differences could be due to condition (experimental versus control) or date. This situation is called a perfect confound. Another concern is that the final product of gene expression is protein (for coding genes), and changes in RNA levels may have little biological significance; we discuss the relatively poor correlation between RNA and protein levels in “The Relationship between DNA, mRNA and Protein Levels” below.

An overview of the procedures used in microarray and RNA-seq experiments is depicted by **Figure 10.13**, divided into five stages. We consider each of these stages in the following sections.

### Stage 1: Experimental Design for Microarrays and RNA-seq

In the first stage, total RNA or mRNA is isolated from samples. Notably, experiments have been performed for organisms as diverse as viruses, bacteria, fungi, and humans. The amount of starting material that is required is typically several hundreds of milligrams (wet weight) or several flasks of cells. For many available microarray or RNA-seq experiments, about 1–3 µg of total RNA is required. With the amplification of RNA or cDNA products it is possible to use substantially less starting material, and even single-cell RNA profiling is feasible. However, the amplified population may not faithfully represent the original RNA population.

The experimental design of a microarray experiment includes biological replicates, technical replicates, and array design (Churchill, 2002). Different sources of variation are associated with each of these three areas.

1. First, the biological samples are selected for comparison, such as a cell line with or without drug treatment. If multiple biological samples are used, these are called “biological replicates.” When experimental subjects are selected for treatment, it is appropriate to assign them to groups randomly. It is critical to have adequate sample size of biological replicates, such as  $n = 3$  to 5 samples for the experimental group and a similar number for the control group. Many experiments are performed with just one biological replicate. Hansen *et al.* (2011) stress the need for biological replicates in RNA-seq studies, noting that the biological variability in datasets generated with microarrays is comparable to that in RNA-seq. For either technology, results with too few replicates are liable to be nonreproducible and not generalizable to the conditions being studied.
2. Second, RNA is extracted and labeled (typically as complementary DNA) with a fluorescent tag (for microarrays). When two RNA extractions are obtained from a biological sample and analyzed, these are called “technical replicates.” Some researchers perform multiple RNA isolations from a single sample (e.g., three independent RNA isolations from a single cultured cell line or a single rat heart). These are not considered biological replicates because they do not capture the variability in expression levels between independent samples.
3. A third aspect of microarray experimental design is the arrangement of array elements on a slide. Ideally, the array elements are arranged in a randomized order on the slide. In some cases, array elements are spotted in duplicate (see **Fig. 10.14**). Artifacts can occur based on the arrangement of elements on an array or because a microarray surface is not washed (or dried) evenly.

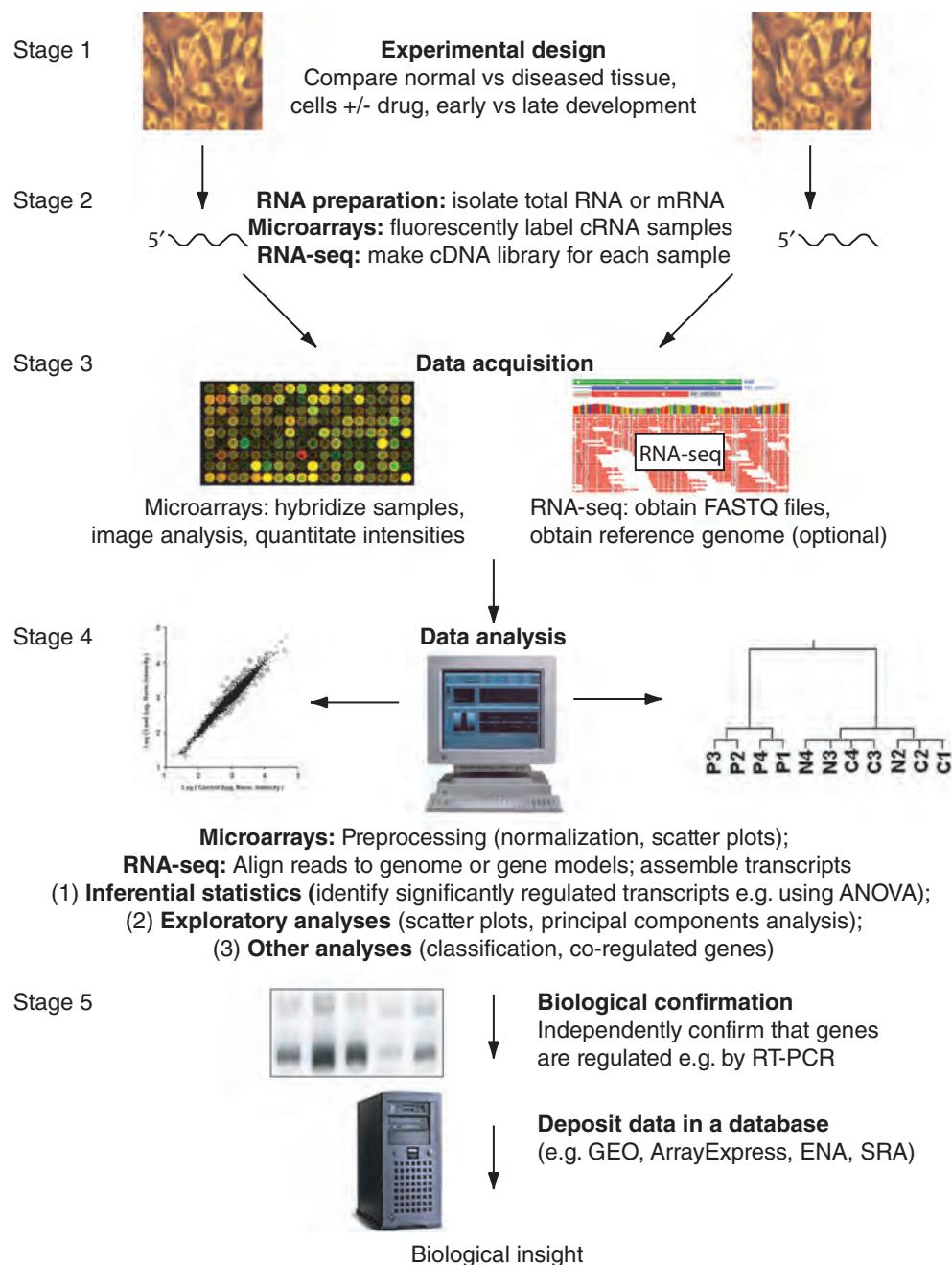
For microarrays from Affymetrix, RNA is converted to cDNA and transcribed to make biotin-labeled complementary RNA (cRNA).

We discuss experimental design further in Chapter 11.

Web Document 10.10 describes competitive microarray hybridization.

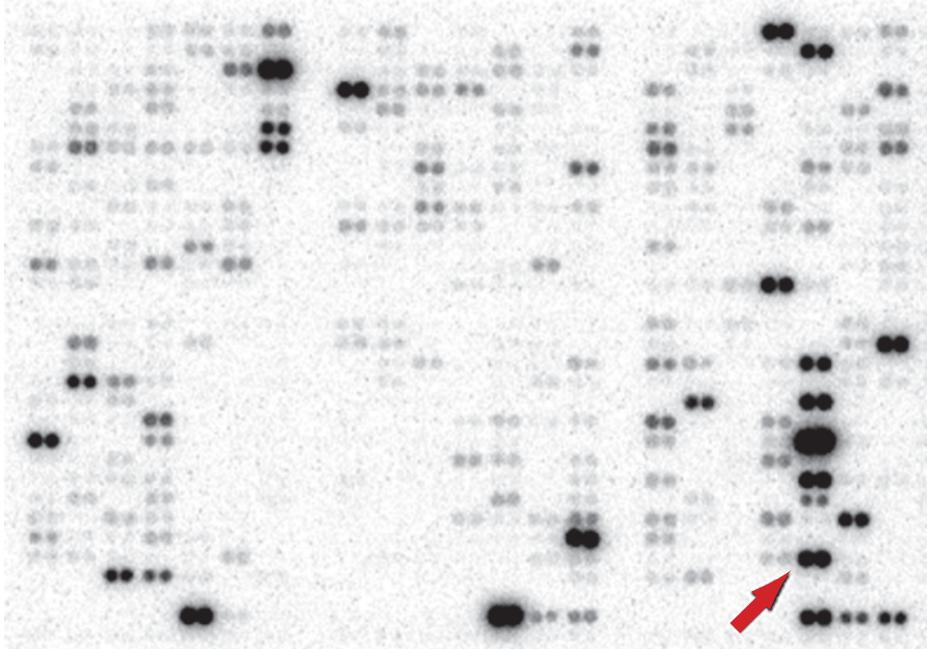
### Stage 2: RNA Preparation and Probe Preparation

RNA can be readily purified from cells or tissues using reagents such as TRIzol (Invitrogen). For some microarray applications, further purification of RNA to mRNA (poly(A)<sup>+</sup>

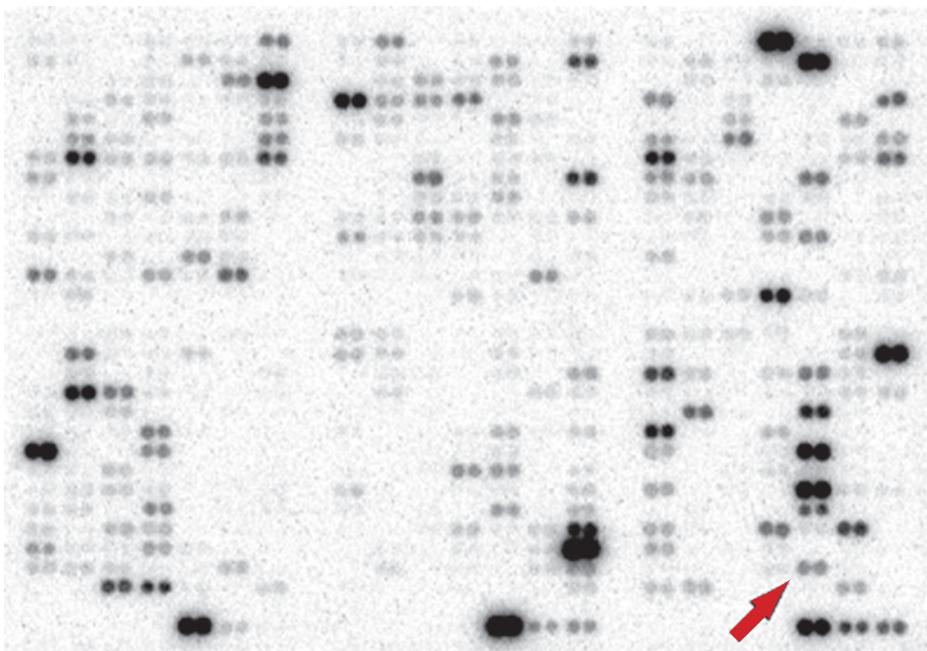


**FIGURE 10.13** Overview of the process of generating high-throughput gene expression data using microarrays or RNA-seq. In stage 1, biological samples are selected for a comparison of gene expression. In stage 2, for microarrays RNA is isolated, converted and labeled, often with fluorescent dyes. For RNA-seq, RNA is converted to cDNA and packaged into a library. In stage 3 data are acquired: samples are hybridized to microarrays, which are solid supports containing complementary DNA or oligonucleotides corresponding to known genes. For RNA-seq, next-generation sequencing is performed. In stage 4, data analysis is performed. Microarray expression data are analyzed to identify differentially regulated genes (e.g., using ANOVA (see Chapter 11) and scatter plots; stage 4, at left) or clustering of genes and/or samples (right). For RNA-seq raw reads are mapped to a reference transcriptome (or genome) and assembled; in some workflows assembly precedes alignment. Read counts are used to infer expression levels of exons and/or transcripts. Based on these findings, independent confirmation of microarray- or RNA-seq-based findings is performed (stage 5). The data (e.g., Affymetrix .cel files or RNA-seq FASTQ and BAM files) are deposited in a database so that data can be shared and further large-scale analyses can be performed.

(a)



(b)



**FIGURE 10.14** Example of a microarray experiment using radioactive probes. While radioactivity is only rarely used today, this figure illustrates the nature of microarrays in which RNA transcripts may be observed at a range of abundances from high (dark spots) to low or absent (corresponding to genes not expressed in this particular body region and/or developmental stage). A total of 588 genes are represented on each array and are spotted in adjacent pairs. The filters were hybridized, washed, and exposed to a phosphorimager screen for 6 hours. The output includes a quantitation (in pixel units) of the signals. (a) Clontech Atlas Neurobiology array probed with cDNA derived from the post-mortem brain of a girl with Rett syndrome; and (b) the profile from a matched control. The arrows point at an RNA transcript ( $\beta$ -crystallin) that is up-regulated in the disease. Note that overall the RNA transcript profiles appear similar in the two brains.

RNA) is necessary. (Purification kits typically remove >95% of ribosomal RNA, and such depletion is essential to detect low-abundance mRNA transcripts or other noncoding RNA transcripts in RNA-seq.) In comparing two samples (e.g., cells with or without a drug), it is essential to purify RNA under closely similar conditions. For example, for cells in culture conditions such as days in culture and percent confluence must be controlled for.

The purity and quality of RNA should also be assessed spectrophotometrically (by measuring the  $a_{260}/a_{280}$  ratio) and by gel electrophoresis. Fluorescent dyes such as RiboGreen (Molecular Probes) can be used to quantitate yields. Purity of RNA may also be confirmed by Northern analysis or PCR. RNA preparations that are contaminated with genomic DNA, rRNA, mitochondrial DNA, carbohydrates, or other macromolecules may be responsible for impure probes that give high backgrounds or other experimental artifacts.

For microarrays, the RNA is converted to cDNA or to complementary RNA, then labeled with fluorescence to permit detection.

In an effort to provide a reference set of RNA transcripts that can serve as a “gold standard,” the External RNA Controls Consortium has been established. This project includes the goals of providing access to clones, protocols, and bioinformatics tools (Baker *et al.*, 2005).

For RNA-seq, the RNA is converted to cDNA and packaged into libraries. For an example of experimental protocols see Nagalakshmi *et al.* (2010).

You can read about the progress of the External RNA Controls Consortium at <http://www.nist.gov> (WebLink 10.48).

Photolithography is a technique with many applications, including the microelectronics industry, in which substances are deposited on a solid support. For microarray technology, oligonucleotides are synthesized *in situ* on a silicon surface by combining standard oligonucleotide synthesis protocols with photolabile nucleotides that permit thousands of specific oligonucleotides to be immobilized to a chip surface.

Many researchers refer to the DNA on a microarray as the probe and the labeled DNA derived from a biological sample as the target. There are therefore opposite definitions of probe and target, and the research community has not reached a consensus. We call the labeled material derived from RNA or mRNA the “probe.” For an image of the density of oligonucleotides on the surface of a chip, see Web Document 10.11.

### Stage 3: Data Acquisition

#### *Hybridization of Labeled Samples to DNA Microarrays*

The immobilized DNA on a microarray sometimes consists of approximately 5 ng of cDNA (length 100–2000 base pairs) arrayed in rows and columns. In other cases, oligonucleotides rather than cDNAs are immobilized (Lipshutz *et al.*, 1999). This has been accomplished by Affymetrix using a modified process of photolithography (Fodor *et al.*, 1991). Depending on the nature of the solid support used to immobilize DNA, the microarray is often called a blot, membrane, chip, or slide. The DNA on a microarray is referred to as “target DNA.” In a typical microarray experiment, the gene expression patterns from two samples are compared. RNA from each sample is labeled with fluorescence or radioactivity to generate a “probe.”

After RNA is converted into cDNA or cRNA labeled with fluorescence, the efficient labeling of probe must be confirmed. This is followed by hybridization of the probe overnight to the filter or slide and washing of the microarray. Image analysis is then performed to obtain a quantitative description of the extent to which each mRNA in the sample is expressed (Duggan *et al.*, 1999). For experiments using radioactive probes (typically using [ $^{33}\text{P}$ ] or [ $^{32}\text{P}$ ] isotopes), image analysis is performed by quantitative phosphorimaging (Fig. 10.14). Image analysis involves aligning the pixels to a grid and manually adjusting the grid to align the spots. Each spot represents the expression level of an individual transcript. The intensity of a spot is presumed to correlate with the amount of mRNA in the sample. However, many artifacts are possible. The spot may not have a uniform shape. An intense signal may “bleed” to a neighboring spot, artifactually lending it added signal intensity. Pixel intensities near background levels may lead to spuriously high ratios. For example, if a control value is 100 units above background levels and an experimental value is 200 units, the experimental condition is up-regulated twofold. However, if the pixel values are 50,100 versus 50,200, then no regulation is described.

For fluorescence-based microarrays, the array is excited by a laser and fluorescence intensities are measured. Data for Cy5 and Cy3 channels may be sequentially obtained and used to obtain gene expression ratios, or a single dye may be used as in the Affymetrix technology.

### Data acquisition for RNA-seq

Libraries are generated for RNA-seq studies, sometimes including barcoded samples, then sequencing is performed as described in Chapter 9.

### Stage 4: Data Analysis

Analysis of microarray data is performed to identify individual genes that have been differentially regulated. It is also used to identify broad patterns of gene expression. In some experiments groups of genes are coregulated, suggesting functional relatedness. Samples (rather than genes) may be analyzed and classified into discrete groups. The analysis of microarray data is described in Chapter 11.

In an effort to standardize microarray data analysis, Alvis Brazma and colleagues (2001) at 17 different institutions proposed a system for storing and sharing microarray data. Minimum Information About a Microarray Experiment (MIAME) provides a framework for researchers to describe information in six areas: the experimental design; the microarray design; the samples (and how they are prepared); the hybridization procedures; the image analysis; and the controls used for normalization. Notably the metadata for RNA-seq experiments are very similar to those for microarray experiments, and MINSEQE (minimum information about a high-throughput nucleotide sequencing experiment) has also been proposed. These efforts are intended to promote research data quality, appropriate annotation, and useful data exchange.

The MIAME project is described at the Microarray Gene Expression Database Group website, which merged with the Functional Genomics Data Society (FGED Society; <http://www.fged.org/>, WebLink 10.49).

### Stage 5: Biological Confirmation

Microarray experiments result in the quantitative measurement of thousands of mRNA transcript values. Data analysis typically reveals that dozens or hundreds of genes are significantly regulated, depending on the particular experimental paradigm and the statistical analysis approach. A list of regulated transcripts may include true positives (those that are authentically regulated) as well as false positives (transcripts reported as significantly regulated even though they were found by chance). It is important to independently confirm the differential regulation of at least some of the most regulated transcripts.

### Microarray and RNA-seq Databases

Raw as well as processed microarray data are routinely deposited in public repositories upon publication. The main public repositories are ArrayExpress and the European Nucleotide Archive (ENA) at the European Bioinformatics Institute and the Gene Expression Omnibus (GEO; Barrett *et al.*, 2013) and Sequence Read Archive (SRA) at NCBI. We describe how to acquire data from these databases in Chapter 11.

ArrayExpress is available at <http://www.ebi.ac.uk/arrayexpress/> (WebLink 10.50), while GEO is at <http://www.ncbi.nlm.nih.gov/geo/> (WebLink 10.51).

While databases of gene expression have been established, it is important to contrast them with DNA databases. A DNA database such as GenBank contains information about the sequence of DNA fragments, ranging in size from small clones to entire chromosomes or entire genomes. The error rate involved in genomic DNA sequencing can be measured (Chapter 9), and independent laboratories can further confirm the quality of DNA sequence data. In general, DNA sequence does not change for an individual organism across time or in different body regions. In contrast, gene expression is context dependent. A database of gene expression contains some quantitative measurement of the expression level of a specified gene. If two laboratories attempt to describe the expression level of beta globin from a cell line, the measurement may vary based on many variables such as: the source of the cell line (e.g., liver or kidney); the cell culture conditions (e.g., cells grown to subconfluent or confluent levels); the cellular environment (e.g., choice of growth media); the age of the cells; the type of RNA that is studied (total RNA versus

mRNA, each with varying amounts of contaminating biomaterials); the measurement technique; and the approach to statistical analysis. While it has been possible to create a project such as RefSeq or VEGA to identify high-quality representative DNA sequences of genes, any similar attempt to describe a standard expression profile for genes must account for many variables related to the context in which transcription occurs.

### Further Analyses

Eventually, it is likely that uniform standards will be adopted for all microarray and RNA-seq experiments (as promoted by the Functional Genomics Data Society). The greatest variables in these studies are likely to be the quality of the RNA isolated by each investigator and the nature of the microarray or short-read alignment that is used to generate data. An ongoing trend in the field of bioinformatics is the unification and cross-referencing of many databases, such as that which has occurred for databases of molecular sequences and for databases of protein domains. In the arena of gene expression, the lack of acceptable standards may limit the extent to which an integrated view of gene expression is obtained. Nonetheless, it is likely that each gene in each organism will be indexed so that in addition to “stable” data on molecular sequence and chromosomal location, “dynamic” information on the mRNA corresponding to each gene will be cataloged. This information will include the abundance level of each transcript, the temporal and regional locations of gene expression, and other information on the behavior of gene expression in a variety of states.

## INTERPRETATION OF RNA ANALYSES

We began this chapter with a description of noncoding RNA, then described coding (messenger) RNA. We conclude with several issues regarding the nature and interpretation of RNA, including insights from large-scale RNA-seq projects.

### The Relationship between DNA, mRNA, and Protein Levels

Many human diseases are associated with changes in the number of chromosomes (termed aneuploidy); the most well-known of these is Down syndrome, associated with a third copy of chromosome 21. Many diseases are caused by the duplication or deletion of a small chromosome region (e.g., several million base pairs), and copy number changes are also commonly associated with cancers. A variety of evidence suggests that an increase in copy number (i.e., of genomic DNA) is associated with a corresponding increase in mRNA transcript levels. My laboratory (Mao *et al.*, 2003, 2005) and others have shown this for Down syndrome brain and heart, and similar findings have been reported in cancers.

Once mRNA levels are present at elevated or reduced levels, are the corresponding proteins differentially expressed in a similar manner? Perhaps surprisingly, there appears to be only a weak positive correlation between mRNA and protein levels. At present, high-throughput protein analyses are technically more difficult to perform (especially protein arrays) than transcriptional profiling studies. We discuss several high-throughput approaches to protein identification and quantitation (e.g., mass spectrometry) in Chapter 12.

Several groups have reported a weak positive correlation between mRNA levels and levels of the corresponding proteins in the yeast *Saccharomyces cerevisiae* and other systems (Futcher *et al.*, 1999; Greenbaum *et al.*, 2002). Greenbaum *et al.* (2002) performed a meta-analysis of gene expression and protein abundance datasets and suggested that there is a broad agreement between mRNA and protein levels. Waters *et al.* (2006) reviewed eight studies and described correlation coefficients that were relatively high when highly

abundant proteins were considered (e.g.,  $r = 0.935$ ,  $r = 0.86$  in two studies) but lower when highly abundant proteins were excluded (e.g.,  $r = 0.36$ ,  $r = 0.49$ ,  $r = 0.21$ ,  $r = 0.18$ ). Maier *et al.* (2009) reviewed the methods used to measure mRNA and protein levels. They discuss various mechanisms that could account for poor correlations including RNA structural effects, regulatory noncoding RNAs, codon bias, variable protein half-lives, and experimental error.

One conclusion from these studies is that it might be appropriate to determine experimentally whether observed changes in RNA correspond to changes in the levels of the corresponding proteins. At present, it is common in the scientific literature for changes in RNA transcripts, derived from genes encoding a category of proteins such as those involved in glycolysis, to be said to provide evidence that glycolysis has changed in the system being studied. Such a finding represents a hypothesis that can be tested experimentally.

The correlation coefficient  $r$  ranges from +1 (perfectly positively correlated) to -1 (negatively correlated), with  $r = 0$  indicating that the two variables are uncorrelated.

## The Pervasive Nature of Transcription

In recent decades, the transcription of DNA to mRNA has been conceptualized in terms of a relatively straightforward model in which protein-coding genes are transcribed into mRNA precursors which are then spliced (to remove introns) and processed (to facilitate export) into mature mRNA. The number of distinct mRNA transcripts was assumed to approximate the number of protein-coding genes, and the exons have been estimated to occupy less than 3% of the human genome. More recently, compelling evidence has emerged that the majority of the genomic DNA (comprising the genome) is transcribed.

Strong evidence for pervasive transcription comes from the ENCODE project (ENCODE Consortium, 2007; Djebali *et al.*, 2012). Transcriptional activity was measured using a series of technologies:

1. Total RNA or poly(A) RNA was hybridized to tiling microarrays. Tiling arrays contain oligonucleotides or PCR products that correspond to positions along each chromosome that are regularly spaced at extremely short intervals such as 5 or 30 base pairs. This contrasts with conventional expression arrays that are targeted to previously annotated exons. Genomic tiling arrays do not depend on prior genome annotations, and they also offer good sensitivity.
2. Cap-selected RNA was tag sequenced at the 5' or joint 5'/3' ends. 5' cap analysis gene expression (CAGE) is a method of enriching for full-length cDNA by priming the first strand cDNA synthesis with an oligo-dT primer (to capture the 3' end of a polyadenylated transcript) or a random primer, and “trapping” the cap that commonly occurs at the 5' end of mRNAs.
3. EST and cDNA sequences were annotated using computational, manual, and experimental approaches.
4. The most recent studies have relied on RNA-seq (Djebali *et al.*, 2012).

The most recent ENCODE conclusions include the following:

- 62.1% and 74.7% of the human genome is spanned by processed or primary transcripts, respectively;
- genes express 10–12 isoforms per cell line;
- coding RNA transcripts tend to be cytosolic, while noncoding transcripts are localized to the nucleus; and
- ~6% of annotated coding and noncoding transcripts overlap small noncoding RNAs.

You can learn more about CAGE at the FANTOM website (<http://fantom3.gsc.riken.jp/>, WebLink 10.52), including access to CAGE databases.

A clear conclusion from the ENCODE project and other studies is that much of the genome is transcribed. Some of this transcription is certain to be biologically relevant, while in other cases it is likely to represent biological “noise” associated with low levels of transcription. We discuss the definitions of function and the meaning of functional

elements in Chapter 14. In Chapter 8 we discussed the proposals by the ENCODE project to propose a novel definition of the gene. Gerstein *et al.* (2007) proposed a novel definition of a gene as “a union of genomic sequences encoding a coherent set of potentially overlapping functional products,” while Djebali *et al.* (2012) proposed the transcript rather than the gene as the “atomic unit of inheritance.”

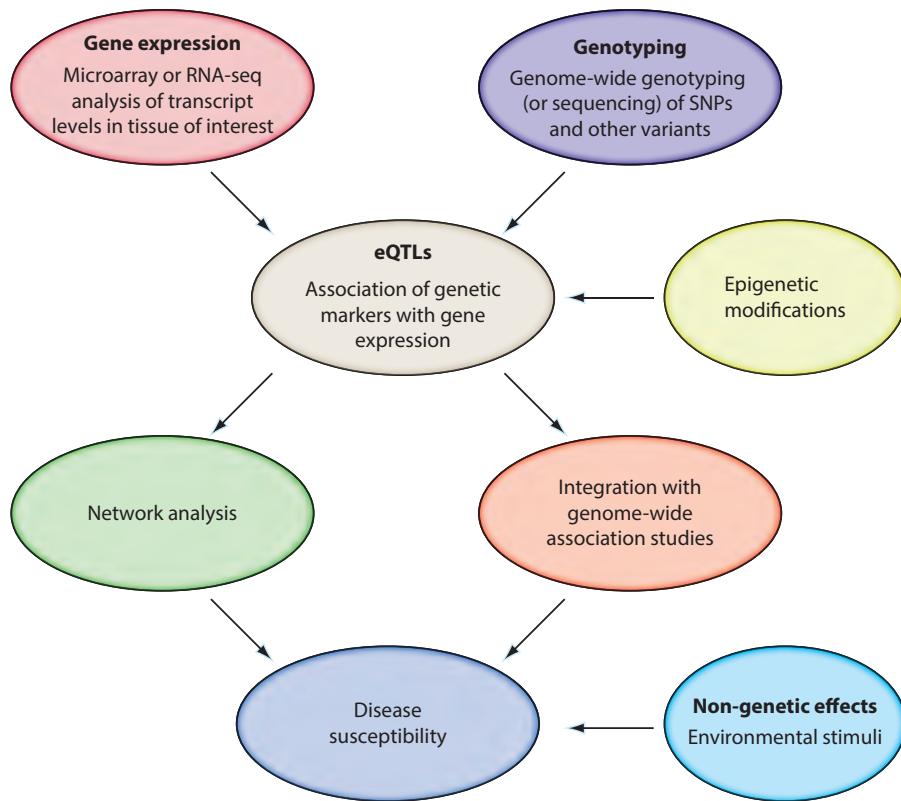
### eQTLs: Understanding the Genetic Basis of Variation in Gene Expression through Combined RNA-seq and DNA-seq

One of the outstanding problems in biology is the relationship between genotype and phenotype (see Chapter 14). It is commonly thought that variation in mRNA (or other RNA) levels may have key consequences for disease susceptibility. mRNA expression is a quantitative trait that can be described for a given cell type and physiological state in an organism. Furthermore, variants in genomic DNA may impact mRNA expression. Expression quantitative trait loci (eQTLs) are genomic loci that control expression levels (Cookson *et al.*, 2009; Majewski and Pastinen, 2011; Wright *et al.*, 2012). Studies in yeast, plants and humans have explored eQTLs. Initially these approaches relied on SNP arrays to measure DNA variation and microarrays to measure gene expression, with subsequent functional analysis applied to determine whether eQTLs are relevant to human disease (Fig. 10.15). Two main types of control regions were found: (1) *cis*-eQTLs are genomic loci that influence the expression of transcripts expressed from neighboring genes within some distance (such as 1 Mb or less), and may undergo allele-specific expression; and (2) *Trans*-eQTLs act on transcripts expressed from genes that are farther away or on another chromosome. eQTLs could affect transcription directly or indirectly, for example by altering the sequence of a transcription factor binding site that controls a gene’s expression proximally or distally.

Several studies have analyzed eQTLs by using RNA-seq (e.g., Pickrell *et al.*, 2010), using genotypes from the HapMap project and expression data from lymphoblastoid cell lines (LCLs, which are immortalized cell lines derived from lymphocytes). In a large-scale project led by the GEUVADIS group, researchers measured mRNA and small RNA transcript levels in 462 individuals from five populations (Lappalainen *et al.*, 2013). Almost all individuals also had whole-exome and/or whole-genome sequencing as part of the 1000 Genomes Project. This effort is significant because of its scale and its integration of DNA and RNA sequencing results. They reported eQTLs affecting gene expression in ~3700 genes; ~7800 genes with an eQTL that affected both gene expression and splicing variation; and 5700 *cis*-eQTLs for repetitive elements (retrotransposon-derived elements) outside genes. For the most significant eQTLs, the regulatory variants themselves tended to be enriched for indels rather than single-nucleotide variants and often occurred in transcription factor loci, enhancers, and DNaseI hypersensitive sites. Lappalainen *et al.* further evaluated the differences in expression between the two haplotypes of an individual, known as allele-specific expression. They report that genetic regulatory variation is a major determinant of allele-specific expression.

These studies catalog variation in expression and catalog associated genomic variation. Understanding variation that occurs outside of gene coding regions will be essential to interpret the findings of genome-wide association studies (GWAS; Chapter 21). This is because the vast majority of disease-associated variants mapped with that approach are localized to intergenic regions. Lactose intolerance provides one of the best-studied examples of this type of genetic variation. Reduced expression of lactase-phlorizin hydrolase (LPH) is associated with lactase nonpersistence (and lactose intolerance). A variant that resides ~14,000 base pairs upstream of the *LCT* gene is able to bind the transcription factor Oct-1 and is responsible for regulating expression of that gene (Lewinsky *et al.*, 2005).

We discuss the HapMap and 1000 Genomes Projects in Chapter 20. The Genetic European Variation in Health and Disease (GEUVADIS) consortium homepage is <http://www.geuvadis.org> (WebLink 10.53). 't Hoen *et al.* (2013) describe the quality control measures used for RNA sequencing in this project.



**FIGURE 10.15** Expression quantitative trait loci (eQTLs). Gene expression and genotype (including DNA sequencing) data are collected from multiple individuals. The association of DNA variants with expression levels of individuals is determined to infer eQTLs. Other forms of variation, such as epigenetic modifications (e.g., CpG methylation or histone modification patterns) may also be mapped. Subsequent network analysis explores connections between transcripts (such as those encoding proteins that participate in a common pathway). eQTLs can be used to identify variants affecting expression, particularly those variants occurring in noncoding regions that are implicated by genome-wide association studies (GWAS; Chapter 21). Nongenetic effects also influence disease susceptibility.

Source: Cookson *et al.* (2009). Reproduced with permission from Macmillan Publishers.

## PERSPECTIVE

Genes in all organisms are expressed in a variety of developmental, environmental, or physiological conditions. The field of functional genomics includes the high-throughput study of gene expression. Before the arrival of this new approach, the expression of one gene at a time was typically studied. Functional genomics may reveal the transcriptional program of entire genomes, allowing a global view of cellular function.

Three major shifts have occurred in recent years in our understanding of genes and their expression. First, complementary DNA microarrays and oligonucleotide-based microarrays were introduced in the mid-1990s, and have emerged as a powerful and popular tool for the rapid, quantitative analysis of RNA transcript levels in a variety of biological systems. The use of microarrays has now been complemented by RNA-seq which promises a broad range of new applications. Second, recent studies including those of the ENCODE project have indicated that much of the genome is transcribed, although the biological significance of this is not yet understood. Third, since the 1990s many small noncoding RNAs such as microRNAs have been identified and

are beginning to be functionally classified. Together, these discoveries and technological advances are leading to a new appreciation of the tremendous structural and functional diversity of RNAs.

## PITFALLS

The recent discovery of the pervasive nature of transcription leads to the question of how many mRNA transcripts have functional roles. For small noncoding RNAs we are only beginning to appreciate the range of possible biological functions. The computational challenge of noncoding RNA identification is great, and many more are likely to be identified.

For studies of gene expression with techniques such as EST analysis, microarrays, or RNA-seq there are many basic concerns. The mRNA molecules are not directly measured; rather, they are converted to cDNA, and that cDNA is analyzed by sequence analysis or by visualization of fluorescent tags. It is important to assess whether the amount of substance that is actually measured corresponds to the amount of mRNA in the biological sample.

- When RNA (or mRNA) is isolated, is it representative of the entire population of mRNA molecules in the cell?
- If two conditions are being compared, was the RNA isolated under exactly the same conditions? Any variations in the experimental protocol may lead to artifactual differences.
- Has degradation of the RNA occurred in any of the samples?
- For microarrays, most researchers cannot confirm the identity of what is immobilized on the surface of a microarray. For RNA-seq there are tremendous data analysis challenges.

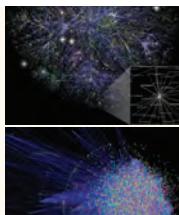
One response to these assorted concerns about microarrays and RNA-seq is that, with appropriate experimental design, results can be obtained with confidence. When data analysis results in the identification of significantly regulated genes (Chapter 11), it is important to perform independent biochemical assays (such as RT-PCR) to validate the findings.

## ADVICE TO STUDENTS

We have described RNA as being expressed in a context-dependent manner (at some time and location and physiological state). Try to get a feel for this by browsing UniGene (or other repositories) to get an idea of the diversity of libraries that have been made and sequenced. Find examples of biological samples (e.g., HapMap or 1000 Genomes cell lines) that have been characterized by microarrays and/or RNA-seq, and decide how reproducible RNA transcript measurements are across replicates and between laboratories.

## WEB RESOURCES

The RNA World Website (<http://www.rna.uni-jena.de/rna.php>, WebLink 10.54) organizes many links related to RNA and is an excellent starting point. RNACentral (<http://rnacentral.org/>, WebLink 10.55) is a new, major portal to RNA sequences. Visit the RNA-seq Blog (<http://www.rna-seqblog.com/>, WebLink 10.56) for a variety of useful resources.



## Discussion Questions

**[10-1]** There has been an explosion of interest in small noncoding RNAs in plant, animal, and other genomes. Why were these small RNAs not identified and studied in earlier decades?

**[10-2]** If you have a human cell line and you want to measure gene expression changes induced by a drug treatment, what are some of the advantages and disadvantages of using RNA-seq versus microarrays? How are your answers different if you want to study gene expression in a less-well-characterized organism such as a parasite?

**[10-3]** When you use a microarray, how can you assess what has been deposited on the surface of the array? How do you know the DNA is of the length and composition that the manufacturer of the array specifies? Suppose your colleague is performing an experiment with four control samples and four experimental samples, and tells you that two of the RNA samples (one control, one experimental) were possibly mixed together by accident. Could you use microarray data or RNA-seq data to figure out whether the mix-up had occurred or not?

### PROBLEMS/COMPUTER LAB

**[10-1]** We introduced the noncoding RNAs *Xist* and *Air*. We also discussed how many noncoding RNAs are poorly conserved. Perform a series of BLAST searches to try to identify human, mouse, and other homologs of *Xist* and *Air*. Try searching the RefSeq, nonredundant, or other nucleotide databases.

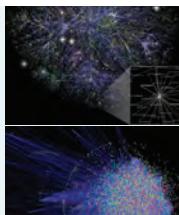
**[10-2]** Choose a human rRNA sequence, then perform BLASTN searches against human genomic DNA databases. How many matches do you find, and to what chromosomes are the rDNA sequences assigned?

**[10-3]** How many noncoding RNAs are in the vicinity of the human beta globin gene? To assess this, go to the UCSC bioinformatics site (<http://genome.ucsc.edu>), select the Genome Browser, set the organism to human, and choose a particular genome build; then enter the search term hbb to find that gene on chromosome 11. Then display annotation tracks related to noncoding RNAs, and set the view to 10 million base pairs surrounding the HBB gene.

**[10-4]** Telomerase is a ribonucleoprotein polymerase that in humans maintains active telomere ends by adding many copies of the repetitive sequence TTAGGG. The enzyme (which is a protein) includes an RNA component that serves as a template for the telomere repeat. To what chromosome is this noncoding RNA gene assigned? As one approach, find the entry in Entrez Nucleotide at NCBI. As another approach, search Rfam with the keyword telomerase.

**[10-5]** Perform digital differential display:

- Go to UniGene (<http://www.ncbi.nlm.nih.gov/UniGene/>).
- Go to *Homo sapiens*.
- Click library differential display.
- Click some brain libraries, then “Accept changes.”
- Choose a second pool of libraries to compare.



## Self-Test Quiz

**[10-1]** Which are the most abundant RNA types?

- rRNA and tRNA;
- rRNA and mRNA;
- tRNA and mRNA; or
- mRNA and microRNA.

**[10-2]** MicroRNAs may be distinguished from other RNAs because of the following properties:

- they are localized to the nucleolus;
- each microRNA is thought to regulate a small number of homologous target messenger RNAs;

(c) they are coding RNAs, each of which is thought to regulate the function of a large number of messenger RNAs to which they are homologous; or

(d) they have a length of about 22 nucleotides, derived from a larger precursor, and regulate messenger RNA function.

**[10-3]** The stages of mRNA processing include all of the following except:

- splicing;
- export;
- methylation; or
- surveillance.

**[10-4]** Digital differential display (DDD) is used to compare the content of expressed sequence tags (ESTs) in UniGene's cDNA libraries. ESTs are also represented on microarrays. Which statement best describes ESTs?

- (a) clusters of nonredundant sequences (approximately 500 base pairs in length);
- (b) stretches of DNA sequence that are repeated many times throughout the genome;
- (c) sequences corresponding to expressed genes that are obtained by sequencing complementary DNAs; or
- (d) a "tag" (i.e., a fragment of DNA) derived from complementary DNA (cDNA) that corresponds to a transcript that has not been identified.

**[10-5]** UniGene has cluster sizes from very small (e.g., 1) to very large (e.g., >10,000). What does it mean for there to be a cluster of size 1?

- (a) one sequence has been identified that has a very large number of EST transcripts (e.g., over 10,000) associated with it;
- (b) one sequence has been identified that corresponds to a gene that has been expressed one time;
- (c) one sequence has been identified (presumably it is an EST) that matches one other known sequence (allowing it to be identified as a UniGene cluster); or
- (d) one sequence has been identified (presumably it is an EST) that is thought to correspond to a known gene, but it matches no other known sequences in UniGene (i.e., it does not align to any other ESTs).

**[10-6]** In analyzing cDNA libraries, a pitfall is that the libraries:

- (a) may be derived from different tissues;

- (b) may contain thousands of sequences;
- (c) may have been normalized differently; or
- (d) may contain many rarely expressed transcripts.

**[10-7]** Most microarrays consist of a solid support on which is immobilized:

- (a) DNA;
- (b) RNA;
- (c) genes; or
- (d) transcripts.

**[10-8]** The purpose of the MIAME project is to provide a unified system for:

- (a) the description of microarray manufacture;
- (b) the description of microarray experiments from design to hybridization to image analysis;
- (c) the description of microarray probe preparation including fluorescence- and radioactivity-based approaches; or
- (d) microarray databases including standards for data storage, analysis, and presentation.

**[10-9]** RNA-seq offers several advantages over DNA microarrays. Which of the following is NOT an advantage for RNA-seq?

- (a) the dynamic range is superior;
- (b) the reproducibility is better, so fewer biological replicates are needed;
- (c) it can be used to characterize previously unannotated transcripts; or
- (d) it can be used to characterize varieties of noncoding RNAs.

## SUGGESTED READING

Alex Bateman and 29 colleagues introduce RNACentral, a relatively new RNA database that centralizes data from many sources. In that paper, Bateman *et al.* (2011) provide brief, excellent overviews of RNA databases and of the relevance of RNA to many disciplines. Washietl *et al.* (2012) and Nawrocki and Eddy (2013) describe methods for identifying functional noncoding RNA elements, and their clearly written reviews describe the benefits of combining structural and sequence information. The Washietl *et al.* paper further explains the use of RNA-seq to identify noncoding RNAs and novel transcripts. For an excellent description of the structure, function, evolution, and phylogenetic distribution of miRNAs, see Berezikov (2011).

Next-generation sequencing of cDNA derived from RNA (RNA-seq) has had a dramatic impact on our ability to characterize many classes of noncoding RNAs, as well as coding RNAs. Morozova *et al.* (2009) and Wang *et al.* (2009) review the impact of this technology.

## REFERENCES

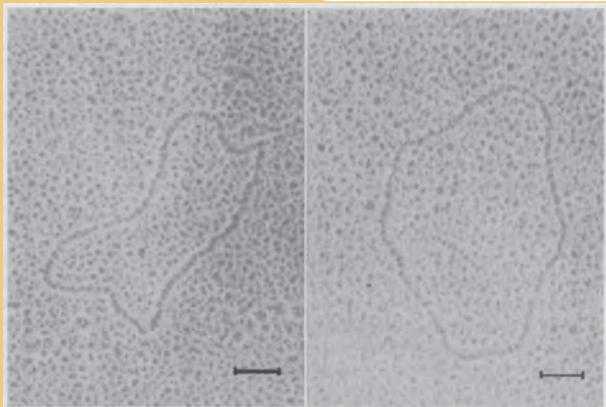
- Ambros, V., Bartel, B., Bartel, D.P. *et al.* 2003. A uniform system for microRNA annotation. *RNA* **9**, 277–279. PMID: 12592000.
- Avery, O.T., MacLeod, C.M., McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. *Journal of Experimental Medicine* **79**, 137–158.
- Baker, S.C., Bauer, S.R., Beyer, R.P. *et al.* 2005. The External RNA Controls Consortium: a progress report. *Nature Methods* **2**, 731–734. PMID: 16179916.
- Bark, C., Weller, P., Zabielski, J., Pettersson, U. 1986. Genes for human U4 small nuclear RNA. *Gene* **50**, 333–344.
- Barrett, T., Wilhite, S.E., Ledoux, P. *et al.* 2013. NCBI GEO: archive for functional genomics data sets: update. *Nucleic Acids Research* **41**(Database issue), D991–995. PMID: 23193258.
- Bateman, A., Agrawal, S., Birney, E. *et al.* 2011. RNAcentral: A vision for an international database of RNA sequences. *RNA* **17**(11), 1941–1946. PMID: 21940779.
- Beadle, G.W., Tatum, E.L. 1941. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Science USA* **27**(11), 499–506. PMID: 16588492.
- Berezikov, E. 2011. Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics* **12**(12), 846–860. PMID: 22094948.
- Betel, D., Wilson, M., Gabow, A., Marks, D.S., Sander, C. 2008. The microRNA.org resource: targets and expression. *Nucleic Acids Research* **36**(Database issue), D149–153. PMID: 18158296.
- Bhartiya, D., Pal, K., Ghosh, S. *et al.* 2013. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database (Oxford)* **2013**, bat034. PMID: 23846593.
- Bonaldo, M. F., Lennon, G., Soares, M. B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Research* **6**, 791–806.
- Borsani, G., Tonlorenzi, R., Simmler, M.C. *et al.* 1991. Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325–329. PMID: 2034278.
- Brazma, A., Hingamp, P., Quackenbush, J. *et al.* 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics* **29**, 365–371. PMID: 11726920.
- Burge, S.W., Daub, J., Eberhardt, R. *et al.* 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research* **41**(Database issue), D226–232. PMID: 23125362.
- Cabili, M.N., Trapnell, C., Goff, L. *et al.* 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development* **25**(18), 1915–1927. PMID: 21890647.
- Carninci, P., Kasukawa, T., Katayama, S. *et al.* 2005. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563. PMID: 16141072.
- Celotto, A. M., Graveley, B. R. 2001. Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* **159**, 599–608.
- Chan, P.P., Lowe, T.M. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* **37**(Database issue), D93–97. PMID: 18984615.
- Cho, S., Jang, I., Jun, Y. *et al.* 2013. MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic Acids Research* **41**(Database issue), D252–257. PMID: 23193297.
- Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495.
- Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics* **8**(10), 1821–1832. PMID: 10469833.
- Cole, J.R., Chai, B., Farris, R.J. *et al.* 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Research* **35**, D169–172. PMID: 17090583.
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M. 2009. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**(3), 184–194. PMID: 19223927.

- Costa, V., Aprile, M., Esposito, R., Ciccodicola, A. 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *European Journal of Human Genetics* **21**(2), 134–142. PMID: 22739340.
- Crick, F.H. 1958. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138–163. PMID: 13580867.
- Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J., Jones, D.D. 1972. Gene duplications in evolution: the globins. In: *Atlas of Protein Sequence and Structure* (ed. Dayhoff, M.O.), Vol. 5. National Biomedical Research Foundation, Washington, DC.
- DeRisi, J., Penland, L., Brown, P.O. *et al.* 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14**, 457–460. PMID: 8944026.
- Derrien, T., Johnson, R., Bussotti, G. *et al.* 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research* **22**(9), 1775–1789. PMID: 22955988.
- Djebali, S., Davis, C.A., Merkel, A. *et al.* 2012. Landscape of transcription in human cells. *Nature* **489**(7414), 101–108. PMID: 22955620.
- Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., Trent, J. M. 1999. Expression profiling using cDNA microarrays. *Nature Genetics* **21**, 10–14.
- Eddy, S.R., Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**, 2079–2088.
- Eggenhofer, F., Hofacker, I.L., Höner Zu Siederdissen, C. 2013. CMCompare webserver: comparing RNA families via covariance models. *Nucleic Acids Research* **41**(Web Server issue), W499–503. PMID: 23640335.
- ENCODE Project Consortium *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816.
- Fichant, G.A., Burks, C. 1991. Identifying potential tRNA genes in genomic DNA sequences. *Journal of Molecular Biology* **220**, 659–671.
- Fire A., Xu S., Montgomery M.K., Kostas S.A., Driver S.E., Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811.
- Fodor, S. P., Read, J.L., Pirrung, M.C. *et al.* 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773. PMID: 1990438.
- Friedman, R.C., Farh, K.K., Burge, C.B., Bartel, D.P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**(1), 92–105. PMID: 18955434.
- Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., Garrels, J. I. 1999. A sampling of the yeast proteome. *Molecular and Cellular Biology* **19**, 7357–7368.
- Gerstein, M.B., Bruce, C., Rozowsky, J.S. *et al.* 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17**(6), 669–681. PMID: 17567988.
- Gonzalez, I.L., Sylvester, J.E. 2001. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73**(3), 255–263. PMID: 11350117.
- Greenbaum, D., Jansen, R., Gerstein, M. 2002. Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18**, 585–596.
- Guenzl, P.M., Barlow, D.P. 2012. Macro lncRNAs: a new layer of cis-regulatory information in the mammalian genome. *RNA Biology* **9**(6), 731–741. PMID: 22617879.
- Hansen, K.D., Wu, Z., Irizarry, R.A., Leek, J.T. 2011. Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29**(7), 572–573. PMID: 21747377.
- Henderson, A.S., Warburton, D., Atwood, K.C. 1972. Location of ribosomal DNA in the human chromosome complement. *Proceedings of the National Academy of Science, USA* **69**, 3394–3398.
- Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Research* **31**, 3429–3431.
- Hofacker, I.L., Lorenz, R. 2014. Predicting RNA structure: advances and limitations. *Methods in Molecular Biology* **1086**, 1–19. PMID: 24136595.

- Humphreys, D.T., Suter, C.M. 2013. miRspring: a compact standalone research tool for analyzing miRNA-seq data. *Nucleic Acids Research* **41**(15), e147. PMID: 23775795.
- Katayama, S., Tomaru, Y., Kasukawa, T. *et al.* 2005. Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566. PMID: 16141073.
- Kikuno, R., Nagase, T., Nakayama, M. *et al.* 2004. HUGE: a database for human KIAA proteins, a 2004 update integrating HUGEPPI and ROUGE. *Nucleic Acids Research* **32**, D502–504.
- Kornienko, A.E., Guenzl, P.M., Barlow, D.P., Paurer, F.M. 2013. Gene regulation by the act of long non-coding RNA transcription. *BMC Biology* **11**, 59. PMID: 23721193.
- Kozomara, A., Griffiths-Jones, S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* **39**(Database issue), D152–157. PMID: 21037258.
- Krek, A., Grün, D., Poy, M.N. *et al.* 2005. Combinatorial microRNA target predictions. *Nature Genetics* **37**(5), 495–500. PMID: 15806104.
- Lagesen, K., Hallin, P., Rodland, E.A. *et al.* 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100–3108. PMID: 17452365.
- Lappalainen, T., Sammeth, M., Friedländer, M.R. *et al.* 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**(7468), 506–511. PMID: 24037378.
- Lee, J.T. 2012. Epigenetic regulation by long noncoding RNAs. *Science* **338**(6113), 1435–1439. PMID: 23239728.
- Lee, R.C., Feinbaum, R.L., Ambros, V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854.
- Lestrade, L., Weber, M.J. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research* **34**(Database issue), D158–162.
- Lewinsky, R.H., Jensen, T.G., Møller, J. *et al.* 2005. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Human Molecular Genetics* **14**(24), 3945–3953. PMID: 16301215.
- Lewis, B.P., Burge, C.B., Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20.
- Liebhäber, S.A., Cash, F.E., Ballas, S.K. 1986. Human alpha-globin gene expression. The dominant role of the alpha 2-locus in mRNA and protein synthesis. *Journal of Biological Chemistry* **261**, 15327–15333.
- Lipshutz, R. J., Fodor, S. P., Gingras, T. R., Lockhart, D. J. 1999. High density synthetic oligonucleotide arrays. *Nature Genetics* **21**, 20–24.
- Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C. *et al.* 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26. PMID: 22115189.
- Lowe, T.M., Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964.
- Lowe, T.M., Eddy, S.R. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171.
- Ludwig, W., Strunk, O., Westram, R. *et al.* 2004. ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363–1371. PMID: 14985472.
- Luteijn, M.J., Ketting, R.F. 2013. PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nature Reviews Genetics* **14**(8), 523–534. PMID: 23797853.
- Maeda, N., Kasukawa, T., Oyama, R. *et al.* 2006. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genetics* **2**, e62. PMID: 16683036.
- Maier, T., Güell, M., Serrano, L. 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* **583**(24), 3966–3973. PMID: 19850042.
- Majewski, J., Pastinen, T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics* **27**(2), 72–79. PMID: 21122937.
- Maniatis, T., Reed, R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416**, 499–506.

- Mao, R., Zielke, C.L., Zielke, H.R., Pevsner, J. 2003. Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. *Genomics* **81**(5), 457–467. PMID: 12706104.
- Mao, R., Wang, X., Spitznagel, E.L. Jr. *et al.* 2005. Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. *Genome Biology* **6**(13), R107. PMID: 16420667.
- Maquat, L. E. 2002. Molecular biology. Skiing toward nonstop mRNA decay. *Science* **295**, 2221–2222.
- Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* **288**, 911–940.
- McGettigan, P.A. 2013. Transcriptomics in the RNA-seq era. *Current Opinion in Chemical Biology* **17**(1), 4–11. PMID: 23290152.
- MGC Project Team, Temple, G., Gerhard, D.S. *et al.* 2009. The completion of the Mammalian Gene Collection (MGC). *Genome Research* **19**(12), 2324–2333. PMID: 19767417.
- Miller, O. L., Hamkalo, B. A., Thomas, C. A. 1970. Visualization of bacterial genes in action. *Science* **169**, 392–395.
- Miranda, K.C., Huynh, T., Tay, Y. *et al.* 2006. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**(6), 1203–1217. PMID: 16990141.
- Modrek, B., Lee, C. 2002. A genomic view of alternative splicing. *Nature Genetics* **30**, 13–19.
- Morozova, O., Hirst, M., Marra, M.A. 2009. Applications of new sequencing technologies for transcriptome analysis. *Annual Reviews in Genomics and Human Genetics* **10**, 135–151. PMID: 19715439.
- Mutz, K.O., Heilkenbrinker, A., Lönne, M., Walter, J.G., Stahl, F. 2013. Transcriptome analysis using next-generation sequencing. *Current Opinion in Biotechnology* **24**(1), 22–30. PMID: 23020966.
- Nagalakshmi, U., Waern, K., Snyder, M. 2010. RNA-Seq: a method for comprehensive transcriptome analysis. *Current Protocols in Molecular Biology Chapter* **4**, Unit 4.11.1–13. PMID: 20069539.
- Nagase, T., Koga, H., Ohara, O. 2006. Kazusa mammalian cDNA resources: towards functional characterization of KIAA gene products. *Briefings in Functional Genomic Proteomics* **5**, 4–7. PMID: 16769670.
- Nawrocki, E.P., Eddy, S.R. 2013. Computational identification of functional RNA homologs in metagenomic data. *RNA Biology* **10**(7), 1170–1179. PMID: 23722291.
- Nawrocki, E.P., Burge, S.W., Bateman, A. *et al.* 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research* **43**(Database issue), D130–137. PMID: 25392425.
- Nirenberg, M. 1965. Protein synthesis and the RNA code. *Harvey Lectures* **59**, 155–185.
- Olsen, G.J., Woese, C.R. 1993. Ribosomal RNA: a key to phylogeny. *FASEB Journal* **7**(1), 113–23. PMID: 8422957.
- Ozsolak, F., Milos, P.M. 2011. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**(2), 87–98. PMID: 21191423.
- Pang, K.C., Stephen, S., Dinger, M.E. *et al.* 2007. RNAdb 2.0: an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research* **35**(Database issue), D178–182. PMID: 17145715.
- Paraskevopoulou, M.D., Georgakilas, G., Kostoulas, N. *et al.* 2013. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research* **41**(Web Server issue), W169–173. PMID: 23680784.
- Pasquinelli, A.E. 2012. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics* **13**(4), 271–282. PMID: 22411466.
- Pasquinelli, A.E., Ruvkun, G. 2002. Control of developmental timing by microRNAs and their targets. *Annual Reviews in Cell Development and Biology* **18**, 495–513.
- Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., Ottonello, S. 1994. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of trnascriptional control regions. *Nucleic Acids Research* **22**, 1247–1256.
- Pedersen, J.S., Bejerano, G., Siepel, A. *et al.* 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology* **2**, e33. PMID: 16628248.
- Pickrell, J.K., Marioni, J.C., Pai, A.A. *et al.* 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**(7289), 768–772. PMID: 20220758.

- Rajewsky, N. 2006. microRNA target predictions in animals. *Nature Genetics* **38**, Suppl: S8–13.
- Ritchie, W., Rasko, J.E., Flamant, S. 2013. MicroRNA target prediction and validation. *Advances in Experimental Medicine and Biology* **774**, 39–53. PMID: 23377967.
- Sakakibara, Y., Brown, M., Hughey, R. *et al.* 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research* **22**, 5112–5120. PMID: 7800507.
- Sayers, E.W., Barrett, T., Benson, D.A. *et al.* 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **40**(Database issue), D13–25. PMID: 22140104.
- Schattner, P., Brooks, A.N., Lowe, T.M. 2005. The tRNAscan-SE, snoScan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**(Web Server issue), W686–689. PMID: 15980563.
- Schloss, P.D., Handelsman, J. 2004. Status of the microbial census. *Microbiology and Molecular Biology Reviews* **68**, 686–691. PMID: 15590780.
- Schmucker, D., Clemens, J.C., Shu, H. *et al.* 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671–684. PMID: 10892653.
- Sleutels, F., Zwart, R., Barlow, D.P. 2002. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813.
- Stekel, D.J., Git, Y., Falciani, F. 2000. The comparison of gene expression from multiple cDNA libraries. *Genome Research* **10**(12), 2055–2061. PMID: 11116099.
- ‘t Hoen, P.A., Friedländer, M.R., Almlöf, J. *et al.* 2013. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology* **31**, 1015–1022. PMID: 24037425.
- Takeda, J., Yamasaki, C., Murakami, K. *et al.* 2013. H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Research* **41**(Database issue), D915–919. PMID: 23197657.
- Tåquist, H., Cui, Y., Ardell, D.H. 2007. TFAM 1.0: an online tRNA function classifier. *Nucleic Acids Research* **35**(Web Server issue), W350–353.
- Temple, G., Lamesch, P., Milstein, S. *et al.* 2006. From genome to proteome: developing expression clone resources for the human genome. *Human Molecular Genetics* **15**, R31–43. PMID: 16651367.
- Valadkhan, S. 2005. snRNAs as the catalysts of pre-mRNA splicing. *Current Opinion in Chemical Biology* **9**, 603–608.
- Waggoner, S.A., Liebhaber, S.A. 2003. Regulation of alpha-globin mRNA stability. *Experimental Biology and Medicine (Maywood)* **228**, 387–395.
- Wang, X. 2008. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**(6), 1012–1017. PMID: 18426918.
- Wang, Z., Gerstein, M., Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63. PMID: 19015660.
- Washietl, S. 2010. Sequence and structure analysis of noncoding RNAs. *Methods in Molecular Biology* **609**, 285–306. PMID: 20221926.
- Washietl, S., Hofacker, I.L. 2010. Nucleic acid sequence and structure databases. *Methods in Molecular Biology* **609**, 3–15. PMID: 20221910.
- Washietl, S., Will, S., Hendrix, D.A. *et al.* 2012. Computational analysis of noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA* **3**(6), 759–778. PMID: 22991327.
- Waters, K.M., Pounds, J.G., Thrall, B.D. 2006. Data merging for integrated microarray and proteomic analysis. *Briefings in Functional Genomic and Proteomics* **5**, 261–272.
- Watson, J.D., Crick, F.H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356), 737–738. PMID: 13054692.
- Westermann, A.J., Gorski, S.A., Vogel, J. 2012. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* **10**(9), 618–630. PMID: 22890146.
- Wright, F.A., Shabalina, A.A., Rusyn, I. 2012. Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics* **13**(3), 343–352. PMID: 22304583.
- Zuker, M., Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **9**, 133–148.



The main idea behind microarrays is that one nucleic acid (DNA) is immobilized on a solid support on a solid surface in a predefined location, and then another nucleic acid (RNA or a derivative such as fluorescently labeled complementary DNA) is hybridized to the surface. Microarrays were first developed in the 1990s by the laboratories of Patrick Brown at Stanford University and Jeffrey Trent, then at the National Institutes of Health (NIH). Beginning in 1950s, Sol Spiegelman (1914–1983) pioneered the study of RNA hybridization to DNA (see <http://profiles.nlm.nih.gov/PX/>). By the early 1960s several groups immobilized DNA on a solid support then hybridized a variety of purified RNA molecules under a variety of conditions. This figure shows electron micrographic images of circular DNA-RNA hybrids by Spiegelman and colleagues (Bassel et al., 1964). The bacteriophage  $\varphi$ X174 was shown to transcribe RNA which bound to DNA in a ribonuclease-resistant complex. Studies such as these established the mechanisms by which DNA is transcribed to RNA, and ultimately led to the development of hybridization-based assays including microarrays. The scale bar is 0.1  $\mu$ m.

Source: Bassel et al. (1964).

# Gene Expression: Microarray and RNA-seq Data Analysis

# CHAPTER 11

*A handful of luck is worth six assloads of learning.*

—Arabic proverb

## LEARNING OBJECTIVES

After completing this chapter you should be able to:

- explain what preprocessing is and how normalization of microarrays is accomplished;
- define a *t*-test and probability values;
- describe different kinds of exploratory statistics (clustering, principal components analysis) and explain how they are used to visualize gene expression data; and
- analyze both microarray and RNA-seq datasets.

## INTRODUCTION

Two powerful experimental approaches have emerged for the large-scale analysis of gene expression (i.e., mRNA transcript levels): microarrays and next-generation sequencing-based transcriptional profiling (RNA-seq). Microarrays became commonly used by the year 2000, and RNA-seq became prominent a decade later. In each case, RNA is extracted from a source of interest (e.g., human fetal brain), and some comparison is sought (e.g., what RNA transcript changes occur across developmental stages, or across brain regions, or in euploid versus trisomic samples). For microarrays, RNA samples are converted to some stable form (complementary DNA or complementary RNA for the Affymetrix platform we focus on), labeled with a fluorescent dye, and hybridized to a microarray surface containing thousands (or millions) of pre-selected DNA elements. For RNA-seq, RNA is converted to cDNA, packaged into libraries, and millions of short reads are obtained by next-generation sequencing.

The main purpose of both technologies is to identify which genes were significantly up- or down-regulated (differential expression is therefore measured). RNA-seq offers the advantage that transcripts to be analyzed are not pre-selected but instead the sequencing determines (in a relatively unbiased manner) all the RNA species that are present in each sample.

RNA-seq is more useful than microarrays for additional purposes:

- measuring transcript abundance;
- identifying transcripts to improve annotation of genes; and
- *de novo* transcript assembly.

The workflow for experiments measuring RNA starts with experimental design (**Fig. 11.1**, lavender-shaded boxes). For biologists it is an excellent idea to collaborate with a biostatistician at the outset for two main practical reasons: to try to ensure there are enough biological replicates in the design to allow meaningful conclusions to be drawn regarding significantly regulated transcripts; and to minimize the interference of inevitable nuisance factors. RNA can be purified in a biology lab or at a core facility. Because of the nature of RNA (it is labile, and its expression varies due to many factors) there are always batch effects in which RNA changes may be attributed to unwanted random variables (such as the date or time of RNA isolation, the method of RNA isolation, the number of people handling the RNA during purification, or whether the samples were processed under comparable conditions). Jeff Leek and colleagues (2010) have reviewed batch effects in high-throughput datasets including microarrays, DNA methylation arrays, and even DNA sequencing from the 1000 Genomes Project. Proper experimental design helps to address these issues, allowing later analyses that identify and correct for different sources of variation, and focus on the changes in RNA transcript levels due to biological causes that the investigator is most interested in identifying.

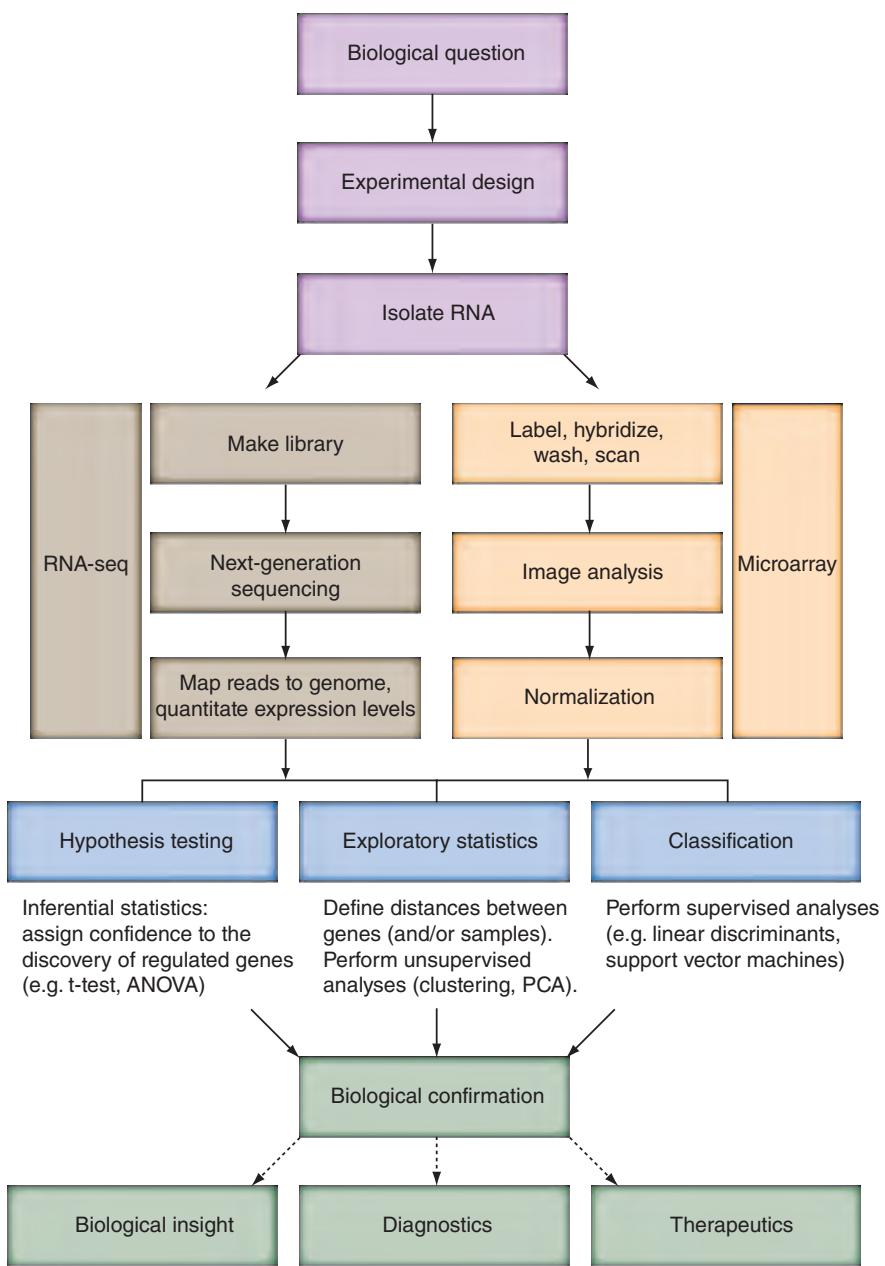
Data analysis for these two technologies is quite different. In this chapter we perform three workflows for microarrays, then one for RNA-seq. Once transcript levels are quantified and normalized, several main analyses are performed (**Fig. 11.1**, blue-shaded boxes). Exploratory statistics allow us to determine whether any samples are outliers (suggesting they might need to be discarded), and techniques such as clustering and principal components analysis are used to search for patterns in the data. Hypothesis testing helps us determine the statistical significance of differentially regulated transcripts: for each individual transcript that appears to be differentially regulated in comparisons of two or more groups, how often are these observations likely to have occurred by chance? Classification can be used to determine how useful the expression patterns of a particular subset of transcripts are, in order to predict whether unknown samples segregate into groups such as disease versus control.

Microarray and RNA-seq experiments typically involve the measurement of the expression levels of tens of thousands of genes in only a few biological samples. There are usually no technical replicates (i.e., measuring gene expression with the same starting material on independent arrays) due to the relatively high cost of performing microarray experiments. There are also few biological replicates (e.g., measuring gene expression from multiple cell lines, each of which has been given an experimental treatment or a control treatment) relative to the large number of transcripts whose expression levels are quantitated. The challenge to the biologist is to apply appropriate statistical techniques to determine which changes are relevant. There is unlikely to be a single best approach to microarray or RNA-seq data analysis, and the tools applied to these workflows are evolving rapidly.

We begin with the analysis of a trisomy 21 (Down syndrome) dataset based on Affymetrix microarrays (a leading platform). In euploid cells each autosome is present in two copies, one from each parent. In trisomy 21 there are three copies of chromosome 21 (in >90% of cases the extra copy is maternal). We select this experiment because we can look to see if the group of RNA transcripts chromosome 21 is present at elevated levels relative to euploid controls. Indeed this is the case, a result published by researchers from my lab (Mao *et al.*, 2003, 2005).

This chapter is organized as follows. We use three methods to analyze the microarray data:

1. The web-based GEO2R tool at NCBI can be run in a matter of minutes, although its capabilities are limited. We use it to introduce basic concepts such as probability values, *t*-tests, the normalization of microarray data, accuracy, and precision. GEO2R uses R scripts so we also introduce those, although the main point of GEO2R is the implementation of these complicated scripts in a simple web-based form.



**FIGURE 11.1** Overview of methods for assessing RNA changes (“gene expression” analysis). Purple-shaded boxes: first, a biological question is formulated and then experimental design is created. After RNA is isolated we consider two technologies. For microarrays (peach boxes) the sample is converted to a set of fluorescently tagged molecules that are hybridized to a solid support, washed, and scanned. Image analysis is performed to provide the raw data of quantified expression levels, typically for >20,000 transcripts. Preprocessing involves normalization and removal of outliers. For Affymetrix arrays, an additional preprocessing step is summarization in which the expression value of a given gene (mRNA transcript) is summarized based on the results from a series of hybridizations to oligonucleotides corresponding to that gene. For RNA-seq (brown boxes), RNA is converted to complementary DNA, packaged into a library, sequenced, and reads are mapped to a genome (or set of DNA regions corresponding to transcripts) to quantitate expression levels of genes including alternatively spliced transcripts. For microarrays or RNA-seq, hypothesis testing is performed (blue boxes) in which t-tests, ANOVA or other statistical tests are applied to determine which transcripts were significantly up- or down-regulated in the experiment. Exploratory (descriptive) statistics may be applied such as clustering of genes (or samples). For supervised approaches, samples (or genes) are associated with labels from a pre-existing classification (such as normal versus diseased tissue) and gene expression measurements are used to predict which unknown samples are diseased. Finally, after microarray or RNA-seq data analysis is performed, biological confirmation experiments (green boxes) may be performed. This may lead to insight about biological processes, or to outcomes relevant to disease such as identifying diagnostic markers or strategies for therapeutic intervention. Adapted in part from Brazma and Vilo (2000).

2. We use the commercial software package Partek® Genomics Suite, demonstrating some of its versatile analysis and plotting features. In this section we introduce scatter plots, volcano plots, and ANOVA.
3. We use the R packages `affy` and `limma`. These are free and open source, and are popular in the biostatistics and bioinformatics communities. They may require a substantial effort to master, but to some these and other R packages are essential for a range of bioinformatics applications from microarrays to proteomics, methylation studies, and next-generation sequence data analysis.

Exploratory analyses (descriptive statistics) of microarray data are our next topic. This area includes hierarchical clustering and principal components analysis (PCA).

We then turn to RNA-seq data and use the Linux operating system to analyze a *Drosophila* dataset with the very popular software tools TopHat, Cufflinks, and the R package `cummeRbund`. The chapter concludes with a brief discussion of the functional annotation of expression data.

The relevant page for GSE1397 is <http://www.ncbi.nlm.nih.gov/gds/?term=GSE1397> (WebLink 11.1). In the analyses below we will not use the three trisomy 13 datasets or their matched controls. This study was performed primarily by Rong Mao, who at the time was a graduate student in the Pevsner lab.

## MICROARRAY ANALYSIS METHOD 1: GEO2R AT NCBI

We now analyze a gene expression dataset using a simple web-based workflow that calls a variety of tools (including R scripts) and databases. First, choose a dataset from GEO. We use a set of human trisomy 21 (TS21, associated with Down syndrome) and euploid (normal chromosome copy number) samples from heart, brain, and astrocytes. The series accession is GSE1397. Enter that accession into a search at the home page of NCBI. There is a single link to BioProjects, leading to a single link to the GEO DataSets. (You can also follow the Entrez search engine result to the GEO DataSets page directly.) Select the option “Analyze with GEO2R.”

### GEO2R Executes a Series of R Scripts

GEO2R performs analyses using well-known libraries (`Biobase`, `GEOquery`, `limma`), provides the accompanying R scripts, and generates plots and tables. First, choose the “define group” option to define TS21 ( $n = 11$ ) and euploid ( $n = 11$ ) groups (Fig. 11.2a).

It is quick and easy to click the “top 250” button to see which transcripts were significantly regulated in this experiment. Before we do that, let’s look at the R script provided by GEO2R. It is not essential to understand this script for users who simply want a quick answer. For biologists who are unfamiliar with R and trying to learn how it works however, the script provides an excellent introduction to the strength of R (as a tool offering libraries and commands that implement advanced software tools in a precise, flexible, validated process) and the limitations of R (these commands are not intuitive and using R packages involves a substantial learning curve). The commands that follow are in blue. Comments are in green, preceded by a hash (#) indicating comments introduced by NCBI (my comments appear with triple hashes).

```
# Version info: R 2.14.1, Biobase 2.15.3, GEOquery 2.23.2, limma 3.10.1
# R scripts generated Thu Apr 3 13:47:04 EDT 2014

#####
# Differential expression analysis with limma
library(BioBase)
library(GEOquery)
library(limma)
### To load a library in R you will need to first install it, e.g.:
### > source("http://bioconductor.org/biocLite.R")
### > biocLite("Biobase")
### > library(Biobase)
### > biocLite("limma")
```

```

#### > library(limma)
#### You can then get information about various functions in these
#### packages, e.g., > limmaUsersGuide()
# load series and platform data from GEO
#### Note that the getGEO command of the GEOquery library is useful to
#### extract GEO datasets

gset <- getGEO("GSE1397", GSEMatrix =TRUE)
if (length(gset) > 1) idx <- grep("GPL96", attr(gset, "names")) else idx
<- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
#### Here the object sml will concatenate (abbreviated c) the samples.
#### The trisomy 13 samples and controls (n=6) are marked "X".
sml <- c("G0","G0","G0","G1","G1","G1","G1","G1","G1","G0","G0",
"G0","G1","G1","G0","G1","G1","G0","G0","X","X","X","X","X");

# eliminate samples marked as "X"
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[,sel]

# log2 transform
#### We will discuss the rationale for log2 transformation below
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100)
(qx[6]-qx[1] > 50 && qx[2] > 0)
(qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
#### model.matrix (from the stats package) creates a design matrix as
#### specified.
colnames(design) <- levels(f1)
fit <- lmFit(gset, design)
#### we will discuss lmFit when we perform analyses with limma (below).
#### lmFit (from the limma package) fits a linear model to the log-
#### transformed expression values for each probe in a series of
#### arrays.
cont.matrix <- makeContrasts(G1-G0, levels=design)
#### makeContrasts determines fold change between groups of samples
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
#### eBayes (from the limma package) uses empirical Bayes statistics to
#### determine differential expression. For usage, details, references, and
#### examples use > ?eBayes
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
#### topTable (from the limma package) extracts a table of the top-
#### ranked genes from a linear model fit that has been processed by
#### eBayes.

# load NCBI platform annotation
gpl <- annotation(gset)
platf <- getGEO(gpl, AnnotGPL=TRUE)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

# replace original platform annotation
tT <- tT[setdiff(colnames(tT), setdiff(fvarLabels(gset), "ID"))]
tT <- merge(tT, ncbifd, by="ID")

```

```
tT <- tT[order(tT$P.Value), ] # restore correct order

tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title","Chromosome.location"))
write.table(tT, file=stdout(), row.names=F, sep="\t")
```

Then click “top 250.” This reports the most regulated transcripts, ranked by smallest *p* value (Fig. 11.2b). By default, the values have been  $\log_2$  transformed.

(a) GEO2R: defining groups for analysis

Group	Accession	Source name
TS21	GSM22509	fetal cerebrum (18 to 19 weeks gestation)
TS21	GSM22510	fetal cerebrum (18 to 19 weeks gestation)
TS21	GSM22511	fetal cerebrum (18 to 19 weeks gestation)
TS21	GSM22512	fetal cerebrum (18 to 19 weeks gestation)
euploid	GSM22523	fetal cerebrum (18 weeks gestation)
euploid	GSM22524	fetal cerebrum (18 weeks gestation)
euploid	GSM22526	fetal cerebrum (18 weeks gestation)
euploid	GSM22527	fetal cerebrum (18 weeks gestation)
euploid	GSM22583	fetal cerebellum (18-19 weeks gestation)
euploid	GSM22584	fetal cerebellum (18-19 weeks gestation)

(b) GEO2R: limma results for differentially expressed transcripts (chromosome 21 genes indicated by →)

ID	adj.P.Val	P.Value	B	logFC	Gene.symbol	Gene.title	Chromosome.loc...
► 206777_s_at	0.0304	0.00000136	-1.54	2.223	CRYBB2P1///CRYBB2	crystallin, beta B2 p...	22q11.2-q12.1///22q...
► 201123_s_at	0.0402	0.00000454	-1.76	1.845	EIF5A	eukaryotic translatio...	17p13-p12
► 201122_x_at	0.0402	0.00000542	-1.79	0.902	EIF5A	eukaryotic translatio...	17p13-p12
► 65588_at	0.3157	0.00005667	-2.29	0.721	SNHG17	small nucleolar RNA...	20q11.23
► 200642_at	0.3545	0.00008342	-2.37	-0.746	SOD1	superoxide dismuta...	21q22.11 →
► 212269_s_at	0.3545	0.00011133	-2.44	-1.147	MCM3AP	minichromosome m...	21q22.3 →
► 212292_at	0.3545	0.00011234	-2.44	1.2	SLC7A1	solute carrier family ...	13q12.3
► 202325_s_at	0.3545	0.00012727	-2.47	-0.731	ATP5J	ATP synthase, H+ tr...	21q21.1 →
► 218386_x_at	0.3577	0.00014739	-2.51	-0.711	USP16	ubiquitin specific pe...	21q22.11 →
► 202671_s_at	0.3577	0.00016052	-2.53	-1.032	PDXK	pyridoxal (pyridoxin...	21q22.3 →
► 200818_at	0.62	0.00033579	-2.71	-0.666	ATP5O	ATP synthase, H+ tr...	21q22.11 →
► 210667_s_at	0.62	0.00035394	-2.72	-0.644	C21orf33	chromosome 21 op...	21q22.3 →
► 203635_at	0.62	0.00040089	-2.75	-0.8	DSCR3	Down syndrome criti...	21q22.2 →
► 202937_x_at	0.62	0.00040754	-2.76	-1.849	RRP7A	ribosomal RNA proc...	22q13.2
► 202217_at	0.62	0.00044204	-2.78	-0.619	C21orf33	chromosome 21 op...	21q22.3 →
► 216954_x_at	0.62	0.00049116	-2.8	-0.838	ATP5O	ATP synthase, H+ tr...	21q22.11 →
► 200740_s_at	0.62	0.0005004	-2.81	-0.649	SUMO3	small ubiquitin-like ...	21q22.3 →

**FIGURE 11.2** GEO2R at NCBI. GEO Datasets may be analyzed using this web-based tool. (a) Two or more groups are defined, and each row is assigned to a group. We enter TS21 (for 11 trisomic samples) and euploid (also  $n = 11$ ). (b) GEO2R produces a list of significantly, differentially expressed transcripts based on results from the limma R package. Arrows indicate regulated transcripts expressed from genes on chromosome 21. ID: Affymetrix probeset identifier; adj.P.Val: adjusted probability value (see text); P.Value: probability value; B: B-statistic, that is, log-odds that the transcript is differentially expressed; logFC:  $\log_2$  fold change between two experimental conditions. Courtesy of GEO2R.

While the “top 250” button gives a convenient view of the most significantly regulated transcripts, you can also choose an option to download all the data. For this experiment there are 22,283 transcripts (available as Web Document 11.1). Of these, 68 have probe names beginning AFFX and are controls you may wish to remove. A total of 37 of these have gene symbols such as *SEPT2* that are irreversibly and inappropriately converted to dates by Microsoft® Excel (it is therefore critical that you work with a text editor rather than Excel or Word). Typical of many microarray platforms, 1207 of these entries lack both gene symbols and gene titles although they do have probeset identifiers.

## GEO2R Identifies the Chromosomal Origin of Regulated Transcripts

Which of the regulated transcripts are derived from chromosome 21 genes? Using the “select column” option we add the chromosome location. Note that of the top entries having unique gene symbols, 10 out of 15 are from genes assigned to chromosome 21 (Fig. 11.2b, arrows). This tremendous overrepresentation can be explained by the presence of an extra copy of chromosome 21 (consisting of three copies rather than the usual two), leading to increased levels of chromosome 21 RNA transcripts relative to euploid controls (Mao *et al.*, 2005).

GEO2R gave us the option to show the chromosome location. An alternative approach would be to export a file of the gene symbols of interest, then upload that to BioMart (or use `biomaRt` in R) to determine the chromosomal origins of the regulated transcripts.

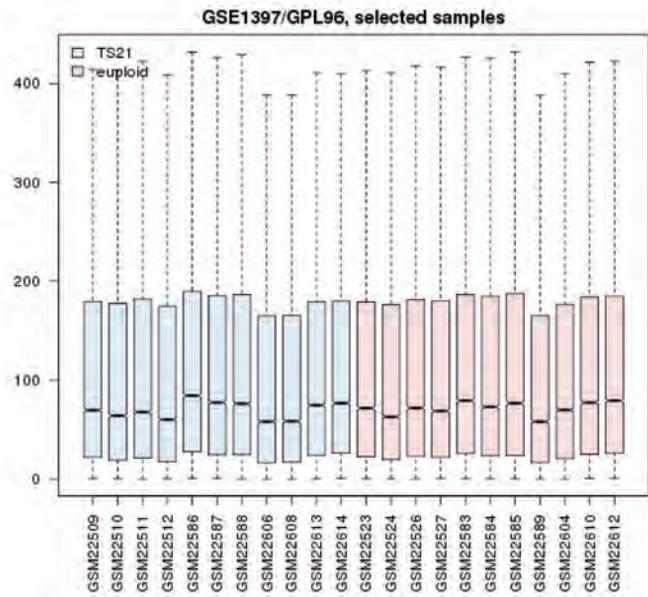
Is the occurrence of 10 chromosome 21-derived transcripts out of 15 statistically significant? Try using a Fisher exact test on a two-by-two matrix. There are 12 chromosome 21-derived transcripts, and 29 non-chromosome 21 transcripts (the total is 41 genes). There are 671 chromosome 21 genes and 53,834 non-chromosome 21 genes in GRCh38. We use the `fisher.test` program in the `stats` R package to evaluate whether the 12 transcripts from chromosome 21 are more than we expect by chance.

```
> mychr21data <- matrix(c(10,5,671,53834),2,2)
> mychr21data
     [,1]   [,2]
[1,]    10     671
[2,]     5   53834
> fisher.test(mychr21data, alternative = "two.sided")
Fisher's Exact Test for Count Data
data: mychr21data
p-value = 2.458e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 49.72036 589.80298
sample estimates:
odds ratio
159.9635
```

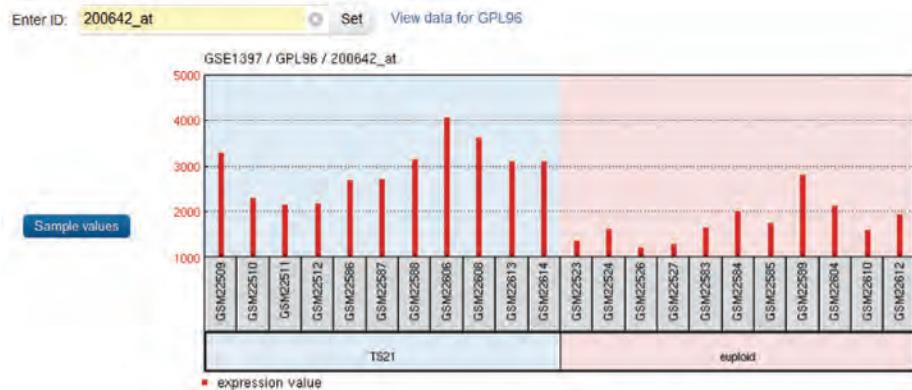
The *p* value is very small ( $2 \times 10^{-16}$ ), indicating that we can reject the null hypothesis. Would the finding have been significant if 2 (instead of 12) out of 15 regulated transcripts had been assigned to chromosome 21?

```
> mychr21data2 <- matrix(c(2,13,671,53834),2,2)
> fisher.test(mychr21data2, alternative = "two.sided")
Fisher's Exact Test for Count Data
data: mychr21data2
p-value = 0.01436
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.349511 54.636938
sample estimates:
odds ratio
12.34114
```

(a) GEO2R: boxplot of normalized expression data for trisomy 21 (n=11) and euploid (n=11) samples



(b) GEO2R: profile of normalized expression data for SOD1 transcripts across 22 samples



**FIGURE 11.3** GEO2R results. (a) GEO2R invokes R scripts that generate a boxplot, here showing that the samples (x axis) have been normalized to have comparable intensity values (y axis). The box-and-whisker plot has upper and lower hinges (at the first and third quartile) and whiskers (corresponding to outlier data points). (b) A gene is selected (*SOD1*). GEO2R displays the expression values for this probeset across trisomy 21 and euploid samples. As a group the TS21 samples have elevated levels of *SOD1* mRNA relative to euploid samples. Courtesy of GEO2R.

Yes, the  $p$  value is 0.01, so this difference is significant. Our test is two-sided: we specify that we look for a difference but we do not say the direction of the effect. Since there is an extra copy of chromosome 21 it would be reasonable to perform a one-sided test, yielding more statistical power.

### GEO2R Normalizes Data

We can view a values distribution plot in the form of a boxplot (Fig. 11.3a). This displays the samples (x axis) and the distribution of values (y axis). These are fairly well normalized across their medians, indicating that they are appropriate for comparisons between groups. (When we use R packages directly we'll see boxplots before and after normalization in “Reading CEL Files and Normalizing with RMA” below.) The R script tab shows the commands used to generate this boxplot in R:

```

# Boxplot for selected GEO samples
library(Bioconductor)
library(GEOquery)
# load series and platform data from GEO
gset <- getGEO("GSE1397", GSEMatrix =TRUE)
if (length(gset) > 1) idx <- grep("GPL96", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]
# group names for all samples in a series
sm1 <- c("G0", "G0", "G0", "G0", "G1", "G1", "G1", "G1", "G1", "G1", "G0", "G0",
"G0", "G1", "G0", "G0", "G1", "G1", "G0", "G0", "X", "X", "X", "X", "X")
# eliminate samples marked as "X"
sel <- which(sm1 != "X")
sm1 <- sm1[sel]
gset <- gset[, sel]
# order samples by group
ex <- exprs(gset)[, order(sm1)]
sm1 <- sm1[order(sm1)]
fl <- as.factor(sm1)
labels <- c("TS21", "euploid")
# set parameters and draw the plot
palette(c("#dfeaf4", "#f4fdf", "#AABCC"))
dev.new(width=4+dim(gset)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(gset))))/2), 4, 2, 1))
title <- paste ("GSE1397", '/', annotation(gset), " selected samples", sep
='')
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)
legend("topleft", labels, fill=palette(), bty="n")

```

These R commands are available as a text file in Web Document 11.2. You can enter them into R and modify them to change the properties of the boxplot (e.g., which samples are plotted or which colors are used), and you can obtain documentation on the libraries and commands that were used. For example, after you load `> library(GEOquery)` enter `> ?getGEO` for help on that function.

We say that the data are “normalized” by GEO2R; what does this mean? Normalization refers to the process of correcting two or more datasets prior to comparing their gene expression values.

As an example of why it is necessary to normalize microarray data, dyes are incorporated into cDNA (or cRNA) samples with different efficiencies. There may also be differences in amounts of input RNA, in DNA quality, washing efficiency, or signal detection. Without normalization, it would not be possible to accurately assess the relative expression of samples. Genes that are actually expressed at comparable levels would have a ratio different from 1 (when considering unlogged data) or different from 0 (for logged data; see below). Normalization is also essential to allow the comparison of gene expression across multiple microarray experiments. Normalization is therefore required for one-channel microarray experiments such as the Affymetrix platform which uses one sample per microarray (chip). Normalization is also needed for two-channel arrays which typically involve separate labeling of samples with Cy3 and Cy5 dyes (interpreted as green and red), then cohybridization on the surface of an array.

There are many approaches to normalization. A simple idea is to measure the background intensity and subtract that from the signal for each probeset (or other element on the surface of a microarray). This background may be constant across the surface of an array or it may vary locally. Another idea is to apply a global normalization to raw array element intensities so that the average ratio for gene expression is 1. The main assumption of microarray data normalization is that the average gene does not change in its expression level in the biological samples being tested. The procedure for global normalization can be applied to two-channel datasets (e.g., Cy3- and Cy5-labeled samples) or one-channel datasets (e.g., Affymetrix chip data). As an example, if the mean expression value for samples in the green channel is 10,000 arbitrary units and the mean value for samples in

RMA was introduced by Terry Speed, Rafael Irizarry and colleagues. Affymetrix arrays include both perfect match oligonucleotide probes, as well as mismatch probes containing a single base mismatch that are used to estimate background. RMA considers only perfect match values, because mismatch values contributed noise and can have values greater than perfect match probes across as much as one-third of a microarray.

RMA is available through the `affy` package of BioConductor, and has also been incorporated into a variety of commercial microarray data analysis packages. GCRMA was developed by Zhijin Wu and Rafael Irizarry.

Sometimes variance present in gene expression data is not constant across the range of element signal intensities. This variation represents an artifact that can be addressed by global and also local normalization processes, which correct bias and variance that are nonuniformly distributed across absolute signal intensity. Many software packages can correct for such variance. One of these, Standardization and Normalization of Microarray Data (SNOMAD), was written by Carlo Colantuoni when he was a graduate student in the Pevsner lab. SNOMAD is a web-based tool written in R and available at <http://pevsnerlab.kennedykrieger.org> (WebLink 11.2); see Colantuoni *et al.* (2002a, b).

the red channel is 5000, then the expression value for each gene in the red channel would be multiplied by 2. If the data are not log transformed, the mean ratio is then 1. Once the data are log transformed, the mean ratio is 0.

Another possible approach is to normalize all expression values to a set of “housekeeping genes” that are represented on the array. Housekeeping genes might include  $\beta$ -actin and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and dozens of others. Each gene expression value in a single array experiment is then divided by the mean expression value of these housekeeping genes. A major assumption of this approach is that such genes do not change in their expression values between two conditions. In many cases, this assumption fails.

Affymetrix introduced the MAS5 normalization method which normalizes each array independently. Perfect match (PM) probesets consist of oligonucleotides affixed to the array surface that bind the fluorescently labeled sample if transcripts corresponding to particular probesets are expressed. Mismatch (MM) oligos are designed with an intentional mismatch and are used to define background signal. MAS5 calculates the difference PM – MM to obtain a robust average, summarizing the signal from a set of probesets that span a gene.

Robust multiarray analysis (RMA) is a method of performing background subtraction, normalization, and averaging of probe-level feature intensities extracted from .CEL files using the Affymetrix platform (Irizarry *et al.*, 2003). It includes steps for background correction, quantile normalization across arrays, a probe-level model fit to each probeset across multiple arrays, and quality assessment.

Quantile normalization is a nonparametric approach that produces the same overall distribution for all the arrays within an experiment (Bolstad *et al.*, 2003). *Parametric* tests are applied to datasets that are sampled from a normal (Gaussian) distribution; common parametric tests include the *t*-test and ANOVA (discussed in “Performing ANOVA in Partek”). *Nonparametric* tests do not make assumptions about the population distribution. They rank the outcome variable (here, gene expression measurements) from high to low and analyze the ranks. In quantile normalization, for each array each signal intensity value is assigned to a quantile. We then consider a pooled distribution of each probe across all chips: for each probe, the average intensity is calculated across all the samples. Normalization is performed for each chip by converting an original probeset value to that quantile’s value.

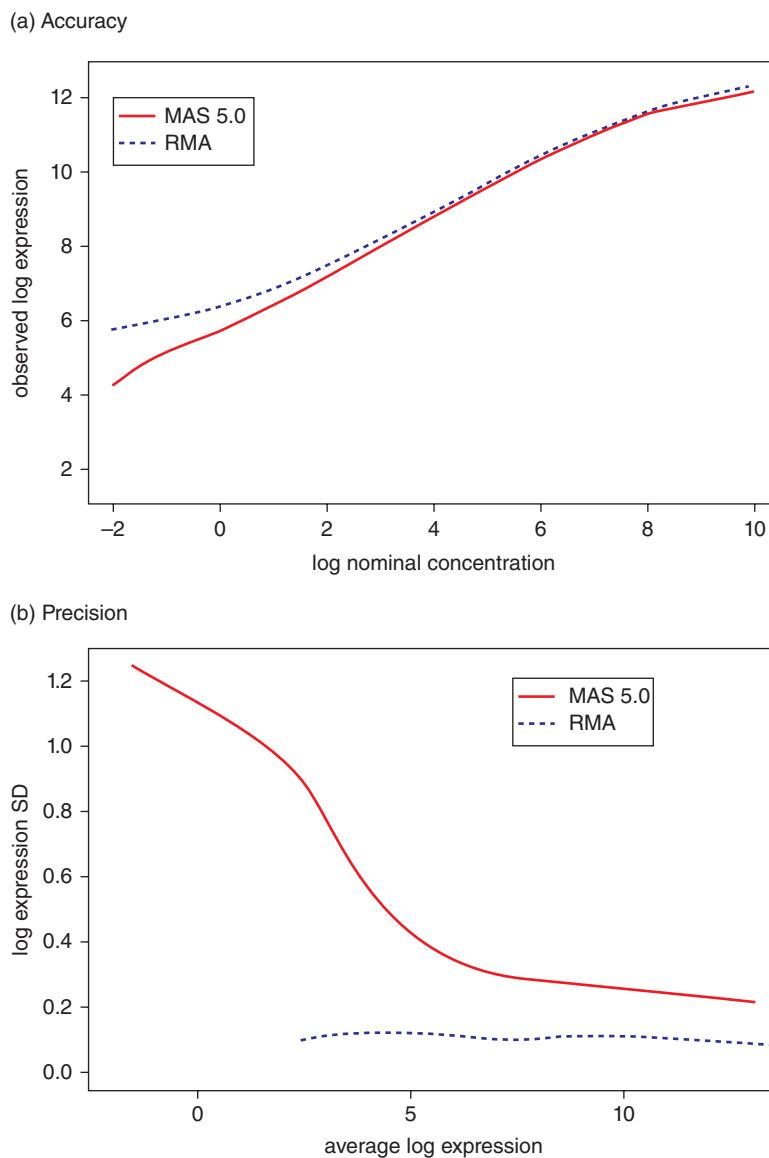
The RMA background correction step includes a convolution model in which the observed signal for each probeset is broken into components of true signal and noise. GCRMA further introduces an adjustment for the presence of nonspecific hybridization that improves accuracy (relative to RMA) while maintaining large gains in precision (relative to other preprocessing techniques). You can use RMA and GCRMA on a trisomy 21 dataset using the `affy` Bioconductor package we introduce in “Microarray Analysis Method 3” below.

## GEO2R uses RMA Normalization for Accuracy and Precision

RMA has accuracy comparable to MAS 5.0 software (Affymetrix; Fig. 11.4a), while its precision is far greater (Fig. 11.4b). We define accuracy and precision next.

Preprocessing steps are designed to improve accuracy of gene expression measurements by lowering bias, and to improve precision by lowering the variance. Accuracy is estimated two ways: by using spike-in samples of known concentrations of RNA, or by diluting known concentrations of RNA. These methods allow an objective assessment of the true positive measurements. A more accurate normalization method produces results that are closer to the “gold standard” truth. The precision is estimated by using replicate measures of the same sample. Samples with great precision have little variance.

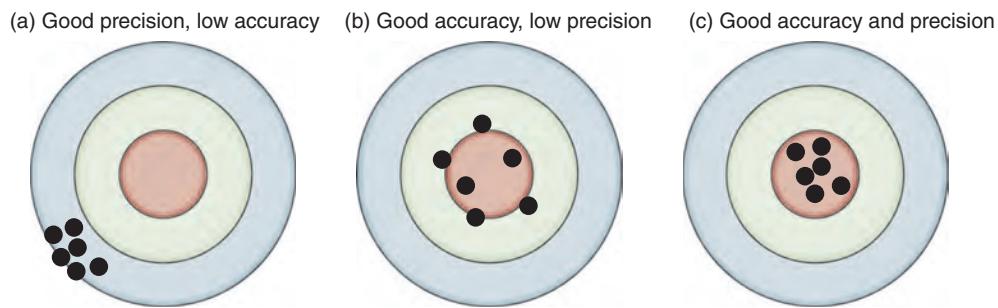
We can think of accuracy and precision in terms of a series of arrows hitting a target: accuracy refers to how close the arrows are to the bull’s-eye, while precision refers to how



**FIGURE 11.4** Improvements in accuracy and precision using RMA (relative to MAS 5.0 software from Affymetrix). (a) Accuracy is measured by plotting known concentrations of RNA ( $x$  axis) versus observed concentrations ( $y$  axis). The two methods are comparable. RMA performs slightly worse at low concentrations, a situation that is improved by the GCRMA algorithm. (b) Precision is measured by plotting the average log expression value ( $x$  axis) versus the log expression standard deviation ( $y$  axis). MAS 5.0 software yields a high standard deviation, particularly for transcripts expressed at low levels, while RMA has a dramatically improved measurement across a broad range of signal intensities.

reliably the arrows hit the same spot (Fig. 11.5; Cope *et al.*, 2004). Irizarry *et al.* (2006) performed a benchmarking study using 31 algorithms for the analysis of Affymetrix probe sets. They concluded that background correction has a large effect on performance, and tends to improve accuracy but worsen precision. The RMA and GCRMA algorithms have consistently performed well in terms of both accuracy and precision and have emerged as leading approaches for the preprocessing of Affymetrix gene expression data. RMA is employed within GEO2R.

Skewing sometimes reflects experimental artifacts such as the contamination of one RNA source with genomic DNA or rRNA. (Such contaminating nucleic acid could bind to elements on the microarray.) Another source of artifact is the use of unequal amounts of fluorescent probes on the microarray.



**FIGURE 11.5** Accuracy and precision. (a) Good precision is characterized by reproducible results. It is assessed by repeated measurements of samples (technical replicates). (b) Good accuracy is characterized by measurements that correspond to an independently known result. It can be assessed by measurement of known (“spiked in”) concentrations of RNA to an experiment, or by measuring dilutions of known concentrations of RNA. (c) A goal of preprocessing algorithms is to achieve both accuracy and precision.

### Fold Change (Expression Ratios)

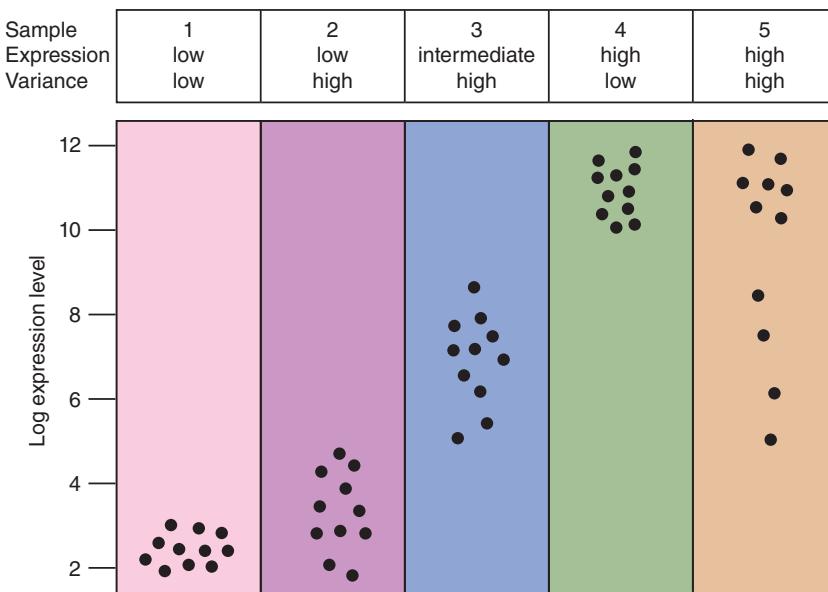
Let’s return to our GEO2R results. The top 250 results are sorted by probability value. Fold change is also reported, reflecting the magnitude of the difference in RNA transcript levels between the mean of our group of euploid samples and the mean of the group of trisomic samples. We can click on any of the top 250 results and see a bar chart showing the expression levels for all 22 samples in two groups. For SOD1 transcripts, the set of trisomy 21 samples clearly have higher overall expression values than the euploid samples (**Fig. 11.2b, 11.3b**).

Some investigators apply an arbitrary minimum fold change to the list of significantly regulated transcripts such as 1.5-fold or 2-fold. This has the benefit of allowing you to focus on the most dramatic changes (in terms of magnitude). It may also make sense from a biological point of view to avoid focusing on expression changes that are minor (e.g., a 1.1-fold up- or down-regulation) even if statistically significant based on probability values. Keep in mind that  $p$  values offer a popular approach to defining statistical significance, but fold change does not. Given a large fold change (even 10-fold) between two groups, there may or may not be statistical significance (depending on the variability in the measurements; see “GEO2R Performs >22,000 Statistical Tests” below; **Fig. 11.6**). Given a statistically significant difference for a given gene, the fold change may be of any magnitude from large to trivially small.

### GEO2R Performs >22,000 Statistical Tests

The goal of inferential statistical analysis of microarray data is to test the hypothesis that some genes are differentially expressed in an experimental comparison of two or more conditions. For each gene (or probeset) on a microarray we perform a statistical test.

Consider our trisomy 21 versus euploid experiment. There are >22,000 transcripts represented on the array. For each transcript there are 11 measurements in the experimental group and 11 measurements in the control group. For each of the >22,000 transcripts, we formulate a null hypothesis  $H_0$  that there is no difference in signal intensity across the two conditions being tested. The alternative hypothesis  $H_1$  is that there are differences in transcript levels. We define and calculate a test statistic which is a value that characterizes the observed gene expression data. We accept or reject the null hypothesis based on the results of the test statistic. The probability of rejecting the null hypothesis when it is true is the significance level  $\alpha$ , which in science is typically set at  $p < 0.05$ . Under the null hypothesis, for a set of gene expression intensity values in two conditions the data are normally distributed with mean 0 and standard deviation  $\sigma$  equal to 1. The standard deviation  $\sigma$  can be estimated using the sample standard deviation  $s$ .



**FIGURE 11.6** Transcript-specific variance is addressed by a t-test. For five hypothetical transcripts the log expression values are plotted (y axis). Transcript 1 has a low absolute expression level and low variance upon repeated measurements in biological replicate samples, while transcript 2 has a low expression level and relatively high variance. Transcript 3 is expressed at intermediate levels, while transcripts 4 and 5 are expressed at high levels, with transcript 3 having low variance and transcript 4 having high variance. Each RNA transcript has a characteristic property of its expression level (although this may vary dramatically across body regions and across developmental stages, or even between experimental conditions). When we compare expression levels for two transcripts, a *t*-test accounts for the difference in mean between the two measurements, and also provides an analysis of the variation in expression levels.

The mean and standard deviation for the expression of each gene represented on the microarray can be calculated. A *t*-test is performed to test the null hypothesis that there is no difference in gene expression levels, considered one gene at a time, between the two populations. The approach is to compute the average expression value for each gene from control ( $x_1$ ) and experimental ( $x_2$ ) conditions and take the absolute value of the difference, providing as a numerator the magnitude of the expression change. We also need to estimate the variance ( $\sigma$ ), providing as a denominator the amount of noise in the measurements. The average for each sample (e.g.,  $\bar{x}_1$ ) is given by:

$$\bar{x}_1 = \frac{1}{M} \sum_{i=1}^M x_i. \quad (11.1)$$

The variance (or square of the standard deviation,  $s^2$ ) for  $x_1$  is given by:

$$s_{x1}^2 = \frac{1}{M-1} \sum_{i=1}^M (x_i - \bar{x})^2. \quad (11.2)$$

The *t*-test essentially measures the signal-to-noise ratio in your experiment by dividing the signal (difference between the means) by the noise (variability estimated in the two groups).

$$t\text{-statistic} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{x1}^2}{M} + \frac{s_{x2}^2}{N}}}. \quad (11.3)$$

From the *t*-statistic we can calculate a *p* value. This allows us to either reject or accept the null hypothesis that the control and experimental conditions have equal gene expression values (i.e., the null hypothesis is that there is no differential expression). For a *t*-test that provides a *p* value of 0.01, this means that one time in 100 the observed difference between

the control and experimental groups will be observed by chance, and we can safely reject the null hypothesis. **Figure 11.2b** presents the results of *t*-tests having the smallest *p* values in our study. For example, SOD1 having a *p* value of 0.00008342 (close to 0.0001) indicates that the observed difference in expression levels between trisomic and euploid samples will occur by chance only 1 time in 10,000. This might lead us to reject the null hypothesis. (There is also an adjusted *p* value of 0.3545, explained in “GEO2R Offers Corrections for Multiple Comparisons” below, which would suggest there is no significant difference between SOD1 expression levels in our trisomy and euploid groups.)

We can think about the usefulness of the *t*-test by considering hypothetical RNA transcripts for which we have gene expression measurements from 11 samples (**Fig. 11.6**). Genes 1 and 2 are expressed at low levels, and there is considerable “noise” (variability) in the measurement of gene 2. In a comparison of genes 1 and 2, both the mean values (which may differ) and the variability in the measurement (which shows overlap) are accounted for in a *t*-test. The difference between sample 3 (with its elevated expression level and moderate variability) and sample 1 might be significant, but sample 3 is not likely to be significantly, differentially regulated in comparison to sample 2 with its higher variance. For genes expressed at high levels, the variance may also be small (gene 4) or large (gene 5). For any sample displaying large variance, a relatively large sample size will be necessary to achieve sufficient statistical power to reject the null hypothesis. Here are three *t*-tests made with the `t.test` program in the `stats` package of R. The first one compares two groups with means of 10 and 20, and the *p* value is very small:

```
> t.test(c(8,12,9,11), y = c(18,22,19,21))
   Welch Two Sample t-test
data: c(8, 12, 9, 11) and c(18, 22, 19, 21) # use c to concatenate
# several numbers
t = -7.746, df = 6, p-value = 0.0002433
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.15895 -6.84105
sample estimates:
mean of x mean of y
10 20
```

The second *t*-test involves a second group with a mean of 12 but, because of the variability in the first group, the *p* value is not significant.

```
> t.test(c(8,12,9,11), y = c(12,12,12))
# Note that we selected n=4 in one group and n=3 in the other
   Welch Two Sample t-test
data: c(8, 12, 9, 11) and c(12, 12, 12)
t = -2.1909, df = 3, p-value = 0.1162
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.9051627 0.9051627
sample estimates:
mean of x mean of y
10 12
```

Finally, in the last set the group means are 10 and 14 and the *p* value is significant.

```
> t.test(c(8,12,9,11), y = c(14,14,14))
   Welch Two Sample t-test
data: c(8, 12, 9, 11) and c(14, 14, 14)
t = -4.3818, df = 3, p-value = 0.02201
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.905163 -1.094837
sample estimates:
mean of x mean of y
10 14
```

The power of a statistical test is the fraction of true positives that will be detected. This is a value between 0 and 1 defined as  $1-\beta$ ;  $\beta$  is the probability of concluding there is no significant difference between two means, when in fact the alternative hypothesis is true. ( $\beta$  is the same as the probability of making a type II error.) The larger the sample size, the larger the power. The R package `pwr` produces power estimates for microarray (or other) experiments, as does the `power.t.test` function in the `stats` package. One parameter is set to null and is determined from the other input parameters. Given a sample size of 11 in each group and a difference in the means of 1, at a significance level of 0.05, what is the power of a *t*-test?

```
> power.t.test(n = 11, delta = 1 , sig.level = 0.05)
Two-sample t test power calculation
n = 11 # note that n is the number in each group
delta = 1
sd = 1
sig.level = 0.05
power = 0.6070844
alternative = two.sided
```

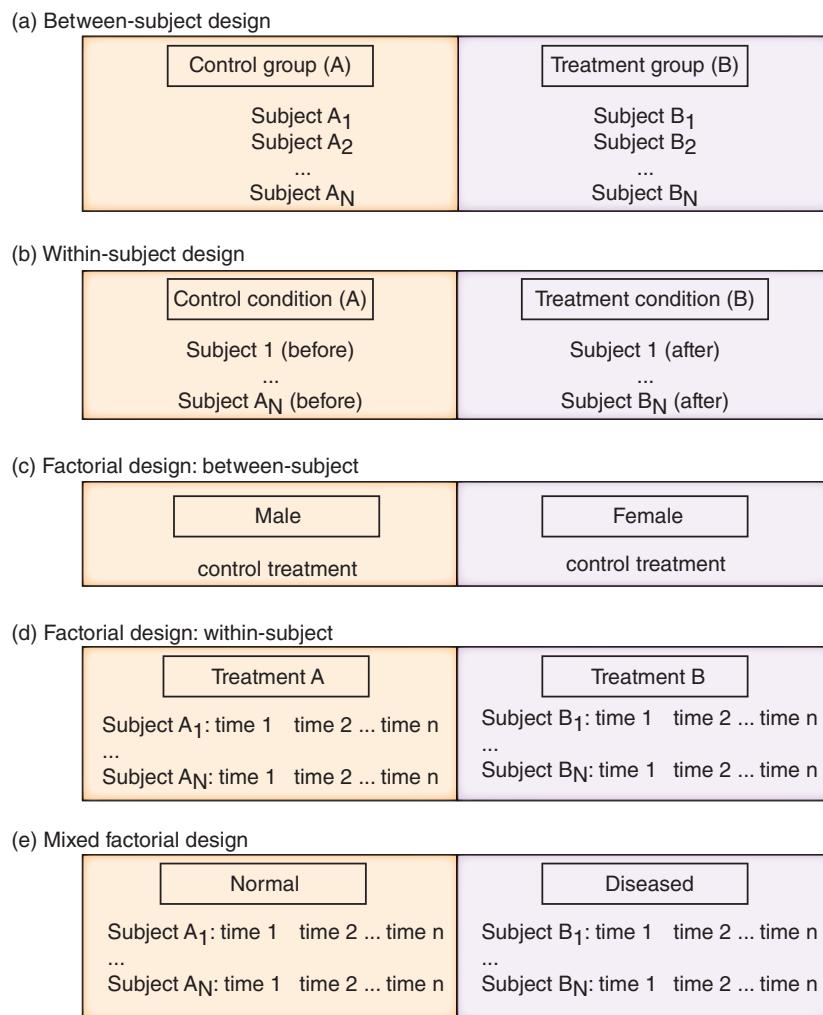
To achieve a power of 0.9 in a *t*-test, how many samples are needed per group? The answer is 22.

```
> power.t.test(power = .90, delta = 1)
Two-sample t test power calculation
n = 22.0211
delta = 1
sd = 1
sig.level = 0.05
power = 0.9
alternative = two.sided
```

An assumption of parametric tests such as the *t*-test approach is that gene expression values are normally distributed. If so, the *t*-statistic follows a distribution that allows us to calculate a set of *p* values. (An alternate assumption is that, for very large numbers of replicates, the *t*-statistic is normally distributed with mean 0 and standard deviation of 1, and again we can compute *p* values. In practice, very larger numbers of replicates are rarely available for microarray studies.)

Nonparametric tests rank the outcome variables and do not assume a normal distribution. These tests, such as the Mann–Whitney and Wilcoxon, are less influenced by data points that are extreme outliers. Such tests are not commonly applied to microarray data. Other approaches have been implemented such as Bayesian analysis of variance (Ishwaran *et al.*, 2006; see also the `limma` package in “Microarray Analysis Method 3” below).

The test that is used depends on the experimental paradigm. Some examples of experimental designs are shown in **Figure 11.7**. For a between-subject design (**Fig. 11.7a**) there are two groups. In this experimental design it is necessary to control for confounding factors such as differences in age, gender, or weight between individuals in the two groups. For a within-subject design (**Fig. 11.7b**), a paired *t*-test would be used to test for the differences in mean values between two sets of measurements on paired samples. An example of this is a study measuring gene expression in cancer biopsy samples before and after drug treatment. Here the covariates (“nuisance variables”) such as age and gender are internally controlled. A biostatistician can help a biologist to choose an appropriate design before an experiment is conducted. The statistician Ronald Fisher (1890–1962) famously stated: “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”



**FIGURE 11.7** Examples of experimental design for microarray experiments involving gene expression profiling. Most such microarray experiments are designed to test the hypothesis that there are significant biological gene expression differences between samples as a function of factors such as tissue type (normal versus diseased or brain versus liver), time, or drug treatment. (a) A between-subject design must control for confounding factors such as age, gender, or weight. (b) A within-subject design removes genetic variability and can be used to measure gene expression before then after some treatment. (c) A two-way between-subject design allows the measurement of differences between both treatment and control conditions, and another factor such as gender. (d) A within-subject factorial design might be used to study two treatments over time. (e) In a mixed factorial design there is both a between-subject design (e.g., normal versus diseased tissue) and a within-subject design (e.g., gene expression measurements over time).

### GEO2R Offers Corrections for Multiple Comparisons

What  $p$  value cutoff is appropriate to establish statistical significance? If you measure the expression values for 20,000 transcripts, you can expect to find differences in 5% of them (1000 transcripts) purely by chance that are nominally significant at the  $p < 0.05$  level. If you hypothesized *a priori* that one specific gene was significantly regulated, then this  $\alpha$  level would be appropriate. However, for 20,000 measurements it is necessary to apply some conservative correction to account for the thousands of repeated, independent measurements you are making. There are two problems we want to avoid. Type I errors (false positive results) involve concluding that a transcript is differentially expressed when it is

not; the null hypothesis is true but is inappropriately rejected. Type II errors (false negative results) involve failing to identify a truly regulated transcript; the null hypothesis that is actually false is not rejected as it should be. A  $p$  value is defined as the minimum false positive rate at which an observed statistic is categorized as significant.

There are several approaches to accounting for the problem of multiple comparisons. At one extreme, some researchers apply a conservative Bonferroni correction in which the  $\alpha$  level for statistical significance is divided by the number of measurements taken (e.g.,  $p < 0.05/20,000$  is set as the criterion for significance). This correction is considered too severe. A more commonly used approach to the multiple comparisons correction problem is to adjust the false discovery rate (FDR), defined:

$$\text{FDR} = \frac{\text{\# false positives}}{\text{\# called significant}}. \quad (11.4)$$

The FDR represents the rate at which genes identified as significantly regulated are not. For an FDR of 0.05, 5% of those transcripts that are called significant are false positives. For 100 significantly regulated genes and an FDR of 8%, 8 genes out of 100 are expected to represent false positive results.

GEO2R offers several multiple test corrections. The Benjamini and Hochberg (1995) FDR is the default: it is the most commonly used adjustment method, and limits the number of false positives while finding true positives (statistically significant transcripts). Five other multiple comparison correction methods are available, including the Bonferroni correction. This divides the threshold for significance  $\alpha$  by the number of measurements  $k$ ; for a microarray experiment with 10,000 measurements  $\alpha$  would be adjusted from 0.05 to  $5 \times 10^{-6}$ . The equation relating a BLAST expect value to a score (see Chapter 4) includes the equivalent of a Bonferroni correction because  $E$  is divided by  $mn$  (i.e., the sizes of the query and the database, corresponding to the number of measurements) to obtain the score.

Note that it is not appropriate to filter by fold change before performing a statistical test such as ANOVA. This may reduce the multiple testing penalty, but it introduces bias (van Iterson *et al.*, 2010). At the same time, if using fold change as a criterion for ranking genes along with a nonstringent  $p$  value cutoff, lists of differentially expressed genes become more reproducible across laboratories. This is the conclusion of Shi *et al.* (2008) who analyzed datasets from the MicroArray Quality Control (MAQC) project, introduced in the following section below.

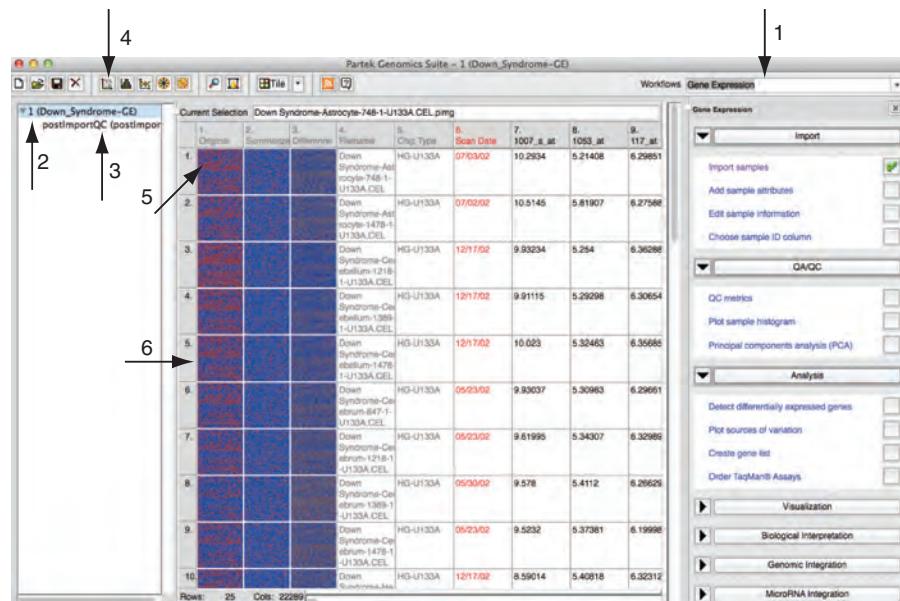
## MICROARRAY ANALYSIS METHOD 2: PARTEK

For our second method we introduce a commercial software package, Partek Genomics Suite. This is a package that requires an annual license, and may be compared to R packages (described in this chapter): some researchers prefer to use free, open-source software such as R. Such software is often designed by academic experts in bioinformatics and/or biostatistics. Advantages of Partek include:

- a minimal learning curve (in contrast to R which most scientists concede has a very steep learning curve);
- packaging of a suite of tools into a user-friendly graphical user interface (GUI), including guided workflows for various tasks such as microarray analysis; and
- dedicated customer support (it should be noted that the culture of the R user community is that authors of software packages tend to be highly responsive to users' queries, and users' forums are available offering excellent support).

In the following analysis of the Down syndrome dataset we select a gene expression workflow within Partek Genomics Suite that provides step-by-step guidance on

Partek software (both Partek Genomics Suite that we describe next and Partek Flow for NGS data) is available from <http://www.partek.com> (WebLink 11.3). Other prominent commercial packages include Nexus Expression and ImaGene from BioDiscovery® (<http://www.biodiscovery.com>, WebLink 11.4), GeneSpring (<http://www.chem.agilent.com/>, WebLink 11.5), GeneTraffic (<http://www.iobion.com/>, WebLink 11.6), and Avadis from Strand Life Sciences (<http://www.avadis-ngs.com/>, WebLink 11.7).



**FIGURE 11.8** A Partek spreadsheet includes rows (here there are 25 rows each with data from a sample) and columns (the first columns include information about each sample, followed in this case by >22,000 columns each corresponding to a microarray probeset). The cells in this spreadsheet consist of log<sub>2</sub> expression values that correspond to the amount of detected signal, interpreted as expression levels. A workflow for gene expression is shown (arrow 1). The main spreadsheet may be selected (arrow 2) or another spreadsheet such as the quality control data (arrow 3). A principal components analysis plot (Fig. 11.10) may be invoked (arrow 4). Images of the microarray surface are clickable (arrow 5; see Fig. 11.9b). Additional functions may be accessed by clicking row or column headers (e.g., arrow 5), such as plotting or annotating row or column elements. Courtesy of Partek Inc.

how to import data, perform quality control, then perform a series of analyses (Fig. 11.8, arrow 1).

## Importing Data

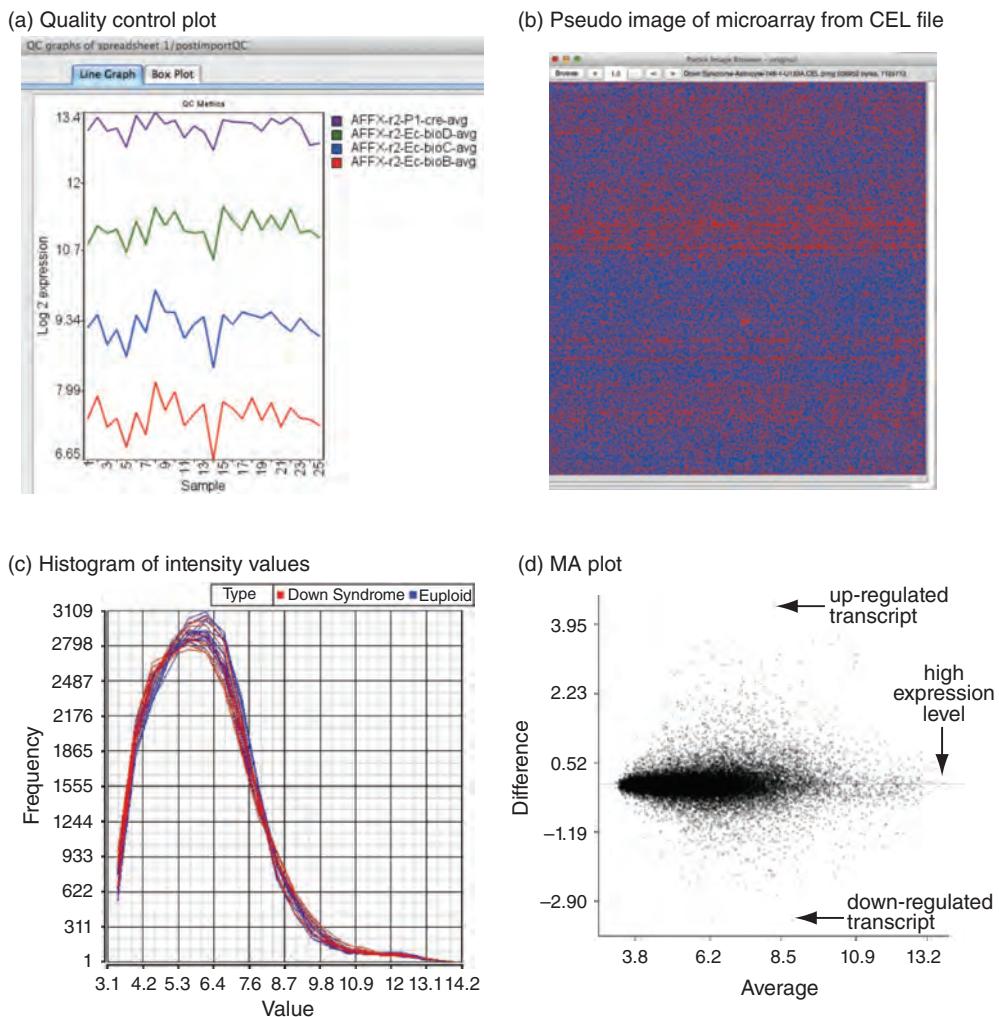
The CEL files are provided at <http://bioinfbook.org/chapter11>. When available at NCBI GEO, you can import them directly using the NCBI GEO download option. CEL files contain measured expression levels for a series of probe sets, each of which corresponds to a gene or portion of a gene. CEL files also contain locations for these measurements.

Partek Genomics Suite can accept over a dozen data formats, ranging from text files and Excel spreadsheets to GEO data (e.g., a GSE series or GSM samples). It also supports next-generation sequence data such as BAM files. For gene expression analysis we will import a set of Affymetrix CEL files.

We browse to a folder containing CEL files, select the ones we wish to analyze, and import them using either RMA or GCRMA for normalization (introduced above). Library files are required, containing annotation information for this particular microarray platform, and Partek identifies whether they are available in a specified library file folder. If not, it automatically downloads and stores the required files from the internet.

## Quality Control

Once CEL files are imported, a postImportQC spreadsheet provides quality control metrics (Fig. 11.9a) including a series of boxplots similar to that shown in Figure 11.3a. The QC data are included in a child spreadsheet (Fig. 11.8, arrow 3). The main data sheet has 25 rows (one per sample) and 22,289 columns (6 with information about these samples and then 22,283 columns containing log<sub>2</sub> expression values measured from the microarray). The first three columns include .CEL file pseudo chip images. By double-clicking one, the image enlarges showing the surface of the array for that



**FIGURE 11.9** Visualization of quality control data from Partek. (a) Quality control plots include log<sub>2</sub> expression (y axis) across samples (x axis) for spike-in controls. (b) Pseudo image of a microarray surface from a CEL file. Artifacts (“squashed bug phenomenon”) may be identified visually and when necessary samples may be discarded. (c) Histogram of intensity values may also reveal outliers. (d) MA plot shows mean intensity values of log<sub>2</sub> transformed data (x axis) and expression change (y axis). Courtesy of Partek Inc.

sample (Fig. 11.9b). This can be useful to explore outliers in the QC step, potentially identifying regions of an array that have defects (e.g., scratches or hybridization errors).

### Adding Sample Information

When we used GEO2R we specified which samples are euploid or trisomic. We do the same with Partek, and can merge a spreadsheet having critical sample information with another spreadsheet with gene expression values. From the gene expression workflow select “Add sample attributes” and specify the type (Trisomy21 or euploid), tissue (astrocyte, cerebellum, cerebrum, heart), and subject (individual). Each column has a header that can be clicked to specify its properties (Fig. 11.8). We can make the subject a random effect (the particular individuals used in this study are a random draw from the total population of trisomy 21 and euploid). The type and the tissue are fixed rather than random effects (those type and tissue categories are invariant in this study).

## Sample Histogram

A sample histogram plots the intensity of the probes ( $x$  axis) and the frequency of the probe intensity ( $y$  axis; **Fig. 11.9c**). This allows us to visually confirm that the samples have been normalized appropriately (or not if there are outliers). Later we will plot histograms using R packages for microarrays (“Reading CEL Files and Normalizing with RMA” below) or RNA-seq data (“CummeRbund to Visualize RNA-seq Results” below).

## Scatter Plots and MA Plots

The scatter plot is a common visualization method for microarray data. This shows the comparison of gene expression values for two samples. Most data points typically fall on a  $45^\circ$  line, but transcripts that are up- or down-regulated are positioned off the line. The scatter plot displays which transcripts are most dramatically and differentially regulated in the experiment.

The MA plot is a type of scatter plot also displaying all expression values from two samples. The average  $\log_2$  expression value is shown on the  $x$  axis, ranging from transcripts expressed at low levels (to the left) to high levels (to the right; **Fig. 11.9d**).

$$M = \log_2(I_1) - \log_2(I_2), \quad (11.5)$$

$$A = \frac{1}{2}(\log_2(I_1) + \log_2(I_2)). \quad (11.6)$$

The  $y$  axis displays difference. For  $\log_2$  transformed data, a value of zero is obtained for two samples with equal expression levels, while up-regulated transcripts are shown higher on the  $y$  axis and down-regulated samples lower.

## Working with $\log_2$ Transformed Microarray Data

We routinely  $\log_2$ -transform microarray data, as is done by both GEO2R and Partek. For scatter plots this creates a more centered distribution in which the properties of the dataset are easier to analyze. Also, it is far easier to describe the fold regulation of genes. Consider three transcripts that are unchanged, up-regulated two-fold, and down-regulated two-fold. The ratios are 1:1, 2:1, and 0.5:1. In  $\log_2$  space, the data points are however conveniently symmetric about 0: for two transcripts expressed at the same level  $\log_2(x/x) = 0$ , while the values are +1 and -1 for the up- and down-regulated cases. Another feature of logarithmic transformations is that, in addition to providing symmetry in expression ratios, they stabilize the variance across a wide range of intensity measurements.

We review some of the basic values and properties of logarithms in **Table 11.1**.

PCA is also called singular-value decomposition (Alter *et al.*, 2000). It is a linear projection method; this means that the data matrix you start with is “projected” or mapped onto lower dimensional space. Projection methods related to PCA include independent components analysis, factor analysis, multidimensional scaling, and correspondence analysis.

## Exploratory Data Analysis with Principal Components Analysis (PCA)

Exploratory analyses are valuable to visualize the relatedness of samples. Principal components analysis (PCA) is a technique to reduce and visualize high-dimensional data in plots having two or three dimensions (Ma and Dai, 2011). The central idea behind PCA is to transform a number of variables into a smaller number of uncorrelated variables called principal components. The variables that are operated on by PCA may be the expression of many genes (e.g., 20,000 gene expression values), or the results of gene expression across various samples. In a typical microarray experiment, PCA detects and removes redundancies in the data (such as genes whose expression values do not change and are therefore not informative about differences in how the samples behave).

We create a PCA plot and note 25 balls (one per sample) appearing in several clusters (**Fig. 11.10a**). These are colored by type and, as expected, there is no major difference

**TABLE 11.1 Common values of logarithms in base 2 and base 10.** Recall that for any positive number  $b$  (where  $b \neq 1$ ),  $\log_b y = x$  when  $y = b^x$ . Thus  $\log_2 8 = 3$  and  $2^3 = 8$ . Note also that  $\log_b b = 1$ ;  $\log_b 1 = 0$ ;  $\log_b xy = \log_b x + \log_b y$ ; and  $\log_b (x/y) = \log_b x - \log_b y$ .

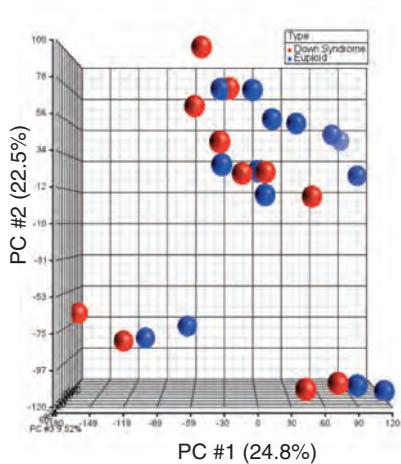
Value	Log <sub>10</sub>	Log <sub>2</sub>
1000	3.00	9.97
100	2.00	6.64
50	1.70	5.64
10	1.00	3.32
5	0.70	2.32
2	0.30	1.00
1	0.00	0.00
0.5	-0.30	-1.00
0.2	-0.70	-2.32
0.1	-1.00	-3.32
0.01	-2.00	-6.64
0.001	-3.00	-9.97

between samples labeled Down syndrome or euploid. A total of 56.9% of the variance is explained in this plot. Principal component (PC) axis #1, which is the  $x$  axis, accounts for 24.0% of the variance, while PC #2 (the  $y$  axis) and PC #3 (the  $z$  axis) account for 22% and 11%, respectively. (By definition, each axis accounts for successively less of the variance.) Since we know that gene expression varies across tissue type, we can label the data points by tissue (Fig. 11.10b). Here we see a clear explanation for the position of the data points in the PCA plot: astrocyte samples are grouped together (bottom left) as are heart samples (bottom right), while the brain regions (cerebrum and cerebellum) are in separate, adjacent clusters. The position of the data points is identical in our two PCA plots (Fig. 11.10a, b) and only the labeling has changed. In the latter plot we include the feature of an ellipse which extends two standard deviations beyond the centroid of each tissue group.

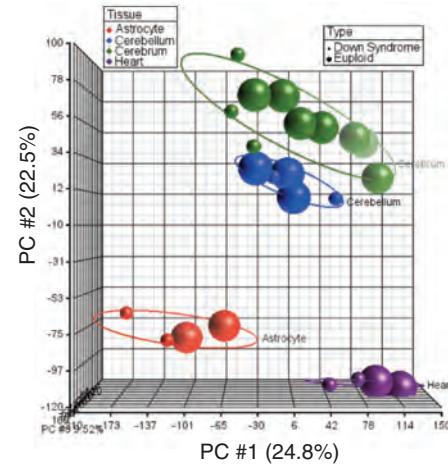
The position of the data points in PCA reflects the relatedness of the objects (samples) in the underlying data matrix. To illustrate this concept, we can return to the data matrix and select Transform > Create transposed spreadsheet. This creates a new spreadsheet having 26 columns (probesets and then the 25 samples) and 22,283 rows. A PCA plot of this dataset shows 22,283 data points each corresponding to a single probeset. We can select a group of three data points (arbitrarily selected based on points that are close together; Fig. 11.10c, arrow). In the main spreadsheet we can click the row number and plot the profile (e.g., refer to Fig. 11.8 arrow 6). These three data points have very similar profiles in terms of expression levels observed across 25 samples (Fig. 11.10d). It is precisely because they have similar profiles that the PCA plot of Figure 11.10c grouped them close together.

The starting point for PCA is any matrix of  $m$  observations (gene expression values) and  $n$  variables (experimental conditions). The goal is to reduce the dimensionality of the data matrix by finding  $r$  new variables (where  $r < n$ ). These  $r$  variables account for as much of the variance in the original data matrix as possible. The first step of PCA algorithms is to create a new matrix of dimensions  $n \times n$ . This may be a covariance matrix or a correlation matrix. (In a study of 25 samples and >22,000 genes there is a  $25 \times 25$  covariance matrix.) The principal components (called eigenvectors) are selected for the biggest variances (called eigenvalues). What this means practically for our example dataset is that, if the expression values of a gene do not vary across the samples, it will not contribute to the formation of the principal components.

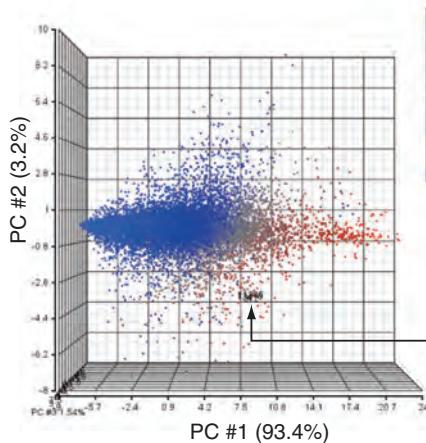
(a) PCA of 25 samples annotated by type



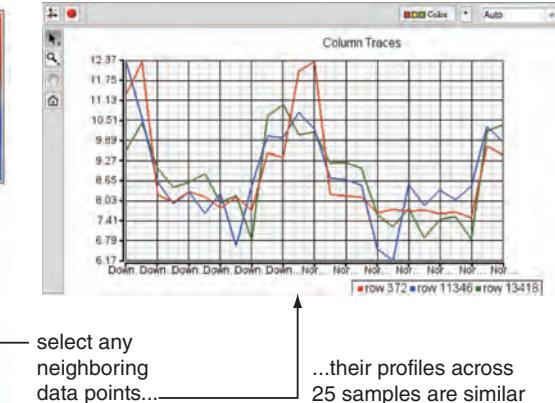
(b) PCA annotated by tissue and type



(c) PCA of 22,000 transcript level values

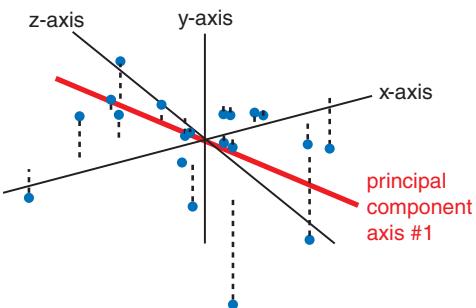


(d) Profile of three transcripts across 25 samples



**FIGURE 11.10** Principal components analysis (PCA) plots. (a) Each dot represents a sample having >20,000 gene expression values. The method is unsupervised, that is, the 25 data points (one per sample) are placed in PCA space based on transformation of the initial data matrix. There is no apparent separation based on type (trisomy 21 versus euploid). (b) The same PCA plot as (a) is annotated by tissue showing separation of heart samples, astrocyte samples, and two brain regions. (c) The data matrix (25 samples, >22,000 expression values) was transposed. The PCA plot shows 22,283 data points (probe-sets). Most of the variance (93.4%) is explained in this plot along PC #1, corresponding to the level of expression. We select three arbitrary, neighboring data points (arrow). (d) Expression data for these three probesets (y axis) are plotted across the 25 samples. The profiles are closely similar. This illustrates that any closely neighboring data points in PCA space have similar properties in the original data matrix. Data analysis performed using Partek software. Courtesy of Partek Inc.

How is the first principal component axis related to our raw data? Take the three-dimensional plot of the raw data and redraw the  $x$ ,  $y$ ,  $z$  coordinate axes so that the origin (“centroid”) is at the center of all the data points (Fig. 11.11). Find the line that best fits the data; this corresponds to the first principal component axis. By rotating this axis, it becomes the  $x$  axis of the plots in Figure 11.10. The second principal component axis must also pass through the origin of the graph in Figure 11.11, and it must be orthogonal to the first axis. In this way, it is uncorrelated. Each axis accounts for successively less of the variability in the data.



**FIGURE 11.11** Principal components analysis. The first principal component axis may be thought of as the best-fit line that traverses the geometric origin of the dataset, accounting for most of the variability in the data. The second principal component (not shown) also passes through the origin and is orthogonal to the first component. Cumulatively, all the principal component axes account for 100% of the variance, with each axis accounting for a successively smaller percentage. A large percentage accounted for by the first and/or second principal component axes indicates that the importance of this axis should be given when inspecting the PCA plot.

## Performing ANOVA in Partek

We can perform ANOVA to generate a list of transcripts there are differentially regulated. Note how analyses in Partek (as in BioConductor R packages described in “Microarray Analysis Method 3” below) offer vastly more flexibility and depth than an online tool such as GEO2R.

It is typical to explore the data with PCA or other visualization methods to decide which factors should be included in an ANOVA model (Fig. 11.12a).

ANOVA is available within a Stat pull-down menu.

- We include type, tissue, and subject as factors.
- We select a type\*tissue interaction. These are both fixed effects. If we include a random effect (such as subject), this becomes a mixed-model ANOVA.
- Partek identifies that some tissue samples (e.g., cerebrum, heart) came from the same subject. Subject is therefore nested within type (in contrast to a situation in which each tissue sample was from a unique individual). The ANOVA model is automatically adjusted for nested relationships.

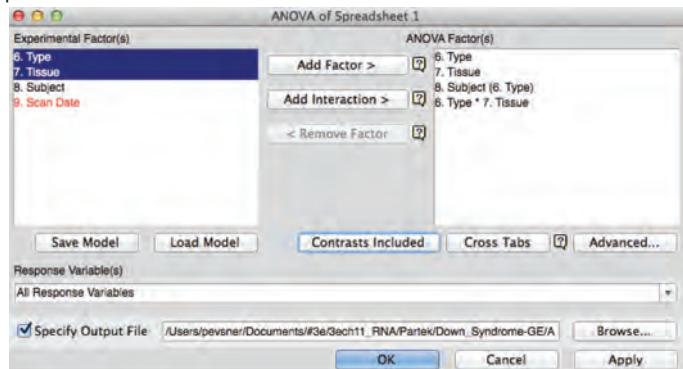
Fold-change calculations are called contrasts. We can select contrast within the ANOVA dialog box to establish Down syndrome (group 1) and euploid (group 2) (Fig. 11.12a). We then perform the ANOVA to generate a spreadsheet of 22,283 rows. Columns include probeset IDs, annotation (such as NCBI Gene identifiers, gene symbols, and RefSeq transcript identifiers), and *p* values according to type, tissue, and the interactions or any other factors selected for the ANOVA.

There are several sources of variation that lead to changes in RNA transcript levels. We can view these with a plot (Fig. 11.12b). This analysis may inform which particular ANOVA models we would like to perform. Adding factors (such as scan date) or factor interactions can offer the benefit of improving the ability of ANOVA to identify changes we are interested in, but at the cost of a loss of power.

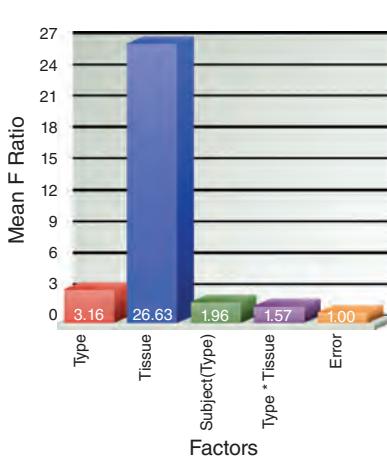
ANOVA results include transcripts that are ranked based on probability value. It is common to filter these based on the criterion of some minimal fold change (such as  $<-1.5$  or  $>1.5$ ). The rationale for filtering is that statistically significant changes having only a subtle fold change (such as 1.1) are unlikely to be biologically meaningful. Within Partek we can use a list manager to create such filtered lists.

A volcano plot presents information about fold change (on the *x* axis) and *p* value (*y* axis) (Fig. 11.12c). Transcripts present in the upper left and upper right sectors have

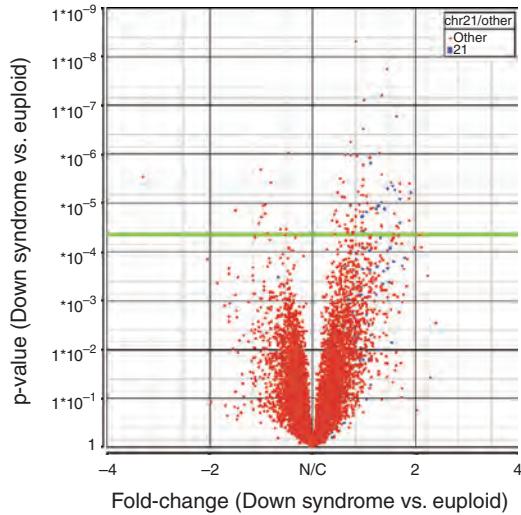
(a) ANOVA input



(b) ANOVA sources of variation



(c) Volcano plot



**FIGURE 11.12** Identification of significantly regulated genes. (a) Select experimental factors to be used in analysis of variance (ANOVA), as well as interacting factors. You can also select “contrasts” to report fold change. Once the ANOVA is executed >22,000 tests are performed: for each probeset on the microarray the mean values of the trisomy 21 and euploid groups are compared, also accounting for the noise in the data measurements. The ANOVA result includes a table of >22,000 genes ranked by lowest probability value. (b) Once the ANOVA is performed, sources of variation may be assessed. The signal-to-noise ratio is shown for all the selected factors (*x* axis) based on mean *F* ratio from ANOVA (*y* axis). Tissue (e.g., heart versus cerebrum) is the dominant factor in this experiment. Type (trisomy 21 versus euploid) accounts for a relatively small effect size. (This outcome is expected since in general the experimental condition does not induce a massive change in RNA transcript levels). (c) A volcano plot depicts fold change (*x* axis) versus *p* value (*y* axis). The green bar is placed at a *p* value threshold, indicating significant values above the bar and nonsignificant values below. This employs a false discovery rate (FDR) of 0.05 at which 1 in 20 of the results called true positives are actually false positives. You can select different FDR thresholds: at a more lenient threshold you will obtain more positive results, but an increased proportion of them will be false positives. Data analysis performed using Partek software. Courtesy of Partek Inc.

particularly low *p* values and large fold changes, and are therefore usually of the greatest interest. In Partek, data points in any sector of the plot can be highlighted and dumped to a separate spreadsheet or plot.

What is the appropriate *p* value threshold for a volcano plot, or for an ANOVA? You can choose one or more FDRs. For example, in our ANOVA (above) we can select three FDR levels. At an FDR of 0.01, which is quite stringent, there are two transcripts that are significantly regulated (i.e., having *p* values below the cutoff value of  $8.9 \times 10^{-7}$ ), and only 1 in 100

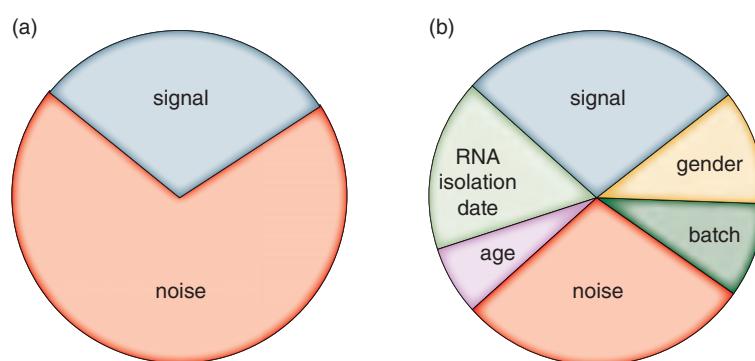
**TABLE 11.2** Test statistics for microarray data. Adapted from Motulsky (1995) with permission from Oxford University Press.

Paradigm	Parametric test	Nonparametric test
Compare one group to a hypothetical value	One-sample t-test	Wilcoxon test
Compare two unpaired groups	Unpaired t-test	Mann–Whitney test
Compare two paired groups	Paired t-test	Wilcoxon test
Compare three or more unmatched groups	One-way ANOVA	Kruskal–Wallis test
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test

such results are likely to be false positives. At an FDR of 0.05 we expect 1 in 20 results to be false positives, and 10 transcripts have  $p$  values below the cutoff value ( $2.2 \times 10^{-5}$ ). At FDR 0.1 there are now 26 regulated transcripts, 2 or 3 of which are likely to be false positives. An investigator must decide which FDR is preferable: would you rather see more results, some of which are false positives, or fewer results, most of which are trusted?

### From *t*-test to ANOVA

A variety of test statistics may be applied to microarray data (e.g., Olshen and Jain, 2002); some of these are listed in **Table 11.2**. These tests are all used to derive  $p$  values that help assess the likelihood that particular genes are regulated. For more than two conditions (e.g., analyzing multiple time points or measuring the effects of several drugs on gene expression), the analysis of variance (ANOVA) method is appropriate rather than a *t*-test. The ANOVA identifies differentially expressed genes while accounting for variance that occurs both within groups and between groups (Zolman, 1993; Ayroles and Gibson, 2006). ANOVA is particularly appropriate when a microarray experiment has multiple classes of treatment (e.g., control samples are compared to two different disease states or to five different time points) or multiple factors for each treatment (e.g., gender, age, date of RNA isolation, hybridization batch) (**Fig. 11.13**).



**FIGURE 11.13** Signal-to-noise ratios in *t*-test and ANOVA. (a) In a *t*-test, the values from a microarray experiment can be thought of as having components of signal (i.e., intensity measurements that reflect a difference between the means of the two groups being compared) and noise (variations in signal intensity that are not attributable to differences in the means of the two groups). If the RNA from control samples is purified on a Monday, and the RNA from experimental samples is purified on a Tuesday, then there is a perfect confound between date and condition. Some or even all of the observed difference between control and experimental samples could be due to date rather than to treatment. (b) In an ANOVA, fixed and/or random effects can be accounted for. The variable due to factors such as date and gender can be analyzed, as well as the main effect of interest (control versus experimental conditions). By partitioning the signal into multiple components, ANOVA improves the signal-to-noise ratio.

ANOVA is a statistical model called a general linear model. This takes the form:

$$Y = \mu + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_j x_j + \varepsilon$$

where  $Y$  is a linear function of  $X$  with slope  $\beta$  and intercept  $\mu$ , and  $x_1, x_2, \dots, x_j$  is a series of independent variables;  $\varepsilon$  is an error term. For expression data, a commonly used statistical model is:

$$Y_{ijk} = \theta_i + \phi_{ij} + \varepsilon_{ijk}$$

where  $Y_{ijk}$  represents a pre-processed probe intensity measurement  $k$  (in the  $\log_2$  scale) of transcript  $i$  measured by platform  $j$ ; if there are 20,000 transcripts represented on a microarray, there will be that many  $Y_{ijk}$  values. The terms  $\phi$  and  $\theta$  are independent variables associated with expression measurement and probe effects.  $\theta_i$  is the absolute gene expression value in the  $\log_2$  scale,  $\phi_{ij}$  is a platform-specific probe effect, and  $\varepsilon_{ijk}$  represents a term for measurement error (residual, unexplained variance). As noted by Irizarry *et al.* (2005), a large probe effect  $\phi_{ij}$  (with a large associated variance) tends to inflate the large correlations that have been reported when comparing absolute gene expression measurements within a given platform, while yielding lower correlations between two platforms that differ in their probe effects. A solution is to evaluate relative (rather than absolute) expression within a platform to cancel the  $\phi_{ij}$  terms.

Both fixed and random factors are independent variables accounted for in the linear model. Fixed factors involve treatment effects systematically selected by the experimenter (such as gender or age) that would remain the same if the experiment were replicated. Fixed factors account for the main conditions that an investigator is interested in, such as the change in signal intensity due to a sample coming from trisomy 21 rather than control individuals. Random factors provide a model of independent variables that are selected randomly or unsystematically from a population. Examples are biological replicates, because when we select a group of 11 trisomy 21 samples we are drawing them in an unbiased manner from the overall population of individuals with trisomy 21. Similarly, array effects are random factors because each microarray is randomly selected from the group of all available arrays.

The idea of ANOVA is that differences in gene expression may be due to main effects (e.g., normal versus diseased sample), while other sources of variation (e.g., gender or age) can be identified and accounted for. Analogous to the  $t$ -statistic of a  $t$ -test, the  $F$ -statistic of an ANOVA consists of a signal-to-noise ratio (Fig. 11.13). However, the ANOVA includes a more detailed estimate of the sources of variation. By partitioning the signal to account for fixed and random effects in the data the ANOVA boosts the signal-to-noise ratio, often allowing you to more effectively identify regulated transcripts.

Visit the Bioconductor website at <http://bioconductor.org/> (Weblink 11.8). There are many excellent books and online guides to using R and BioConductor packages including Gentleman *et al.* (2005) and Zuur *et al.* (2009). Sean and Meltzer (2007) describe GEOquery, an R tool that facilitates the import and analysis of data files from GEO.

The CEL files are available at <http://bioinfbook.org/chapter11>. You can download R from <http://www.r-project.org/> (Weblink 11.9) and RStudio from <https://www.rstudio.com/> (Weblink 11.10).

## MICROARRAY ANALYSIS METHOD 3: ANALYZING A GEO DATASET WITH R

The BioConductor project has emerged as a collection of over 1000 bioinformatics software packages that are run in R. There is tremendous enthusiasm for R and BioConductor, especially in the biostatistics community.

### Setting up the Analyses

Let's begin by creating a directory (in PC, Mac, or Linux environments) and placing the same 25 CEL files there that we used earlier. You should install R and (optionally) install the excellent user interface RStudio.

Start R, and either use a pull-down menu to browse to your working directory or set working directory with the `setwd()` command. While we used `$` to indicate the symbol for Linux (or Unix) commands, `>` indicates the prompt for R commands.

```
> getwd()
[1] "/Users/pevsner/Documents/#3e/3ech11_RNA/ch11_R"
> source("http://bioconductor.org/biocLite.R")
> biocLite("affy")
> biocLite("limma")
# Next we load the affy and limma libraries.
> library(affy)
> library(limma)
```

We describe the `affy` and `limma` libraries below. To get help on these (or any other) R packages, consider the following:

- There is extensive documentation on the BioConductor website. This usually includes vignettes and R code.
- You can join the BioConductor mailing list.
- Get help on any function within a package (e.g., `lmFit` within `limma`) by typing in `R > ?lmFit` or equivalently `> help("lmFit")`.
- Biostars features questions and answers from the bioinformatics community.

Loading `affy` automatically results in the download of the required CEL definition file (CDF). We need to load the phenotype data to specify what type each file corresponds to (trisomy 21 or euploid) and what region (cerebrum, cerebellum, heart, or astrocyte). This information is written in a tab-delimited text file for this example. If you create such a file in Microsoft Excel or Word be careful because these programs may reformat your data, introducing errors. Alternative text editors include NotePad or Crimson Editor (for PC) orTextEdit (Mac).

```
> phenoData <- read.AnnotatedDataFrame("pheno.txt", header=TRUE, sep="\t")
```

We create the object `phenoData` by using the function `read.AnnotatedDataFrame`. This reads our text document `pheno.txt` and creates an object of the class `AnnotatedDataFrame`. We specify that the file does have a header and its separator is tab. The `<-` symbols indicate an object we will create; for example we can create the variable `x` as the sum of `2 + 2`, then type `x` to output its result.

```
> x <- 2 + 2
> x
[1] 4
> show(x) # Equivalent way to display x
[1] 4
> 3*x
[1] 12
```

Now let's look at information about the contents of `phenoData`.

```
> phenoData
An object of class 'AnnotatedDataFrame'
  rowNames: Down Syndrome-Astrocyte-1478-1-U133A.CEL Down
              Syndrome-Astrocyte-748-1-U133A.CEL ...
              Normal-Heart-1411-1-U133A.CEL (25 total)
  varLabels: diagnosis tissue
  varMetadata: labelDescription
> dim(phenoData) # report the dimensions of the file
  rowNames columnNames
      25          2
> summary(phenoData)
  Length     Class           Mode
      1     AnnotatedDataFrame     S4
```

This confirms that there are 25 rows (samples) and 2 columns (diagnosis and tissue).

Join the Bioconductor mailing list  
by contacting [bioconductor@stat.math.ethz.ch](mailto:bioconductor@stat.math.ethz.ch).

Visit Biostars at <http://www.biostars.org> (WebLink 11.11).

The phenotype data are  
available as `pheno.txt` at Web  
Document 11.3.

## Reading CEL Files and Normalizing with RMA

Expression data are represented as a matrix of rows (corresponding to probes) and columns (corresponding to samples on separate arrays). As one option the `affy` package (Gautier *et al.*, 2004) includes `justRMA`, a function that reads CEL files, performs RMA, and computes expression measures. Its arguments include `phenoData`. Optionally, widgets can be used to facilitate data input; in our workflow we will instead read in CEL files from our working directory. Type `> ?justRMA` for a help page describing the arguments and usage details.

We will use the `ReadAffy` function of the `affy` package to read the CEL files and `phenoData` in our working directory. (It can also read MIAME data (Chapter 10); furthermore, it can read `zip` and `gzip` compressed CEL files.) For more details, type

```
> ?read.affybatch
> MyBioinfData <- ReadAffy()
```

What is in the `MyBioinfData` object?

```
> MyBioinfData
AffyBatch object
size of arrays=712x712 features (28 kb)
cdf=HG-U133A (22283 affyids)
number of samples=25
number of genes=22283
annotation=hgul33a
notes=</p><p>> rownames(MyBioinfData)[1:10]
[1] "1007_s_at" "1053_at" "117_at" "121_at" "1255_g_at"
[6] "1294_at" "1316_at" "1320_at" "1405_i_at" "1431_at"</p><p>>
colnames(MyBioinfData)[1:5]
[1] "Down Syndrome-Astrocyte-1478-1-U133A.CEL"
[2] "Down Syndrome-Astrocyte-748-1-U133A.CEL"
[3] "Down Syndrome-Cerebellum-1218-1-U133A.CEL"
[4] "Down Syndrome-Cerebellum-1389-1-U133A.CEL"
[5] "Down Syndrome-Cerebellum-1478-1-U133A.CEL"
> summary(MyBioinfData)
      Length   Class    Mode
      25      AffyBatch   S4
```

Typing its name shows that there are 22,283 genes, 25 samples, and annotation from the Affymetrix U133a microarray. We can further look at the first few rows and columns. For additional information try `dim(MyBioinfData)` (showing the dimensions of the rows and columns) and `str(MyBioinfData)` for the structure of the file.

We then employ `rma`, a function that converts an `AffyBatch` object into an `ExpressionSet` object.

```
> eset <- rma(MyBioinfData)
Background correcting
Normalizing
Calculating Expression
```

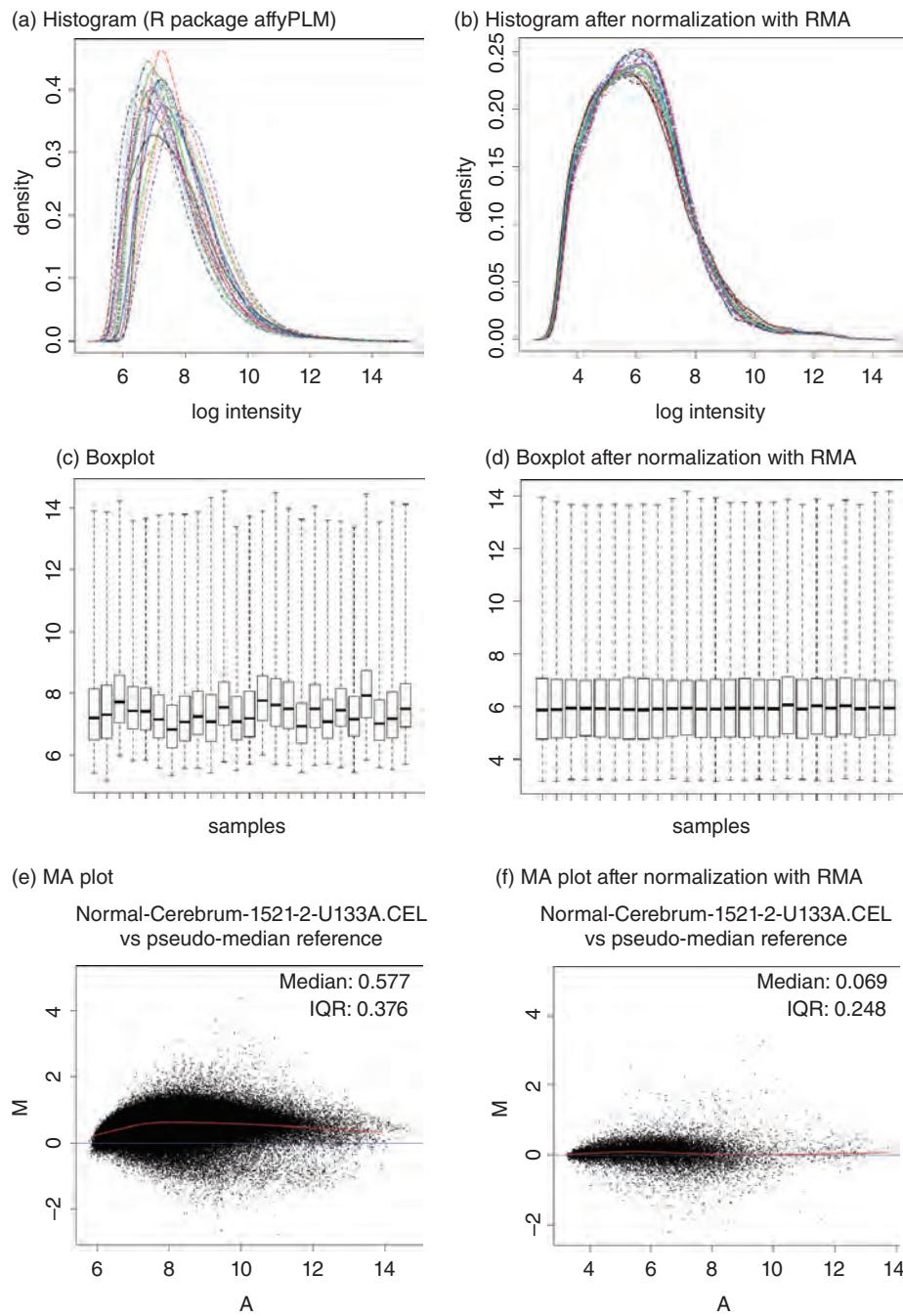
The `rma` function implements RMA by: (1) probe-specific correction of perfect match probes; (2) normalization of corrected perfect match probes by quantile normalization (Bolstad *et al.*, 2003); and (3) calculation of expression measures using median polish. We can view the effects of these steps before and after normalization for three kinds of plots (Fig. 11.14).

```
> hist(MyBioinfData)
> hist(eset)
> boxplot(MyBioinfData)
> boxplot(eset)
> MAplot(MyBioinfData)
> MAplot(eset)
> MAplot(Dilution,pairs=TRUE,plot.method="smoothScatter")
```

Note that a briefer, alternative workflow could use `justRMA`:

```
> eset <- justRMA()
(phenoData=phenoData)
```

See the `affy` package documentation on the BioConductor website for more details.



**FIGURE 11.14** Plotting microarray analyses in R. The R packages `affy` and `limma` are used to import, preprocess, and analyze Affymetrix CEL files. Plots show (a, b) histograms, (c, d) boxplots, and (e, f) MA plots before and after normalization.

Source: R.

A strength of R is its versatility as a plotting tool. Beyond making a simple boxplot we can add colors and labels. For example, we can define the first 11 (trisomy 21) samples in sienna and the remaining 14 euploid samples in dark green. We then add a title, label, and notches to the boxplot:

```
> colors = c(rep("sienna",11),rep("darkgreen",14))
> boxplot(eset, ylab = "log2 intensities", col=colors)
```

Notches can be drawn in the sides of a boxplot. They extend  $\pm 1.58$  interquartile range/sqrt(n). See `boxplot` in the `Graphics` package for details, or begin with `?boxplot`.

When we imported CEL files into Partek, several quality control plots were generated. We can create many similar plots within the `affy` package such as measures of RNA degradation:

```
> deg <- AffyRNADeg(MyBioinfData)
> names(deg)
> summaryAffyRNADeg(deg)
> mean(mm(MyBioinfData)>pm(MyBioinfData))
```

`limma` was developed by  
Gordon Smyth.

## Identifying Differentially Expressed Genes (Limma)

Next we use the `limma` package to analyze gene expression (Smyth, 2004, 2005). `limma` requires a design matrix representing the different RNA targets that have been hybridized to the array, and a contrast matrix that allows analysis of contrasts of interest based on coefficients defined by the design matrix. We use `model.matrix` (from the `stats` package) to create a design matrix from the description given in `eset`. Then we use `lmFit` to fit a linear model for each gene (i.e., probeset) across our series of microarrays.

```
> design <- model.matrix(~diagnosis, phenoData(eset))
> fit <- lmFit(eset, design) # fit each probeset to model
```

Let's take a look at the `fit` object.

```
> dim(fit) # dim shows us dimensions in rows and columns
[1] 22283 2
> colnames(fit)
[1] "(Intercept)" "diagnosisEuploid"
# We next look at the first 10 rows of fit.
# Without this limit all 22,283 rows would be printed.
> rownames(fit)[1:10]
[1] "1007_s_at"          "1053_at"        "117_at"         "121_at"        "1255_g_at"
[6] "1294_at"            "1316_at"        "1320_at"        "1405_i_at"     "1431_at"
# We can use tail to display the last rows of the file.
> tail(rownames(fit))
[1] "AFFX-ThrX-3_at"      "AFFX-ThrX-5_at"      "AFFX-ThrX-M_at"
[4] "AFFX-TrpnX-3_at"     "AFFX-TrpnX-5_at"     "AFFX-TrpnX-M_at"
> names(fit)
[1] "coefficients"        "rank"              "assign"
[4] "qr"                  "df.residual"       "sigma"
[7] "cov.coefficients"    "stdev.unscaled"   "pivot"
[10] "Amean"               "method"            "design"
> summary(fit)
           Length Class Mode
coefficients 44566 -none- numeric
rank          1     -none- numeric
assign         2     -none- numeric
qr             5     qr   list
df.residual  22283 -none- numeric
sigma          22283 -none- numeric
cov.coefficients 4     -none- numeric
stdev.unscaled 44566 -none- numeric
pivot          2     -none- numeric
Amean          22283 -none- numeric
method          1     -none- character
design          50    -none- numeric
```

We next use `eBayes` from the `limma` package to make an empirical Bayes adjustment. Given a linear model fit, `eBayes` will compute moderated *t*-statistics. A series of ordinary *t*-statistics are generated and then the standard errors are moderated across all genes (shrunk using a Bayesian model).

```
> efit <- eBayes(fit) # empirical Bayes adjustment
> tt <- topTable(efit, coef=2)
> fix(tt)
```

**TABLE 11.3 Results of `topTable` (limma analysis of differential gene expression).**  
**log FC:** log<sub>2</sub> fold change; **Ave.expr:** average expression; **t:** moderated t-statistic  
(but available when two groups of samples are defined); **P.value:** raw p value; **Adj.P.value:** p value after adjustment for multiple testing; **B:** B-statistic or log-odds that the gene is differentially expressed.

Row names	log FC	Ave.expr	t	P.value	Adj.P.value	B
200818_at	-0.71	10.16	-6.95	2.85×10 <sup>-7</sup>	0.0063	5.95
206777_s_at	0.81	6.42	6.60	6.58×10 <sup>-7</sup>	0.0073	5.29
200642_at	-0.84	10.12	-6.23	1.63×10 <sup>-6</sup>	0.0080	4.57
201123_s_at	1.80	7.06	6.22	1.70×10 <sup>-6</sup>	0.0080	4.54
202217_at	-0.60	8.63	-6.20	1.80×10 <sup>-6</sup>	0.0080	4.50
221677_s_at	-0.60	4.90	-6.02	2.78×10 <sup>-6</sup>	0.0103	4.15
201086_x_at	-0.45	9.01	-5.84	4.34×10 <sup>-6</sup>	0.0135	3.79
202325_s_at	-0.86	9.02	-5.80	4.86×10 <sup>-6</sup>	0.0135	3.70
203635_at	-0.35	6.17	-5.69	6.49×10 <sup>-6</sup>	0.0142	3.47
216954_x_at	-0.41	7.46	-5.68	6.59×10 <sup>-6</sup>	0.0142	3.45

`topTable` generates a table of differentially expressed probesets. The `fix` command lets us view the results by calling a data editor for a data frame (**Table 11.3**). We can also export these results, or annotate them for further analysis.

What are the gene symbols and chromosome locations of these top 10 hits? Let's use `biomaRt` to find out.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("biomaRt")
> library("annotate")
> library("biomaRt")
> ensembl=useMart("ensembl")
> ensembl = useDataset("hsapiens_gene_ensembl",mart=ensembl)
```

We have installed `biomaRt` and specified the `mart` and the dataset we want to use. Next we define the object `affyids` as the concatenation of the ten Affymetrix identifiers from the top table list of **Table 11.3**.

We introduced `biomaRt` in Chapter 8.

```
> affyids = c("200818_at", "206777_s_at", "200642_at", "201123_s_at", "202217_at",
  "221677_s_at", "201086_x_at", "202325_s_at", "203635_a", "216954_x_at")
> getBM(attributes=c('affy_hg_u133_plus_2', 'hgnc_symbol', 'chromosome_name',
  'start_position', 'end_position', 'band'), filters = 'affy_hg_u133_plus_2',
  values = affyids, mart = ensembl)
affy_hg_u133_plus_2 hgnc_symbol chromosome_name start_position end_position
band
1    202325_s_at      ATP5J        21     27088815    27107984    q21.3
2    200642_at       SOD1        21     33031935    33041244    q22.11
3    206777_s_at     CRYBB2       22     25615489    25627836    q11.23
4    206777_s_at   CRYBB2P1      22     25844072    25916821    q11.23
5    201086_x_at      SON         21     34914924    34949812    q22.11
6    221677_s_at     DONSON       21     34931848    34961014    q22.11
7    200818_at          NA        21     34956993    35284635    q22.11
8    201123_s_at     EIF5AP4       10     82006975    82007439    q23.1
9    201123_s_at     EIF5AL1       10     81272357    81276188    q22.3
10   200818_at      ATP5O        21     35275757    35288284    q22.11
11   202217_at     C21orf33      21     45553487    45565605    q22.3
12   201123_s_at     EIF5A         17     7210318     7215774    p13.1
```

Seven of the top 10 regulated genes are assigned to chromosome 21. (In the `biomaRt` output probeset `201123_s_at` maps to three EIF5A-related genes at three loci on two

chromosomes, and we count that as one of 10 total genes on the output list.) Most of the same probesets are in the top 12 list from GEO2R (this is expected since both use `limma`, although with different settings).

We saw a volcano plot produced by Partek (Fig. 11.12c); you can also generate one in R using `> volcanoplot(fit)`.

### Microarray Analysis and Reproducibility

The same raw microarray dataset can yield entirely different results based on the analysis method, spanning all steps such as normalization and implementation of ANOVA. The same basic experiment (such as defining differentially regulated transcripts in the post-mortem brains of individuals with schizophrenia versus controls) may yield substantially different results between laboratories. Tan *et al.* (2003) compared gene expression measurements from three commercial platforms (Affymetrix, Agilent, and Amersham) using the same RNA as starting material, and included both biological and technical replicates. They reported that there was only limited overlap in the RNA transcripts identified by the three platforms, with an average Pearson's correlation coefficient  $r$  for measurements between the three platforms of only 0.53 (see Box 11.1). Others have raised concerns about microarray data reproducibility and broader issues regarding data analysis (Draghici *et al.*, 2006; Miron and Nadon, 2006; Shields, 2006), with accompanying responses (Quackenbush and Irizarry, 2006).

### BOX 11.1 PEARSON CORRELATION COEFFICIENT $r$

When two variables vary together they are said to correlate. The Pearson correlation coefficient  $r$  has values ranging from  $-1$  (a perfect negative correlation) to  $0$  (no correlation) to  $1$  (perfect positive correlation). It is possible to state a null hypothesis that two variables are not correlated, and an alternative hypothesis that they are correlated. A probability  $p$  value can be derived to test the significance of the correlation. The Pearson correlation coefficient is perhaps the most common metric used to define similarity between gene expression data points. It is used by tree-building programs such as Cluster. For any two series of numbers  $X = \{X_1, X_2, \dots, X_N\}$  and  $Y = \{Y_1, Y_2, \dots, Y_N\}$ ,

$$r = \frac{\sum_{i=1}^N \left[ \frac{(X_i - \bar{X})}{\sigma_x} \cdot \frac{(Y_i - \bar{Y})}{\sigma_y} \right]}{N - 1} \quad (11.9)$$

where  $\bar{X}$  is the average of the values in  $X$  and  $\sigma_x$  is the standard deviation of these values. For a scatter plot,  $r$  describes how well a line fits the values. The Pearson correlation coefficient always has a value between  $+1$  (two series are identical) and  $-1$  (two sets are perfectly uncorrelated).

The square of the correlation coefficient,  $r^2$ , has a value between  $0$  and  $1$ . It is also smaller than  $r$ ;  $r^2 \leq |r|$ . For two variables having a correlation coefficient  $r = 0.9$  (such as two microarray datasets measured in different laboratories using the same RNA starting material),  $r^2$  is  $0.81$ . This means that  $81\%$  of the variability in the gene expression measurements in the two datasets can be explained by the correspondence of the results between the two laboratories, while just  $19\%$  can be explained by other factors such as error.

Correlation coefficients have been widely misused (Bland and Altman, 1986, 1999).  $r$  measures the strength of a relation between two variables, but it does not measure how well those variables agree. Picture a scatter plot showing the correlation of two measures; a perfect correlation occurs if the points fall on any straight line, but perfect agreement occurs only if the points fall on a  $45^\circ$  line. See Bland and Altman (1986, 1999) for additional caveats in interpreting  $r$  values.

*Source:* Motulsky (1995).

An optimistic assessment was provided by the MicroArray Quality Consortium (MAQC *et al.*, 2006). This project was established to evaluate the performance of a broad set of microarray platforms and data analysis techniques using identical RNA samples. A total of 20 microarray products and three technologies were evaluated for 12,000 RNA transcripts expressed in human tumor cell lines or brain. There was substantial agreement between sites and platforms for regulated transcripts, with various measures of concordance ranging from 60% to over 90% and a median rank correlation of 0.87 for comparability across platforms based on a log ratio measurement. Microarray data were also validated using polymerase chain-reaction-based methods, again showing a high correlation (Canales *et al.*, 2006). MAQC has been extended to problems of classification (Shi *et al.*, 2010) and RNA-seq technology (Mane *et al.*, 2009).

Many in the community appreciate the demonstrated ability of microarray experiments to produce reproducible results. Many factors can strongly influence the results, however. These factors include appropriate experimental design (e.g., avoiding confounding variables), consistent approach to preparation of RNA through the hybridization steps, appropriate image analysis (in which it is determined which pixels are part of the transcript-associated features), preprocessing (including global and local background signal correction), identification and removal of batch effects, appropriate identification of differentially expressed transcripts, application of multiple comparison correction, and other downstream analyses.

The MAQC project involved over 100 researchers at over 50 institutions. The MAQC website is <http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/> (WebLink 11.12).

## MICROARRAY DATA ANALYSIS: DESCRIPTIVE STATISTICS

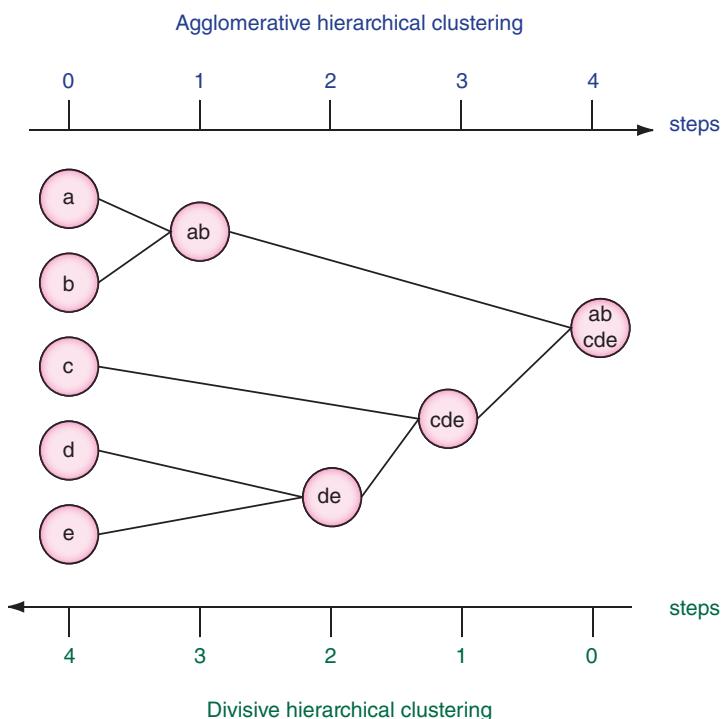
One of the most fundamental features of microarray experiments is that they generate large amounts of data. There are far more measurements (gene expression values) than samples. How can we evaluate the results of an experiment in which 20,000 gene expression values are obtained in (at most) several dozen samples? Each gene expression value can be conceptualized as a point in 20,000-dimensional space. The brain is not equipped to visualize highly dimensional space, and so we need to apply mathematical techniques that reduce the dimensionality of the data.

Mathematicians refer to the problems associated with the study of very large numbers of variables as the “curse of dimensionality.” In highly dimensional space, the distances between any two points are very large and approximately equal. Descriptive statistics are useful to explore such data. These mathematical approaches typically do not yield statistically significant results because they are not used for hypothesis testing. Rather, they are used to explore the dataset and to try to find biologically meaningful patterns. We have seen that PCA can show how genes (or samples) form groups. We next examine several other main descriptive techniques. These can be explored using software such as R or Partek. In each case, we begin with a matrix of genes (typically arranged in rows) and samples (typically arranged in columns). Appropriate global and/or local normalizations are applied to the data. Some metric is then defined to describe the similarity (or alternatively to describe the distance) between all the data points.

The approaches are unsupervised: prior assumptions about the genes and/or samples are not made, and the data are explored to identify groups with similar gene expression behaviors. You can choose a variety of distance functions in clustering, principal components analysis, multidimensional scaling, and other visualization techniques. These can produce strikingly different outputs. If you want to report your findings to others it is a good idea to clearly describe the choices you make and, if they are unusual, justify their selection.

### Hierarchical Cluster Analysis of Microarray Data

Clustering is the representation of distance measurements between objects (Kaufman and Rousseeuw, 1990). It is a commonly used method to find patterns of gene expression in



**FIGURE 11.15** Two main kinds of hierarchical clustering are agglomerative and divisive. In agglomerative clustering, the data points (genes or samples, represented as the letters a–e) are considered individually (step 0). The two most related data points are joined (circle ab, step 1). The relationship between all the data points is defined by a metric such as Euclidean distance. The next two closest data points are identified (step 2, de). This process continues (steps 3, 4) until all data points have been combined (agglomerated). The path taken to achieve this structure defines a clustering tree. Divisive hierarchical clustering involves the same process in reverse. The data points are considered as a combined group (step 0, abcde). The most dissimilar object is removed from the cluster. This process is continued until all the objects have been separated. Again, a tree is defined. In practice, agglomerative and divisive clustering strategies often result in similar trees. Adapted from Kaufman and Rousseeuw (1990) with permission from Wiley.

microarray experiments (Gollub and Sherlock, 2006; Thalamuthu *et al.*, 2006). Clustered trees may consist of genes, samples, or both. Clusters are commonly represented in scatter plots or in dendograms, such as those used for phylogenetic analysis (Chapter 7) or for microarray data. The main goal of clustering is to use similarity (or distance) measurements between objects to represent them. Data points within a cluster are more similar, and those in separate clusters are less similar. It is common to use a distance matrix for clustering based upon Euclidean distances.

There are several kinds of clustering techniques. The most common form for microarray analysis is hierarchical clustering, in which a sequence of nested partitions is identified resulting in a dendrogram (tree). Hierarchical clustering can be performed using agglomerative or divisive approaches (Fig. 11.15). In each case, the result is a tree that depicts the relationships between the objects (genes, samples, or both). In divisive clustering, the algorithm begins at step 1 with all the data in one cluster ( $k = 1$ ). In each subsequent step a cluster is split off, until there are  $n$  clusters. In agglomerative clustering, all the objects start apart. There are therefore  $n$  clusters at step 0; each object forms a separate cluster. In each subsequent step two clusters are merged, until only one cluster is left.

Agglomerative and divisive clustering techniques generally produce similar results, although large differences can occur in their representation of the data. Agglomerative techniques tend to give more precision at the bottom of a tree, while divisive techniques

Agglomerative clustering is sometimes called “bottom up” while divisive clustering is “top down.”

## BOX 11.2 EUCLIDEAN DISTANCE

Euclidean distance is defined as the distance  $d_{12}$  between two points in three-dimensional space (with coordinates  $x_1, x_2, x_3$  and  $y_1, y_2, y_3$ ) as follows:

$$d_{12} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}. \quad (11.10)$$

Euclidean distance is therefore the square root of the sum of the squared differences between two features. For  $n$ -dimensional expression data, the Euclidean distance is given by:

$$d = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}. \quad (11.11)$$

offer more precision at the top of a tree and may be better suited for finding relatively few, large clusters. Another feature of a clustering tree is that it may be highly sensitive to the choice of which genes (or samples) to include or exclude.

Clustering requires two basic operations. One is the creation of a distance matrix (or in some cases a similarity matrix). The two most commonly used metrics used to define the distance between gene expression data points are Euclidean distance (Box 11.2) and the Pearson coefficient of correlation (Box 11.1). Many software packages that perform microarray data analysis allow you to choose between these and other distance measures (such as manhattan, canberra, binary or minkowski) that describe the relatedness between gene expression values. In R you can use the `hclust` command within the `stats` package. In Partek you can select Euclidean distance as shown in **Figure 11.16a**. That dataset consists of the 25 trisomy 21 and euploid samples we studied earlier, annotated by chromosome to select only chromosome 21 genes. Two-way hierarchical clustering was performed of genes ( $x$  axis) and samples ( $y$  axis). Consistent with the PCA results, the astrocyte and heart samples form distinct clusters, while the samples from two brain regions are intermixed. We show just the sample dendrogram using three other distance metrics (Canberra, Pearson's dissimilarity, and City Block; **Fig. 11.16b-d**). These different metrics can dramatically alter the tree topology.

Row methods include average linkage, single linkage, complete linkage, centroid method, and Ward's method.

Given a distance metric, a second operation is the construction of a tree. We can select a variety of methods to calculate the proximity between a single object and a group containing several objects (or to calculate the proximity between two groups). In Partek we used the default approach of average linkage (**Fig. 11.16a-d**). The distance between clusters is defined using the average distance between all the points in one cluster and all the points in another cluster. This is used in the unweighted pair-group method average (UPGMA) procedure in the context of phylogenetic trees (Chapter 7).

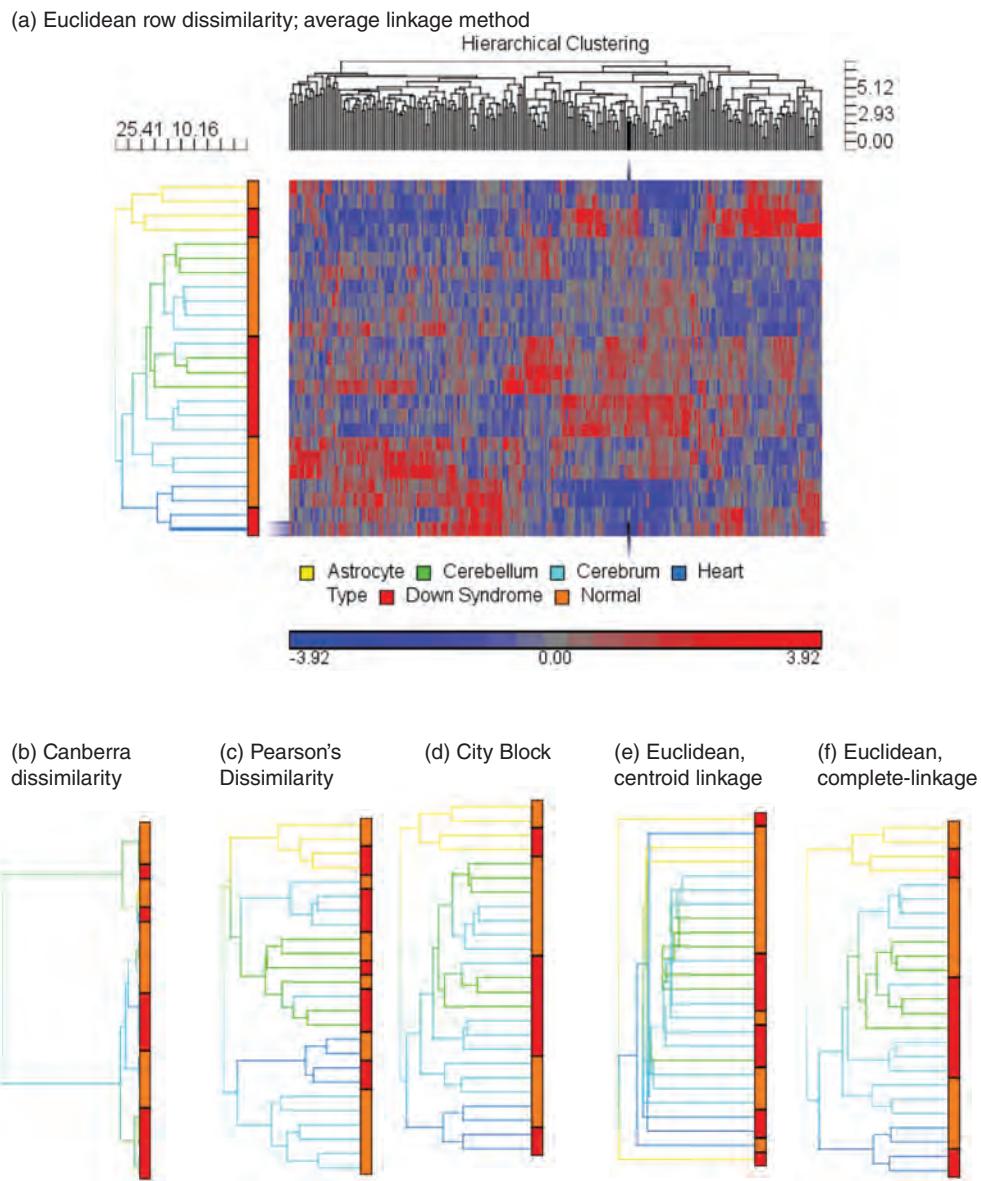
In single linkage clustering, an object that is a candidate to be placed into a cluster has a similarity that is defined as its relatedness to the closest member within that cluster (**Fig. 11.17a**). This method has also been called the minimum method or the nearest neighbor method. It is subject to an artifact called chaining in which “long straggly clusters” form (Sneath and Sokal, 1973, p. 218) as shown in **Figure 11.17b**. This can obscure the production of discrete clusters. In complete linkage clustering, the most distant OTUs in two groups are joined (**Fig. 11.17c**); the effect is to tend to form tight, discrete clusters that join other clusters relatively rarely. In centroid clustering, the central or median object is selected (**Fig. 11.17d**). These methods often produce different clustering patterns. Many alternative strategies exist (see Sneath and Sokal, 1973).

What is the significance of these different approaches to making a clustering tree? We can consider the general problem involved in defining a cluster. Objects that are clustered

The Canberra distance metric is calculated in R by

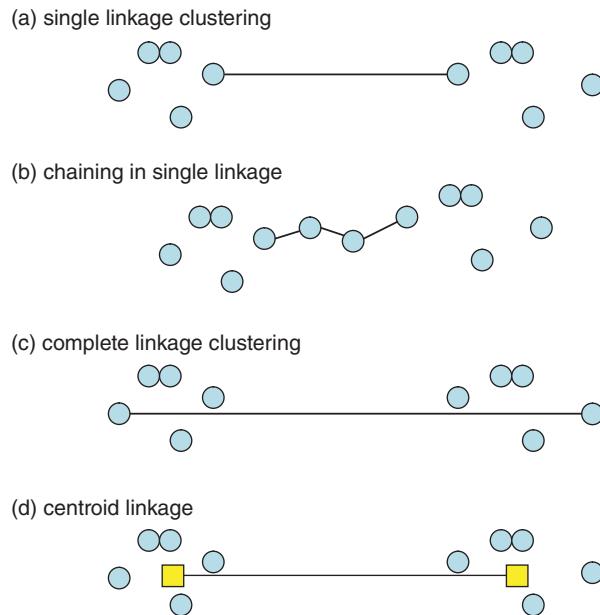
$$\sum \left( \frac{|x_i - y_i|}{|x_i + y_i|} \right).$$

Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing.

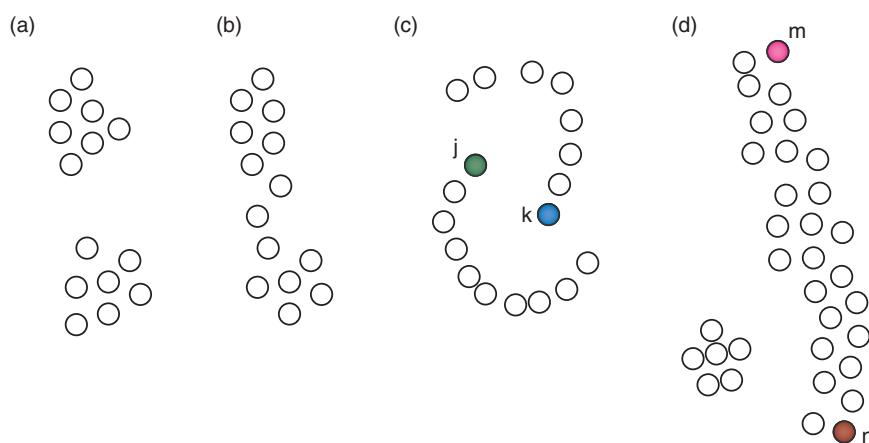


**FIGURE 11.16** Hierarchical clustering of 250 chromosome 21 transcripts in 25 samples using Partek software. (a) Hierarchical clustering of microarray data using the default settings of Euclidean dissimilarity for rows (samples) and columns (transcripts). Colors correspond to expression intensity values. For (b–f) the clustering was repeated and only the dendograms of 25 samples are shown. These use metrics of (b) Canberra, (c) Pearson's dissimilarity, and (d) city block (d). The clustering methods are (a–d) average linkage, (e) centroid linkage, and (f) complete linkage (f). Courtesy of Partek Inc.

form groups that have homogeneity (internal cohesion) and separation (external isolation) (Sneath and Sokal, 1973; Everitt *et al.*, 2001). The relationships between objects being studied, whether intensity measurements from microarray data or operational taxonomic units (OTUs) in phylogeny, are assessed by similarity or dissimilarity measures. Intuitively, the objects in **Figure 11.18a** form two distinct clusters. However, after shifting just two of the data points (**Fig. 11.18b**) it is not clear whether there are two clusters or not. Other challenges to identifying the nature of clusters are depicted in **Figure 11.18c, d**. Each figure shows two apparent clusters that demonstrate both homogeneity and separation. However, if we identify a central point in each cluster (the centroid) and calculate the



**FIGURE 11.17** Defining the relatedness between clusters. (a) Single linkage clustering identifies the nearest neighboring objects between clusters. (b) The single linkage approach is sometimes subject to the artifact of chaining, in which clusters that might reasonably be expected to remain separate are instead connected and merged. (c) Complete linkage clustering identifies the farthest members of each cluster. This approach tends to generate tight, well-separated clusters that exclude objects from clusters. (d) Centroid linkage can represent a compromise approach to placing objects in clusters.



**FIGURE 11.18** Examples of the nature of clusters and clustering approaches. (a) Two clusters are intuitively apparent in a group of 14 data points (circles). Good clusters are characterized by internal cohesion and by separation. (b) Two data points are shifted relative to (a), making the assignment of two clusters more questionable. (c) Two clusters are clearly present, by inspection (“c” shapes). However, the separation between each cluster is not robust. For example, point j in the lower cluster may be closer to the center of the upper cluster than point k, even though j is not a member of the upper cluster. (d) Two clusters are again intuitively apparent. The great distance from the long cluster (e.g., points m to n) presents a challenge to finding a rule that distinguishes that cluster from the small one to its left. Such challenges motivate the development of algorithms to define distances between objects and clusters. (a, c) Adapted from Gordon (1980).

distance to the farthest points within a cluster, that distance will also result in overlap with the adjacent cluster.

Two-way clustering of both genes and samples is used to define patterns of genes that are expressed across a variety of samples (Fig. 11.16a). A dramatic example is provided by Alizadeh *et al.* (2000), who defined subtypes of malignant lymphocytes based upon gene expression profiling (Web Document 11.4).

We can draw several conclusions about hierarchical clustering.

- While hierarchical clustering is commonly used in microarray data analysis, the same underlying dataset can produce vastly different results. Datasets with a relatively small number of samples (typically 4–20) and a large number of transcripts (typically 5000–30,000) occupy high-dimensional space, and different methods summarize the relationships of genes and/or samples as influenced by the distance metric that is chosen as well as the strategy for producing a tree.
- Clustering is an exploratory tool, and is used to identify associations between genes and/or between samples. However, clustering is not used inferentially.
- Clustering is not a classification method (see “Classification of Genes or Samples” below). It is unsupervised in that information about classes (e.g., trisomy 21 versus control) is not used to generate the clustering tree.

### Partitioning Methods for Clustering: *k*-Means Clustering

Sometimes we know into how many clusters our data should fit. For example, we may have treatment conditions we are evaluating, or a set number of time points. An alternative type of unsupervised clustering algorithm is a partitioning method that constructs  $k$  clusters (Tavazoie *et al.*, 1999). The steps are as follows. (1) Choose samples and/or genes to be analyzed. (2) Choose a distance metric such as Euclidean. (3) Choose  $k$ ; data are classified into  $k$  groups as specified by the user. Each group must contain at least one object  $n$  (e.g., gene expression value), and each object must belong to exactly one group. (In all cases,  $k \leq n$ .) Two different clusters cannot have any objects in common, and the  $k$  groups together constitute the full dataset. (4) Perform clustering. (5) Assess cluster fit.

How is the value of  $k$  selected? If you perform a microarray experiment with two different kinds of diseased samples and one control sample, you might choose  $k = 3$ . Also,  $k$  may be selected by a computer program that assesses many possible values of  $k$ . The output of  $k$ -means clustering does not include a dendrogram because the data are partitioned into groups, but without a hierarchical structure.

The  $k$ -means clustering algorithm is iterative. It begins by randomly assigning each object (e.g., gene) to a cluster. The center (“centroid”) of each cluster is calculated (defined using a distance metric). Other cluster centers are identified by finding the data point farthest from the center(s) already chosen. Each data point is assigned to its nearest cluster. In successive iterations, the objects are reassigned to clusters in a process that minimizes the within-cluster sum of squared distances from the cluster mean. After a large number of iterations, each cluster contains genes with similar expression profiles. Tavazoie *et al.* (1999) described the use of  $k$ -means clustering to discover transcriptional regulatory networks in yeast.

A concern with using  $k$ -means clustering is that the cluster structure is not necessarily stable in that it can be sensitive to outliers. Cluster fit has been assessed using a variety of strategies such as measuring the effect of adding random noise to a dataset.

We can select partition clustering in Partek using a Euclidean distance function and the  $k$ -means clustering method. You can select a number of clusters (e.g., 3 anticipating heart, astrocyte, and brain clusters) or check a range of possible cluster sizes (e.g., from 2 to 10). For each cluster a Davies–Bouldin metric is plotted, indicating from its profile that

using 3 clusters is reasonable (Fig. 11.19a). When these are plotted a PCA plot results with added data points (black spheres) at the center of each cluster (Fig. 11.19b).

### Multidimensional Scaling Compared to Principal Components Analysis

We have seen that principal components analysis (PCA) is an exploratory technique used to find patterns in gene expression data from microarray experiments. It involves a linear projection of data from a high-dimensional space to two or three dimensions.

Multidimensional scaling (MDS) is a related dimensional reduction technique that uses nonlinear projection. MDS plots represent the relationships between objects from a similarity (or dissimilarity) matrix in a manner comparable to PCA. We may compare and contrast these techniques.

- Both reduce the dimensionality of datasets to easily interpretable plots.
- Both represent the relatedness of objects (e.g., samples) in an unsupervised fashion. You must therefore interpret the meaning of separation of samples (e.g., based on tissue, expression value, or some other attribute) based on your knowledge of the underlying dataset.
- PCA reports information content in terms of the percent of variance explained along each principal component axis; MDS does not.
- MDS may more accurately reflect small dissimilarities between samples.

We can use MDS in R with `cmdscale` (in the `stats` package), `plotMDS` (in `limma`), or `plotMDS.DGEList` (in the `edgeR` package for RNA-seq data). MDS of our 250 genes in 25 samples in Partek shows separation of the 25 samples into 4 clusters (Fig. 11.19c).

### Clustering Strategies: Self-Organizing Maps

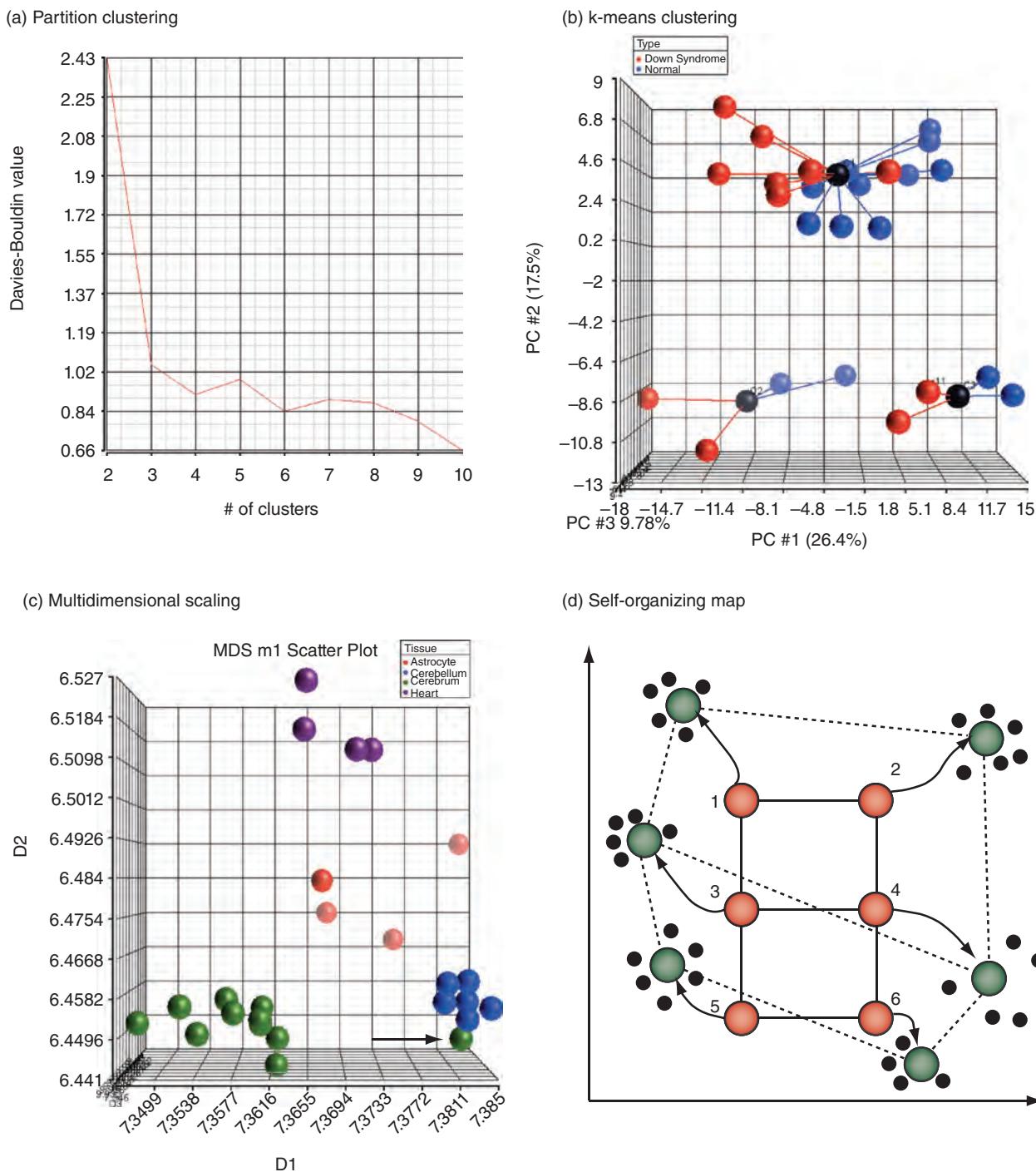
The self-organizing map (SOM) algorithm resembles  $k$ -means clustering in that it partitions data into a two-dimensional matrix. For SOMs and other structured clustering techniques, you can estimate the number of clusters you expect (e.g., based on the number of experimental conditions) in order to decide on the initial number of clusters to use.

Unlike  $k$ -means clustering which is unstructured, SOMs impose a partial structure on the clusters (Tamayo *et al.*, 1999). Also in contrast to  $k$ -means clustering, adjacent partitions in SOMs can influence each other's structure. The principle of SOMs is as follows (Fig. 11.19d). A number of “nodes” (similar to a value  $k$ ) and an initial geometry of nodes such as a  $3 \times 2$  rectangular grid (indicated by solid lines in the figure connecting the nodes) are chosen. Clusters are calculated in an iterative process, as in  $k$ -means clustering, with additional information from the profiles in adjacent clusters. Nodes migrate to fit the data during successive iterations. The result is a clustering tree with an appearance similar to those produced by hierarchical clustering.

The SOM approach to microarray data analysis has been championed by Todd Golub, Eric Lander, and colleagues from the Whitehead Institute.

### Classification of Genes or Samples

The distances and similarities among gene expression values can be described using two types of analysis: supervised or unsupervised. The unsupervised approaches we have described so far are especially useful for finding patterns in large datasets. In supervised analyses, the approach is different because the experimenter assumes some prior knowledge of the genes and/or samples in the experiment. For example, transcriptional profiling has been performed on cell lines or biopsy samples that are either normal or cancerous (e.g., prominent early studies were by Alizadeh *et al.*, 2000; Perou *et al.*, 1999). In some cases, the cancerous samples are further subdivided into those that are relatively malignant or relatively benign. Some of these studies apply unsupervised approaches.



**FIGURE 11.19** Data visualization methods. (a) In partition clustering it is possible to select an optimal number of clusters based on a Davies–Bouldin statistic. Here a plateau at a value of 3 (for 25 samples and 250 chromosome 21 transcripts) suggests an appropriate number of clusters. (b) k-means clustering resembles PCA output, but includes a sphere corresponding to cluster centroids. (c) Multidimensional scaling (MDS) produces a clustering pattern that resembles PCA but does not include information on the percent of variance that is explained. However, MDS can more accurately depict the relationships of objects. Here, note that the cerebrum and cerebellum samples are well separated (except for a single cerebrum sample; see arrow). (d) Self-organizing maps (SOMs) allow partial structuring to be imposed on clusters. This contrasts with k-means clustering, which imposes a fixed number of clusters. An initial set of nodes (numbered 1–6) forms a rectangular grid. During iterations of the self-organizing map algorithm, the nodes migrate to new positions (arrows) to better fit the data. Black dots represent data points.

Source for (d): Tamayo *et al.* (1999). Reproduced with permission from the National Academy of Sciences.

The goal of supervised microarray data analysis algorithms is to define a rule that can be used to assign genes (or conditions) into groups. In each case, we begin with gene expression values from known groups (e.g., normal versus cancerous) and “train” an algorithm to learn a rule. Positive and negative examples are used to train the algorithm. The algorithm is then applied to unknown samples, and its accuracy as a predictor or classifier is assessed. It is critical that the data used for building a classifier are entirely separate from the data used to assess its predictive accuracy.

Some of the most commonly applied supervised data analysis algorithms are support vector machines, supervised machine learning, neural networks, and linear discriminant analysis. As an example of a supervised approach, Brown *et al.* (2000) used support vector machines to classify six functional classes of yeast genes: tricarboxylic acid cycle, respiration, cytoplasmic ribosomes, proteasome, histones, and helix–turn–helix proteins. They used a threefold cross-validation method: the dataset is divided into thirds (sets 1, 2, and 3). Sets 1 and 2 are used to train the support vector machine, then the algorithm is tested on set 3 as the “unknowns.” Next, sets 1 and 3 are used for training and set 2 is tested as the unknowns. Finally, sets 2 and 3 are used for training, and set 1 is tested. They measured the false positive rate and found that support vector machines outperform both unsupervised clustering and alternative supervised clustering approaches.

Dupuy and Simon (2007) described many strategies for properly performing supervised analyses, and also listed many of the common data analysis errors. For example, improperly performing cross-validation leads to overly optimistic prediction accuracy. It is also essential to have an adequate sample size for both the training and the test sets.

Luigi Marchionni, Jeff Leek and colleagues (2013) stress the importance of reproducible tests for clinical applications such as measuring gene expression to classify samples as having cancer or not. Hundreds of thousands of patients receive diagnoses from the leading clinical tests. Marchionni *et al.* developed a predictor from training data and evaluated it with independent test data.

We can perform classification using Partek, asking the question if we can build a classifier that can discriminate tissue of origin (heart, astrocyte, cerebrum, cerebellum) from expression data. Instead of using an external dataset for independent validation, we employ leave-one-out cross-validation. In this approach, the dataset is divided into ten random partitions (**Fig. 11.20a**). At each pass, nine-tenths of the data are used for training and one-tenth is withheld for testing. We can decide to create one classifier or study a set of possible classifiers (using two-level nested cross-validation; **Fig. 11.20b**). Variable selection can be performed, using ANOVA or other approaches to improve the accuracy of classification (**Fig. 11.20c**). We can select  $k$ -nearest neighbor or other classification approaches, including over a dozen distance measures (e.g., Euclidean, Canberra, or Pearson’s as described above). After running the model we can assess its accuracy with a variety of metrics. A confusion matrix shows the number of samples that are truly from a region compared to the predicted tissue source (**Table 11.4**). For a perfect classifier, all values would occur on the diagonal. In our example 17 samples are correctly classified, while 8 samples are misclassified (e.g., four cerebrum samples are incorrectly assigned to cerebellum).

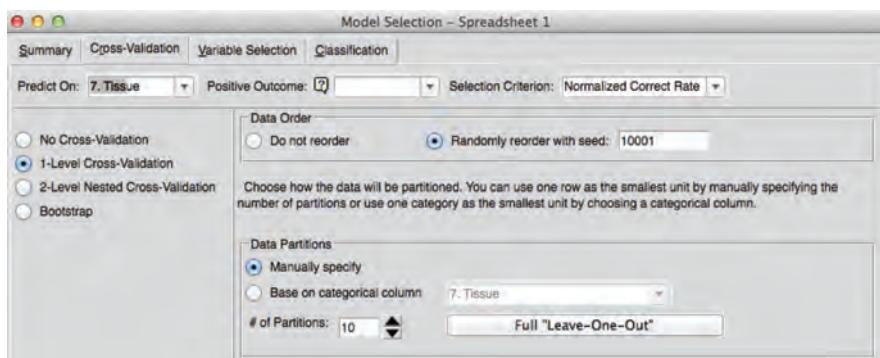
## RNA-SEQ

RNA-seq is used to accomplish the same goal as microarrays: quantifying RNA transcript levels. However, it is considered revolutionary because it allows the measurement of essentially all RNA transcripts (rather than only those pre-selected on a microarray surface), it has a broader dynamic range, it allows identification of novel transcripts and transcript isoforms, and it is able to quantify alternative splicing events. For reviews including

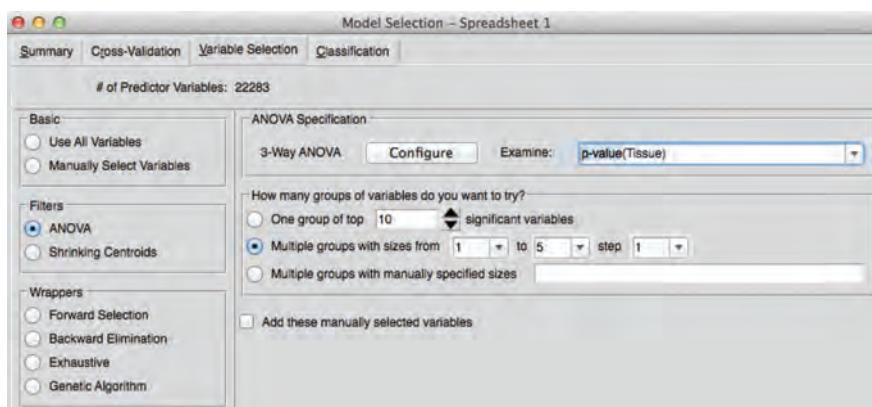
(a) Ten-fold leave one out cross-validation procedure

Pass 1	Pass 2	Pass 3	Pass 4	Pass 5	Pass 6	Pass 7	Pass 8	Pass 9	Pass 10
test	train								
train	test	train							
train	train	test	train						
train	train	train	test	train	train	train	train	train	train
train	train	train	train	test	train	train	train	train	train
train	train	train	train	train	test	train	train	train	train
train	train	train	train	train	train	test	train	train	train
train	test	train	train						
train	test	train							
train	test								

(b) Cross-validation



(c) Variable selection



**FIGURE 11.20** Classification of microarray data. (a) A leave-one-out cross-validation approach divides a dataset into randomly selected partitions ( $n = 10$  in this case). For each pass at building a classifier, one part is withheld for subsequent testing. This provides an alternative to testing classifiers on independent datasets. (b) Cross-validation options include the simpler one-level method, or two-level nested cross-validation to evaluate multiple classifiers in parallel. (c) Variable selection includes ANOVA and other approaches. For example, forward selection each variable (e.g., gene expression measurement) is evaluated separately and paired with remaining variables according to optimization criteria. Courtesy of Partek Inc.

**TABLE 11.4 Confusion matrix from classification of tissue types from microarray data. Gene expression data were analyzed in Partek using  $K$ -nearest neighbor with Euclidean distance measure. The number of samples is 25.**

Real/Predicted	Cerebellum	Heart	Cerebrum	Astrocyte
Cerebellum	3	0	2	1
Heart	0	4	0	0
Cerebrum	4	0	7	0
Astrocyte	0	0	1	3

details of analysis methods see Wang *et al.* (2009), Nagalakshmi *et al.* (2010), Garber *et al.* (2011), and Ozsolak and Milos (2011).

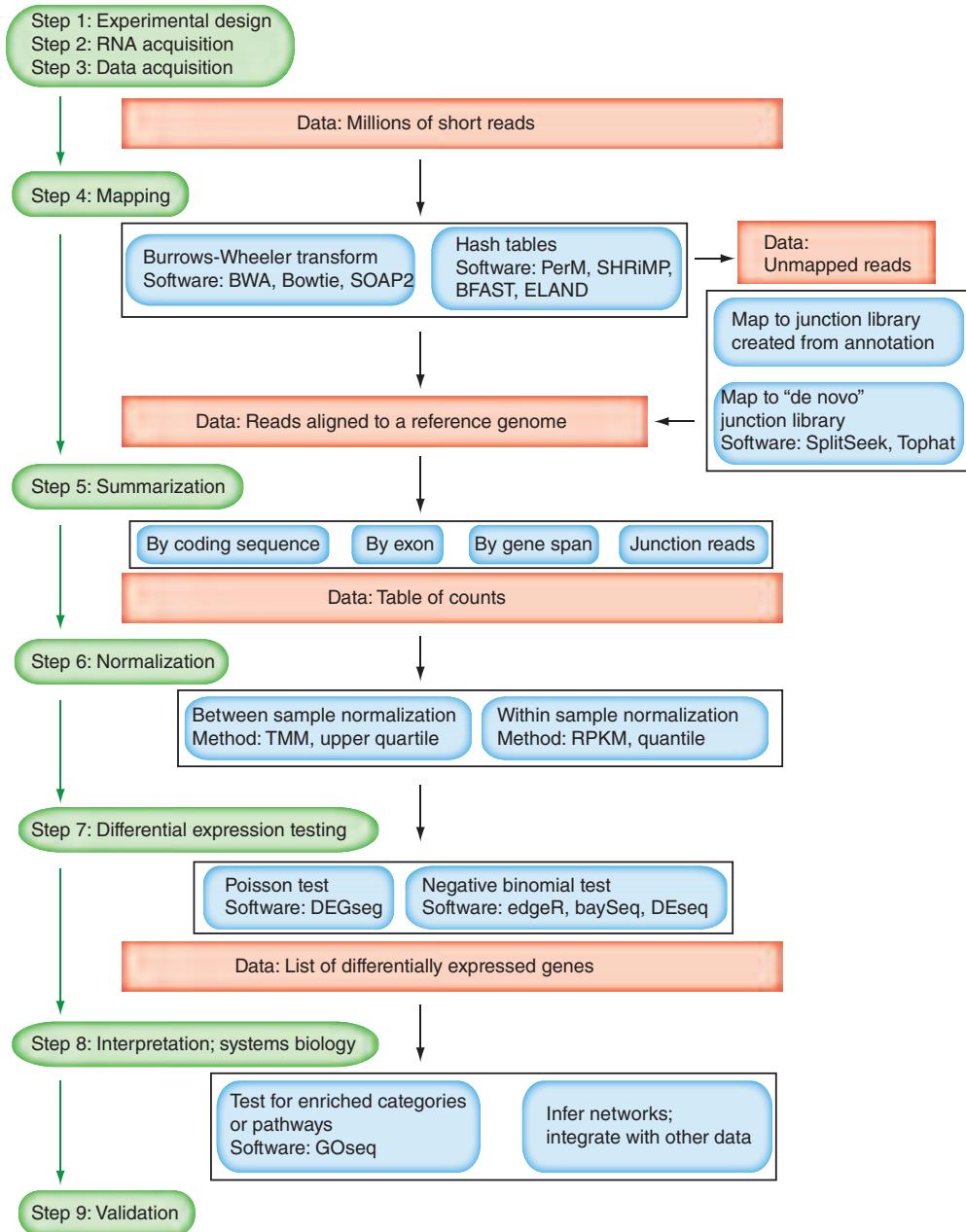
There are extraordinary data analysis challenges associated with this new technology. Some consider RNA-seq analysis to be vastly more challenging than either microarray data analysis or even the analysis of next-generation sequence analysis for DNA. Soneson and Delorenzi (2013) compare 11 methods, finding that from ~200 to ~3200 differentially expressed genes are identified for one particular dataset, with highly variable overlap between methods. Why does this occur? No single method is optimal, methods vary in how they are influenced by outlier data points, they vary in the sample sizes required to achieve adequate statistical power, and they vary in accuracy. All methods are affected by variations in read coverage across the expressed portions of the genome.

We can consider a workflow for RNA-seq experiments (Fig. 11.21; Oshlack *et al.*, 2010). Experimental design (step 1) needs to include sufficient replicates to measure the biological variability between samples (Hansen *et al.*, 2011). Hansen *et al.* note that the variability in RNA transcript levels is similar in microarray and RNA-seq technologies, and that individual transcripts vary greatly in their biological variability. It is common for biologists to design RNA-seq experiments with a sample size of just one or two per group, but more are needed to enable meaningful conclusions about the findings.

You can do power calculations for RNA-seq data using R. We'll load the package `RNASeqPower` and assess four effect sizes (1.25 to 2) at read depths of 20 then 200 and a power of 0.9 (i.e., the fraction of true positives that will be detected).

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("RNASeqPower")
> library(RNASeqPower)
> rnapower(depth=20, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
+ alpha=.05, power=.9)
  1.25      1.5      1.75      2
 88.629200 26.843463 14.091771 9.185326
> rnapower(depth=200, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
+ alpha=.05, power=.9)
  1.25      1.5      1.75      2
 69.637228 21.091292 11.072106 7.217042
```

RNA acquisition (step 2 of the workflow) often involves isolation of messenger RNA and enrichment of complementary DNA corresponding to exons using oligonucleotide baits (Fig. 11.22). Other techniques are also available (Ozsolak and Milos, 2011). For the next steps in the RNAseq workflow – mapping, summarization, normalization, and differential expression testing – a key aspect of the mapping of RNA-seq reads is transcriptome assembly. This can be performed *de novo* or using a reference

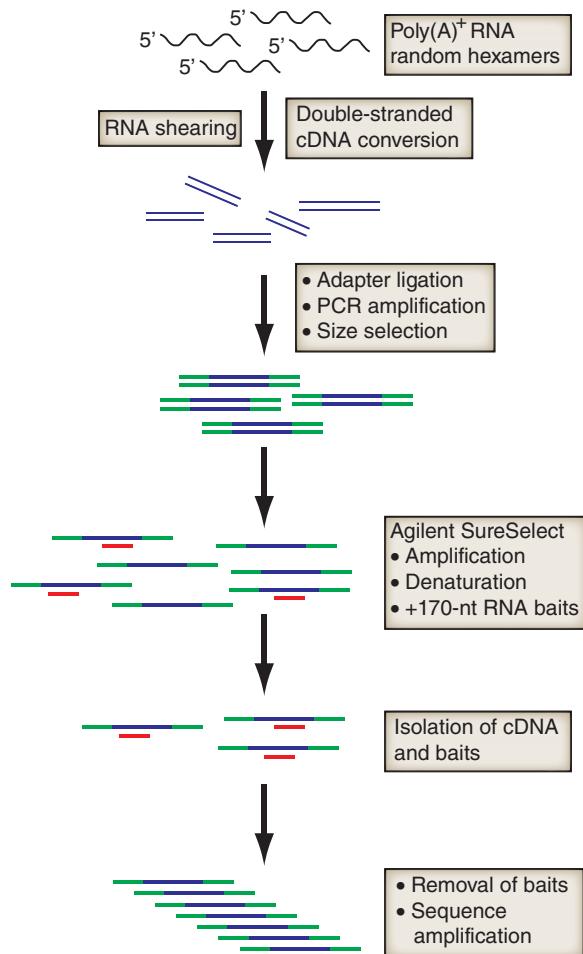


**FIGURE 11.21** Workflow for RNA-seq data analysis. Steps in the pipeline are shown in green-shaded ellipses; datasets are in peach rectangles; methods are shown in blue boxes.

Source: Oshlack *et al.* (2010). Licensed under Creative Commons Attribution License 4.0.

genome (Robertson *et al.*, 2010; Li and Dewey, 2011; Martin and Wang, 2011; Steijger *et al.*, 2013).

We next perform a practical RNA-seq analysis using the popular tools TopHat and Cufflinks. These are free, open-source software packages; they can be used to identify novel splice variants (and novel genes, depending on how well the organism of interest has already been annotated), and they can be used to measure differential expression levels of RNA transcripts. Many other RNA-seq software packages are available (listed in Oshlack *et al.*, 2010). Guo *et al.* (2013), Kvam *et al.* (2013) and Soneson and Delorenzi (2013) each evaluate a variety of recent software packages, assessing factors such as



**FIGURE 11.22** Method for targeted RNA-seq. The Agilent SureSelect workflow is shown. Adapted from Ozsolak and Milos (2011) with permission from Macmillan Publishers.

speed and accuracy. Gordon Smyth and colleagues introduced Voom, a method for RNA-seq analysis that uses a limma empirical Bayes analysis pipeline (Law *et al.*, 2014).

### Setting up a TopHat and CuffLinks Sample Protocol

Cole Trapnell, Lior Pachter and colleagues provide a description of `TopHat` and `Cufflinks` as well as a detailed protocol for using the software to analyze differential expressions (Trapnell *et al.*, 2012). We follow that exact protocol using a Linux server (you should preferably use a machine with more than 4 GB of RAM); refer to their paper for additional details. We perform the following set-up tasks: (1) organize our directories; (2) download a *Drosophila* reference genome; and (3) download sequence data in the FASTQ format. We then proceed with the analysis using `TopHat` then `CuffLinks` then `CuffMerge`, downloading required software packages at each stage. We conclude by plotting the results with the R package `cummeRbund`.

We begin by organizing our directories. First, open terminal and navigate to your home directory with `cd ~`. Make directories for this tutorial and (optionally) for your data files.

```
$ mkdir rnaseq_tutorial
$ mkdir data
```

The Cufflinks website (<http://cole-trapnell-lab.github.io/cufflinks/>, WebLink 11.13) includes links to many available genomes. The one we will use is [ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila\\_melanogaster/Ensembl/BDGP5.25/Drosophila\\_melanogaster\\_Ensembl\\_BDGP5.25.tar.gz](ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila_melanogaster/Ensembl/BDGP5.25/Drosophila_melanogaster_Ensembl_BDGP5.25.tar.gz) (WebLink 11.14). GTF files from many dozens of organisms are also available from Ensembl at <http://www.ensembl.org/info/data/ftp/> (WebLink 11.15).

In a Linux environment, type `$ tar -zxvf myfile.tar.gz` to open a compressed file.

The GTF format is described at <http://mblab.wustl.edu/GTF2.html> (WebLink 11.16).

The GEO page (<http://www.ncbi.nlm.nih.gov/geo/>, WebLink 11.17) includes a link with instructions on how to download data, and whether the data are original GEO records or curated DataSets and Profiles. You can begin by entering GSE32038 as a query of the home page of NCBI, and search GEO DataSets; alternatively, visit <http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE32038> (WebLink 11.18) directly.

The TopHat website is <http://ccb.jhu.edu/software/tophat/index.shtml> (WebLink 11.19). We can download the binary release. Also, we can copy the `tophat` binary to the `~/bin` directory. The `-p` argument refers to the number of threads and depends on the configuration of your computer. `-G` refers to the GTF file name. Bowtie's website is <http://bowtie-bio.sourceforge.net/index.shtml> (WebLink 11.20).

We need a reference *Drosophila* genome. We work with a file called `genes.gtf`, which can obtain via an iGenome website or (optionally) from Ensembl. After you download the file, navigate to your data directory, copy the tar compressed file, and unpack it:

```
$ cd ~/data # cd is the command to change directory.  
# ~/data refers to a directory called data under the home directory.  
$ cp ~/Downloads/Drosophila_melanogaster_Ensembl_BDGP5.25.tar ~/data/  
$ tar xzpf Drosophila_melanogaster_Ensembl_BDGP5.25.tar
```

As we use `cd` to change directory and see the *Drosophila* genome contents, we encounter three directories: `Annotation`, `GenomeStudio`, and `Sequence`. The `Annotation` directory contains a `Genes` folder which includes the file `genes.gtf`. We can inspect the first two lines of the GTF file using `head`, and using `wc -l` we can see it has about 270,000 rows. Then (working from our `rnaseq_tutorial` directory) we can create a symbolic link (`ln -s`) to it and to a set of Bowtie index files in a neighboring directory. The links allow us to later specify a file (or set of files) in other directories by simply pointing to them. The final `.` symbol indicates that the link should be placed here in the current directory.

```
$ ln -s ~/data/Drosophila_melanogaster/Ensembl/BDGP5.25/Sequence/  
Bowtie2Index/genome* . # the * specifies all files in that directory  
# beginning with genome  
$ ln -s ~/data/Drosophila_melanogaster/Ensembl/BDGP5.25/Annotation/Genes/  
genes.gtf .  
$ head -2 genes.gtf  
2L protein_coding exon 75298116 . + . exon_number "1";  
gene_id "FBgn0031208"; gene_name "CG11023"; p_id "P9062"; transcript_id  
"FBtr0300689"; transcript_name "CG11023-RB"; tss_id "TSS8382";  
2L protein_coding exon 75298116 . + . exon_number "1";  
gene_id "FBgn0031208"; gene_name "CG11023"; p_id "P8862"; transcript_id  
"FBtr0300690"; transcript_name "CG11023-RC"; tss_id "TSS8382";
```

Our next task is to download RNA-seq data. We use GSE32038 from GEO. I prefer to store files such as these in a directory called `data`; after unpacking the FASTQ files, copy them to the `rnaseq_tutorial` directory with `cp` and shorten their names as shown below. (To rename a file from an old name to a new name use `mv old.txt new.txt`). They are named C1 and C2 (for conditions 1 and 2); R1, R2, and R3 (for three biological replicates); and `1.fq` or `2.fq` for the forward or reverse reads of paired end sequencing. There are 12 files in total, each 1.8 GB in size and, from `$ grep -c '@' GSM794483_C1_R1_1.fq`, have 11.6 million reads.

## TopHat to Map Reads to a Reference Genome

TopHat is a fast splice junction mapper for RNA-seq data (Kim *et al.*, 2013). It uses Bowtie (Langmead and Salzberg, 2012) to align reads to a reference genome. We download, unpack, and install those two programs.

To use TopHat we first map the reads for each sample to a reference genome.

```
$ tophat -p 8 -G genes.gtf -o C1_R1_thout genome C1_R1_1.fq C1_R1_2.fq  
$ tophat -p 8 -G genes.gtf -o C1_R2_thout genome C1_R2_1.fq C1_R2_2.fq  
$ tophat -p 8 -G genes.gtf -o C1_R3_thout genome C1_R3_1.fq C1_R3_2.fq  
$ tophat -p 8 -G genes.gtf -o C2_R1_thout genome C2_R1_1.fq C2_R1_2.fq  
$ tophat -p 8 -G genes.gtf -o C2_R2_thout genome C2_R2_1.fq C2_R2_2.fq  
$ tophat -p 8 -G genes.gtf -o C2_R3_thout genome C2_R3_1.fq C2_R3_2.fq
```

The output includes a folder for each run (e.g., `-o C1_R1_thout` specifies the TopHat output folder for condition 1, biological replicate 1). This includes BAM files (`accepted_hits.bam` and `unmapped.bam`), BED files, and a set of log files. To

assess the quality of the mapping try using SAMtools (this step is not in the Trapnell *et al.* protocol):

```
$ samtools flagstat accepted_hits.bam
```

The output includes the percent of reads that map to the reference. It also describes how many reads mapped with a mate to a different chromosome.

Note that the Trapnell *et al.* paper protocol mistakenly labels the FASTQ files in the C2 group. See Web Document 11.5 for a correct version. Each TopHat run takes about a half hour using several cores.

## Cufflinks to Assemble Transcripts

Cufflinks determines the fragment length distributions of reads from a BAM file, that is, from RNA-seq reads that have been aligned to the human genome. It assembles transcripts for each sample, and estimates abundances. In the following commands we invoke `CuffLinks`, specify that we want to use 8 processors (you may have more or fewer), define the output filename, and specify the inputs as `accepted_hits` BAMfiles that are located within the various `thout` (TopHat output) folders that were created in the previous TopHat run.

```
$ cufflinks -p 8 -o C1_R1_clout C1_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R2_clout C1_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C1_R3_clout C1_R3_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R1_clout C2_R1_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R2_clout C2_R2_thout/accepted_hits.bam
$ cufflinks -p 8 -o C2_R3_clout C2_R3_thout/accepted_hits.bam
```

The Cufflinks outputs are sent to a folder with text files listing the loci and lengths of genes, transcripts, and isoforms.

Next, create a file called `assemblies.txt`. This lists the assembly file for each sample.

```
$ nano assemblies.txt
$ less assemblies.txt
./C1_R1_clout/transcripts.gtf
./C1_R1_clout/transcripts.gtf
./C1_R1_clout/transcripts.gtf
./C1_R1_clout/transcripts.gtf
./C1_R1_clout/transcripts.gtf
./C1_R1_clout/transcripts.gtf
```

We next run Cuffmerge on all the assemblies. This generates a single merged transcriptome annotation. The `-s` option specifies the genomic DNA sequences for the reference, while `-g genes.gtf` is an optional reference GTF file.

```
$ cuffmerge -g genes.gtf -s genome.fa -p 8 assemblies.txt
```

The output is a merged transcriptome annotation (in a file called `merged_asm`, placed in a new subfolder). Word count (`wc -l`) tells us it has 143,569 rows.

## Cuffdiff to Determine Differential Expression

Cuffdiff is used to identify differentially expressed genes and transcripts. We use the merged transcriptome assembly and the BAM files from TopHat. In the following command, the argument `-o` specifies the output directory; `-b` uses a bias correction; `-p` specifies the number of processors we use; `-L` indicates a comma-separated list of condition labels (ours are C1, C2, i.e., conditions 1 and 2); and `-u` is for a read correction method. All options are listed by simply entering `$ cuffdiff`.

```
$ cuffdiff -o diff_out -b genome.fa -p 8 -L C1,C2 -u merged_asm/merged.gtf ./C1_R1_thout/accepted_hits.bam,./C1_R2_thout/accepted_hits.bam,./C1_R3_thout/accepted_hits.bam ./C2_R1_thout/accepted_hits.bam,./C2_R3_thout/accepted_hits.bam,./C2_R2_thout/accepted_hits.bam
```

The output consists of 18 files in the `diff_out` folder, which we explore in R in the following section.

### CummeRbund to Visualize RNA-seq Results

We use the R package `cummeRbund` to visualize our results. We can work in R on the command line, and it is also convenient to use the RStudio environment on a Mac (or PC). First we load the `cummeRbund` package into R. In this environment the command prompt is `>` (rather than `$` for Unix).

```
$ R
> source("http://bioconductor.org/biocLite.R")
> biocLite("cummeRbund")
> library(cummeRbund)
```

Next, take the `CuffDiff` output and create a `cummeRbund` database called `cuff_data`. Before we make plots, let's look at the database and explore the transcripts that are most regulated based on *p* value and based on fold change.

```
> cuff_data <- readCufflinks('diff_out')
> ?cuff_data # you can get more details about usage here
> cuff_data
CuffSet instance with:
  2 samples
  14410 genes
  25077 isoforms
  17360 TSS # these are transcription start sites
  18175 CDS # these are coding sequences
  14410 promoters
  17360 splicing
  13270 relCDS
```

Create the file `gene_diff_data` using the `diffData` function, and then select the subset of significantly regulated transcripts. We see the number of rows (271), the dimensions ( $271 \times 11$  columns), and the first few values.

```
> gene_diff_data <- diffData(genes(cuff_data))
> sig_gene_data <- subset(gene_diff_data, (significant == 'yes'))
> nrow(sig_gene_data)
[1] 271
> dim(sig_gene_data)
[1] 271 11
> head(sig_gene_data)
  gene_id sample_1 sample_2 status value_1 value_2 log2_fold_change
3   XLOC_000003   C1      C2    OK    48.4754  82.4077    0.765526
59  XLOC_000059   C1      C2    OK    65.3518 113.1420    0.791835
133 XLOC_000133   C1      C2    OK    84.3472 148.3190    0.814293
180 XLOC_000180   C1      C2    OK    39.4686  59.3858    0.589412
241 XLOC_000241   C1      C2    OK    19.9367  35.7757    0.843553
249 XLOC_000249   C1      C2    OK    24.4575  44.6019    0.866825
  test_stat p_value q_value significant
3   3.91602   5e-05  0.00160278    yes
59  4.81631   5e-05  0.00160278    yes
133 3.94607   5e-05  0.00160278    yes
180 3.59121   5e-05  0.00160278    yes
241 4.48772   5e-05  0.00160278    yes
249 4.08778   5e-05  0.00160278    yes
```

The gene identifier of the most significantly regulated transcript is 3 (from the first entry in the table above). You can pursue this in `biomaRT`, or enter it into a web-based search of NCBI Entrez with 3 AND “Drosophila melanogaster”[`porgn:txid7227`]. For

the search 5752 AND “Drosophila melanogaster”[porgn:\_\_txid7227], the official gene symbol is *Msp-300*.

Which transcript is up-regulated the most? We can take the `sig_gene_data` table and sort it by the column `log2_fold_change`.

```
> attach(sig_gene_data)
> sig_fc <- sig_gene_data[order(-log2_fold_change),]
> head(sig_fc)
   gene_id      sample_1 sample_2 status value_1  value_2 log2_fold_
change
5752 XLOC_005752 C1       C2       OK    398.219 1060.680 1.41335
1272 XLOC_001272 C1       C2       OK    411.755 1064.820 1.37075
2660 XLOC_002660 C1       C2       OK    513.880 1308.490 1.34840
4677 XLOC_004677 C1       C2       OK    1527.160 3881.330 1.34570
678  XLOC_000678 C1       C2       OK    244.958 621.402  1.34299
4609 XLOC_004609 C1       C2       OK    122.289 306.945  1.32768
   test_stat    p_value   q_value significant
5752 9.62010  5e-05  0.00160278  yes
1272 9.07087  5e-05  0.00160278  yes
2660 8.61505  5e-05  0.00160278  yes
4677 9.61730  5e-05  0.00160278  yes
678   8.77837  5e-05  0.00160278  yes
4609 8.00396  5e-05  0.00160278  yes
```

XLOC\_005752 therefore has a fold change of 1.41 and is the most up-regulated of the significantly, differentially expressed transcripts. Next we plot the data.

```
> csDensity(genes(cuff_data))
> csScatter(genes(cuff_data), 'C1', 'C2')
> csVolcano(genes(cuff_data), 'C1', 'C2')
```

The outputs of these plots are shown in **Figure 11.23a–c**. We continue by looking at a specific gene, the *Drosophila* globin glob1:

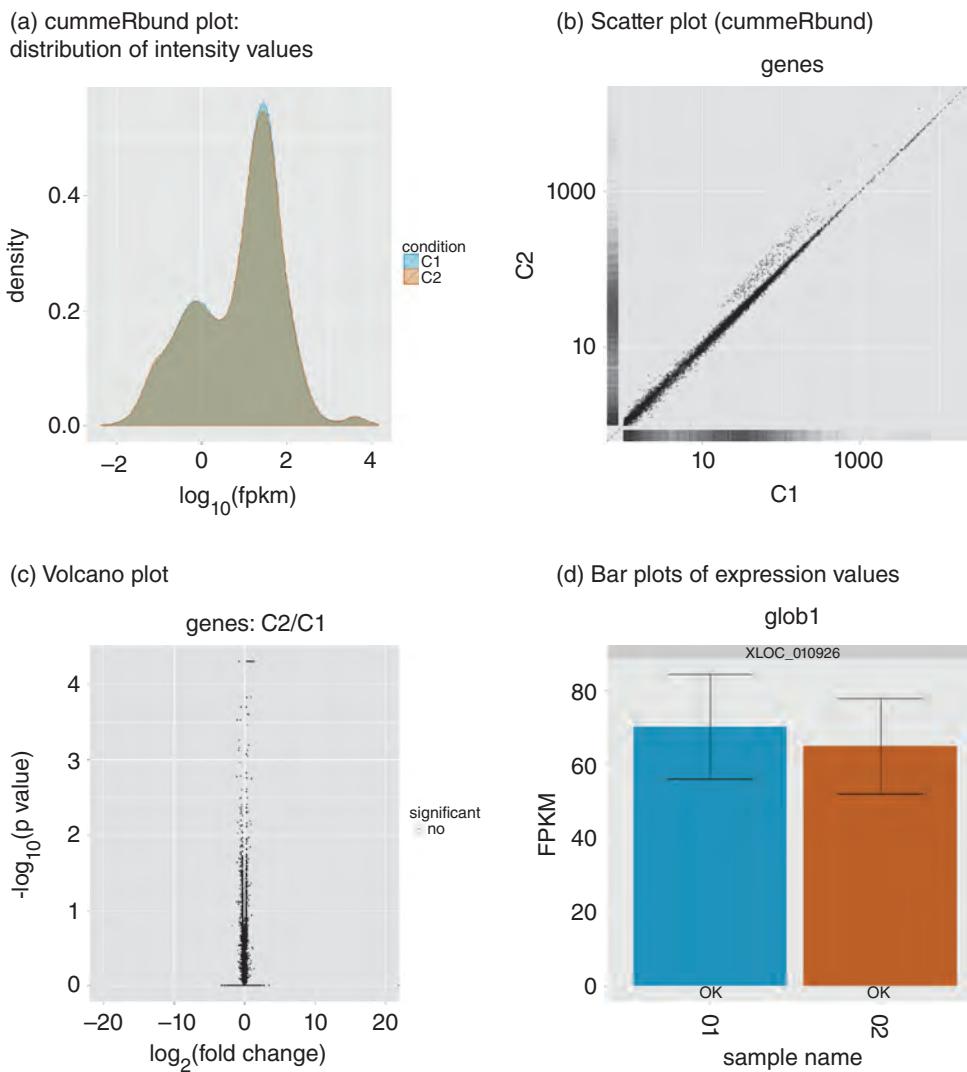
```
> globin <- getGene(cuff_data, 'glob1')
> expressionBarplot(globin)
```

This produces a barplot showing the expression levels in our two groups (**Fig. 11.23d**). Additional analyses offered in the protocol of Trapnell *et al.* (2012) include analyzing particular isoforms and plotting read coverage in IGV (Chapter 9).

## RNA-seq Genome Annotation Assessment Project (RGASP)

RGASP was designed to evaluate computational methods to predict and quantify expressed transcripts from RNA-seq data. Developers of 14 software programs were invited to analyze RNA-seq data to assess methods for exon identification, transcript reconstruction, and expression-level quantification (Steijger *et al.*, 2013). Performance was lower for *Homo sapiens* data than for *Drosophila* or *C. elegans* datasets. There are many computational challenges. For example, identifying all exons cannot be accomplished, and valid assembly of exons into transcript isoforms was accomplished for just 41% of human genes. Methods also vary substantially in their estimates of expression levels from the same gene loci.

In another RGASP project, Engström *et al.* (2013) described the ability of 26 protocols (involving 11 programs) to align transcript reads to a reference genome. There were dramatic differences involving alignment yield (68–95%), concordance of paired end reads, mismatches of number and position, basewise accuracy, indel frequency and accuracy, and spliced alignment (involving both splices detected in individual reads and genomic splice sites).



**FIGURE 11.23** Visualizing RNA-seq data with the R package `cummeRbund`: (a) distribution of intensity values; (b) scatter plot; (c) volcano plot; and (d) bar plots of expression values for a single gene (*Glob1* from *Drosophila*).

Source: R.

Assessments such as RGASP are useful to compare the performance of various software tools, and to identify aspects that need to be improved.

## FUNCTIONAL ANNOTATION OF MICROARRAY DATA

A major task confronting the user of microarrays is to learn the biological significance of the observed gene expression patterns. Often researchers rely on manual literature searches and expert knowledge to interpret microarray results. Several software tools accept lists of accession numbers (corresponding to genes that are represented on microarrays) and provide annotation.

When Christopher Bouton was a graduate student in the Pevsner lab back in 2000, he developed the Database Referencing of Array Genes Online (DRAGON) database. This includes a website that allows microarray data to be annotated with data from publicly available databases such as UniGene, Pfam, SwissProt, and KEGG (Bouton and Pevsner, 2000; Bouton *et al.*, 2003). DRAGON offers a suite of visualization tools allowing the

user to identify gene expression changes that occur in gene or protein families. The goal of annotation tools such as DRAGON is to provide insight into the biological significance of gene expression findings. Today DRAGON is obsolete because of the emergence of richer databases as well as projects such as BioMart (or biomaRt; Chapter 8) that enable comprehensive searches.

An active area of research is the annotation of microarray data based on functional groups such as Gene Ontology categories (we introduce Gene Ontology in Chapter 12). The premise is that, in addition to considering individual transcripts that are significantly regulated, groups that are functionally related (such as transcripts that encode kinases or function in mitochondrial biogenesis) can be identified. Tools to analyze datasets based on annotation groups include GOMiner (Zeeberg *et al.*, 2005) and are reviewed by Osborne *et al.* (2007).

Gene Set Enrichment Analysis (GSEA), introduced by Jill Mesirov and colleagues, represents an increasingly popular approach to identifying regulated sets of genes (Subramanian *et al.*, 2005; reviewed in Hung *et al.*, 2012). Suppose you measure genome-wide expression in two classes (e.g., control and wildtype), obtaining  $\log_2$  ratios of samples in these conditions for >20,000 transcripts. GSEA examines predefined groups, such as 75 genes defined as relevant to heart development or 750 genes defined as relevant to the regulation of transcription. GSEA tests whether the set of genes in each of those groups are randomly distributed among all 20,000 measurements (null hypothesis) or not (alternate hypothesis). GSEA: (1) calculates an enrichment score; (2) estimates significance with a permutation test (the class labels are permuted randomly as part of the null model, and the enrichment score from scrambled labels is calculated 1000 times); and (3) performs a multiple test correction. The false discovery rate is the estimated probability that a gene set with some enrichment score is a false positive result. The GSEA developers have suggested that an FDR of 25% is reasonable, although the user must decide what is appropriate for generating hypotheses about biological function.

GSEA software is available from the Broad Institute at <http://www.broadinstitute.org/gsea/index.jsp> (WebLink 11.21).

With all these annotation procedures, it is important to keep in mind that the product of mRNAs is protein. Identification of a set of mRNAs encoding proteins in a particular cellular pathway does not mean that the proteins themselves are present in altered levels, nor does it mean that the function of that pathway has been perturbed. Such conclusions can only be drawn from experiments on proteins and pathways performed at the cellular level.

## PERSPECTIVE

DNA microarray technology allows the experimenter to rapidly and quantitatively measure the expression levels of thousands of genes in a biological sample. This technology emerged in the late 1990s as a tool to study diverse biological questions. Thousands to millions of data points are generated in microarray experiments. Microarray data analysis therefore employs mathematical tools that have been established in other data-rich branches of science. These tools include cluster analysis, principal components analysis, and other approaches to reduce highly dimensional data to a useful form. The main questions that microarray data analysis seeks to answer are as follows:

- For a comparison of two conditions (e.g., cell lines treated with and without a drug), which genes are dramatically and significantly regulated?
- For comparisons across multiple conditions (e.g., analyzing gene expression in 100 cell lines from normal and diseased individuals), which genes are consistently and significantly regulated?
- Is it possible to cluster data as a function of sample and/or as a function of genes?

RNA-seq emerged more recently as a complementary approach to measuring steady-state RNA levels. A challenge is to translate the discoveries from these experiments into further insight about biological mechanisms.

Finally, while DNA microarrays and RNA-seq have been used to measure gene expression in biological samples, they have also been used in a variety of alternative applications. Microarrays have been used as a tool to detect genomic DNA (e.g., to identify polymorphisms, to obtain DNA sequence, to identify regulatory DNA sequence, to identify deletions and duplications, and to determine the methylation status of DNA). RNA-seq has been applied to small non-coding RNAs and to variant detection (complementing DNA sequencing studies). Such diverse applications are likely to expand in the near future.

## PITFALLS

A key study by John Ioannidis and colleagues (2009) showed that of 20 gene expression microarray papers published in *Nature Genetics*, the results of only 2 could be reproduced; 6 could be reproduced with discrepancies; and 10 could not be reproduced. Problems included lack of availability of the raw data; requirement for software that is not available; or unclear analysis methods. Roger Peng (2011) described approaches to performing reproducible research including access to the original data, code used for analysis, and everything needed for full replication of the study.

Dupuy and Simon (2007) reviewed 90 publications in which gene expression profiles were related to cancer outcome. Half of the studies they reviewed in detail had at least one of three flaws:

1. Controls for multiple testing were not properly described or performed.
2. In class discovery a correlation was claimed between clusters and clinical outcomes. However, such correlation is spurious because differentially expressed genes were identified and then used to define clusters.
3. Supervised predictions included estimates of accuracy that were biased because of incorrect cross-validation procedures.

Dupuy and Simon (2007) offer a useful and practical list of 40 guidelines for the statistical analysis of microarray experiments, spanning topics from data acquisition to identifying differentially regulated genes, class discovery, and class prediction.

For RNA-seq studies, a major problem is that vast numbers of experiments include either no replicates or just duplicates. This is problematic because biological variability cannot be assessed. It is not possible to generalize the meaning of the results; instead, for each condition we can only know what RNA changes occurred in the particular one or two samples that were studied.

Errors occur in a variety of stages of microarray and RNA-seq experiments:

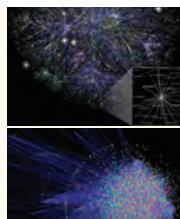
- Experimental design is a critical but often overlooked stage of a microarray experiment. It is essential to study an adequate number of experimental and control samples. The appropriate number of replicates must also be employed. While there is no consensus on what this number is for every experiment, there must be adequate statistical power and using one to three biological replicates is often insufficient.
- It is difficult to relate intensity values from gene expression experiments to actual copies of mRNA transcripts in a cell. This situation arises because each step of the experiment occurs with some level of efficiency, from total RNA extraction to conversion to a probe labeled with fluorescence and from hybridization efficiency to variability in image analysis. Some groups have introduced universal standards for analysis of a uniform set of RNA molecules, but these have not yet been widely adopted.
- Data analysis requires appropriate attention to global and local background correction. Benchmark studies suggest that while excellent approaches have been developed (such as GCRMA), applying different normalization procedures will lead to different outcomes (such as differing lists of regulated transcripts).

- For exploratory analyses, the choice of distance metric, such as Pearson's correlation coefficient, can have a tremendous influence on outcomes such as clustering of samples.
- Each data analysis approach has advantages and limitations. For example, popular unsupervised methods (such as cluster analysis) sacrifice information about the classes of samples that are studied (such as cell lines derived from patients with different subtypes of cancer). Supervised methods make assumptions about classes that could be false.
- Many experimental artifacts can be revealed through careful data analysis. Skewing of scatter plots may occur because of contamination of the biological sample being studied. Cluster analysis may reveal consistent differences, not between control and experimental conditions, but between samples analyzed as a function of day or operator.

## ADVICE FOR STUDENTS

For work with RNA-seq, as for any high throughput or next-generation sequencing technology, I believe that biologists need to actively collaborate with biostatisticians regarding experimental design and data analysis. As you begin to analyze RNA-seq data here are several suggestions. (1) Make a goal of gaining experience working in the Linux environment. In this chapter we followed a protocol by Trapnell *et al.* (2012); try it yourself. (2) Once you learn a single workflow, try to deepen your understanding by reading the documentation for the packages you choose; reading papers in the literature that use that workflow; and running the workflow multiple times in order to understand the effects of changing various parameters. (3) Many people begin with RNA-seq analysis in Galaxy. It is user-friendly and offers excellent tutorials and documentation. Some researchers rely on Galaxy entirely, while for many others it provides initial exposure to a variety of RNA-seq analysis tools and it can serve as a stepping-stone using those tools in Linux. (4) Try to develop a feel for the current state of RNA-seq data analysis. Read review papers cited in this chapter; note some of the main challenges at each step such as the choice of aligner, or the difficulty of assembly. New tools are continuously developed and you should actively follow the literature. Read broadly and note how the authors assess the performance of their software and perform benchmarking to show how and when it outperforms other software. (5) Join forums such as Biostars and Seqanswers to keep in touch with new developments in the community.

Biostars is online at <http://www.biostars.org> (WebLink 11.22). For a tutorial on analyzing microarray data using Bioconductor, visit <https://www.biostars.org/p/53870/>.



## Discussion Questions

**[11.1]** A microarray dataset can be clustered using multiple approaches, yielding different results. How can you decide which clustering results are “correct” (most biologically relevant)? For microarray data normalization we described the concepts of precision and accuracy; do these apply to clustering as well?

**[11.2]** What are the best criteria to use to decide if a gene is significantly regulated? If you apply fold change as a criterion, will there be situations in which a fold change is statistically significant but not likely to be significant in a biological sense? If you apply a conservative correction

and find that no genes change significantly in their expression levels in a microarray experiment, is this a biologically plausible outcome?

**[11.3]** In this chapter we examined a dataset comparing trisomy 21 samples versus euploid controls, and observed an increase in the levels of RNA transcripts assigned to chromosome 21 genes. What other microarray datasets involve experiments for which you can hypothesize what changes might occur? Consider cancer studies (e.g., tumor/normal), wildtype versus knockout experiments, pharmacological treatments (e.g., cells  $\pm$  a drug), or studies of physiological states.

## PROBLEMS/COMPUTER LAB

**[11.1]** We described GEO2R. Visit NCBI's GEO and select another dataset to analyze. (For example, search GEO DataSets with the term encode or "1000 Genomes" and select a dataset having a GEO2R link.) Inspect the R code; copy the R code to a text editor. Open R (or RStudio) and repeat the commands line-by-line.

**[11.2]** Obtain a set of CEL files from NCBI GEO. You can download them from the website to your working directory, or use the `getGEO` function (from the `GEOquery` package) to import them.

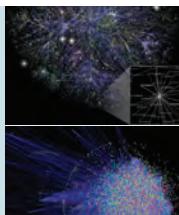
```
> source("http://bioconductor.org/biocLite.R")
> biocLite("GEOquery")
```

Repeat the `affy` and `limma` analyses described in this chapter on your dataset.

**[11.3]** Perform hierarchical clustering using R. Obtain a matrix of genes ( $n = 8$ ) and samples ( $n = 14$ ) from Web Document 11.6 at <http://www.bioinfbook.org/chapter11>. Copy this as a text file into an R working directory. Then use the following commands (# indicates a comment line).

```
> dir()
#view the contents of your directory; this
should include the file myarraydata.txt
> z=read.delim("myarraydata.txt")
#read.delim is a principal way of reading a
table of data into R. This creates a new file
called z with 8 rows (genes) columns including
gene name, chromosomal locus, and 14 samples.
> z
#view the data matrix z consisting of 8 genes
and 14 samples
> row.names(z)=z[,1]
> clust=hclust(dist(z[,3:16]),method="complete")
#create a distance matrix using columns 3 to
16; perform hierarchical clustering using the
complete linkage agglomeration method
```

```
> plot(clust)
#generate a plot of the clustering tree, such as
a figure shown in this chapter
#Note that you can repeat this using a variety
of different methods (e.g., method="single" or
method="median". Type ?hclust for more options.
> z.back=z[,-c(1,2)]
#create a version of matrix z called z.back in
which two columns containing the gene names and
chromosomal loci are removed.
> z.back
#view this matrix
> w=t(z.back)
#create a new file called w by transposing z.back.
> w
#view matrix w. There are now 4 rows (samples)
and 8 columns (genes).
> clust=hclust(dist(w[,1:8]),method="complete")
> plot(clust)
#perform clustering. The cluster dendrogram now
shows 14 samples (rather than 8 genes).
> clust=hclust(dist(z[,3:16]),method="euclidean",
method="complete")
> plot(clust)
> clust=hclust(dist(z[,3:16]),method="manhattan"),
method="complete")
> plot(clust)
> clust=hclust(dist(z[,3:16]),method="minkowski"),
method="complete")
> plot(clust)
> clust=hclust(dist(z[,3:16]),method="binary"),
method="complete")
> plot(clust)
> clust=hclust(dist(z[,3:16]),method="maximum"),
method="complete")
> plot(clust)
> clust=hclust(dist(z[,3:16]),method="canberra"),
method="complete")
> plot(clust)
#You can vary the metric by which you create a
distance matrix (e.g., Euclidean, manhattan,
minkowski, binary, maximum, canberra) as well as
varying the clustering method ("ward", "single",
"complete", "average", "mcquitty", "median" or
"centroid").
```



## Self-Test Quiz

**[11.1]** It is necessary to normalize microarray data because:

- (a) gene expression values are not normally distributed;
- (b) some experiments use cDNA labeled with fluorescence while others employ cDNA labeled with radioactivity;
- (c) the efficiency of dye incorporation may vary for different samples; or
- (d) housekeeping genes (such as actin) may be expressed at varying levels between samples.

**[11.2]** Microarray data analysis can be performed with scatter plots. Which of the following pieces of information do you *not* get from a scatter plot:

- (a) whether a gene is expressed at a relatively high or low level;
- (b) whether a gene has been up- or down-regulated;
- (c) if a gene is expressed in a region that suggests it is skewing data points; or
- (d) if a gene is statistically significantly regulated in that experiment.

**[11.3]** Log<sub>2</sub> ratios of gene expression values are often used rather than raw ratios because:

- (a) two-fold up-regulation or two-fold down-regulation log<sub>2</sub> ratios each have the same absolute value;
- (b) two-fold up-regulation or two-fold down-regulation log ratios each have the same relative value;
- (c) the scale of log<sub>2</sub> ratios is hypergeometrically compressed relative to the scale of raw ratios; or
- (d) a plot of log<sub>2</sub> ratios compresses the expression values to reduce the number of outliers.

**[11.4]** Inferential statistics can be applied to expression datasets to perform hypothesis testing:

- (a) in which the probability is assessed that any individual transcript is significantly regulated in a comparison of two samples;
- (b) in which the probability is assessed that any individual transcript is significantly regulated in a comparison of two or more samples;
- (c) by clustering of array data; or
- (d) by either supervised or unsupervised analyses.

**[11.5]** Which one of the following statements is FALSE?

- (a) Clustering of expression data produces a tree that can resemble a phylogenetic tree.
- (b) Clustering of expression data can be performed on genes and/or samples.
- (c) Clustering of expression data can be performed with partitioning methods (such as *k*-means) or hierarchical methods (such as agglomerative or divisive clustering).
- (d) Clustering of expression data is always performed using principal components analysis.

**[11.6]** Clustering techniques rely on distance metrics to:

- (a) describe whether a clustering tree is agglomerative or divisive;

- (b) reduce the dimensionality of a highly dimensional dataset;
- (c) identify the absolute values of gene expression measurements in a matrix of gene expression values versus samples; or
- (d) define the relatedness of gene expression values from a matrix of gene expression values versus samples.

**[11.7]** A self-organizing map:

- (a) imposes some structure on the formation of clusters;
- (b) is unstructured, like *k*-means clustering;
- (c) has neighboring nodes that represent dissimilar clusters; or
- (d) cannot be represented as a clustering tree.

**[11.8]** Principal components analysis (PCA):

- (a) minimizes entropy to visualize the relationships among genes and proteins;
- (b) can be applied to gene expression data from microarrays but not to protein analyses;
- (c) can be performed by agglomerative or divisive strategies; or
- (d) reduces highly dimensional data to show the relationships among genes or among samples.

**[11.9]** The main difference between supervised and unsupervised analyses is:

- (a) Supervised approaches assign some prior knowledge about function to the genes and/or samples, while unsupervised analyses do not.
- (b) Supervised approaches assign a fixed number of clusters, while unsupervised analyses do not.
- (c) Supervised approaches cluster genes and/or samples, while unsupervised approaches cluster only genes.
- (d) Supervised approaches include algorithms such as support vector machines and decision trees, while unsupervised approaches use clustering algorithms.

## SUGGESTED READING

Miron and Nadon (2006) provide a review of key concepts in microarray data analysis. See Leek *et al.* (2010) for an important overview of batch effects. For RNA-seq methodology see Garber *et al.* (2011).

There are many introductions to R, including an introduction to statistics using R by Verzani (2005) and a book on R and Bioconductor for bioinformatics by Gentleman *et al.* (2005).

Ma and Dai (2011) review PCA. For cluster analysis of microarray data, Gollub and Sherlock (2006) provide an excellent overview. Michael Eisen and colleagues (1998) describe the clustering of 8600 human genes as a function of time. This classic paper

includes an excellent description of the metric used to define the relationships of gene expression values and also a discussion of the usefulness of clustering in defining functional relationships among expressed genes.

## REFERENCES

- Alizadeh, A.A., Eisen, M.B., Davis, R.E. *et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511. PMID: 10676951.
- Alter, O., Brown, P. O., Botstein, D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Science USA* **97**, 10101–10106.
- Ayroles, J.F., Gibson, G. 2006. Analysis of variance of microarray data. *Methods in Enzymology* **411**, 214–33.
- Bassel, A., Hayashi, M., Spiegelman, S. 1964. The enzymatic synthesis of a circular DNA–RNA hybrid. *Proceedings of the National Academy of Science USA* **52**, 796–804.
- Benjamini, Y., Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B* **57**, 289–300.
- Bland, J.M., Altman, D.G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–310.
- Bland, J.M., Altman, D.G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**, 135–160.
- Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193. PMID: 12538238.
- Bouton, C. M., Pevsner, J. 2000. DRAGON: Database Referencing of Array Genes Online. *Bioinformatics* **16**, 1038–1039.
- Bouton, C. M., Henry, G., Colantuoni, C., Pevsner, J. 2003. DRAGON and DRAGON view: methods for the annotation, analysis, and visualization of large-scale gene expression data. In: *The Analysis of Gene Expression Data: Methods and Software* (eds G. Parmigiani, E. S. Garrett, R. A. Irizarry, S. L. Zeger). Springer, New York, pp. 185–209.
- Brazma, A., Vilo, J. 2000. Gene expression data analysis. *FEBS Letters* **480**, 17–24.
- Brown, M.P., Grundy, W.N., Lin, D. *et al.* 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Science USA* **97**, 262–267. PMID: 10618406.
- Canales, R.D., Luo, Y., Willey, J.C. *et al.* 2006. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology* **24**, 1115–1122.
- Colantuoni, C., Henry, G., Zeger, S., Pevsner, J. 2002a. SNOMAD (Standardization and NOrmalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics* **18**, 1540–1541.
- Colantuoni, C., Henry, G., Zeger, S., Pevsner, J. 2002b. Local mean normalization of microarray element signal intensities across an array surface: Quality control and correction of spatially systematic artifacts. *Biotechniques* **32**, 1316–1320.
- Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z., Speed, T.P. 2004. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**, 323–331. PMID: 14960458.
- Draghici, S., Khatri, P., Eklund, A.C., Szallasi, Z. 2006. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics* **22**, 101–109.
- Dupuy, A., Simon, R.M. 2007. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of National Cancer Institute* **99**, 147–157.
- Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA* **95**, 14863–14868.
- Engström, P.G., Steijger, T., Sipos, B. *et al.* 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **10**(12), 1185–1191. PMID: 24185836.
- Everitt, B.S., Landau, S., Leese, M. 2001. *Cluster Analysis*. Fourth edition. Arnold, London.

- Garber, M., Grabherr, M.G., Guttman, M., Trapnell, C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**(6), 469–477. PMID: 21623353.
- Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. 2004. Affy: analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**(3), 307–315. PMID: 14960456.
- Gentleman, R., Carey, V., Huber, W., Irizarry, R., Sandrine Dudoit, S. (eds.) 2005. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.
- Gollub, J., Sherlock, G. 2006. Clustering microarray data. *Methods in Enzymology* **411**, 194–213.
- Gordon, A.D. 1980. *Classification*. Chapman & Hall CRC, London.
- Guo, Y., Li, C.I., Ye, F., Shyr, Y. 2013. Evaluation of read count based RNAseq analysis methods. *BMC Genomics* **14** Suppl 8, S2. PMID: 24564449.
- Hansen, K.D., Wu, Z., Irizarry, R.A., Leek, J.T. 2011. Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29**(7), 572–573. PMID: 21747377.
- Hung, J.H., Yang, T.H., Hu, Z., Weng, Z., DeLisi, C. 2012. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics* **13**(3), 281–291. PMID: 21900207.
- Ioannidis, J.P., Allison, D.B., Ball, C.A. et al. 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41**(2), 149–155. PMID: 19174838.
- Irizarry, R.A., Bolstad, B.M., Collin, F. et al. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**(4), e15. PMID: 12582260.
- Irizarry, R.A., Warren, D., Spencer, F. et al. 2005. Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2**, 345–350.
- Irizarry, R.A., Wu, Z., Jaffee, H.A. 2006. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* **22**, 789–794.
- Ishwaran, H., Rao, J.S., Kogalur, U.B. 2006. BAMarray: Java software for Bayesian analysis of variance for microarray data. *BMC Bioinformatics* **7**, 59.
- Kaufman, L., Rousseeuw, P. J. 1990. *Finding groups in data. An Introduction to Cluster Analysis*, Wiley, New York.
- Kim, D., Pertea, G., Trapnell, C. et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**(4), R36. PMID: 23618408.
- Kvam, V.M., Liu, P., Si, Y. 2013. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany* **99**(2), 248–256. PMID: 22268221.
- Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359. PMID: 22388286.
- Law, C.W., Chen, Y., Shi, W., Smyth, G.K. 2014. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**(2), R29. PMID: 24485249.
- Leek, J.T., Scharpf, R.B., Bravo, H.C. et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**(10), 733–739. PMID: 20838408.
- Li, B., Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323. PMID: 21816040.
- Ma, S., Dai, Y. 2011. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics* **12**(6), 714–722. PMID: 21242203.
- Mane, S.P., Evans, C., Cooper, K.L. et al. 2009. Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC Genomics* **10**, 264). PMID: 19523228.
- Mao, R., Zielke, C.L., Zielke, H.R., Pevsner, J. 2003. Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. *Genomics* **81**(5), 457–467. PMID: 12706104.
- Mao, R., Wang, X., Spitznagel, E.L. Jr. et al. 2005. Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. *Genome Biology* **6**(13), R107. PMID: 16420667.
- MAQC Consortium, Shi, L., Reid, L.H. et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151–1161. PMID: 16964229.

- Marchionni, L., Afsari, B., Geman, D., Leek, J.T. 2013. A simple and reproducible breast cancer prognostic test. *BMC Genomics* **14**, 336. PMID: 23682826.
- Martin, J.A., Wang, Z. 2011. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**(10), 671–682. PMID: 21897427.
- Miron, M., Nadon, R. 2006. Inferential literacy for experimental high-throughput biology. *Trends in Genetics* **22**, 84–89.
- Motulsky, H. 1995. *Intuitive Biostatistics*. Oxford University Press, New York.
- Nagalakshmi, U., Waern, K., Snyder, M. 2010. RNA-Seq: a method for comprehensive transcriptome analysis. *Current Protocols in Molecular Biology Chapter 4*, Unit 4.11.1–13. PMID: 20069539.
- Olshen, A. B., Jain, A. N. 2002. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* **18**, 961–970.
- Osborne, J.D., Zhu, L.J., Lin, S.M., Kibbe, W.A. 2007. Interpreting microarray results with gene ontology and MeSH. *Methods in Molecular Biology* **377**, 223–242.
- Oshlack, A., Robinson, M.D., Young, M.D. 2010. From RNA-seq reads to differential expression results. *Genome Biology* **11**(12), 220. PMID: 21176179.
- Ozsolak, F., Milos, P.M. 2011. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**(2), 87–98. PMID: 21191423.
- Peng, R.D. 2011. Reproducible research in computational science. *Science* **334**(6060), 1226–1227. PMID: 22144613.
- Perou, C.M., Jeffrey, S.S., van de Rijn, M. *et al.* 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Science USA* **96**, 9212–9217. PMID: 10430922.
- Quackenbush, J., Irizarry, R.A. 2006. Response to Shields: ‘MIAME, we have a problem’. *Trends in Genetics* **22**, 471–472.
- Robertson, G., Schein, J., Chiu, R. *et al.* 2010. De novo assembly and analysis of RNA-seq data. *Nature Methods* **7**(11), 909–912. PMID: 20935650.
- Sean, D., Meltzer, P.S. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847.
- Shi, L., Jones, W.D., Jensen, R.V. *et al.* 2008. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics* **9** Suppl 9, S10. PMID: 18793455.
- Shi, L., Campbell, G., Jones, W.D. *et al.* 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28**(8), 827–838. PMID: 20676074.
- Shields, R. 2006. MIAME, we have a problem. *Trends in Genetics* **22**, 65–66.
- Smyth, G. K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1), Article 3.
- Smyth, G. K. 2005. Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds R.Gentleman, V.Carey, S.Dudoit, R.Irizarry, W.Huber), Springer, New York, pp. 397–420.
- Sneath, P.H.A., Sokal, R.R. 1973. *Numerical Taxonomy*. W.H. Freeman and Co., San Francisco.
- Soneson, C., Delorenzi, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91). PMID: 23497356.
- Steijger, T., Abril, J.F., Engström, P.G. *et al.* 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**(12), 1177–1184. PMID: 24185837.
- Subramanian, A., Tamayo, P., Mootha, V.K. *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Science USA* **102**, 15545–15550.
- Tamayo, P., Slonim, D., Mesirov, J. *et al.* 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science USA* **96**, 2907–2912. PMID: 10077610.

- Tan, P.K., Downey, T.J., Spitznagel, E.L. Jr. *et al.* 2003. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research* **31**, 5676–5684.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., Church, G. M. 1999. Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.
- Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C. 2006. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* **22**, 2405–2412.
- Trapnell, C., Roberts, A., Goff, L. *et al.* 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**(3), 562–578. PMID: 22383036.
- van Iterson, M., Boer, J.M., Menezes, R.X. 2010. Filtering, FDR and power. *BMC Bioinformatics* **11**, 450. PMID: 20822518.
- Verzani, J. 2005. *Using R for Introductory Statistics*. Chapman and Hall, New York.
- Wang, Z., Gerstein, M., Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**(1), 57–63. PMID: 19015660.
- Zeeberg, B.R., Qin, H., Narasimhan, S. *et al.* 2005. High-throughput GoMiner, an ‘industrial-strength’ integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* **6**, 168.
- Zolman, J. F. 1993. *Biostatistics*. Oxford University Press, New York.
- Zuur, A.F., Ieno, E.N., Meesters, E.H.W.G. 2009. *A Beginner’s Guide to R*. Springer, New York.

	1 Globin of the Oxyhemoglobin of Horse's Blood.	2 Serum-albumin of Horse's Blood.	3 Serum-globulin.	4 Egg-white.	5 Egg-albumin crystallised.	6 Albumin of Yolk.
Glycocol	0·9	0·10	3·52 <sup>96</sup>	...	...	...
Alanin	4·19 <sup>9</sup>	2·68 <sup>10</sup>	2·22 <sup>96</sup>	...	...	...
Leucin	29·04 <sup>9</sup>	20·48 <sup>10</sup>	18·7 <sup>96</sup>	22·6 <sup>40</sup>	...	...
Phenylalanin	4·24 <sup>9</sup>	3·08 <sup>10</sup>	3·84 <sup>96</sup>	+ <sup>3</sup>	...	...
$\alpha$ -Pyrrolidin-carboxylic acid	2·34 <sup>9</sup>	1·04 <sup>10</sup>	2·76 <sup>96</sup>	+ <sup>3</sup>	...	...
Glutaminic acid	1·73 <sup>9</sup>	1·52 <sup>10</sup>	2·20 <sup>96</sup>	+ <sup>40</sup>	+ <sup>48</sup>	...
Aspartic acid	4·43 <sup>9</sup>	3·12 <sup>10</sup>	2·54 <sup>96</sup>	+ <sup>37</sup>	+ <sup>48</sup>	...
Cystin	0·31 <sup>9</sup>	2·53 <sup>43</sup>	1·51 <sup>43</sup>	0·4 <sup>43</sup>	0·29 <sup>43</sup>	...
Serin	0·56 <sup>9</sup>	0·6 <sup>10</sup>	...	...	...	...
Oxy- $\alpha$ -Pyrrolidin-carboxylic acid	1·04 <sup>9</sup>	...	...	...	...	...
Tyrosin	1·33 <sup>9</sup>	2·1 <sup>10</sup>	...	0·58 <sup>66</sup>	1·5 <sup>31</sup>	...
Lysin	4·28 <sup>9</sup>	...	...	+ <sup>16</sup> 75	...	...
Histidin	10·96 <sup>9</sup>	...	...	+ <sup>18</sup>	+ <sup>18</sup>	+ <sup>18</sup>
Arginin	5·42 <sup>9</sup>	...	...	+ <sup>17</sup>	+ <sup>16</sup> 17	+ <sup>17</sup>
Tryptophane	+	+ <sup>9</sup>	+	+ <sup>80</sup>	...	...
Ammonia	0·93 <sup>50</sup>	1·2 <sup>49</sup>	1·75 <sup>49</sup>	...	1·5 <sup>49</sup>	...
Cystein	...	+	...	+ <sup>44</sup>	...	...
Amino-valerianic acid	...	...	...	...	...	...

While it is obvious to us that most proteins are composed of 20 amino acids, chemists in the late nineteenth century struggled to understand protein composition. At the turn of the century only several dozen proteins were known, including so-called albumins (including serum albumins, lactoglobulins, fibrinogen, myosin, and histones), proteids (e.g., hemoglobin and mucins), and albuminoids (e.g., collagen, keratin, elastin, and amyloid). Of these proteins only a very small group were available in pure form as crystals (e.g., hemoglobin and serum albumin from horse, ovalbumin, and ichthulin (salmon albumin)). Gustav Mann (1906, p. 70–75) described the dissociation products of 51 assorted proteins into their fundamental units. The results are shown for seven proteins (see columns). The rows indicate various compounds found upon dissolving the proteins. Most of these are amino acids; for example, glycocol is a name formerly given to glycine. This table shows that, from when proteins could first be analyzed, scientists made an effort to understand both the nature of individual protein molecules and the relationships of related proteins from different species.

Source: Mann (1906).

# Protein Analysis and Proteomics

# CHAPTER 12

*The egg-white like bodies (albumin) have occupied a considerable part of my life. All kinds of difficulties had to be surmounted, difficulties not met with other bodies; and whatever may have been said about this, one point is certain, that I have been the first who has shown (in 1838) that the meat is present in the bread and the cheese in the grass; that the whole organic kingdom is endowed with one and the same group, which is transferred from plants to animals and from one animal to another: one group, which is the first and foremost, and which I therefore still wish to call protein, a word derived from the Greek πρωτοζ [of first rank] suggested to me by Berzelius.*

—G. J. Mulder, *Levensschenets van G. J. Mulder door Hemzelen geschreven en door drie zijner vrienden uitgegeven* (1881). Translation by Westerbrink (1966) p. 154.

*Quantitative proteomics is a broad term, but it began with a specific meaning that should persist. It is the use of mass spectrometry (MS)-based technologies to detect, identify and quantitate changes in proteins and their post-translational modifications in biological systems. The key approach is MS: a highly sensitive, accurate and precise technology for measuring very small amounts of molecules, such as proteins and peptides. Importantly, quantitative proteomics is predominantly a tool for biological discovery.*

—Michael Washburn (2011) p. 170.

## LEARNING OBJECTIVES

After reading this chapter you should be able to:

- describe techniques to identify proteins including Edman degradation and mass spectrometry;
- define protein domains, motifs, signatures, and patterns;
- describe physical properties of proteins from a bioinformatics perspective;
- describe how protein localization is captured by bioinformatics tools; and
- provide definitions of protein function.

## INTRODUCTION

A living organism consists primarily of five substances: proteins, nucleic acids, lipids, water, and carbohydrates. Of these essential ingredients, it is the proteins that most define the character of each cell. DNA has often been described as a substance that corresponds to the blueprints of a house, specifying the materials used to build the house. These materials are the proteins, and they perform an astonishing range of biological functions. This

includes structural roles (e.g., actin contributes to the cytoskeleton), roles as enzymes (proteins that catalyze biochemical reactions, typically increasing a reaction rate by several orders of magnitude), and roles in transport of materials within and between cells. If DNA is the blueprint of the house, proteins form primary components not just of the walls and floors of the house but also of the plumbing system, the system for generating and transmitting electricity, and the trash removal system.

The history of protein studies is briefly discussed in Web Document 12.1. UniProtKB release 2014\_09 of October 2014 has ~550,000 sequence entries comprising ~200 million amino acids (<http://web.expasy.org/docs/relnotes/relstat.html>, WebLink 12.1). Another 84 million automatically annotated (i.e., not reviewed) sequences are in TrEMBL.

Proteins are polypeptide polymers consisting of a linear arrangement of amino acids. There is a rich history of attempts to purify proteins and identify their constituent amino acids. By 1850 a series of proteins had been identified (albumin, hemoglobin, casein, pepsin, fibrin, crystallin) and partially purified. It was not until 1950s that the complete amino acid sequences of several small proteins were determined. Today, we have access to 85 million protein sequences.

Earlier we learned how to access proteins from databases (Chapter 2), we aligned them and searched them against databases (Chapters 3–6), and we visualized multiple sequence alignments as phylogenetic trees (Chapter 7). In this chapter, we discuss techniques to identify proteins (direct sequencing, gel electrophoresis, and mass spectrometry). We then present four perspectives on individual proteins: domains and motifs, physical properties, localization, and function. We consider the structure of proteins in Chapter 13; in Chapter 14 we address functional genomics, the genome-wide assessment of gene function. Functional genomics encompasses large-scale studies of protein function both in normal conditions and following genetic or environmental perturbations.

## Protein Databases

Protein sequences were initially obtained directly from purified proteins (starting in the 1950s), but the vast majority of newly identified proteins are predicted from genomic DNA sequence. GenBank/DDBJ/EMBL, the separate whole-genome shotgun (WGS) division and the Short Read Archive/European Nucleotide Archive include vast amounts of nucleotide sequence data (Chapter 2). For proteins, one major resource is the NCBI nonredundant protein database. Perhaps the most prominent resource is UniProt, consisting of a series of databases (UniProt Consortium, 2013):

- UniProtKB is a protein knowledgebase that includes UniProtKB/Swiss-Prot (~500,000 reviewed protein entries given expert manual annotation) and UniProtKB/TrEMBL (~84 million unreviewed sequences, most of which are predicted from DNA sequencing projects).
- UniRef offers sets of sequence clusters which generally hide redundant sequences. For UniRef100, identical sequences (and subfragments with  $\geq 11$  residues) are merged into a single UniRef entry; UniRef90 and UniRef50 datasets are also available.
- UniMES includes metagenomic and environmental sequences. There are also UniMES Clusters merging entries with 100% or 90% identity.
- UniParc is a UniProt archive.

You can access UniProt or other protein databases by querying their websites, or by using tools such as BioMart at Ensembl. We can also use the R package `biomaRt` to accomplish a variety of tasks. (Refer to Chapter 8 where we performed other `biomaRt` tasks.) Open R (or it may be convenient to use RStudio), set the working directory to a convenient location, and install `biomaRt`.

Example 1: consider a list of gene symbols. For the proteins they encode, what are the InterPro database identifiers and descriptions?

```
> getwd() # Confirm which directory you are working in
> source("http://bioconductor.org/biocLite.R")
> biocLite("biomaRt") # the package is now installed
```

UniProt is available at <http://www.uniprot.org> (WebLink 12.2). It is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR).

```

> library("biomaRt") # load the package
> listMarts() # This displays >60 available databases
  biomart
  1 ensembl
  2 snp
  3 functional_genomics
  4 vega
# additional Marts from this list of 60 are truncated.
> ensembl = useMart("ensembl")
> listDatasets(ensembl)
      dataset          description
 1   oanatinus_gene_ensembl Ornithorhynchus anatinus genes (OANAS5)
 2   cporcellus_gene_ensembl Cavia porcellus genes (cavPor3)
# This list is truncated.
> ensembl = useDataset("hsapiens_gene_ensembl", mart=ensembl)
> filters = listFilters(ensembl)
> filters
> attributes = listAttributes(ensembl)
> attributes
# Browse the attributes to find protein-related topics!
# Let's select a small set of globin gene symbols
> globinsymbols <- c(HBB,HBA2,HBE,HBF)
# Next let's do the search, sending the results to a file
# called myinterpro:
> myinterpro <-
getBM(attributes=c("interpro","interpro_description"),
filters="hgnc_symbol",values=globinsymbols, mart=ensembl)
> myinterpro # we print the results
      interpro      interpro_description
 1 IPR000971      Globin
 2 IPR002338  Haemoglobin, alpha
 3 IPR002339  Haemoglobin, pi
 4 IPR009050  Globin-like
 5 IPR002337  Haemoglobin, beta

```

Example 2: Given a region of interest (e.g., 100,000 base pairs on chromosome 11) what are the gene symbols? For the genes that are protein-coding, which have predicted transmembrane regions?

```

> getBM(c("hgnc_symbol","transmembrane_domain"),
filters=c("chromosome_name","start","end"),
values=list(11,5200000,5300000), mart=ensembl)
      hgnc_symbol      transmembrane_domain
 1    OR52A1           Tmhmm
 2    OR51V1           Tmhmm
 3    HBB
 4    HBD
 5    HBD           Tmhmm
 6    HBG1
 7    HBG2
 8    HBE1

```

As another approach to accessing protein data, you can obtain sequences in the FASTA format from the command line using EDirect utilities from NCBI (Chapter 2).

Metagenomics (introduced in Chapter 15) has had a major impact on the discovery of genome sequences and the inference of protein sequences. For example, Craig Venter and colleagues assembled 7.7 million genomic DNA sequence reads as part of the Global Ocean Sampling (GOS) project as well as an earlier Sargasso Sea project (Venter *et al.*, 2004; Yooseph *et al.*, 2007). They used shotgun sequencing to randomly sample the DNA of microorganisms, including bacteria, archaea, and viruses, in seawater. They predicted the existence of 6.1 million proteins; a single publication therefore doubled the number of proteins known at that time. We discuss other metagenomics projects in which

The GOS project accession number at NCBI is AACY00000000. Note that many of the GOS project predicted proteins were not full-length, that is, they were not derived from a DNA segment that included both a start and a stop codon.

HPRD is available at <http://www.hprd.org> (WebLink 12.3). Currently (February 2015) HPRD includes over 30,000 protein entries. Human Proteinpedia is at <http://www.humanproteinpedia.org/> (WebLink 12.4) and currently has ~15,000 protein entries and ~250 contributing laboratories.

A controlled vocabulary is a set of predefined terms that are used to annotate data.

To learn about PSICQUIC visit <http://code.google.com/p/psicquic/> (WebLink 12.5). To use a PSICQUIC web-based tool visit PSICQUIC View at the European Bioinformatics Institute (<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/>, WebLink 12.6). Enter a query for GNAQ to see >1100 interactions defined in several dozen databases.

The HUPO Proteomics Standards Initiative website is <http://www.psiweb.info/> (WebLink 12.7).

The ABRF website is <http://www.abrf.org/> (WebLink 12.8).

microorganisms are sequenced from environmental samples in Chapters 15–17. Such projects are intended to explore the relationship between communities of microorganisms and their ecosystems, and will continue to greatly expand the number of known proteins.

In addition to cataloguing protein sequences, a variety of databases provide annotation of proteomics data such as protein–protein interactions, subcellular localization, post-translational modifications of proteins, and regional expression. The Human Protein Reference Database (HPRD) features expert curation on thousands of proteins (Mishra *et al.*, 2006; Goel *et al.*, 2011). Human Proteinpedia, established by Akhilesh Pandey and colleagues as for HPRD, is another broad, expertly curated proteomics resource (Muthusamy *et al.*, 2013). It serves as a community portal for the sharing, annotation, and integration of proteomics data.

## Community Standards for Proteomics Research

In all areas of bioinformatics, efforts are underway to standardize the way biological models are formulated and experimental data are generated and described. The Human Proteome Organization (HUPO) supports a Proteomics Standards Initiative (PSI) with the goals of defining standards for proteomic data representation to facilitate the comparison, exchange, and verification of data (Martens *et al.*, 2007). HUPO-PSI currently has working groups in three areas, with each group issuing community guidelines, data formats, and controlled vocabularies.

These three areas are:

- *Mass spectrometry and proteomics informatics.* Guidelines have been issued for the topics of mass spectrometry (e.g., defining the minimum information necessary to describe a proteomics experiment, identification, and quantitation (Taylor *et al.*, 2007; Martínez-Bartolomé *et al.*, 2013, 2014; Mayer *et al.*, 2013). For example, mass spectra are stored and exchanged with a Mass Spectrometry Markup Language (mzML; Turewicz and Deutsch, 2011).
- *Protein separation,* with guidelines for gel electrophoresis, gel informatics, column chromatography, capillary electrophoresis, and phosphoproteomics.
- *Molecular interactions,* with guidelines for Minimum Information about a Molecular Interaction eXperiment (MIMIx), information about a bioactive entity (MIABE), and standard formats for protein affinity reagents (MIAPAR). A practical example is the PSI Common Query InterfaCe (PSICQUIC) which allows you to access multiple molecular interaction databases with a single query (Orchard, 2012; del-Toro *et al.*, 2013) (Fig. 14.21).

These guidelines for reporting data, as well as data exchange formats and controlled vocabularies, require effort on the part of researchers but offer the great benefit of providing guidance for producing reproducible research (Orchard and Hermjakob, 2011; Orchard *et al.*, 2012; Gonzalez-Galarza *et al.*, 2014; Orchard, 2014). This particular sector of the research community has placed tremendous effort into preparing for the next phase of research, namely acquiring and cataloguing large datasets in an organized way that will maximize its utility.

## Evaluating the State-of-the-Art: ABRF analytic challenges

The Association of Biomolecular Resource Facilities (ABRF) is a professional society whose 600 members organize community experiments, often at core facilities, to perform research in proteomics, genomics, and other areas. ABRF research groups distribute (or request) test samples and have participating laboratories try to solve tasks such as determining the composition of a protein mixture or identifying the phosphorylation site of a phosphopeptide. The successes and failures of the various labs inform the community

about the current limits of accuracy, precision, and efficiencies in commonly performed experiments. This process of self-evaluation by the community illuminates the current state of the art and helps develop best practices in all aspects of proteomics from sample preparation to data analysis. We refer to several ABRF studies in this chapter.

## TECHNIQUES FOR IDENTIFYING PROTEINS

In this section we introduce three fundamental approaches to protein identification: direct protein sequencing; gel electrophoresis; and mass spectrometry.

### Direct Protein Sequencing

The first protein sequencing was by Frederick Sanger and Hans Tuppy (1951) who hydrolyzed insulin with acid, fractionated the resulting peptides by paper chromatography, and labeled them with dinitrophenyl (DNP) to identify the amino acid residues. The methods of Sanger and Tuppy were laborious, and soon an approach pioneered by Pehr Edman (1949) became established. Edman systematically degraded proteins, beginning with the amino terminal residue (which is derivatized, cleaved, and identified) and proceeding toward the carboxy terminus.

The Edman degradation procedure requires purification of a protein to relative homogeneity. This can be achieved by conventional biochemical means such as purification on ion exchange, size exclusion, other columns, or by electrophoresis. A portion of the amino acid sequence of a protein is obtained by transferring it to a specialized polyvinylidene fluoride (or PVDF) membrane, then performing microsequencing by sequential Edman degradations (Fig. 12.1). About 60–85% of the time, the amino terminus of yeast and other eukaryotic proteins is blocked (e.g., acetylated and unavailable for Edman degradations). A standard procedure is to proteolyze (e.g., trypsinize) the protein, purify the proteolytic fragments by reverse-phase high-performance liquid chromatography (HPLC), confirm the purity of the fragments, and then perform Edman degradations.

The Edman degradation method has been reviewed by Shively (2000). It remains a fundamental method of protein identification, and is useful to identify sequences of 1–10 picomoles of a protein. It is well suited to identifying the amino terminus of an intact protein (when unblocked), in contrast to mass spectrometry techniques that only analyze peptide fragments. It can be used for carboxy-terminal sequencing (Nakazawa *et al.*, 2008). However, the Edman technique has several limitations.

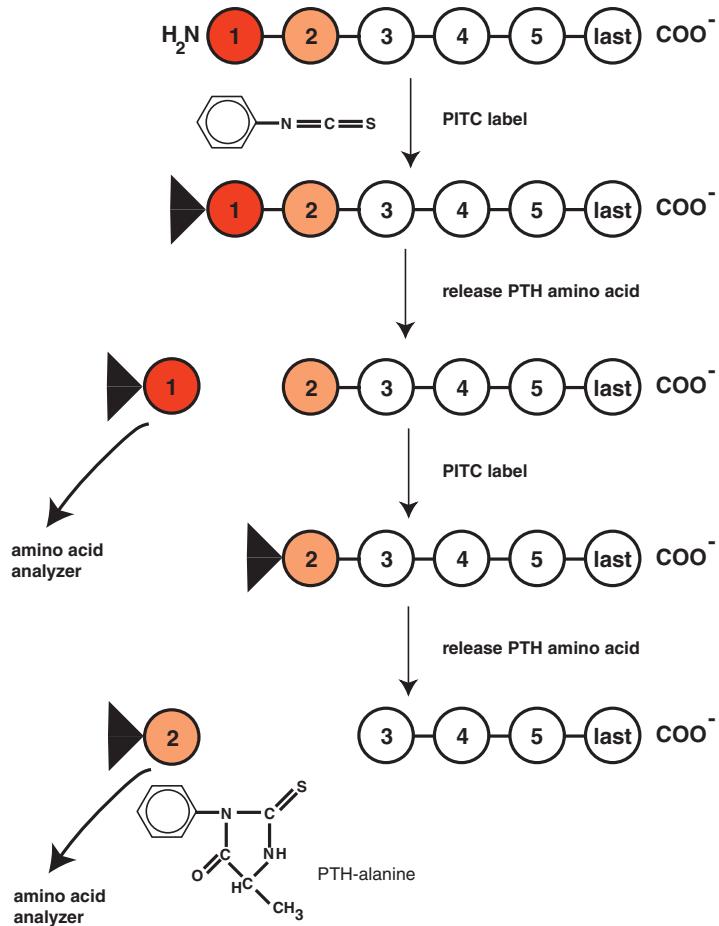
- It is laborious and not amenable to high-throughput analyses.
- While it is sensitive, mass spectrometry techniques can be 10–100 times more sensitive.
- Direct sequence is not useful for the analysis of post-translational modifications, unless combined with two-dimensional gel electrophoresis and mass spectrometry.

The ABRF has conducted 19 studies of Edman degradation including one by Brune *et al.* (2007) that also reviews earlier studies. Three synthetic peptides were synthesized including one with a modified residue (an acetyl lysine). Amino acid assignments were highly accurate for the peptides, but quantification of acetylated peptide was inaccurate (1.49/1 ratio for two peptides that were actually 1/1). This could be due to the lack of commercial PTH standards for many modified peptides.

Consider a 10 kilodalton protein; given a molecular weight of ~115 daltons per residue, such a protein consists of about 87 amino acids. To obtain 1 pmol, just 10 ng or  $10^{-8}$  g of protein is required.

### Gel Electrophoresis

Polyacrylamide gel electrophoresis (PAGE) is a premier tool for the analysis of protein molecular weight (for reviews of recent advances see Curreem *et al.*, 2012; Righetti, 2013). Proteins (like nucleic acids) possess a charge and thus migrate when introduced



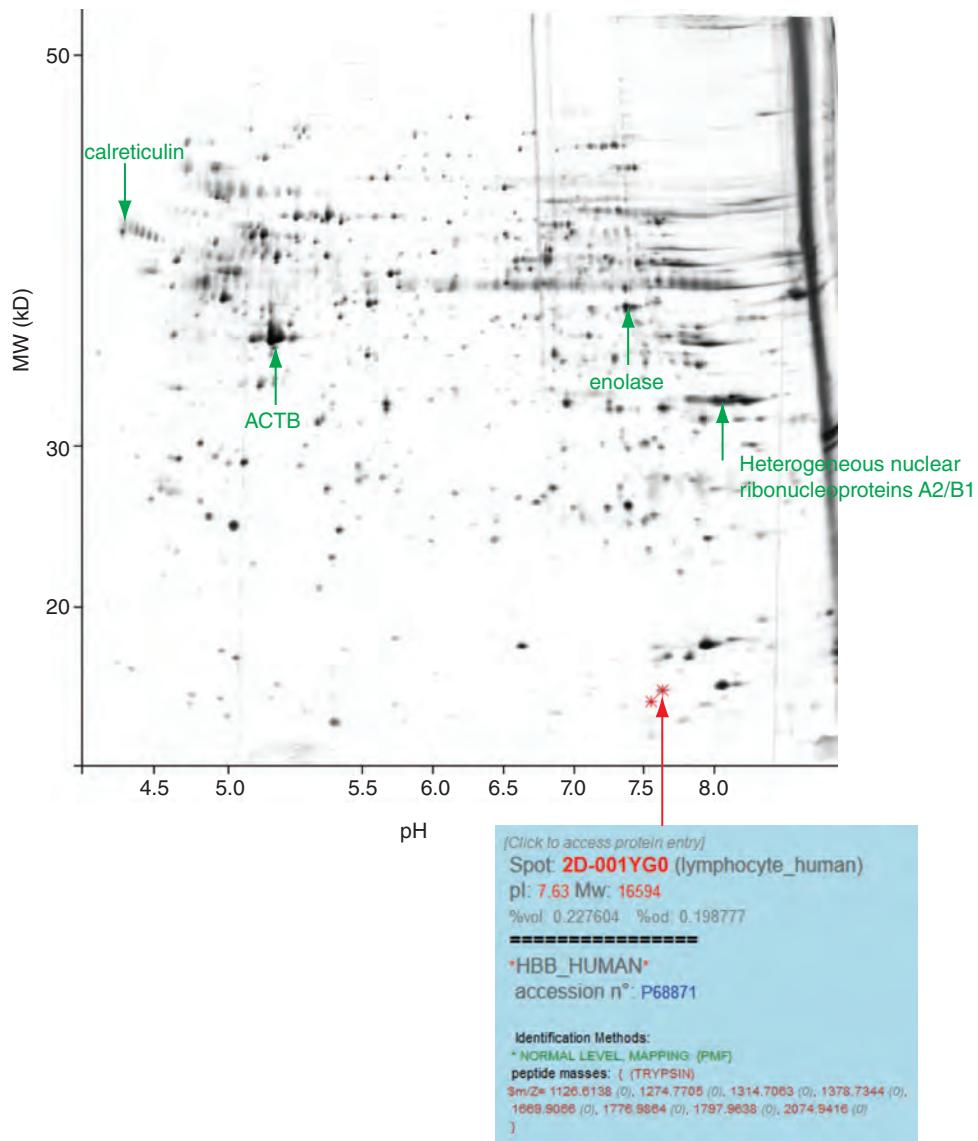
**FIGURE 12.1** Protein sequencing by Edman degradation. The Edman process is illustrated for a protein fragment of six amino acids. The first amino acid reacts through its amino terminus with phenylisothiocyanate (PITC). Under acidic conditions this amino acid residue, derivitized with phenylthiohydantoin (PTH), is cleaved and can be identified in an amino acid analyzer. The peptide now has five amino acid residues, and the cycle is repeated with successive amino-terminal amino acids. The structure of PTH-alanine is shown as an example. The typical result is a readout of 10–20 amino acids. The corresponding protein and gene can be evaluated by performing BLAST searches (Chapter 4).

into an electric field. Proteins are denatured and electrophoresed through a matrix of acrylamide that is inert (so it does not interact with the protein) and porous (so that proteins can move through it). The velocity of a protein as it migrates through an acrylamide gel is inversely proportional to its size; a complex mixture of proteins can therefore be separated in a single experiment. Proteins are almost always electrophoresed through acrylamide under denaturing conditions in the presence of the detergent sodium dodecyl sulfate (SDS), so this technique is commonly abbreviated SDS-PAGE.

O'Farrell (1975) greatly extended the capabilities of this technology by combining it with an initial separation of proteins based on their charge. In the first step, proteins are separated by isoelectric focusing. A gel matrix (or strip) is produced that contains ampholytes spanning a continuous range of pH values, usually between pH 3 and 11. Each protein is zwitterionic (having both positive and negative ions) and, when electrophoresed, it migrates to the position at which its total net charge is zero. This is the isoelectric point (abbreviated pI) at which the protein stops migrating. A complex mixture of proteins may therefore be separated based upon charge, and this corresponds to the first

dimension of two-dimensional gel electrophoresis. In the second dimension, proteins are separated by SDS-PAGE.

The technique of two-dimensional gel electrophoresis has matured into an important technology used to analyze proteomes (Görg *et al.*, 2004; Carrette *et al.*, 2006; Curreem *et al.*, 2012). An example of a two-dimensional gel profile is shown in **Figure 12.2**. Several hundred micrograms of protein from human lymphocytes were separated by



**FIGURE 12.2** Example of a two-dimensional protein gel result. The ExPASy two-dimensional gel resource was searched for beta globin. This profile is of several hundred proteins from human lymphocytes. The x axis corresponds to pH; proteins migrate to their isoelectric point (pI) where the net charges are zero. The y axis corresponds to molecular weight. On this particular gel, relatively low molecular weight proteins (10–50 kilodaltons) are well resolved, while other gels resolve larger proteins. The highly abundant proteins include two beta globin isoforms at molecular weights of about 17 kilodaltons (arrow). Several other identified proteins are indicated (beta actin (ACTB), calreticulin, enolase, ribonucleoproteins). By mousing over each identified spot, a dialog box appears with information on the protein (here, HBB with accession P68871) as well as an identifier, a statement of the molecular weight and pI, and a link to further information at ExPASy.

Source: ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

Resource	Hits	Category	Comment
OMA	657 hits	ge, ph	
PROSITE	7 hits	pr	PROSITE documentation entries
STRING	415 hits	pr	
SWISS-MODEL Repository	1348 hits	bi, pr, st	
UniProtKB		pr	No valid response from server.
ViralZone	0 hits	pr	ViralZone pages (for given virus name/family)
ENZYME	7 hits	pr	ENZYME entries
GPSDB	78 hits	ge, pr	gene synonyms
HAMAP		pr	No valid response from server.
miROrtho	0 hits	ge, ph	orthologous group(s) returned by your request
MyHits	1000 hits	pr	Protein found: >1000; Motif found: 0; Other node found: 0;
OpenFlu	0 hits	ge	
OrthoDB	43 hits	ge, ph	in Metazoa; 28 in Bacteria; 23 in Arthropods; 10 in Vertebrates; 5 in Fungi
Protein Spotlight	5 hits	pr	Spotlight/Prolune articles
Selectome	157 hits	ph	results in Selectome
SWISS-2DPAGE	04 hits	pr	SWISS-2DPAGE entries
SwissVar	13 hits	pr	Swiss-Prot protein - variant - disease groups

**FIGURE 12.3** ExPASy offers a premier web server for protein analysis as well as genomics, imaging and other analyses (<http://www.expasy.ch/>). You can input a query (such as hemoglobin, as shown at the top). The site provides a gateway to a two-dimensional gel database, to a large, well-organized list of links to databases, and to a vast variety of tools for protein analysis.

Source: ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

pH (on the *x* axis) by isoelectric focusing, then by molecular mass (on the *y* axis) by SDS-PAGE. Thousands of proteins may be visualized with a protein-binding dye such as silver nitrate or Coomassie blue (Panfoli *et al.*, 2012). Note that several proteins are especially abundant, including alpha and beta globin as well as the structural proteins actin and spectrin. Many proteins have a characteristic pattern of spots that spread along the first dimension. This is a “charge train” that usually represents a series of variants of a protein with differing amounts of charged groups such as phosphates that are covalently attached.

A central website for proteomics is the Expert Protein Analysis System (ExPASy; Artimo *et al.*, 2012; Fig. 12.3). ExPASy includes the main public database for information on two-dimensional gel electrophoresis (Hoogland *et al.*, 2004). Information is available for gels from a variety of organisms and experimental conditions, including the experiment shown in Figure 12.2. These profiles may be queried by choosing a two-dimensional gel map by other criteria such as keyword.

A key property of two-dimensional protein gels is that the individual proteins may be identified by direct protein microsequencing or by sensitive mass spectroscopy techniques (see following section). The ExPASy Swiss-2DPAGE site includes reference maps from organisms such as human, mouse, the plant *Arabidopsis thaliana*, the slime mold

ExPASy is located at <http://www.expasy.ch/> (WebLink 12.9). Many of the tools we explore in this chapter are available at ExPASy, which is part of the Swiss Institute of Bioinformatics.

*Dictyostelium discoideum*, and several bacteria. There have been thousands of applications of two-dimensional SDS-PAGE. These studies range across diverse species, cell types, and physiological states. Examples include descriptions of hundreds of proteins in human and rat brain (Langen *et al.*, 1999), characterizing traits of beer (Iimure *et al.*, 2014), studying vaginal, nasal, or other secretions, and characterizing changes in cancer or during the cell cycle of bacteria.

There have been many improvements to two-dimensional gel technology. Jonathan Minden and colleagues introduced difference gel electrophoresis (DIGE), a technique in which two (or sometimes three) samples are labeled with amine-reactive, fluorescent dyes (Viswanathan *et al.*, 2006; Minden, 2012). These samples are mixed, electrophoresed, and then the relative abundance of many proteins is determined based on fluorescence imaging. In some cases, DIGE has been used to detect 0.5 femtmoles of protein (for a 10 kilodalton protein, this corresponds to just 5 pg).

We may summarize the strengths of two-dimensional gel electrophoresis as follows:

- It offers the ability to describe both isoelectric point and molecular mass of intact proteins; this contrasts with mass spectrometry methods that identify molecular mass based on peptide fragments and that further lose information on pI.
- Several thousand proteins can be resolved and visualized with an appropriate stain.
- It is possible to detect and quantitate less than 1 ng per spot on the gel. A variety of sensitive stains (dyes) are available to detect proteins.
- Mass spectrometry is commonly used in conjunction with two-dimensional gels for protein identification, as discussed below.

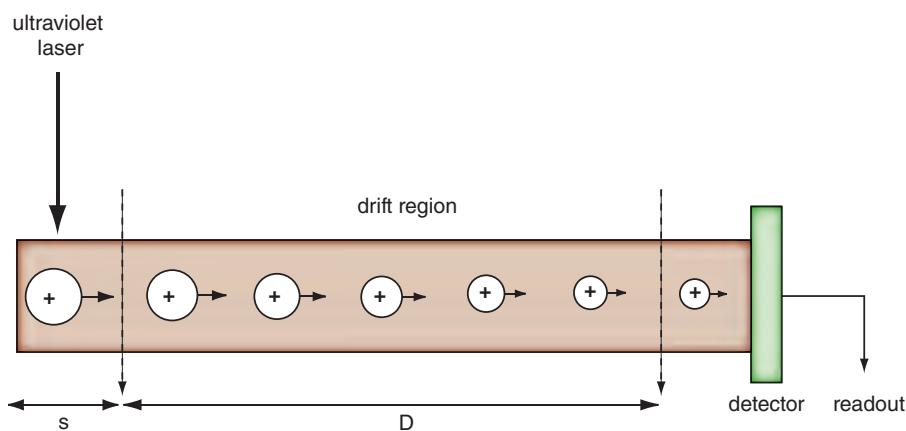
The two-dimensional gel approach has several limitations, however.

- It is not amenable to high-throughput processing of many samples in parallel.
- Sample preparation is a critical step and often requires a great deal of optimization. However, this is true of essentially all proteomics methods.
- Only the most abundant proteins in a sample are usually detected. Hydrophobic proteins, including proteins with transmembrane regions, are underrepresented on two-dimensional gels. Similarly, highly basic or acidic proteins are often excluded.
- It requires considerable expertise to reliably generate consistent results. In comparing two gel profiles, if the polyacrylamide gels vary even slightly in composition or if the samples are electrophoresed under differing conditions, it can be difficult to accurately align the protein spots. An important technical advance in the reproducibility of 2DG electrophoresis was the introduction of immobilized pH gradients preformed on dry strips, replacing an older system of pH gradient formation with ampholytes.

## Mass Spectrometry

Mass spectrometry techniques have revolutionized the field of proteomics by allowing proteins to be identified with extraordinary sensitivity. There are many excellent reviews of the technology (Gstaiger and Aebersold, 2009; Kumar and Mann, 2009; Washburn, 2011; Bruce *et al.*, 2013) and discussions on its future (Walsh *et al.*, 2010; Roepstorff, 2012; Thelen and Miernyk, 2012). Mass spectrometry is useful for: (1) identifying proteins (e.g., for identifying protein spots from two-dimensional gels, complex mixtures such as extracts of cells, or other biochemical purification approaches); (2) quantifying proteins; and (3) characterizing post-translational modifications of proteins. The ability of mass spectrometry to measure the mass of a protein with extremely high accuracy and precision allows it to distinguish even subtle changes in proteins such as the addition of a single phosphate group.

Mass spectrometers analyze charged protein or peptide molecules in the gaseous state. A key step is to transfer proteins into the gas phase and ionize them. This is accomplished



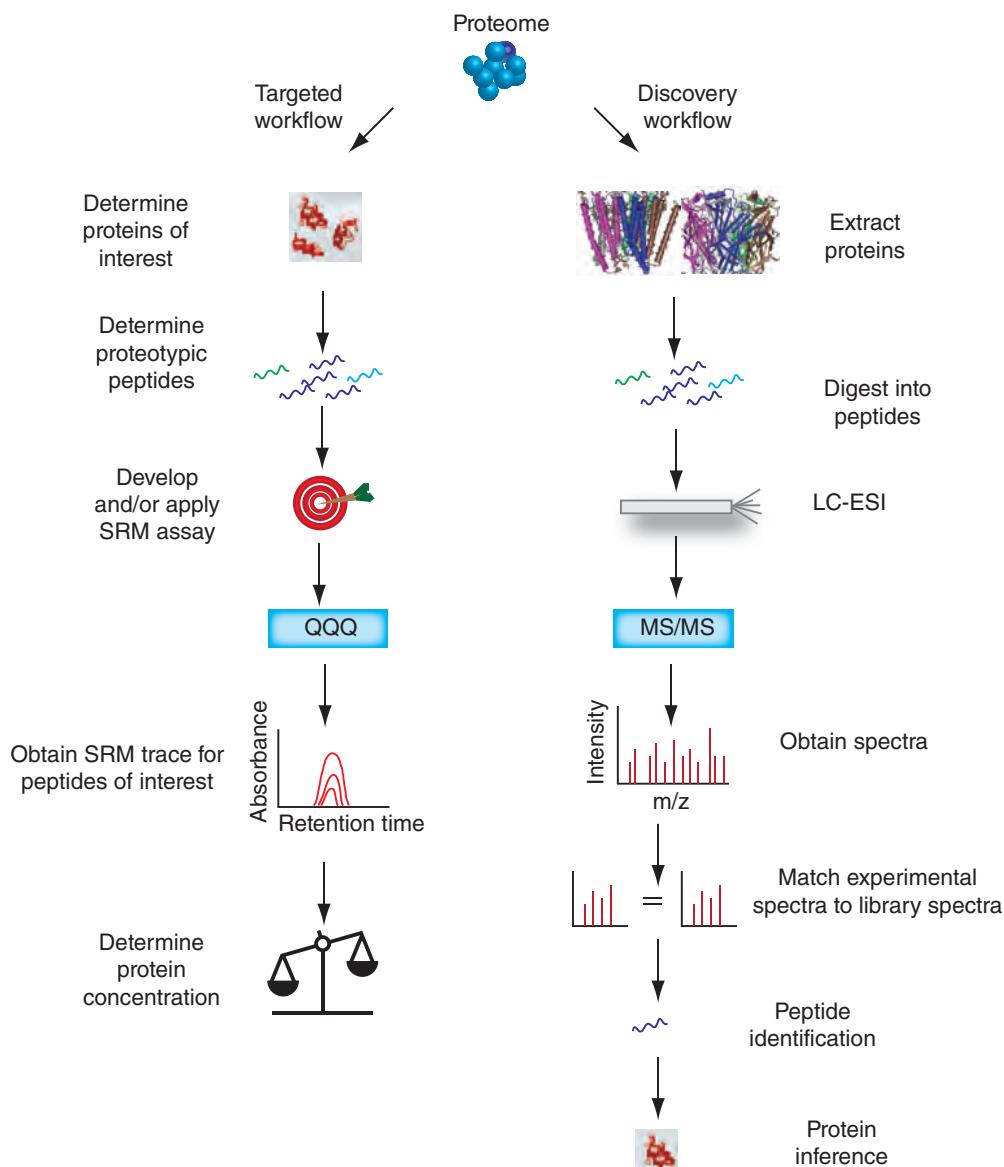
**FIGURE 12.4** Matrix-assisted laser desorption/ionization time-of-flight spectroscopy (MALDI-TOF). Spectroscopy is a technique to measure the mass of protein samples and other macromolecules. A sample is placed in a matrix of material that absorbs ultraviolet light. A laser is fired at the sample in the source region(s), and in the context of the matrix the sample becomes ionized. Some of the protein samples evaporate (i.e., desorption occurs). The ionization occurs in the presence of an electric field that accelerates the ions into a long drift region (D). The acceleration of each protein fragment is proportional to the mass of the ion. A detector records a time-of-flight spectrum that can be analyzed to determine the mass of each fragment. Peptide fragments are then searched against a protein database to determine the identity of the analyte (protein).

John Fenn and Koichi Tanaka shared half the Nobel Prize in Chemistry 2002 for their development of “soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules” ([http://nobelprize.org/nobel\\_prizes/chemistry/laureates/2002/](http://nobelprize.org/nobel_prizes/chemistry/laureates/2002/), WebLink 12.10).

using either matrix-assisted laser desorption ionization (MALDI) or electrospray ionization. In MALDI-TOF (MALDI with time-of-flight spectroscopy), the analyte molecules (i.e., the material to be analyzed) are dried on a metal substrate, irradiated with a laser, and fragmented (Fig. 12.4). The resulting ions are accelerated in a field that imparts a fixed kinetic energy. The ions traverse a path, are reflected in an ion mirror, and are then detected by a channeltron electron multiplier. The mass-to-charge ratio ( $m/z$ ) of an ion determines the time it takes to reach the detector; lighter ions (smaller analytes) have a higher velocity and are detected first. A time-of-flight spectrum is recorded from which the amino acid composition of even one femtomole of peptide can be deduced.

We can consider two common applications of mass spectrometry (Fig. 12.5; Doerr, 2013). First, discovery-based proteomics involves extracting proteins from a source of interest (such as a cell line, an organelle, a region excised with a razor from a two-dimensional gel, or other biological specimens). Proteins are digested with a protease such as trypsin to produce a set of peptides. These are fractionated to reduce the complexity of the sample, using techniques such as liquid chromatography with electrospray ionization. A mass spectrometer then identifies the  $m/z$  ratio of the peptide ions. The experimentally derived spectra are compared to a library of known peptide spectra, and the peptides in the sample can be identified. This allows inference of the original proteins in the sample. The databases that are searched for matches to mass spectrometry spectra typically include RefSeq and dbEST at NCBI, UniProt, and the Mass Spectrometry protein sequence DataBase (MSDB). The resolution of MALDI-TOF is excellent (about 0.1–0.2 daltons), allowing identification of the protein especially when multiple peptides correspond to the same protein.

Second, in a targeted workflow (Fig. 12.4, left side) selected reaction monitoring (SRM) is applied. The mass spectrometer identifies a pre-specified protein of interest. This can involve a triple quadrupole mass spectrometer (QQQ) in which peptide ions having a particular  $m/z$  ratio are selected in a first filter, resulting ions are selected in a second filter, and SRM traces are obtained showing absorbance (signal intensity) versus retention time for particular peptide ions.



**FIGURE 12.5** Targeted workflow and discovery workflow analyses of a proteome using mass spectrometry. In a discovery workflow (right side of figure), a goal is to identify as many proteins as possible. In a typical approach a set of proteins (proteome) is extracted from a sample, enriched, and cleaved with selective proteases to generate a set of peptides. These are separated by techniques such as liquid chromatography-electrospray ionization (LC-ESI) and tandem mass spectrometry (MS/MS). By matching observed spectra to a library of known spectra many peptides can be identified, allowing inference of the set of proteins present in the original sample. Some discovery workflows are sensitive enough to identify thousands of proteins in a sample. In a targeted workflow (left side of figure), a goal is to identify and quantify a set of proteins of interest. Here the mass spectrometer may be set to detect particular ions derived from selected proteins. In one approach using a triple quadrupole mass spectrometer (QQQ) a mass filter selects ions of interest (based on their  $m/z$  ratio) for quantitation.

Source: Doerr (2013). Reproduced with permission from Macmillan Publishers.

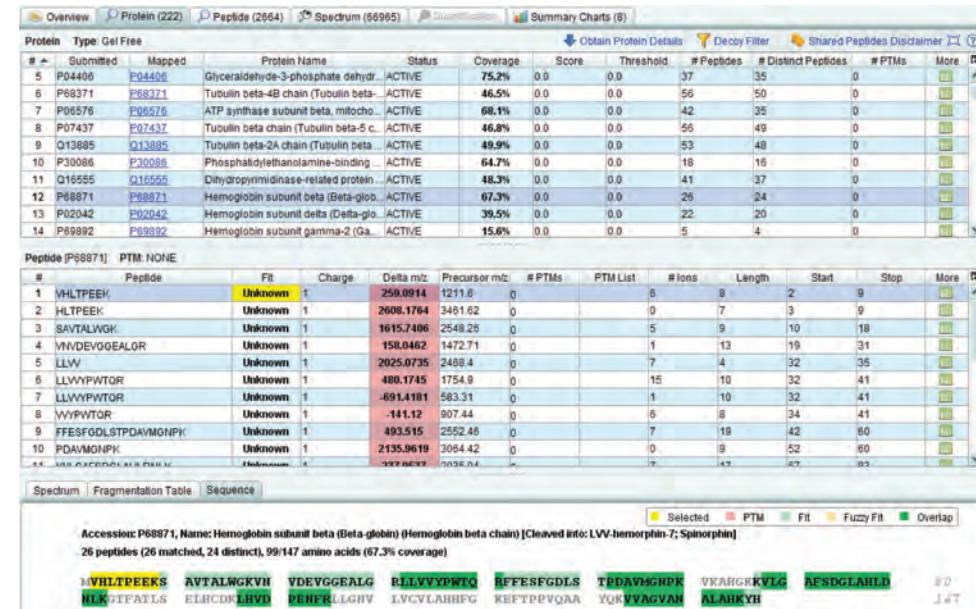
A key step in mass spectrometry experiments is the identification of proteins by matching of observed mass spectra to the theoretical spectral profiles of peptide fragments obtained from protein databases (Marcotte, 2007; Malik *et al.*, 2010). A variety of software tools are available to do this. An example is the PRoteomics IDEntifications (PRIDE) database at the European Bioinformatics Institute website. PRIDE is a

central public repository for mass spectrometry-based proteomics data (Vizcaíno *et al.*, 2013). We can search PRIDE with a UniProt accession for human beta globin (P68871, obtained from the NCBI Gene page for HBB); currently there are over 900 experiments to choose from. Viewing a HUPO brain proteome project from Martins-de-Souza *et al.* (2012; Fig. 12.6a) we can examine results in PRIDE Inspector software (downloaded as

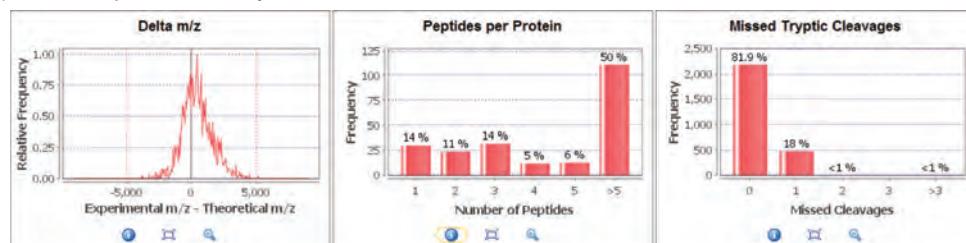
(a) PRIDE search results for mass spectrometry datasets including P68871 (beta globin)

Accession	Title	Species	Tissue	Cell Type	GO Term	Disease	Protein Count	Peptide Count	Spectra Count	Retrieve Details (View in web browser or download as XML file)
193	Plasma Proteome (GPM10100000689)	Homo sapiens (Human)	-	-	-	-	1	4	0	<a href="#">Web View</a> <a href="#">PRIDE Inspector</a> <a href="#">Download</a>
8959	Human Hep3B cells, untreated; cytoplasmic fraction	Homo sapiens (Human)	HEP-3B cell, liver	hepatocyte	cytoplasm	-	1	1	1	<a href="#">Web View</a> <a href="#">PRIDE Inspector</a> <a href="#">Download</a>
19112	Human Occipital Lobe (BA17)	Homo sapiens (Human)	-	-	-	-	1	26	22	<a href="#">Web View</a> <a href="#">PRIDE Inspector</a> <a href="#">Download</a>
26907	The proteome of mononuclear cells from human blood 2	Homo sapiens (Human)	mononuclear cell, blood	-	-	-	1	10	10	<a href="#">Web View</a> <a href="#">PRIDE Inspector</a> <a href="#">Download</a>

(b) PRIDE Inspector software 1.3.2



(c) PRIDE Inspector summary charts



**FIGURE 12.6** The PRoteomics IDEntifications database (PRIDE) database at EBI is a central repository for mass spectrometry-based proteomics data. (a) A search for beta globin protein produces >900 rows of results, several of which are shown here. One of the results links is to PRIDE Inspector software (b). This includes lists of peptides and their overlap with the protein sequence (green boxes; the yellow region is selected). Individual *m/z* spectra may also be seen. (c) A variety of summary statistics are available. Three shown here are Delta *m/z* (a measure of quality control that should be symmetrical and centered on zero); peptides per protein (50% of identified proteins have >5 matching peptides, indicating high confidence in the assignment); and missed tryptic cleavages.

a Java application). This shows metadata (including the experiment, instrument, species, authors, references) and lists of identified peptides and proteins (**Fig. 12.6b**), here showing an overlap between 22 peptides that match beta globin. Summary plots describe the quality of the experiment (**Fig. 12.6c**), and MS spectra are also available.

Among the most commonly used software is MASCOT® (Perkins *et al.*, 1999). Like other tools it provides a scoring algorithm to evaluate the false positive rate, and an *E* value similar to that used in BLAST (Chapter 4). The main strength of MASCOT® is its integration of three different search methods: peptide mass fingerprinting (in which peptide mass values are obtained); sequence queries (in which peptide mass data are combined with amino acid sequence data and compositional information); and MS/MS data obtained from peptides. Other prominent software includes ProteinPilot and Sequest.

How can we assess the accuracy of protein identification by mass spectrometry? The Association of Biomolecular Resource Facilities (ABRF) has conducted several studies to address this question. Falick *et al.* (2011) added 12 known proteins at varying, defined ratios to several complex mixtures of *E. coli* lysate. There were 43 study participants; each used mass spectrometry to identify proteins, and for quantification most used iTRAQ. The results showed how challenging this exercise was: only one-third of the participants could identify and detect differences in the five most abundant of the added proteins. The experience of the personnel was a key factor, a finding also seen in the realm of gene expression microarrays.

In an earlier ABRF study, Arnott *et al.* (2002) prepared five purified proteins at quantities of either 2 picomoles or 200 femtomoles: bovine protein disulfide isomerase (PDI); serum albumin (BSA); superoxide dismutase; *Escherichia coli* GroEL; and *Schistosoma japonicum* glutathione-S-transferase (GST). They digested the samples with trypsin, mixed them, and sent them “blind” to 41 participating laboratories that performed a total of 55 mass spectrometric analyses. The laboratories tended to use MALDI-TOF or microliquid chromatography with nanospray ionization (μLC-NSI). At the 2 picomole level, 96% (53/55) of the analyses correctly identified PDI, while 80% correctly identified GST. At the 200 femtomole level, 44% identified GroEL, 27% identified BSA, and 11% identified SOD. From one perspective, this is an enormous improvement over earlier mass spectrometry performance; from another perspective, this indicates that it is challenging for many laboratories to detect quantities below one or two picomoles.

There are dozens of important applications of mass spectrometry. We discuss some applications in this chapter and others in Chapter 14 when we describe functional genomics as applied to protein–protein interactions.

## FOUR PERSPECTIVES ON PROTEINS

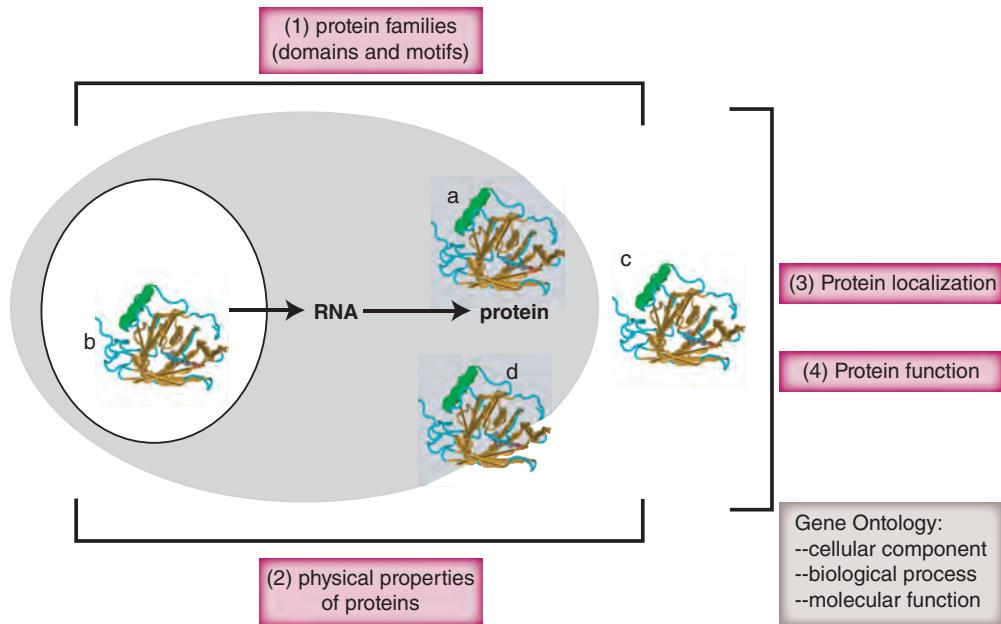
We will next describe four different perspectives on proteins (summarized in **Fig. 12.7**): (1) protein families (domains and motifs); (2) physical properties of proteins; (3) protein localization; and (4) protein function.

The first perspective we consider is the protein family. We define terms such as family, domain, and motif. Next, we consider the physical properties of proteins and how we can assess them. These properties include molecular weight, isoelectric point, and post-translational modifications (of which several hundred have been described).

The third and fourth perspectives, protein localization and function, complete our approach to proteins. These views are loosely related to a conceptual framework provided by the Gene Ontology (GO) Consortium. We therefore introduce GO as well, including its organizing principles for describing proteins (cellular component, biological process, and molecular function).

PRIDE is available from the European Bioinformatics Institute at <http://www.ebi.ac.uk/pride/> (WebLink 12.11) along with extensive documentation, search features, tools, and support. It might seem surprising that this particular brain proteome project includes abundant hemoglobin. However, blood is almost always present during dissections of brain.

MASCOT® software is available from Matrix Science (<http://www.matrixscience.com/>, WebLink 12.12). ProteinPilot is from AB Sciex (<http://www.absciex.com/>, WebLink 12.13) and Sequest is from the laboratory of John Yates III (<http://fields.scripps.edu/index.php>, WebLink 12.14).



**FIGURE 12.7** Overview of proteins. A protein is composed of a series of amino acids specified by a gene. Proteins can be classified by a variety of criteria, including family, localization, physical properties, and function. (1) Protein families are defined by the homology of a protein to other proteins; the proteins may be homologous over a partial region. Databases of protein families and motifs (discussed in Chapter 12) allow hundreds of thousands of proteins to be classified in groups that may be functionally related. (2) Proteins may be described in terms of their physical properties, such as size (molecular weight), shape (e.g., Stokes radius and frictional coefficient), charge (isoelectric point), post-translational modifications (see text), or the existence of isoforms due to proteolytic processing or alternative mRNA splicing. (3) A protein is depicted in several possible locations: it may be soluble in the cytosol (a), in an intracellular organelle such as the nucleus (b), or extracellular as a secreted protein (c). A protein may be bound to membranes on the cell surface (d) or on an intracellular organelle (not shown); membrane localization may be via transmembrane domains or by peripheral attachment. (4) Proteins may be categorized according to function. The Gene Ontology (GO) Consortium classifies proteins according to cellular component (i.e., localization), biological process (e.g., transcription or endocytosis), and molecular function (e.g., enzyme or transporter). A protein can belong to multiple categories of any of these groups. The GO system provides a dynamic, controlled vocabulary that can be applied to all eukaryotic proteins.

### Perspective 1: Protein Domains and Motifs: Modular Nature of Proteins

We begin our discussion of protein domains by considering several types of proteins. In the simplest case, a protein (or gene) has no matches to any other sequences in the available databases. This situation occurs less frequently as increasing numbers of genomes are sequenced, and yet it is not unusual to find that substantial numbers of predicted proteins have no identifiable homologs (e.g., Chapters 15–17). Even if there are no known homologs, a protein may have features such as a transmembrane domain, potential sites for phosphorylation, or some predicted secondary structure (see following section and Chapter 13). Such features may provide clues to the structure and/or function of the protein.

For proteins that do have orthologs and/or paralogs, there are regions of significant amino acid identity between at least two proteins (or DNA sequences). Such regions of proteins that share significant structural features and/or sequence identity have a variety of names: signatures, domains, modules, modular elements, folds, motifs, patterns, or

repeats. These terms have varied definitions, but all refer to the idea that there are closely related amino acid sequences shared by multiple proteins (Bork and Gibson, 1996; Bork and Koonin, 1996). Such regions may be considered in terms of protein structure and/or function (Copley *et al.*, 2002). We will primarily adopt the definitions provided by the InterPro Consortium (Hunter *et al.*, 2012). InterPro is an integrated documentation resource that encompasses a group of databases of protein families, domains, and functional sites.

A *signature* is a broad term that denotes a protein category, such as a domain or family or motif. When you consider a single protein sequence in isolation, there is only a limited amount of information you can infer about its structure or function. However, when you align related sequences, a consensus sequence may be identified. There are two principal kinds of signatures, and each is identified with its own methodology.

1. A domain is a region of a protein that can adopt a particular three-dimensional structure (Doolittle, 1995). Domains are also called modules (Sonnhammer and Kahn, 1994; Henikoff *et al.*, 1997). The term *fold* is commonly used in the context of three-dimensional structure (Jones, 2001). Together, a group of proteins that share a domain is called a family. Many protein domains are further classified based upon the subcellular localization of the domain (e.g., intracellular domains of proteins occur in the cytoplasm; extracellular domains are oriented outside the cell) or in terms of the structure of the domain (e.g., zinc finger domains bind the divalent cation zinc).

There are many databases of protein families, such as Pfam and SMART, that we explored in Chapter 6. The definitions of the terms *family*, *domain*, *repeat*, and related terms in the InterPro and SMART databases are given in **Tables 12.1** and **12.2**.

2. Motifs (or fingerprints) are short, conserved regions of proteins (discussed in “Protein Patterns”). A motif typically consists of a pattern of amino acids that characterizes a protein family (Bork and Gibson, 1996). The size of a defined motif is often 10–20 contiguous amino acid residues, although it can be smaller or larger. Some simple and common motifs, such as a stretch of amino acids that form a transmembrane region or a consensus phosphorylation site, do not imply homology when found in a group of proteins. In other cases, a small motif may provide a characteristic signature for a protein family.

To introduce specific examples of domains, **Table 12.3** lists the 10 most common domains in the proteins encoded by the human genome. Similar lists are available for

InterPro is accessed at <http://www.ebi.ac.uk/interpro/> (WebLink 12.15). It includes 11 consortium member databases: PROSITE (described in “Protein Patterns”); PRINTS (which uses position-specific scoring matrices); ProDom (which uses automatic sequence clustering); HAMAP (high-quality automated and manual annotation of proteins); and seven databases that use hidden Markov models (CATH-Gene3D, Panther, Pfam, PIRSF, SMART, SUPERFAMILY, TIGRFAMs). InterPro further links to dozens of additional resources including UniProt.

Wong *et al.* (2010) noted that transmembrane regions or signal peptides, commonly shared by vast numbers of proteins that are not homologous, are the source of over a thousand false positive entries in databases such as Pfam and iProClass (Chapter 6).

**TABLE 12.1 Definitions of protein families and related terms from InterPro database.**

Term	Definition
Family	A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.
Domain	Domains are distinct functional, structural, or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain.
Repeat	A match to an InterPro entry of this type identifies a short sequence that is typically repeated within a protein.
Site	A match to an InterPro entry of this type indicates a short sequence that contains one or more conserved residues. The type of sites covered by InterPro are active sites, binding sites, post-translational modification sites, and conserved sites.

Source: <http://www.ebi.ac.uk/interpro/>.

**TABLE 12.2 Definitions of protein domains and motifs from SMART database (a tool to allow automatic identification and annotation of domains in user-supplied protein sequences; see Chapter 6). Adapted from [http://smart.embl-heidelberg.de/help/smart\\_glossary.shtml](http://smart.embl-heidelberg.de/help/smart_glossary.shtml) (accessed November 2013); Gribskov *et al.* (1987); Lüthy *et al.* (1994); Thompson *et al.* (1994a, b); Bork and Gibson (1996); Gribskov and Veretnik (1996); Higgins *et al.* (1996).**

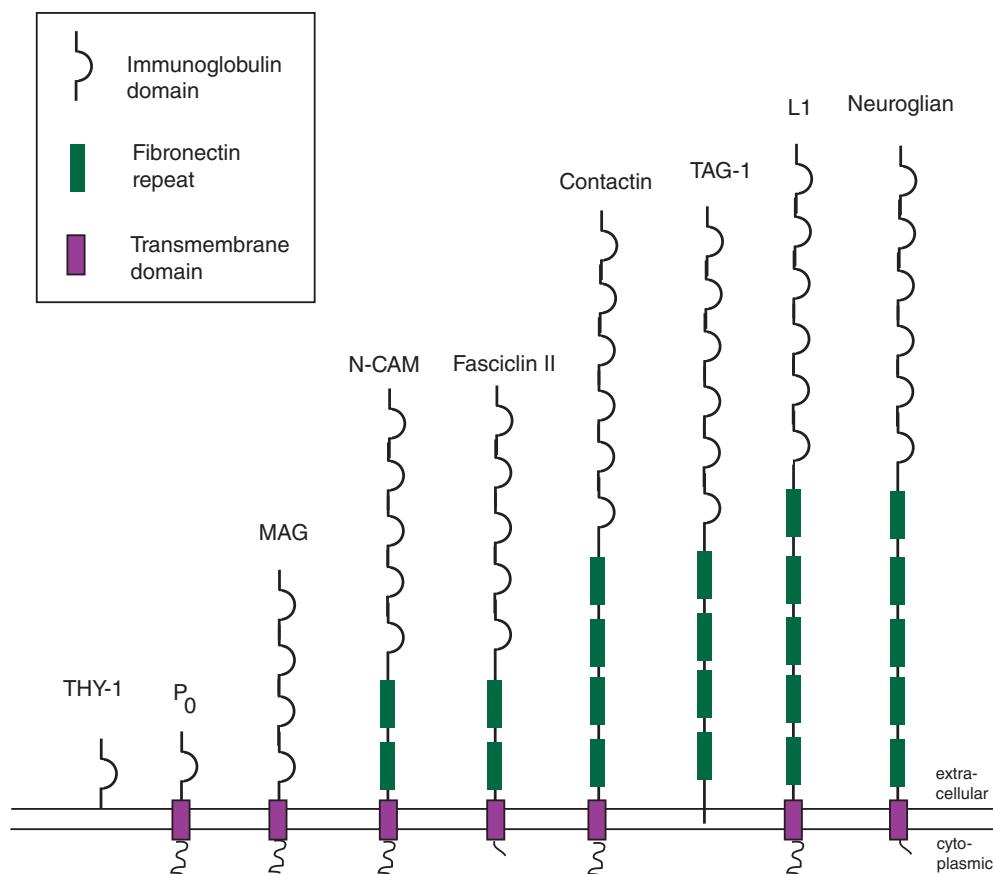
Term	Definition
Domain	Conserved structural entities with distinctive secondary structure content and a hydrophobic core. In small disulfide-rich and Zn <sup>2+</sup> -binding or Ca <sup>2+</sup> -binding domains, the hydrophobic core may be provided by cystines and metal ions, respectively. Homologous domains with common functions usually show sequence similarities.
Domain composition	Proteins with the same domain composition have at least one copy of each domain of the query.
Domain organization	Proteins having all the domains as the query in the same order (additional domains are allowed).
Motif	Sequence motifs are short conserved regions of polypeptides. Sets of sequence motifs need not necessarily represent homologs.
Profile	A profile is a table of position-specific scores and gap penalties, representing an homologous family, that may be used to search sequence databases (Gribskov <i>et al.</i> , 1987; Lüthy <i>et al.</i> , 1994; Gribskov and Veretnik, 1996). In CLUSTAL-W-derived profiles those sequences that are more distantly related are assigned higher weights (Thompson <i>et al.</i> , 1994a, b; Higgins <i>et al.</i> , 1996). Issues in profile-based database searching are discussed in Bork and Gibson (1996).

the abundant protein domains of other organisms (Chapters 16–19). In many cases, two proteins that share a domain also share a common function. For example, the immunoglobulin-like domain (InterPro accession IPR007110 with over 1000 members) is one of the most common domains encoded by the human genome. Many proteins having this domain have roles in extracellular signaling (Fig. 12.8). As another example, in humans there are hundreds of small guanosine triphosphate- (GTP-) binding proteins (InterPro

**TABLE 12.3 Most common domains of *Homo sapiens*.**

InterPro accession	Proteins matched	Name of domain
IPR027417	1022	P-loop containing nucleoside triphosphate hydrolase
IPR007110	1015	Immunoglobulin-like domain
IPR007087	806	Zinc finger; C2H2
IPR015880	801	Zinc finger; C2H2-like
IPR017452	796	GPCR; rhodopsin-like; 7TM
IPR000276	789	G protein-coupled receptor; rhodopsin-like
IPR003599	623	Immunoglobulin subtype
IPR013106	619	Immunoglobulin V-set
IPR011009	560	Protein kinase-like domain
IPR000719	513	Protein kinase; catalytic domain

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

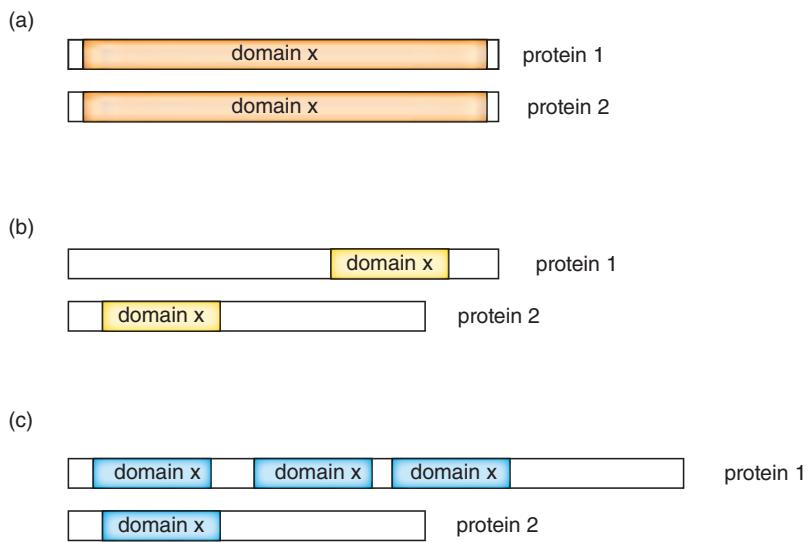


**FIGURE 12.8** Many proteins have multiple copies of distinct domains. The most common domain in humans is the immunoglobulin (Ig) domain, and the fibronectin repeat also commonly occurs. These domains are especially prevalent in the extracellular regions of proteins. Information about domains such as these is summarized in the InterPro database.

IPR005225). Many dozens are thought to regulate the intracellular docking and fusion of transport vesicles through a cycle of GTP binding and hydrolysis (Geppert *et al.*, 1997). Other related low-molecular-weight GTP-binding proteins function in cell cycle control and cytoskeletal organization (reviewed in Takai *et al.*, 2001). This superfamily is organized into related subfamilies that are usually presumed to share common functions.

Focusing our attention on a single domain, there are many ways in which proteins can share that domain in common. The entire protein may consist of one domain, such as the lipocalin domain or globin domain (Fig. 12.9a). Many other small, globular proteins also consist of a single domain.

It is even more common for a domain to form a subset of a protein. A comparison of two proteins often indicates that the domains occupy different regions of each protein (Fig. 12.9b). A group of six proteins contain a domain that confers the ability of each protein to bind methylated DNA. One of these proteins, methyl-CpG-binding protein 2 (MeCP2), is a transcriptional repressor that binds the regulatory region of a variety of genes. (Mutations in the *MECP2* gene cause Rett syndrome, a neurological disorder that affects girls and is one of the most common causes of intellectual disability in females; see Box 21.2.) We can perform a BLASTP search with the MeCP2 protein sequence to illustrate the concept of protein domains. The BLAST formatting page shows that the methyl-CpG-binding domain (MBD) is present in several databases of protein domains (Fig. 12.10a). The BLAST search result shows that a portion of MeCP2 matches four other MBD proteins (Fig. 12.10b).



**FIGURE 12.9** Proteins can share a common domain in a number of ways. (a) A domain may essentially extend across the length of a protein. An example of this format is the lipocalin family. (b) Domains may contain highly related stretches of amino acids that form only a subset of each protein's sequence. An example of this situation is found in the family of transcriptional regulators that bind methylated DNA. (c) A domain may be repeated within a single protein (sometimes with many copies). Such a domain may occur in homologous proteins any number of times. An example is the family of proteins containing a fibronectin III-like repeat.

Furthermore, examination of the MeCP2/MBD family shows that the proteins are various different sizes, only having the MBD domain in common (Fig. 12.10c).

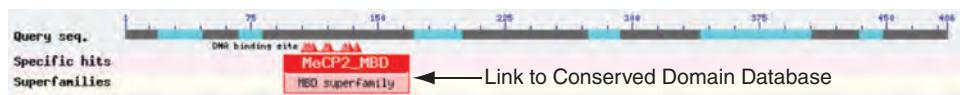
What is the definition of a family? Is a group of proteins homologous if they share only one domain in common? The MBD domains are clearly homologous (descended from a common ancestor), defining this group of proteins as a family; the regions outside the MBD domain share no significant amino acid identity, however. A family is a group of evolutionarily related proteins that share one (or more) regions of homology.

A third scenario for proteins containing individual domains is that the domain may be repeated many times (Fig. 12.9c). Two of the most common protein domains in *H. sapiens* are immunoglobulin domains (Table 12.3) and fibronectin repeats. Both of these domains are present in variable numbers in a group of proteins having extracellular domains (Fig. 12.8). Notably, these and other extracellular domains are highly abundant in humans and the multicellular nematode *Caenorhabditis elegans*, but nearly absent in the single-celled eukaryote *Saccharomyces cerevisiae* (Copley *et al.*, 1999). Comparison of protein families that are encoded by various genomes sheds light on the biological processes that each organism performs (Chapters 16–20).

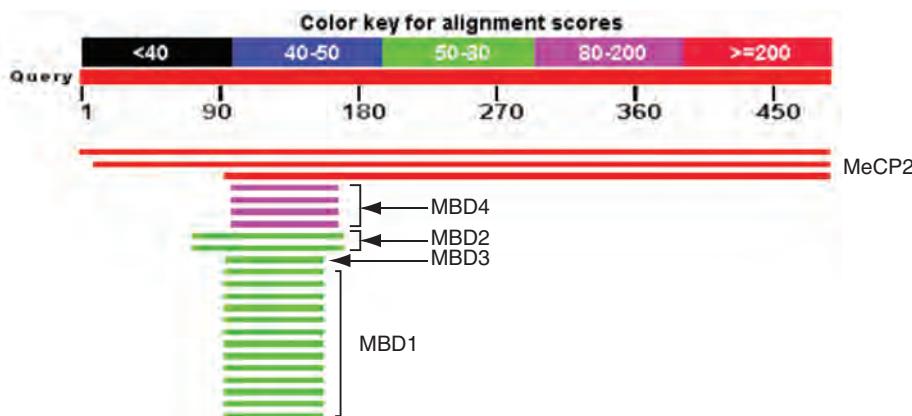
#### *Added Complexity of Multidomain Proteins*

Multidomain proteins provide a common, more complicated scenario than single-domain proteins. HIV-1 gag-pol is an example of such a protein (Frankel and Young, 1998). The *gag-pol* gene encodes a single large polypeptide that is cleaved into several independent proteins with distinct biochemical activities including an aspartyl protease, a reverse transcriptase (RNA-dependent DNA polymerase), and an integrase. Note that other multidomain proteins, such as the immunoglobulin domain proteins depicted in Figure 12.8, maintain separate domains within a mature polypeptide without cleaving them into separate proteins.

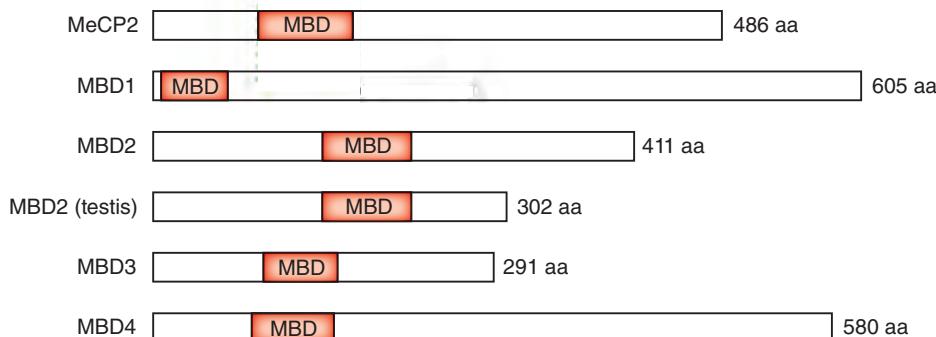
(a) BLAST result links



(b) BLAST alignments



(c) Domain structure



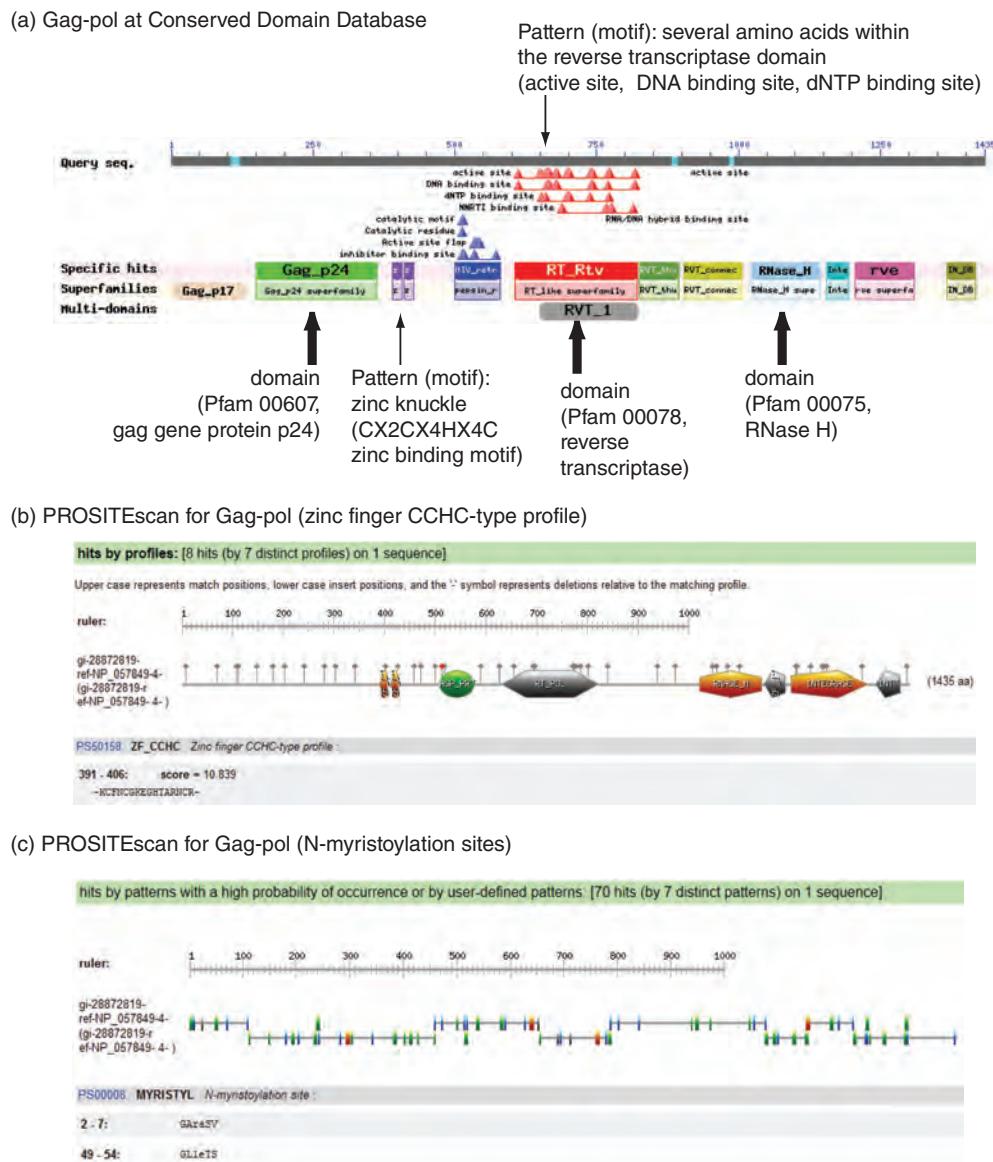
**FIGURE 12.10** A methyl-binding domain is found in several human proteins. To illustrate the concept of domains, methyl-CpG-binding protein 2 (MeCP2; NP\_004983.1) was used as a query in a BLASTP search restricted to human RefSeq proteins. (a) The formatting BLAST web page shows that this protein has a domain that is present in the Conserved Domain Database. (b) The BLAST search reveals there are separate MeCP2 entries that match the query (top alignments). Additionally, there is a region of about 80 amino acids in MeCP2 that matches other methyl-CpG-binding proteins: MBD1 (NP\_056671), MBD2 (NP\_003918), a testis-specific isoform of MBD2 (NP\_056647), MBD3 (NP\_003917), and MBD4 (NP\_003916). (c) These proteins have different sizes. Also, the methylated DNA-binding domain that these proteins share occurs in different regions of the proteins. Further BLAST searches confirm that together these six proteins share no significant amino acid identity at any region other than the methyl-binding domain.

Source: BLASTP, NCBI.

To examine the sequence of gag-pol, we first go to NCBI Gene. That entry shows that the protein accession is NP\_057849.4 (corresponding to a protein of 1435 amino acid residues), associated with at least six mature proteins, each with a RefSeq identifier. A link to the Conserved Domain Database shows assorted domains graphically (Fig. 12.11a). These domains include links to the authoritative Pfam database. There are also assorted motifs which we discuss in the following section.

#### Protein Patterns: Motifs or Fingerprints Characteristic of Proteins

Within a domain or outside a domain there may be a small number of characteristic amino acid residues that occur consistently. These are called motifs (or fingerprints). Several are



**FIGURE 12.11** Searches for a multidomain protein. (a) The NCBI Gene entry for HIV-1 gag-pol provides RefSeq accession numbers for the precursor protein (NP\_057849.4, 1435 amino acids) and for six predicted mature protein products. The protein includes both domains (thick arrows) and patterns (thin arrows). (b) A search of PROSITEcan using the FASTA-formatted gag-pol protein sequence reveals a variety of profiles, for example a zinc-finger profile, and (c) patterns, for example, N-myristoylation sites. Although a pattern or motif may not adopt a known three-dimensional structural conformation, it may nonetheless contain an amino acid sequence that is characteristic of a protein family.

Source: NCBI Gene, NCBI.

indicated in our Gag-pol protein graphic from the Conserved Domain Database, including a zinc knuckle (i.e., a CX2CX4HX4C motif) and an active site within a reverse transcriptase domain (Fig. 12.11a). PROSITE is a dictionary of protein motifs (Sigrist *et al.*, 2002, 2013). Following the link from ExPASy (Fig. 12.3) or searching the site directly, we can paste the FASTA format of the gag-pol protein into a search box and identify a variety of motifs (Fig. 12.11b, c). A zinc finger profile is characteristic of many proteins but its occurrence does not imply homology. Similarly there are 70 patterns corresponding to potential N-myristoylation sites (defined in the following section), although none of these sites is

necessarily modified *in vivo*. Another example of a motif is the amino acids that are reliably found at the active site of an enzyme. In the aspartyl protease domain of HIV-1 pol, an aspartate residue is crucial for the proteolytic reaction. The motif is defined by a string of 12 amino acid residues: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA]. This format is identical to that used by PHI-BLAST (Chapter 5).

Motifs are typically subsets of protein domains. A short motif that is found in almost all lipocalins is GXW. The consensus pattern defined in PROSITE (document PDOC00187) incorporates several additional amino acids surrounding GXW. That motif is [DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-x-[LIVMTA]. The GXW sequence is represented as G-{C}-W, where the curly brackets indicate that any amino acid other than cysteine is accepted at that position. Some motifs are extremely short and very common, such as the sequence surrounding a serine or threonine that is a substrate for many kinases. Such motifs are not specific to a particular protein family, and their occurrence in multiple proteins does not reflect homology. A search of PROSITE for “kinase” reveals >170 entries, including both kinase and kinase substrate signatures. One of these entries is for the protein kinase C (PKC) consensus phosphorylation site, [ST]-x-[RK] (S or T is the phosphorylation site and x is any residue; PROSITE document PDOC00005). This simple motif occurs in proteins many thousands of times.

An important aspect of regular expressions (or patterns) in the PROSITE database is that they are qualitative (i.e., either matching or not) and not quantitative (i.e., we do not recognize partial matches to a pattern). While patterns can accommodate complex definitions, such as having one of several different amino acid residues in a given position, mismatches are not tolerated when a protein sequence is compared to a pattern. In contrast to such rigid patterns, many databases such as Pfam, ProDom, and SMART (described in Chapter 6) use profiles. Profiles, like patterns, are built from multiple sequence alignments, but they employ position-specific scoring matrices. They also span larger stretches of protein sequence than do patterns. For both patterns and profiles we can define true positives (e.g., >1000 globin proteins match the globin family profile of PROSITE family PS01033), as well as false negatives (10 proteins that are known to be globins are not included in that globin family profile).

## Perspective 2: Physical Properties of Proteins

Proteins are characterized by a variety of physical properties that derive both from their essential nature as an amino acid polymer and from a variety of post-translational modifications (**Table 12.4**). Over 200 post-translational modifications are known, occurring on 15 of the 20 amino acids (all but Leu, Ile, Val, Ala, and Phe; Walsh, 2006). Some of these modifications allow the covalent attachment of a hydrophobic group to a protein to promote insertion into a lipid bilayer. Examples include palmitoylation, farnesylation, myristoylation, and inositol glycolipid attachment (**Fig. 12.12**). Other prominent modifications include phosphorylation and glycosylation (Temporini *et al.*, 2008; Amoresano *et al.*, 2009; Eisenhaber and Eisenhaber, 2010). The InterPro database also lists categories of post-translational domains (**Table 12.5**).

A variety of web-based services are available to evaluate the predicted physical properties of proteins (Blom *et al.*, 2004; Trost and Kusalik, 2011). Resources are available to input an individual protein sequence and to predict its physical properties such as: mass and isoelectric point (pI; **Fig. 12.13**; see also “Web Resources”); amino acid composition; glycosylation sites (see “Web Resources”); phosphorylation sites in which kinases reversibly add a phosphate group to individual serine, threonine, or tyrosine residues (**Fig. 12.14**); and tyrosine sulfation. Many programs predict secondary-structure features

PROSITE is accessed at <http://www.expasy.org/prosite/>

(WebLink 12.16). In PROSITE, the term *profile* refers to a quantitative motif description based on a generalized profile syntax. The term *pattern* refers to a qualitative motif description based on a regular expression-like syntax such as those described below. The term *motif* refers to the biological object approximated by a pattern or a profile. See Web Document 12.2 at <http://www.bioinfbook.org/chapter12> for these definitions.

You can use the ScanProsite tool to search a pattern against the PROSITE database, and the PRATT tool to generate a pattern based on an input of unaligned sequences. PRATT is available at <http://web.expasy.org/pratt/> (WebLink 12.17). See computer lab exercise (12.1).

For websites offering protein motif analysis tools, see “Web Resources” below.

The COILS server is available at [http://www.ch.embnet.org/software/COILS\\_form.html](http://www.ch.embnet.org/software/COILS_form.html) (WebLink 12.18).

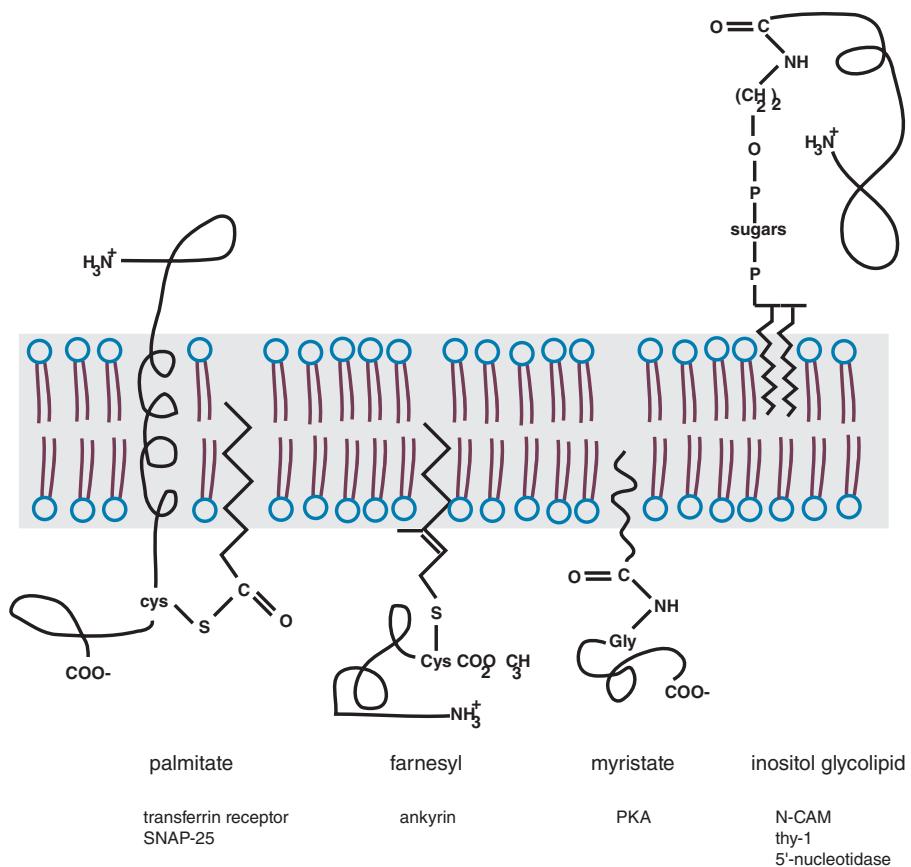
**TABLE 12.4 Some physical properties of proteins.** G protein: guanosine triphosphate-binding protein; GAP-43: growth-associated protein of 43 kD; MARCKS: myristoylated alanine-rich C-kinase substrate; nAChR: nicotinic acetylcholine receptor; PDZ domain: post-synaptic density protein; PSD-95: the *Drosophila* tumor suppressor discs-large, tight-junction protein ZO-1; PKA: protein kinase A; SNAP-25: synaptosomal-associated protein of 25 kD; Rab3A: rat brain GTP-binding protein 3A; thy-1: thymocyte- 1.

Property	Classical method	Example
Amino acid motifs	–	PDZ domain (e.g., nitric oxide synthase), coiled-coil domain (e.g., hemagglutinin, syntaxin, SNAP-25, myosin)
Isoelectric point (pI)	Derived from isoelectric focusing	–
Molecular weight	Derived from Stokes radius and sedimentation coefficient	–
Post-translational modifications: phosphorylation	Enzymatic analyses	Synapsin
Post-translational modifications: glycosylation	Enzymatic analyses	Nerve growth factor, neural cell adhesion molecule
Post-translational modifications: isoprenylation	Biochemical analyses	Lamin B, G protein $\gamma$ subunits, rab3A
Post-translational modifications: palmitoylation	Biochemical analyses	$\beta$ -Adrenergic receptor, GAP-43, insulin receptor, rhodopsin, nAChR
Post-translational modifications: myristoylation	Biochemical analyses	PKA, $G_{\alpha}$ -subunit, MARCKS protein, calcineurin
Post-translational modifications: GPI-anchored proteins	Enzymatic analyses	Alkaline phosphatase, thy-1, prion protein, 5'-nucleotidase, uromodulin
Sedimentation coefficient	Derived from sucrose density gradients	–
Stokes radius	Derived from gel filtration	–
Transmembrane domain	Derived from subcellular fractionation	–

**TABLE 12.5 Post-translational modifications at InterPro.**

Accession	Post-translational modification site
IPR000152	EGF-type aspartate/asparagine hydroxylation site
IPR001020	Phosphotransferase system, HPr histidine phosphorylation site
IPR002114	Phosphotransferase system, HPr serine phosphorylation site
IPR002332	Nitrogen regulatory protein P-II, urydylation site
IPR004091	Chemotaxis methyl-accepting receptor, methyl-accepting site
IPR006141	Intein splice site
IPR006162	Phosphopantetheine attachment site
IPR012902	Prokaryotic N-terminal methylation site
IPR018051	Surfactant-associated polypeptide, palmitoylation site
IPR018070	Neuromedin U, amidation site
IPR018243	Neuromodulin, palmitoylation/phosphorylation site
IPR018303	P-type ATPase, phosphorylation site
IPR019736	Synapsin, phosphorylation site
IPR019769	Translation elongation factor, IF5A, hypusine site
IPR021020	Adhesin, Dr family, signal peptide

Source: InterPro, <http://www.ebi.ac.uk/interpro/>.



**FIGURE 12.12** A variety of post-translational modifications are added to proteins. Examples are palmitoylation (e.g., to the transferrin receptor and SNAP-25), farnesylation (e.g., to ankyrin), myristylation (e.g., to protein kinase A), and inositol glycolipid anchoring to a membrane (e.g., neural cell adhesion molecule, thy-1, and 5'-nucleotidase). While these covalent modifications can be studied biochemically, a variety of websites offer predictions of possible sites of covalent modification to proteins. Adapted from Austen and Westwood (1991), with permission from Oxford University Press.



**FIGURE 12.13** The Compute pI/Mw server at ExPASy calculates the predicted molecular weight and isoelectric point of input proteins. Here, the values for beta globin are calculated. Programs at ExPASy do not accept RefSeq accession numbers as input (e.g., NP\_000509 for beta globin), but do accept raw sequence or UniProt accessions (e.g., P68871).

Source: ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

## 147 Sequence

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	80
NLKGTFTALSELHCDKLHVDPENFRLLGVLCVLAHHFGKEFTPPVQAAQKVVAGVANALAHKYH	160
.....T.....	S.....S.....
.....T.....	.....

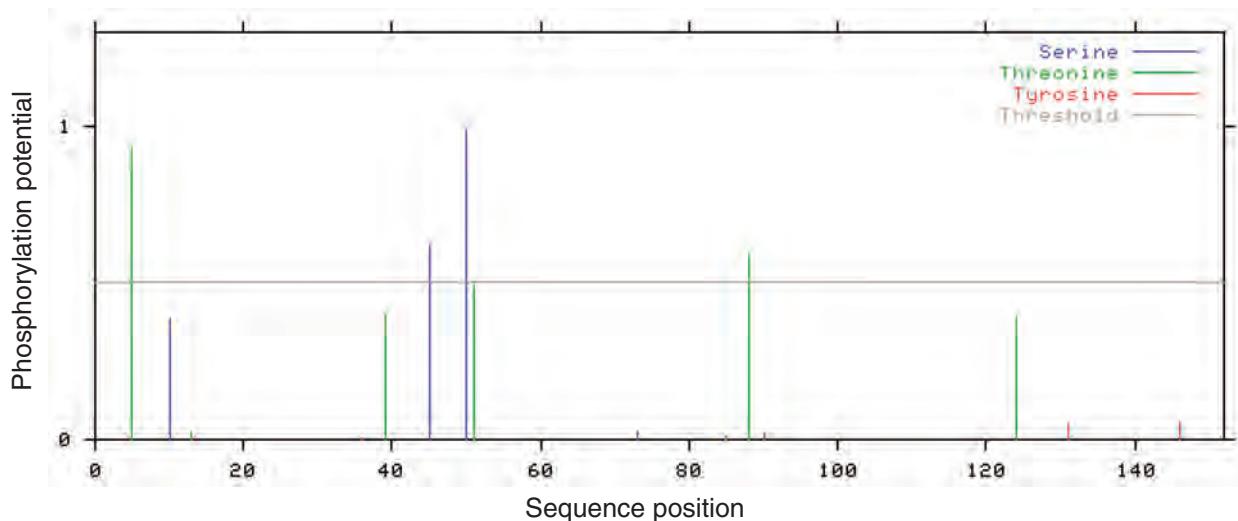
Phosphorylation sites predicted: Ser: 2 Thr: 2 Tyr: 0

## Serine predictions

Name	Pos	Context	Score	Pred
Sequence	10	PEEKSAVTA	0.389	.
Sequence	45	RFFESFGDL	0.621	*S*
Sequence	50	FGDLSTPDA	0.987	*S*
Sequence	73	LGAFLSDGLA	0.026	.
Sequence	90	FATLSELHC	0.020	.

## Threonine predictions

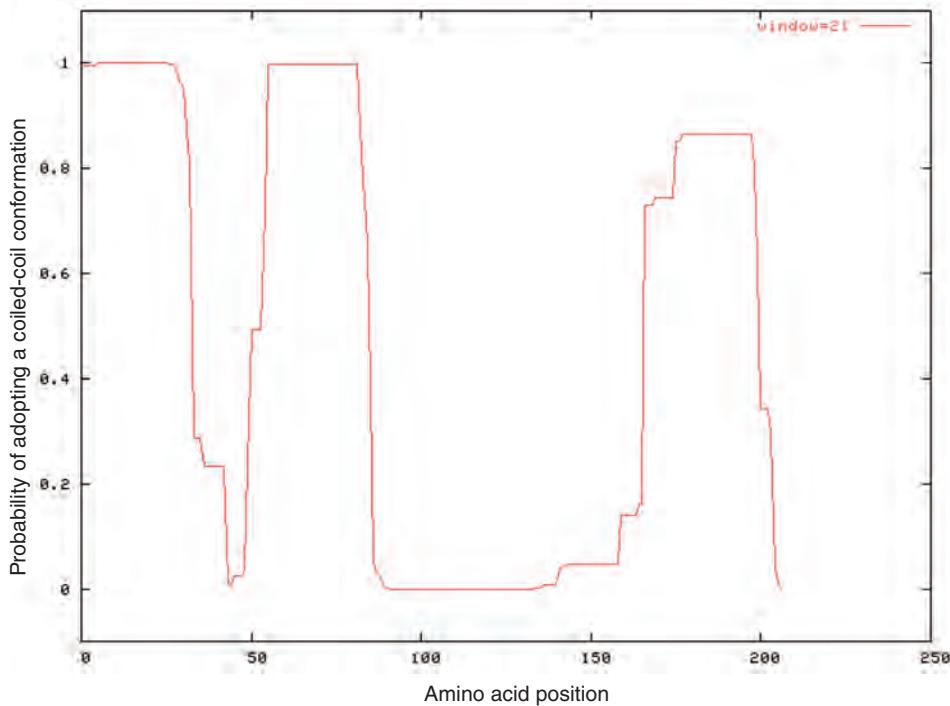
Name	Pos	Context	Score	Pred
Sequence	5	MVHLTPEEK	0.930	*T*
Sequence	13	KSAVTALWG	0.022	.
Sequence	39	VYPWTQRFF	0.398	.
Sequence	51	GDLSTPDAV	0.489	.
Sequence	85	NLKGTFTAL	0.012	.
Sequence	88	GTFATLSEL	0.587	*T*
Sequence	124	GKEFTPVQ	0.393	.



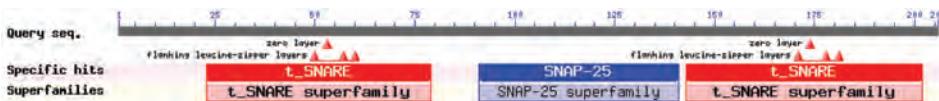
**FIGURE 12.14** The ExPASy web server offers a large group of protein analysis tools such as the NetPhos server for prediction of phosphorylation sites (<http://www.cbs.dtu.dk/services/NetPhos/>). Beta globin protein sequence was input and the output included two likely sites for phosphorylation on serines, two on threonines, and none on tyrosines based on scores exceeding a threshold value of 0.5. Such information on sulfation, phosphorylation, glycosylation, or other post-translational modifications may be fundamental in designing experiments to test the function of a protein.

Source: NetPhos: <http://www.cbs.dtu.dk/services/NetPhos/> ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

(a) COILS output for SNAP-25



(b) Domains from Conserved Domain Database (NCBI)



**FIGURE 12.15** The COILS program of Lupas *et al.* (1991) assesses the likelihood that a protein sequence forms a coiled-coil structure. (a) Output of the coils program using human SNAP-25 protein (NP\_003072.2) as input. The result depicts the probability that the protein will form a coiled-coil secondary structure motif (y axis) across the length of the protein (x axis). Coiled-coils often represent protein–protein interaction domains. In this case, the coiled-coils of SNAP-25, a peripherally associated plasma membrane protein, allow it to bind tightly to two other proteins (syntaxin and vesicle-associated membrane protein (synaptobrevin)) to coordinate synaptic vesicle docking and neurotransmitter release at the presynaptic nerve terminal. (b) According to the Conserved Domain Database (CDD at NCBI), SNAP-25 has two t-SNARE domains that are known to coordinate binding to syntaxin and synaptobrevin. These domains partially overlap the predicted coiled-coil domains.

Source: COILS software at ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

of proteins (see Chapter 13). One such feature is coiled-coil regions, which are typically associated with protein–protein interaction domains (Lupas *et al.*, 1991; Lupas, 1997; Fig. 12.15).

For experimental studies of post-translational modifications, mass spectrometry has a critical role because of its accuracy, broad dynamic range, and sensitivity (Choudhary and Mann, 2010; Sabidó *et al.*, 2012). At the same time, mass spectrometry can be limited: obtaining high-quality results requires considerable expertise, and when a modification such as a phosphorylated amino acid residue is detected we do not necessarily know which kinase was responsible for the activity. Most post-translational modifications require highly specific enzymes to recognize a span of approximately 10 amino acids, including the residue(s) that are physically modified. Computational prediction of such modifications, based on primary sequence data (or primary data further informed by tertiary structure data), complements experimental approaches.

Some proteins with unusual occurrences of particular amino acids are given in “Web Resources”. We provided other examples in Web Documents 4.1–4.4 at <http://www.bioinfbook.org/chapter4>. These proteins may have physical properties (such as pI) that are difficult to predict.

### Accuracy of Prediction Programs

For each of these various prediction programs, it is important to assess the accuracy. This is typically done by measuring sensitivity and specificity relative to a “gold standard” of a set of proteins known to have a particular modification. In recent decades, the physical properties of proteins were assessed at the laboratory bench, one protein at a time (Cooper, 1977). The molecular mass of a protein can be estimated by gel filtration chromatography or by polyacrylamide gel electrophoresis (PAGE). Its shape can be estimated by calculating the frictional coefficient, obtained through a combination of gel filtration and sucrose density gradient centrifugation. Such techniques cannot be applied to large numbers of proteins. Almost all proteins that are studied using the tools of bioinformatics have not been purified, but instead the protein sequence is predicted from genomic DNA or cDNA sequence data.

Prediction programs vary in their accuracy. For proteins with typical amino acid compositions, the prediction of the molecular weight and pI (Fig. 12.13) is likely to be accurate. These protein features can also be experimentally confirmed using techniques such as gel electrophoresis and isoelectric focusing. A prediction algorithm may accurately specify that a protein has a consensus site for phosphorylation or sulfation, but these modifications are not necessarily made in living cells and their regulation is likely to be dynamic. Whether or not a protein has a potential site for modification can be asked; a separate issue is the conditions under which such modification occurs.

### Proteomic Approaches to Phosphorylation

Reversible protein phosphorylation occurs when a kinase mediates the covalent attachment of a phosphate moiety from adenosine triphosphate (ATP) to an acceptor site on a serine, threonine, or tyrosine residue (Dissmeyer and Schnittger, 2011; Derouiche *et al.*, 2012). It has been estimated that one-third of all proteins are phosphorylated, affording an important mechanism for regulating their function. There are also nearly 1000 kinases encoded by the human genome.

Over three dozen software programs have been introduced to predict phosphorylation sites (reviewed in Miller and Blom, 2009; Trost and Kusalik, 2011). Blom *et al.* (1999) introduced NetPhos, the first *in silico* program that predicts phosphorylation (Fig. 12.14). (Available via ExPASy, it continues to be commonly used.) The authors analyzed a large number of amino acid sequences surrounding known acceptor residues on substrate proteins. They applied an artificial neural network to classify sequence patterns in a training set, and then examined a test set. This allowed them to determine the sensitivity (proportion of positive sites correctly predicted) and specificity (proportion of all positive classifications that are correct). A challenge they addressed is that the sequence databases include sites incorrectly annotated as nonphosphorylated (i.e., the false positive rate of their program was inappropriately high). Some methods Blom *et al.* (1999) tested surpassed 95% sensitivity and specificity for predictions of phosphorylation on serine, with less accuracy for predictions on threonine or tyrosine.

Many machine learning methods have been employed to predict phosphorylation sites. Some software packages (such as Scansite) use a position-specific scoring matrix (PSSM) such as those we described in Chapter 5. For phosphorylation site prediction these PSSMs describe the frequency of amino acids occurring at positions surrounding a phosphorylation site. A major limitation of PSSMs is that they do not detect patterns of co-occurring amino acid residues. Other software such as NetPhos use artificial neural networks or support vector machines which are able to model more complex patterns of residues. In addition to the basic algorithm used to predict phosphorylation sites, prediction algorithms also vary in other ways including the following (Trost and Kusalik, 2011):

- the number of residues they examine flanking the phosphorylated residue;
- the properties of surrounding residues such as hydrophobicity;

- the inclusion of secondary and tertiary structural information, as well as the inclusion of information about intrinsic disorder (described in Chapter 13);
- the specification of which particular kinase is responsible for the phosphorylation event; and
- the types of true positive and true negative training data that are employed.

Mass spectrometry is the method of choice for experimental determination of phosphorylation. A competition from the Association of Biomolecular Resource Facilities (ABRF) was used to assess the ability of 54 laboratories to detect phosphorylation sites (Arnott *et al.*, 2003). They prepared a sample consisting of bovine protein disulfide isomerase (PDI; 5 picomoles), two phosphopeptides corresponding to PDI (length 8 and 17 amino acids; 1 picomole each), and bovine serum albumin (BSA; 200 femtomoles). After proteolytic digestion with trypsin, the samples were distributed blind to the research community and 54 laboratories reported 67 analyses. A total of 96% of the laboratories identified PDI, but only 10% detected BSA. There was a surprisingly low success rate for detecting the phosphopeptides and assigning the phosphorylation site; only 3 of 54 laboratories did so for both phosphopeptides. This study highlights the enormous challenges of experimental protein analyses. Most of the laboratories employed MALDI-TOF or LC-MS.

In addition to considering the phosphorylation of individual proteins, many investigators have examined the total collection of phosphorylated sites in a biological sample (the “phosphoproteome”; Kalume *et al.*, 2003; Ptacek and Snyder, 2006). Advances have occurred in the ability to enrich complex mixtures for phosphoproteins and in mass spectrometry approaches (e.g., Ptacek *et al.*, 2005).

A variety of databases provide annotation of post-translational modifications of proteins. The Human Protein Reference Database (HPRD) and Proteinpedia feature expert curation on thousands of proteins, including information on phosphoproteins (Mishra *et al.*, 2006; Goel *et al.*, 2011). Phospho3D specifically focuses on three-dimensional structures of phosphorylation sites (Zanzoni *et al.*, 2011).

### **Proteomic Approaches to Transmembrane Regions**

Cells and intracellular compartments are bordered by phospholipid bilayers. These bilayers include polar heads facing aqueous compartments and lipid tails oriented to the interior of a ~3 nm hydrophobic core. Perhaps 25% of all proteins include transmembrane regions that are capable of spanning the membrane, minimizing the energetically unfavorable interactions of polar amino acid residues with the hydrophobic core. The secondary structure features of the membrane-spanning regions of these proteins include transmembrane  $\alpha$ -helices (typically having a length of 20–25 residues) or transmembrane  $\beta$ -strands organized into  $\beta$ -sheets (typically 9–11 residues). We discuss these aspects of protein secondary structure in Chapter 13, where we also introduce the Protein Data Bank as the main repository of protein structural data. Currently PDB has >100,000 three-dimensional protein structures. Membrane-spanning proteins are notoriously difficult to crystallize and characterize structurally. At present, only ~2000 structures in PDB include transmembrane proteins.

Algorithms can predict the number of transmembrane spans in a protein, their boundaries, and their orientation with respect to the membrane (Punta *et al.*, 2007; Tusnády and Simon, 2010; Nugent and Jones, 2012). One simple approach is to measure hydrophobicity in sliding windows. The hydropathy index of Kyte and Doolittle (1982) was a predominant method in the 1980s and 1990s, aided by the “positive-inside” rule of von Heijne (1992) who noted the tendency of positively charged amino acids to localize to the cytoplasmic face of the membrane in bacteria. More recently it has been possible to apply machine-learning algorithms such as neural networks, support vector machines, or hidden Markov models to transmembrane prediction.

HPRD is available at <http://www.hprd.org> (WebLink 12.3). It includes a PhosphoMotif Finder. Phospho3D can be viewed at <http://www.phospho3d.org/> (WebLink 12.19). It includes P3Dscan in which a Protein Data Bank-formatted structure is searched for a set of phosphorylation site 3D zones present in the phospho3D database.

The PDBTM database collects PDB structures having transmembrane regions. Visit <http://pdtbm.enzim.hu/> (WebLink 12.20). The transmembrane-containing proteins in this database are predicted using TMDET software available at <http://tmdet.enzim.hu/> (WebLink 12.21). The Orientations of Proteins in Membranes (OPM) database lists >2200 membrane proteins at <http://opm.phar.umich.edu/> (WebLink 12.22) (Lomize *et al.*, 2011).

One prominent program for transmembrane domain prediction, TMHMM, employs a hidden Markov model whose states include regions spanning the membrane (the core of a transmembrane helix as well as cytoplasmic and noncytoplasmic caps) and globular regions and loops on the cytoplasmic and noncytoplasmic sides of the membrane (Krogh *et al.*, 2001). The accuracy of this program in predicting the topology of 160 proteins was about 78%. A further advance comes from incorporating information about transmembrane spans with signal peptide predictions (Käll *et al.*, 2007). In analyses of various eukaryotic, bacterial, and archaeal genomes, about 5–10% of all proteins had predicted transmembrane segments that overlap predicted signal peptides (as predicted by software such as SignalP). The Phobius server improves its accuracy by accounting for this.

The TMHMM server is available at <http://www.cbs.dtu.dk/services/TMHMM/> (WebLink 12.23). The Phobius web server is at <http://phobius.sbc.su.se/> (WebLink 12.24). SignalP (Emanuelsson *et al.*, 2007) has a server at <http://www.cbs.dtu.dk/services/SignalP/> (WebLink 12.25).

What is the accuracy of a program that predicts transmembrane topology? It is easy to use a search tool to find a prediction, and large-scale predictions are extremely valuable. However, this is fundamentally a cell biological question which requires the tools of cell biology to obtain a clear answer. Many proteins have stretches of 10–25 hydrophobic amino acid residues that may form transmembrane regions. The most rigorous assessment of the true number of transmembrane spans comes from experimental approaches such as immunocytochemistry. Specific antisera can be raised in rabbits, mice, or other species and used to detect an antigen (such as a stretch of amino acids) in a sample affixed to a microscope slide. In unpermeabilized cells, the antisera can be used to visualize protein regions that are oriented outside the cell. However, when cells are permeabilized with detergent, the antisera can gain access to the cytosol and can therefore visualize intracellular (cytoplasmic) regions. Cell biological analyses such as these have been used to experimentally determine the number of transmembrane regions; in some cases, these results contradict the predictions of hydropathy plots (e.g., Ratnam *et al.*, 1986). As reviewed by Punta *et al.* (2007) and Nugent and Jones (2012), recent high-resolution X-ray crystallographic structures reveal the complexity of topologies:

- Some  $\alpha$ -helices are re-entrant: they enter and exit the membrane on the same side.
- There is an enrichment of interfacial  $\alpha$ -helices at the membrane–water interface; these helices could have functions such as gating channels.
- Some transmembrane helices have kinks and coils, often caused by prolines; half of these helices for which PDB structures are known have kinks,  $\beta_{10}$  helices, or  $\pi$ -helix turns (Chapter 13). Such changes can cause the helical backbone to deviate.
- Many transmembrane helices have a tilted orientation. Other helices include unexpected polar residues, potentially involved in ligand binding or channel gating.

The biological complexity therefore adds to the challenge of *in silico* predictions of transmembrane regions.

### Introduction to Perspectives 3 and 4: Gene Ontology Consortium

The Gene Ontology Consortium main web site is <http://www.geneontology.org/> (WebLink 12.26).

An ontology is a description of concepts. The GO Consortium is a project that compiles a dynamic, controlled vocabulary of terms related to different aspects of genes and gene products (proteins) (Thomas *et al.*, 2007; Gene Ontology Consortium, 2010). A prominent use of this vocabulary is to annotate and interpret the results of microarray experiments that profile RNA transcripts, although many other kinds of high-throughput assays are also annotated using GO (Beissbarth, 2006; Whetzel *et al.*, 2006). The consortium was begun by scientists associated with three model organism databases: the *Saccharomyces* Genome Database (SGD), the *Drosophila* genome database (FlyBase), and the Mouse Genome Informatics databases (MGD/GXD) (Ashburner *et al.*, 2000; Gene Ontology Consortium, 2001). Subsequently, databases associated with many other organisms have joined the GO Consortium (Table 12.6). The GO database is not centralized *per se*, but instead relies on external databases (such as a mouse database) in which each gene or

**TABLE 12.6** Participating organizations and databases in Gene Ontology Consortium.

Database or organization	Organism (or comment)	Common name	URL
Berkeley Bioinformatics Opensource Projects	Various	–	<a href="http://www.berkeleybop.org/">http://www.berkeleybop.org/</a>
DictyBase	<i>Dictyostelium discoideum</i>	Slime mold	<a href="http://dictybase.org/">http://dictybase.org/</a>
EcoliWiki	<i>Escherichia coli</i>	<i>E. coli</i>	<a href="http://ecoliwiki.net/colipedia/">http://ecoliwiki.net/colipedia/</a>
European Bioinformatics Institute (EBI)	GO Editorial office	–	<a href="http://www.ebi.ac.uk/GOA/">http://www.ebi.ac.uk/GOA/</a>
FlyBase	<i>D. melanogaster</i>	Fly	<a href="http://flybase.org/">http://flybase.org/</a>
GeneDB (Wellcome Trust Sanger Institute)	protozoans, fungi		<a href="http://www.genedb.org/">http://www.genedb.org/</a>
Gramene	<i>Oryza sativa</i> ; other grains, monocots	Rice	<a href="http://www.gramene.org/">http://www.gramene.org/</a>
Institute of Genome Sciences, University of Maryland	Various	–	<a href="http://igs.umaryland.edu/">http://igs.umaryland.edu/</a>
InterPro	Various	–	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
J. Craig Venter Institute	Various	–	<a href="http://www.jcvi.org/cms/home/">http://www.jcvi.org/cms/home/</a>
Mouse Genome Informatics	<i>Mus musculus</i>	Mouse	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
Pombase	<i>Schizosaccharomyces pombe</i>	Fission yeast	<a href="http://www.pombase.org/">http://www.pombase.org/</a>
Rat Genome Database (RGD)	<i>Rattus norvegicus</i>	Rat	<a href="http://rgd.mcw.edu/">http://rgd.mcw.edu/</a>
Reactome			<a href="http://www.reactome.org/">http://www.reactome.org/</a>
Saccharomyces Genome Database (SGD)	<i>Saccharomyces cerevisiae</i>	Baker's yeast	<a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a>
The Arabidopsis Information Resource (TAIR)	<i>Arabidopsis thaliana</i>	Thale cress	<a href="http://www.arabidopsis.org/">http://www.arabidopsis.org/</a>
UniProtKB-Gene Ontology Annotation	Various	–	<a href="http://www.ebi.ac.uk/GOA">http://www.ebi.ac.uk/GOA</a>
WormBase	<i>Caenorhabditis elegans</i>	Worm	<a href="http://www.wormbase.org/">http://www.wormbase.org/</a>
Zebrafish Information Network	<i>Danio rerio</i>	zebrafish	<a href="http://zfin.org/">http://zfin.org/</a>

Source: Gene Ontology Consortium (2001), licenced under the Creative Commons Attribution 4.0 Unported License, CC-BY-4.0.

gene product is annotated with GO terms. It therefore represents an ongoing, cooperative effort to unify the way genes and gene products are described. There are several web browsers that serve as principal gateways to search GO terms (**Table 12.7**). Additionally, EBI, Ensembl, and NCBI gene and protein entries (Chapter 2) contain GO terms.

There are three main organizing principles of GO: (1) molecular function; (2) biological process; and (3) cellular component. Molecular function refers to the tasks performed by individual gene products. For example, a protein can be a transcription factor or a carrier protein. Biological process refers to the broad biological goals that a gene product (protein) is associated with, such as mitosis or purine metabolism. Cellular component refers to the subcellular localization of a protein. Examples include nucleus and lysosome. Any protein may participate in more than one molecular function, biological process, and/or cellular component.

Genes and gene products are assigned to GO categories through a process of annotation. The author of each GO annotation supplies an evidence code that indicates the basis for that annotation (**Table 12.8**). As an example of a GO-annotated protein, look at the NCBI Gene entry for human beta globin (HBB; **Fig. 12.16**). NCBI Gene entries include a section on function that includes information from OMIM (Chapter 21), Enzyme Commission nomenclature (see “Perspective 4”), and GO terms. For HBB, the GO terms include heme binding, oxygen binding, and oxygen transporter activity (molecular functions); oxygen transport (a biological process); and hemoglobin complex (a cellular component).

**TABLE 12.7** Websites useful to access gene ontology data.

Browser	Description	URL
AmiGO	GO browser from the Berkeley Drosophila Genome Project	<a href="http://amigo.geneontology.org/">http://amigo.geneontology.org/</a>
Mouse Genome Informatics (MGI) GO Browser	From Jackson Laboratories	<a href="http://www.informatics.jax.org/searches/GO_form.shtml">http://www.informatics.jax.org/searches/GO_form.shtml</a>
"QuickGO" at EBI	From the EMBL and European Bioinformatics Institute; integrated with InterPro (Chapter 10)	<a href="http://www.ebi.ac.uk/QuickGO/">http://www.ebi.ac.uk/QuickGO/</a>
Cancer Gene Anatomy Project (CGAP) GO Browser	From the National Cancer Institute, NIH	<a href="http://cgap.nci.nih.gov/Genes/AllAboutGO">http://cgap.nci.nih.gov/Genes/AllAboutGO</a>

**TABLE 12.8** Evidence Codes for Gene Ontology Project.

Code	Evidence code	Example(s), notes
EXP	Inferred from experiment <sup>a</sup>	Parent of IDA, IEP, IGI, IMP, IPI
IDA	Inferred from direct assay <sup>a</sup>	An enzyme assay (for function); immunofluorescence microscopy (for cellular component)
IEP	Inferred from expression pattern <sup>a</sup>	Transcripts levels (e.g., based on Northern blotting or microarrays) or protein levels (e.g., Western blots)
IGI	Inferred from genetic interaction <sup>a</sup>	Suppressors; genetic lethals; complementation assays; experiments in which one gene provides information about the function, process, or component of another gene
IMP	Inferred from mutant phenotype <sup>a</sup>	Gene mutation; gene knockout; overexpression; antisense assays
IPI	Inferred from physical interaction <sup>a</sup>	Yeast two-hybrid assays; copurification; co-immunoprecipitation; binding assays
IGC	Inferred from genomic context <sup>b</sup>	Identity of the genes neighboring the gene product in question (i.e., synteny), operon structure, and phylogenetic or other whole-genome analysis
IRD	Inferred from rapid divergence <sup>b</sup>	A type of phylogenetic evidence
ISA	Inferred from sequence alignment <sup>b</sup>	Note that quantitative criteria for sequence alignment are not employed
ISO	Inferred from sequence orthology <sup>b</sup>	Note that orthology is defined permissively, with nonquantitative sequence comparisons
ISS	Inferred from sequence or structural similarity <sup>b</sup>	Sequence similarity; domains; BLAST results that are reviewed for accuracy by a curator
RCA	Inferred from reviewed computational analysis <sup>b</sup>	Predictions based on large-scale experiments (e.g., genome-wide two-hybrid, genome-wide synthetic interactions); predictions based on integration of large-scale datasets of several types; text mining)
NAS	Nontraceable author statement <sup>c</sup>	Database entries such as a SwissProt record that does not cite a published paper
TAS	Traceable author statement <sup>c</sup>	Information in a review article or dictionary
IC	Inferred by curator <sup>d</sup>	A protein is annotated as having the function of a "transcription factor"; a curator may then infer that the localization is "nucleus"
ND	No biological data available <sup>d</sup>	Corresponds to "unknown" molecular function, biological process, or cellular compartment
IEA	Inferred from electronic annotation <sup>e</sup>	Annotations based on "hits" in searches such as BLAST (but without confirmation by a curator; compare ISS)

<sup>a</sup>Experimental evidence codes; <sup>b</sup>Computational analysis evidence codes; <sup>c</sup>Author statement evidence codes; <sup>d</sup>Curator statement evidence codes; <sup>e</sup>Automatically-assigned evidence code. Not all evidence codes are shown.

**GeneOntology**Provided by [GOA](#)

Function	Evidence
<a href="#">heme binding</a>	IEA
<a href="#">hemoglobin binding</a>	IDA <a href="#">PubMed</a>
<a href="#">iron ion binding</a>	IEA
<a href="#">metal ion binding</a>	IEA
<a href="#">molecular function</a>	ND
<a href="#">oxygen binding</a>	IDA <a href="#">PubMed</a>
<a href="#">oxygen binding</a>	IEA
<a href="#">oxygen transporter activity</a>	IEA
<a href="#">oxygen transporter activity</a>	NAS <a href="#">PubMed</a>
<a href="#">selenium binding</a>	IDA <a href="#">PubMed</a>

Process	Evidence
<a href="#">biological process</a>	ND
<a href="#">nitric oxide transport</a>	NAS <a href="#">PubMed</a>
<a href="#">oxygen transport</a>	IEA
<a href="#">oxygen transport</a>	NAS <a href="#">PubMed</a>
<a href="#">oxygen transport</a>	TAS <a href="#">PubMed</a>
<a href="#">positive regulation of nitric oxide biosynthetic process</a>	NAS <a href="#">PubMed</a>
<a href="#">transport</a>	IEA

Component	Evidence
<a href="#">hemoglobin complex</a>	IEA
<a href="#">hemoglobin complex</a>	NAS <a href="#">PubMed</a>
<a href="#">hemoglobin complex</a>	TAS <a href="#">PubMed</a>

**FIGURE 12.16** The GO Consortium provides a dynamic, controlled vocabulary that describes genes and gene products from a variety of organisms. Its three organizing principles are molecular function, biological process, and cellular component. GO terms can be accessed through a variety of browsers or through NCBI Gene, as shown for human beta globin. These GO terms are obtained from the Gene Ontology Annotation (GOA) Database at the European Bioinformatics Institute.

Source: Gene Ontology Consortium (2001), licenced under the Creative Commons Attribution 4.0 Unported License, CC-BY-4.0.

You can also access gene ontology information by entering a query term such as “HBB” or “lipocalin” into a GO web browser. In some cases the output includes a graphical tree view. This displays the relationships between the different levels of GO terms, which have the form of a “directed acyclic graph” or network. This differs from a hierarchy; in a hierarchy each child term can have only one parent, while in a directed acyclic graph it is possible for a child to have more than one parent. A child term may be an instance of its parent term, in which case the graph is labeled “isa,” or the child term may be component of the parent term (a “partof” relationship). This complicates the structure of the terms in GO and the evaluation of their biological and statistical significance. Some statistical tests assess the likelihood that each GO category is under- or overrepresented more than is expected by chance. However, a concept such as “mitochondria” occurs in

all the three categories (biological process, molecular function, cellular compartment) and at multiple levels.

We next consider protein localization and protein function. These topics loosely correspond to the GO categories “cellular component” and “molecular function.” In Chapter 14 we discuss protein pathways, although the GO category “biological process” does not refer specifically to pathways.

### Perspective 3: Protein Localization

The cellular localization of a protein is one of its fundamental properties. Proteins are synthesized on ribosomes from mRNA, and some are synthesized in the cytosol. Other proteins, destined for secretion or insertion in the plasma membrane, are inserted into the endoplasmic reticulum (in eukaryotes) or into the plasma membrane (in bacteria and archaea). This insertion, which occurs either cotranslationally or post-translationally, is mediated by the signal recognition particle, an RNA–multiprotein complex (Stroud and Walter, 1999). In the endoplasmic reticulum, proteins may be transported through the secretory pathway to the Golgi apparatus and then to further destinations such as intracellular organelles (e.g., endosomes, lysosomes) or to the cell surface.

Proteins may further be secreted into the extracellular milieu. The trafficking of a protein to its appropriate destination is achieved by transport in secretory vesicles. These vesicles are typically 75–100 nm in diameter, and they transport soluble or membrane-bound cargo to specific compartments.

We may also distinguish two main categories of proteins based upon their relationships to phospholipid bilayers: (1) those that are soluble and exist in the cytoplasm, in the lumen of an organelle, or in the extracellular environment; and (2) those that are membrane attached, associated with a lipid bilayer. Those proteins associated with membranes may be integral membrane proteins (having a span of 10–25 hydrophobic amino acid residues that traverse the lipid bilayer) or they may be peripherally associated with membranes (attached via a variety of anchors such as those shown in Fig. 12.12).

Many proteins defy categorization into one static location in the cell. For example, the annexins and the low-molecular-weight GTP-binding proteins are families of proteins that migrate between the cytosol and a membrane compartment. This movement typically depends on the presence of dynamically regulated cellular signals such as calcium or transient phosphorylation.

Proteins are often targeted to their appropriate cellular location because of intrinsic signals embedded in their primary amino acid sequence. For example, the sequence KDEL (lysine–aspartic acid–glutamic acid–leucine) at the carboxy terminus of a soluble protein specifies that it is selectively retained in the endoplasmic reticulum. Other targeting motifs have been identified for import into mitochondria, lysosomes, or peroxisomes and for endocytosis. However, these motifs are typically not as invariant as KDEL.

Several web-based programs predict the intracellular localization of any individual protein sequence (Casadio *et al.*, 2008; Imai and Nakai, 2010; see also “Web Resources”). For example, WoLF PSORT accurately predicts the signal sequence at the amino terminus of retinol-binding protein (Fig. 12.17). This signal peptide is characteristic of proteins that enter the secretory pathway in the endoplasmic reticulum. WoLF PSORT analyzes a protein query for localization features based on sorting signals, amino acid composition, and functional motifs (Horton *et al.*, 2007). It then uses a *k*-nearest neighbor classifier to predict the localization.

### Perspective 4: Protein Function

We have described bioinformatics tools to describe protein families, their physical properties, and the cellular localization of proteins. A fourth aspect of proteins is their function

In eukaryotic cells, the intracellular organelles account for up to 95% of the cell's membranes.

You can access WoLF PSORT server at <http://wolfpsort.org/> (WebLink 12.27).

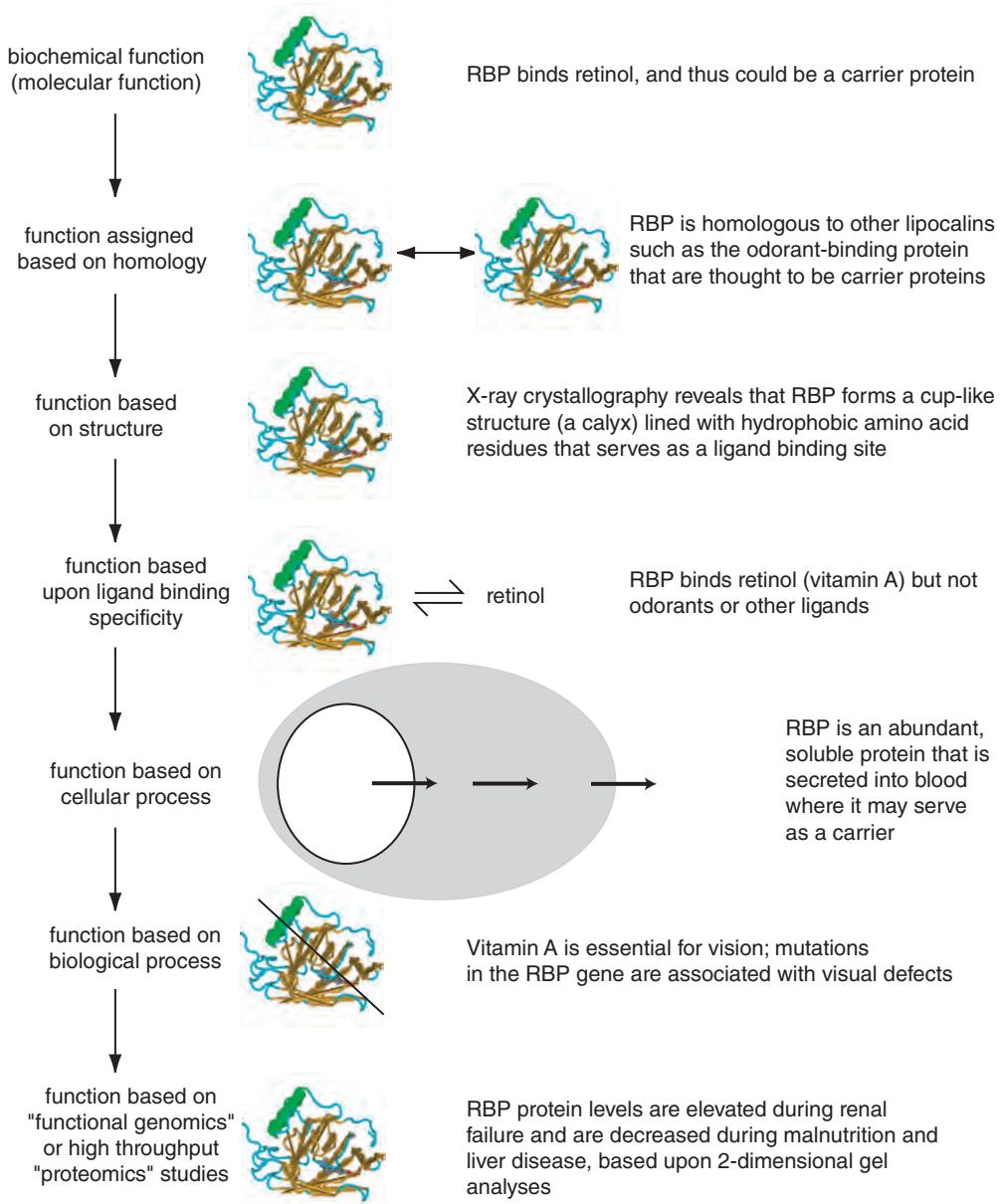
		Normalized Feature Values																						
id	site	iPSORT					PSORT Features										Amino Acid Content				Misc.			
		-1	25	MxHy1	30	act	alm	dna	gvh	leu	mNt	mp	mit	myr	nuc	rib	rmp	tms	tyr	vac	C	I	K	S
queryProtein	extr?	78	75	50	76	44	96	46	49	71	59	49	29	50	50	27	49	48	68	10	51	27	26	
CASP_CHICK	extr	78	70	50	82	44	90	46	49	55	80	49	29	50	50	27	49	48	60	26	34	17	36	
IL10_HUMAN	extr	64	89	50	93	44	100	46	49	71	61	49	29	50	50	27	49	48	74	35	76	31	23	
IL10_MACNE	extr	64	89	50	93	44	100	46	49	84	61	49	29	50	50	27	49	48	74	35	70	41	23	
A2HS_HUMAN	extr	64	75	50	57	44	97	46	49	81	56	49	70	50	50	27	49	48	80	12	32	28	53	
IL10_CERTO	extr	64	89	50	93	44	100	46	49	84	61	49	29	50	50	27	49	48	74	35	76	41	23	
IL10_MACFA	extr	64	89	50	93	44	100	46	49	84	61	49	29	50	50	27	49	48	74	35	76	41	23	
IL10_MACMU	extr	64	89	50	93	44	100	46	49	84	61	49	29	50	50	27	49	48	74	35	76	41	23	
IBP2_BRARE	extr	46	80	50	50	44	84	46	49	75	67	49	29	50	50	27	49	48	92	15	56	37	36	
PPT1_HUMAN	lyso	64	75	50	52	44	96	46	49	76	63	49	29	50	50	27	49	48	61	61	52	37	40	
NDDB_CAVPO	extr	46	74	50	47	44	86	46	49	84	77	49	29	50	50	27	49	48	65	7	48	37	32	

**FIGURE 12.17** The WoLF PSORT server provides a web-based query form to predict the subcellular location of a protein. The program searches for sorting signals and other features that are characteristic of proteins localized to particular compartments. The output of a search using retinol-binding protein protein sequence (NP\_006735) includes 32 nearest neighbors (of which ten are shown here in rows along with the query). The columns include features analyzed including site (proposing correctly that the query is extracellular; see column labeled site), results of the iPSORT program (including calculations of the negative charge and hydrophobicity of the initial 25–30 amino acid residues), and results from the PSORT program (including the presence of motifs typical of proteins localized to various subcellular compartments). The output shows that there is strong evidence for a signal peptide with a cleavage site between amino acid residues 16 and 17. Such a signal peptide characterizes proteins that enter the secretory pathway where some (such as RBP) are secreted outside the cell.

Source: WoLF PSORT. Courtesy of K. Nakai.

(Raes *et al.*, 2007). Function is defined as the role of a protein in a cell (Jacq, 2001). Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function. We can consider the concept of protein function from seven perspectives (Fig. 12.18) as follows.

1. A protein has a biochemical function synonymous with its molecular function.  
For an enzyme, the biochemical function is to catalyze the conversion of one or more substrates to product(s). For a structural protein such as actin or tubulin, the biochemical function is to influence the shape of a cell. For a transport protein, the biochemical function is to carry a ligand from one location to another. (Such a transport role may even occur in the absence of a requirement for an energy source such as ATP; in such a way, retinol-binding protein transports retinol through serum, and hemoglobin transports oxygen.) For a hypothetical protein that is predicted to be encoded by a gene, the biochemical function is unknown but is presumed to exist. There are thought to be no proteins that exist without a biochemical function.
2. Functional assignment is often made based upon homology (Ponting, 2001; Lee *et al.*, 2007; Emes *et al.*, 2008; Mazumder *et al.*, 2008). Currently, when a genome is sequenced the great majority of its predicted proteins can be functionally assigned based on orthology. If a hypothetical protein is homologous to an enzyme, it is often provisionally assigned that enzymatic function. This is best viewed as a hypothesis that must be tested experimentally. As an example, many globin-like proteins occur in bacteria, protozoa, and fungi having biochemical properties distinct from those of vertebrate globins (Poole and Hughes, 2000).



**FIGURE 12.18** Protein function may be analyzed from several perspectives. Retinol-binding protein (RBP) is used as an example.

- Function may be assigned based upon structure (Chapter 13). If a protein has a three-dimensional fold that is related to that of a protein with a known function, this may be the basis for functional assignment. Note, however, that structural similarity does not necessarily imply homology, and homology does not necessarily imply functional equivalence.
- All proteins function in the context of other proteins and molecules. A definition of a protein's function may include its ligand (if the protein is a receptor), its substrate (if the protein is an enzyme), its lipid partner (if the protein interacts with membrane), or any other molecule with which it interacts. The odorant-binding protein (OBP) is a lipocalin that binds a variety of odorants in nasal mucus, suggesting that the binding properties of the protein are central to its function (Pevsner *et al.*, 1990). However, the biological function of OBP is not known from its ligand-binding properties alone.

- The protein could transport odorants toward the olfactory epithelium to promote sensory perception, it could carry odorants from the olfactory epithelium to facilitate odorant clearance, or it could metabolize odorants.
5. Many proteins function as part of a distinct biochemical pathway such as the Krebs cycle, in which discrete steps allow the cell to perform a complex task. Other examples are fatty acid oxidation in peroxisomes or proteolytic degradation that is accomplished by the proteasome.
  6. Proteins function as part of some broad cell biological process. Cells divide, grow, and senesce; neurons have axons that display outgrowth, pathfinding, target recognition, and synapse formation; and all cells secrete molecules through discrete pathways. All cellular processes require proteins in order to function, and each individual protein can be defined in the context of the broad cellular function it serves. The Gene Ontology Consortium (Ashburner *et al.*, 2000, p. 27) defines a biological process as “a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation, in the sense that something goes into a process and something different comes out of it.”
  7. Protein function can be considered in the context of all the proteins that are encoded by a genome, that is, in terms of the proteome. The term *functional genomics* includes the use of experimental approaches and/or computational tools to analyze the role of many hundreds or thousands of expressed genes (i.e., RNA transcripts). Since the ultimate product of transcription is a protein, the term functional genomics is sometimes applied to large-scale studies of protein function. Chapter 14 addresses the topic of functional genomics.

Protein function can therefore be defined in many ways. Many proteins are enzymes (Alderson *et al.*, 2012). The Enzyme Commission (EC) system provides a standardized nomenclature for almost 4000 enzymes (**Table 12.9**). When a genome is sequenced and a potential protein-coding sequence is identified, homology of that protein to an enzyme with a defined EC listing provides a specific, testable hypothesis about the biochemical function of that hypothetical protein.

Another broader approach to the functional assignment of proteins is provided by the Clusters of Orthologous Groups (COGs) database developed by Eugene Koonin and colleagues (Tatusov *et al.*, 1997, 2003; Kristensen *et al.*, 2010). The functional groups defined by this system are listed in **Table 12.10**. While the COGs database initially focused on bacterial and archaeal genomes, the general categories are relevant to basic cellular processes in all living organisms. Many other functions that are unique to eukaryotes, such as apoptosis and complex developmental processes, are represented in the eukaryotic portion of the COGs scheme (Tatusov *et al.*, 2003). Peer Bork and colleagues developed the Evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG) database which extends the COGs concept (although without its manual annotation) to >1100 organisms and >700,000 orthologous groups (Powell *et al.*, 2012).

Apoptosis is programmed cell death. It occurs in a variety of multicellular organisms, both as a normal process in development and as a homeostatic mechanism in adult tissues. Apoptosis can be triggered by external stimuli (such as infectious agents or toxins) or by internal agents such as those causing oxidative stress. You can visit the COGs database at <http://www.ncbi.nlm.nih.gov/COG/> (WebLink 12.28). EggNOG is online at [http://eggnog.embl.de/version\\_3.0/](http://eggnog.embl.de/version_3.0/) (WebLink 12.29).

## PERSPECTIVE

In this chapter we have considered bioinformatics approaches to individual proteins. In Chapter 13 we consider protein structure, which provides us with deeper insight into the nature of proteins including their domains, physical properties, and function. In Chapter 14 (functional genomics) we explore high-throughput approaches to studying sets of proteins (e.g., techniques employing gel electrophoresis and mass spectrometry) as well as protein–protein interactions and networks.

**TABLE 12.9 Functional assignment of 5276 proteins based upon their enzymatic activity: partial list of enzyme commission classification system. Number of enzymes refers to the UniProtKB entries matching that query (e.g., “ec:2.-.-reviewed:yes”).**

EC number	Description of class	Number of enzymes	Example of subclass
1. -.-.-	Oxidoreductases	38,216	
1. 1. -.-	-	-	Acting on the CH-OH group of donors
1. 2. -.-	-	-	Acting on the aldehyde or oxo group of donors
2. -.-.-	Transferases	89,624	
2. 1. -.-	-	-	Transferring one-carbon groups
3. -.-.-	Hydrolases	62,574	
4. -.-.-	Lyases	23,427	
5. -.-.-	Isomerases	14,163	
6. -.-.-	Ligases	30,569	

Source: <http://www.expasy.org/enzyme/> ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

In the past decade, our understanding of the properties of proteins has advanced dramatically, from the level of biochemical function to the role of proteins in cellular processes. Advances in instrumentation have propelled mass spectrometry into a leading role for many proteomics applications.

Many web-based tools are available to evaluate the biochemical features of individual proteins. Such programs can predict the existence of glycosylation, phosphorylation, or other sites. These predictions can be extremely valuable in guiding the biologist to experimentally test the possible post-translational modifications of a protein.

High-throughput approaches have been used in an effort to define the function of all proteins. Large numbers of proteins still have no known function because they lack detectable homology to other characterized proteins. We will continue to obtain a more comprehensive description of protein function as distinct high-throughput strategies are applied to model organisms, such as large-scale analyses of protein localization and protein interactions.

## PITFALLS

Many of the experimental and computational strategies used to study proteins have limitations. Two-dimensional protein gels are most useful for studying relatively abundant proteins, but thousands of proteins expressed at low levels are harder to characterize. Experimental approaches are extremely challenging in practice, as shown by the ABRF critical assessments. Many computational approaches suffer from high false positive error rates, reflecting the difficulty of obtaining adequate training sets.

## ADVICE FOR STUDENTS

In my experience many students have a single protein they are studying in depth. My suggestion is to try to learn all that can be known about it (including its domains, physical properties, localization, and function as outlined in this chapter). At the same time, try to

**TABLE 12.10 Functional classification of proteins in clusters of orthologous groups database.**

General category	Function	Clusters of orthologous groups	Domains
Information storage and processing	Translation, ribosomal structure, and biogenesis	245	10,572
	RNA processing and modification	25	137
	Transcription	231	11,271
	Replication, recombination, and repair	238	10,338
	Chromatin structure and dynamics	19	228
Cellular processes and signaling	Cell cycle control, cell division chromosome partitioning	72	1,678
	Defense mechanisms	46	2,380
	Signal transduction mechanisms	152	7,683
	Cell wall/membrane/envelope biogenesis	188	7,858
	Cell motility	96	2,747
	Cytoskeleton	12	128
	Extracellular structures	1	25
	Intracellular trafficking, secretion, vesicular transport	159	3,743
	Post-translational modification, protein turnover, chaperones	203	6,206
	Energy production and conversion	223	5,584
	Carbohydrate transport and metabolism	170	5,257
Metabolism	Energy production and conversion	258	9,830
	Carbohydrate transport and metabolism	230	10,816
	Amino acid transport and metabolism	270	14,939
	Nucleotide transport and metabolism	95	3,922
	Coenzyme transport and metabolism	179	6,582
	Lipid transport and metabolism	94	5,201
	Inorganic ion transport and metabolism	212	9,232
	Secondary metabolites biosynthesis, transport and catabolism	88	4,055
Poorly characterized	General function prediction only	702	22,721
	Function unknown	1,346	13,883

Source: Clusters of Orthologous Groups Database, NCBI. <http://www.ncbi.nlm.nih.gov/COG/>

understand what questions you want to ask about this protein, and what techniques are appropriate. We have taken the example of a predicted transmembrane spanning protein: while computational predictions are easy to make, their predictive value may be relatively low. In such cases let the predictions guide which biological experiments need to be performed. Let bioinformatics serve biology, rather than the other way around.

## WEB RESOURCES

I suggest proteomics web resources including tools to analyze protein motifs (**Table 12.11**), secondary structure analysis (**Table 12.12**), glycosylation analysis (**Table 12.13**), post-translational modifications (**Table 12.14**), proteins with unusual occurrences of particular amino acids (**Table 12.15**), prediction of protein localization (**Table 12.16**), and prediction of transmembrane regions (**Table 12.17**).

**TABLE 12.11 Tools to analyze protein motifs.**

Program	Comment	URL
Source of many tools	ExPASy	<a href="http://www.expasy.org/proteomics">http://www.expasy.org/proteomics</a>
InterProScan	At EBI	<a href="http://www.ebi.ac.uk/Tools/pfa/iprscan/">http://www.ebi.ac.uk/Tools/pfa/iprscan/</a>
PROSITE Scan	At EBI	<a href="http://www.ebi.ac.uk/Tools/pfa/ps_scan/">http://www.ebi.ac.uk/Tools/pfa/ps_scan/</a>
PRATT	At EBI	<a href="http://www.ebi.ac.uk/Tools/pfa/pratt/">http://www.ebi.ac.uk/Tools/pfa/pratt/</a>
Motif Scan	At SIB	<a href="http://hits.isb-sib.ch/cgi-bin/PFSCAN">http://hits.isb-sib.ch/cgi-bin/PFSCAN</a>
Source of many tools	Pôle Bio-Informatique Lyonnais	<a href="http://pbil.univ-lyon1.fr/">http://pbil.univ-lyon1.fr/</a>
SMART	At EMBL	<a href="http://smart.embl-heidelberg.de/">http://smart.embl-heidelberg.de/</a>
TEIRESIAS	At IBM	<a href="http://cbcdrv.watson.ibm.com/Tspd.html">http://cbcdrv.watson.ibm.com/Tspd.html</a>

**TABLE 12.12 Tools to analyze primary and/or secondary structure features of proteins.**

Program	Source/comment	URL
COILS	Prediction of coiled-coil regions in proteins	<a href="http://www.ch.embnet.org/software/COILS_form.html">http://www.ch.embnet.org/software/COILS_form.html</a>
Compute pI/Mw	From ExPASy	<a href="http://web.expasy.org/compute_pi/">http://web.expasy.org/compute_pi/</a>
Helical wheel	Draws an helical wheel (i.e., an axial projection of a regular alpha helix)	<a href="http://www-nmr.cabm.rutgers.edu/bioinformatics/Proteomic_tools/Helical_wheel/">http://www-nmr.cabm.rutgers.edu/bioinformatics/Proteomic_tools/Helical_wheel/</a>
M.M., pl, composition, titrage	Many tools from the Atelier Bio Informatique de Marseille	<a href="http://sites.univ-provence.fr/wabim/english/logligne.html">http://sites.univ-provence.fr/wabim/english/logligne.html</a>
Paircoil	Prediction of coiled-coil regions in proteins	<a href="http://groups.csail.mit.edu/cb/paircoil/paircoil.html">http://groups.csail.mit.edu/cb/paircoil/paircoil.html</a>
Peptidemass	From ExPASy	<a href="http://web.expasy.org/peptide_mass/">http://web.expasy.org/peptide_mass/</a>

*Source:* ExPASy, <http://www.expasy.org/tools/> ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

**TABLE 12.13 Web resources for characterization of glycosylation sites on proteins**

Program	Comment/source	URL
DictyOGlyc 1.1 Prediction Server	Neural network predictions for GlcNAc O-glycosylation sites in <i>Dictyostelium discoideum</i> proteins	<a href="http://www.cbs.dtu.dk/services/DictyOGlyc/">http://www.cbs.dtu.dk/services/DictyOGlyc/</a>
NetGlycate	Prediction of glycation of ε amino groups of lysines in mammalian proteins	<a href="http://www.cbs.dtu.dk/services/NetGlycate/">http://www.cbs.dtu.dk/services/NetGlycate/</a>
NetOGlyc	Prediction of type O-glycosylation sites in mammalian proteins	<a href="http://www.cbs.dtu.dk/services/NetOGlyc/">http://www.cbs.dtu.dk/services/NetOGlyc/</a>
YinOYang 1.2	Produces neural network predictions for O-β-GlcNAc attachment sites in eukaryotic protein sequences	<a href="http://www.cbs.dtu.dk/services/YinOYang/">http://www.cbs.dtu.dk/services/YinOYang/</a>

**TABLE 12.14 Tools to analyze post-translational modifications.**

Program	Comment	URL
big-PI Predictor	GPI modification site prediction	<a href="http://mendel.imp.ac.at/gpi/gpi_server.html">http://mendel.imp.ac.at/gpi/gpi_server.html</a>
NetPhos 2.0 Prediction Server	Produces neural network predictions for serine, threonine, and tyrosine phosphorylation sites in eukaryotic proteins	<a href="http://www.cbs.dtu.dk/services/NetPhos/">http://www.cbs.dtu.dk/services/NetPhos/</a>
Sulfinator	Prediction of tyrosine sulfation sites	<a href="http://web.expasy.org/sulfinator/">http://web.expasy.org/sulfinator/</a>

Source: ExPASy, <http://www.expasy.org/tools/> ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

**TABLE 12.15 Examples of proteins with unusually high occurrences of specific amino acids. The hydrophobic residues characteristic of transmembrane helices are from Tanford (1980). Adapted from Ponting (2001), with permission from Oxford University Press.**

Amino acid(s)	Proteins
C	Disulfide-rich proteins; metallothioneins; zinc finger proteins
D, E	Acidic proteins (e.g., NP_033802.2)
G	Collagens (e.g., NP_000079)
H	Hisactophilin; histidine-rich glycoprotein (e.g., XP_629852)
W, L, P, Y, I, V, M, A	Transmembrane domains (e.g., NP_004594, NP_062098)
K, R	Nuclear proteins (nuclear localization signals)
N	<i>Dictyostelium</i> proteins
P	Collagens (e.g., NP_000079.2); filaments; SH3/WW/EVHI binding sites
Q	Proteins encoded by genes mutated in triplet repeat disorders (Chapter 21; e.g., huntingtin, NP_002102.4)
S, R	Some RNA-binding motifs
S, T	Mucins; oligosaccharide attachment sites (e.g., XP_855042)
abcdefg	Heptad coiled coils ( <b>a</b> and <b>d</b> are hydrophobic residues; e.g., myosin NP_005370)

**TABLE 12.16 Web-based programs for prediction of protein localization**

Program	Comment	URL
ChloroP	Predicts presence of chloroplast transit peptides (cTP) in protein sequences	<a href="http://www.cbs.dtu.dk/services/ChloroP/">http://www.cbs.dtu.dk/services/ChloroP/</a>
MITOPROT	Calculates the N-terminal protein region that can support a mitochondrial targeting sequence and the cleavage site	<a href="http://ihg.gsf.de/ihg/mitoprot.html">http://ihg.gsf.de/ihg/mitoprot.html</a>
PSORT	Prediction of protein-sorting signals and localization sites; access to PSORT II, WoLF PSORT	<a href="http://psort.hgc.jp/">http://psort.hgc.jp/</a>
SignalP	Predicts presence and location of signal peptide cleavage sites in prokaryotes and eukaryotes	<a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a>
TargetP	Predicts subcellular location of eukaryotic protein sequences	<a href="http://www.cbs.dtu.dk/services/TargetP/">http://www.cbs.dtu.dk/services/TargetP/</a>

**TABLE 12.17** Web servers for prediction of transmembrane regions in protein sequences. From ExPASy web server.

Program	Comment/source	URL
DAS server	Prediction of transmembrane regions	<a href="http://www.sbc.su.se/~miklos/DAS/">http://www.sbc.su.se/~miklos/DAS/</a>
APSSP	Advanced protein secondary structure prediction server	<a href="http://imtech.res.in/raghava/apssp/">http://imtech.res.in/raghava/apssp/</a>
HMMTOP	Prediction of transmembrane helices and topology of proteins	<a href="http://www.enzim.hu/hmmtop/">http://www.enzim.hu/hmmtop/</a>
Phobius	Combined transmembrane topology and signal peptide predictor	<a href="http://phobius.sbc.su.se/">http://phobius.sbc.su.se/</a> <a href="http://www.ebi.ac.uk/Tools/pfa/phobius/">http://www.ebi.ac.uk/Tools/pfa/phobius/</a>
PredictProtein server	Prediction of transmembrane helix location and topology	<a href="https://www.predictprotein.org/">https://www.predictprotein.org/</a>
TMpred	Prediction of membrane-spanning regions and their orientation	<a href="http://www.ch.embnet.org/software/TMPRED_form.html">http://www.ch.embnet.org/software/TMPRED_form.html</a> <a href="http://embnet.vital-it.ch/software/TMPRED_form.html">http://embnet.vital-it.ch/software/TMPRED_form.html</a>
TopPred2	Topology prediction of membrane proteins	<a href="http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::toppred">http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::toppred</a>



## Discussion Questions

**[12-1]** InterPro is an important resource that coordinates information about protein signatures from a variety of databases.

When these databases all describe a particular protein family or a particular signature, what different kinds of information can you obtain? Is the information in InterPro redundant?

**[12-2]** How do you define the function of a protein? Does the function, physiological state, or other condition change over time?

### PROBLEMS/COMPUTER LAB

**[12-1]** Use biomaRt to extract a protein sequence starting with the HGNC gene symbol for hemoglobin beta (*HBB*). (1) Install R and RStudio and load `biomaRt` (Chapter 8 and this chapter). (2) Retrieve the sequence with the following four commands.

```
> library(biomaRt)
> mart <- useMart(biomart="ensembl",
dataset="hsapiens_gene_ensembl")
> seq = getSequence(id="HBB",
type="hgnc_symbol", seqType="peptide",
mart=mart)
> seq
```

**[12-2]** We introduced EDirect (Chapter 2, “Accessing NCBI Databases with EDirect”) as a way to access NCBI Entrez databases from the command line. First, set it up

(on Linux, Windows, or Mac machines). Then extract the largest human proteins in the FASTA format. Try the following command, in which the \$ symbol indicates a UNIX prompt.

```
$ esearch -db protein -query
"1000000:1500000 [MLWT] AND human [ORGN]"
| efetch -format fasta
```

Here the `esearch` program searches through the database we specify (protein database), and the query is restricted to a particular molecular weight (MLWT) and species (human). We then use the pipe command (`|`) to send the result to the `efetch` utility that downloads the data of interest. We specify that we want the data in the FASTA format. You can also send the results to a file called `myresults.txt`:

```
$ esearch -db protein -query "1000000:1500000
[MLWT] AND human [ORGN]"
| efetch -format fasta > myresults.txt
```

Next try searching for different information using additional queries and filters.

**[12-3]** Select a group of unaligned, divergent globins (Web Document 6.3 at <http://www.bioinfbook.org/chapter6>). Use them as input to the PRATT program at Prosite (<http://www.expasy.ch/prosite/>) in order to find a representative pattern. Scan this pattern against the PROSITE database using the ScanPROSITE tool. Do you identify globin proteins? Are there nonglobin proteins as well?

**[12-4]** InterPro has a BioMart (available at <http://www.ebi.ac.uk/interpro/biomart/martview/>). Use it to find how many human proteins have sequence lengths greater than 20,000 amino acids. Currently, there are five having the following UniProtKB protein accessions, names, and lengths: (1) Q8WXI7, MUC16\_HUMAN, 22,152; (2) D3DPF9\_HUMAN, 26,926; (3) C0JYZ2\_HUMAN, 33,423; (4) Q8WZ42, TITIN\_HUMAN, 34,350; (5) D3DPG0\_HUMAN, 34,942.

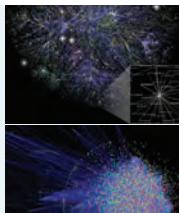
**[12-5]** Salmon has a pinkish color and some lobsters are blue (but turn red when boiled) because a chromophore called astaxanthin binds to a carrier protein called crustacyanin. Examine the protein sequence of crustacyanin from the European lobster *Homarus gammarus*. What are some of its physical properties (e.g., molecular weight, isoelectric point)? Does it have any known domains or

motifs that might explain how or why it binds to the chromophore? Use the tools at the ExPASy site. (For more information about this protein, read the article at ExPASy: [http://www.expasy.org/spotlight/back\\_issues/sptl026.shtml](http://www.expasy.org/spotlight/back_issues/sptl026.shtml).)

**[12-6]** Evaluate human syntaxin at the ExPASy site. Does it have coiled-coil regions? How many predicted transmembrane regions does it have? What is its function?

**[12-7]** Olfactory receptors are related to the rhodopsin-like G-protein coupled receptor (GPCR) superfamily. What percent of the mouse proteome is formed of these receptors? What percent of the human proteome is formed of these receptors?

**[12-8]** Are any of the 15 most common protein domains in *E. coli* K12 also present in humans?



## Self-Test Quiz

**[12-1]** Can a domain be at the amino terminus of one protein and the carboxy terminus of another protein?

- (a) yes; or
- (b) no.

**[12-2]** In general, if you compare the size of a pattern (also called a motif or fingerprint) and a domain:

- (a) they are about the same size;
- (b) the pattern is larger;
- (c) the pattern is smaller; or
- (d) the comparison always depends on the particular proteins in question.

**[12-3]** The amino acid sequence [ST]-X-[RK] is the consensus for phosphorylation of a substrate by protein kinase C. This sequence is an example of:

- (a) a motif that is characteristic of proteins that are homologous to each other;
- (b) a motif that is characteristic of proteins that are not necessarily homologous to each other;
- (c) a domain that is characteristic of proteins that are homologous to each other; or
- (d) a domain that is characteristic of proteins that are not necessarily homologous to each other.

**[12-4]** If you analyze a single, previously uncharacterized protein using programs that predict glycosylation,

sulfation, phosphorylation, or other post-translational modifications:

- (a) the predictions of the programs are not likely to be accurate;
- (b) the accuracy of the predictions is unknown and difficult to assess;
- (c) the predictions of the programs are likely to be accurate concerning the possible presence of particular modifications, but their biological relevance is unknown until you assess the protein's properties experimentally; or
- (d) the predictions of the programs are likely to be accurate concerning the possible presence of particular modifications, but it is not feasible to assess the protein's properties experimentally.

**[12-5]** An underlying assumption of the Gene Ontology Consortium is that the description of a gene or gene product according to three categories (molecular function, biological process, and cellular component):

- (a) is likely to be identical across many species, from plants to worms to human;
- (b) is likely to vary greatly across many species, from plants to worms to human;
- (c) may or may not be identical across many species, and therefore must be assessed for each gene or gene product individually; or

- (d) may or may not be identical across many species and therefore must be assessed for each gene or gene product individually by an expert curator.

**[12-6]** Protein localization is described primarily in which Gene Ontology category?

- (a) molecular function;
- (b) cellular component;
- (c) cellular localization; or
- (d) biological process.

**[12-7]** Which of the following is a means of assessing protein function?

- (a) finding structural homologs;
- (b) studying bait-prey interactions;

- (c) determining the isoelectric point; or
- (d) all of the above.

**[12-8]** A major advantage of two-dimensional protein gels as a high-throughput technology for protein analysis is that:

- (a) sample preparation and the process of running two-dimensional gels is straightforward and can be automated;
- (b) the result of two-dimensional gels includes data on both the size and the charge of thousands of proteins;
- (c) the technique is well suited to the detection of low-abundance proteins; or
- (d) the technique is well suited to the detection of hydrophobic proteins.

## SUGGESTED READING

Reviews on proteomics include Becnel *et al.* (2012), Bruce *et al.* (2013), and Ivanov *et al.* (2013). Bernard Jacq (2001) and Raes *et al.* (2007) have reviewed protein function. Both articles discuss the complexity of protein function and the use of bioinformatic tools to dissect function. Jacq proposes to consider function from six structural levels, from the structure of a protein to its role in a population of organisms. Raes *et al.* include a discussion of the impact of genomics projects on assessing protein function.

Reviews on mass spectrometry in proteomics include Gstaiger and Aebersold (2009), Kumar and Mann (2009), and Washburn (2011). Trost and Kusalik (2011) have written a highly recommended review on computational prediction of phosphorylation sites.

## REFERENCES

- Alderson, R.G., De Ferrari, L., Mavridis, L. *et al.* 2012. Enzyme informatics. *Current Topics in Medicinal Chemistry* **12**(17), 1911–1912. PMID: 23116471.
- Amoresano, A., Carpentieri, A., Giangrande, C. *et al.* 2009. Technical advances in proteomics mass spectrometry: identification of post-translational modifications. *Clinical Chemistry and Laboratory Medicine* **47**(6), 647–665. PMID: 19426139.
- Arnott, D.P., Gawinowicz, M., Grant, R.A. *et al.* 2002. Proteomics in mixtures: study results of ABRF-PRG02. *Journal of Biomolecular Techniques* **13**, 179–186.
- Arnott, D., Gawinowicz, M.A., Grant, R.A. *et al.* 2003. ABRF-PRG03: phosphorylation site determination. *Journal of Biomolecular Techniques* **14**(3), 205–215. PMID: 13678151.
- Artimo, P., Jonnalagedda, M., Arnold, K. *et al.* 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* **40**(Web Server issue), W597–603. PMID: 22661580.
- Ashburner, M., Ball, C.A., Blake, J.A. *et al.* 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25–29. PMID: 10802651.
- Austen, B. M., Westwood, O. M. 1991. *Protein Targeting and Secretion*. IRL Press, Oxford.
- Becnel, L.B., McKenna, N.J. 2012. Minireview: progress and challenges in proteomics data management, sharing, and integration. *Journal of Molecular Endocrinology* **26**(10), 1660–1674. PMID: 22902541.
- Beissbarth, T. 2006. Interpreting experimental results using gene ontologies. *Methods in Enzymology* **411**, 340–352.

- Blom, N., Gammeltoft, S., Brunak, S. 1999. Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology* **294**, 1351–1362.
- Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., Brunak, S. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**(6), 1633–1649. PMID: 15174133.
- Bork, P., Gibson, T. J. 1996. Applying motif and profile searches. *Methods in Enzymology* **266**, 162–184.
- Bork, P., Koonin, E. V. 1996. Protein sequence motifs. *Current Opinion in Structural Biology* **6**, 366–376.
- Bruce, C., Stone, K., Gulcicek, E., Williams, K. 2013. Proteomics and the analysis of proteomic data: 2013 overview of current protein-profiling technologies. *Current Protocols in Bioinformatics Chapter* 13, Unit 13.21. PMID: 23504934.
- Brune, D.C., Hampton, B., Kobayashi, R. et al. 2007. ABRF ESRG 2006 study: Edman sequencing as a method for polypeptide quantitation. *Journal of Biomolecular Techniques* **18**(5), 306–320. PMID: 18166674.
- Carrette, O., Burkhard, P.R., Sanchez, J.C., Hochstrasser, D.F. 2006. State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research. *Nature Protocols* **1**, 812–823. PMID: 17406312.
- Casadio, R., Martelli, P.L., Pierleoni, A. 2008. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings in Functional Genomics and Proteomics* **7**(1), 63–73. PMID: 18283051.
- Choudhary, C., Mann, M. 2010. Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology* **11**(6), 427–439. PMID: 20461098.
- Cooper, T. G. 1977. *The Tools of Biochemistry*. Wiley, New York.
- Copley, R. R., Schultz, J., Ponting, C. P., Bork, P. 1999. Protein families in multicellular organisms. *Current Opinion in Structural Biology* **9**, 408–415.
- Copley, R.R., Doerks, T., Letunic, I., Bork, P. 2002. Protein domain analysis in the era of complete genomes. *FEBS Letters* **513**, 129–134.
- Curreem, S.O., Watt, R.M., Lau, S.K., Woo, P.C. 2012. Two-dimensional gel electrophoresis in bacterial proteomics. *Protein and Cell* **3**(5), 346–363. PMID: 22610887.
- del-Toro, N., Dumousseau, M., Orchard, S. et al. 2013. A new reference implementation of the PSIC-QUIC web service. *Nucleic Acids Research* **41**(Web Server issue), W601–606. PMID: 23671334.
- Derouiche, A., Cousin, C., Mijakovic, I. 2012. Protein phosphorylation from the perspective of systems biology. *Current Opinion in Biotechnology* **23**(4), 585–590. PMID: 22119098.
- Dissmeyer, N., Schnittger, A. 2011. The age of protein kinases. *Methods in Molecular Biology* **779**, 7–52. PMID: 21837559.
- Doerr, A. 2013. Mass spectrometry-based targeted proteomics. *Nature Methods* **10**(1), 23. PMID: 23547294.
- Doolittle, R. F. 1995. The multiplicity of domains in proteins. *Annual Reviews of Biochemistry* **64**, 287–314.
- Edman, P. 1949. A method for the determination of amino acid sequence in peptides. *Arch. Biochem.* **22**(3), 475. PMID: 18134557.
- Eisenhaber, B., Eisenhaber, F. 2010. Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods in Molecular Biology* **609**, 365–384. PMID: 20221930.
- Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**, 953–971. PMID: 17446895.
- Emes, R.D. 2008. Inferring function from homology. *Methods in Molecular Biology* **453**, 149–168. PMID: 18712301.
- Falick, A.M., Lane, W.S., Lilley, K.S. et al. 2011. ABRF-PRG07: advanced quantitative proteomics study. *Journal of Biomolecular Techniques* **22**(1), 21–26. PMID: 21455478.
- Flicek, P., Amode, M.R., Barrell, D. et al. 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.

- Frankel, A. D., Young, J. A. 1998. HIV-1: Fifteen proteins and an RNA. *Annual Reviews of Biochemistry* **67**, 1–25.
- Gene Ontology Consortium. 2001. Creating the gene ontology resource: Design and implementation. *Genome Research* **11**, 1425–1433.
- Gene Ontology Consortium. 2010. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Research* **38**(Database issue), D331–335. PMID: 19920128.
- Geppert, M., Goda, Y., Stevens, C. F., Sudhof, T. C. 1997. The small GTP-binding protein Rab3A regulates a late step in synaptic vesicle fusion. *Nature* **387**, 810–814.
- Goel, R., Muthusamy, B., Pandey, A., Prasad, T.S. 2011. Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology. *Molecular Biotechnology* **48**(1), 87–95. PMID: 20927658.
- Gonzalez-Galarza, F.F., Qi, D., Fan, J., Bessant, C., Jones, A.R. 2014. A tutorial for software development in quantitative proteomics using PSI standard formats. *Biochimica et Biophysica Acta* **1844**(1 Pt A), 88–97. PMID: 23584085.
- Görg, A., Weiss, W., Dunn, M.J. 2004. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4**, 3665–3685.
- Gstaiger, M., Aebersold, R. 2009. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics* **10**(9), 617–627. PMID: 19687803.
- Gribskov, M., Veretnik, S. 1996. Identification of sequence pattern with profile analysis. *Methods in Enzymology* **266**, 198–212. PMID: 8743686.
- Gribskov, M., McLachlan, A.D., Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Science, USA* **84**(13), 4355–4358. PMID: 3474607.
- Henikoff, S., Greene, E.A., Pietrovski, S. *et al.* 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614. PMID: 9381171.
- Higgins, D.G., Thompson, J.D., Gibson, T.J. 1996. Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology* **266**, 383–402. PMID: 8743695.
- Hoogland, C., Mostaguir, K., Sanchez, J.C., Hochstrasser, D.F., Appel, R.D. 2004. SWISS-2DPAGE, ten years later. *Proteomics* **4**(8):2352–2356. PMID: 15274128.
- Horton, P., Park, K.J., Obayashi, T. *et al.* 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **35**(Web Server issue), W585–587. PMID: 17517783.
- Hunter, S., Jones, P., Mitchell, A. *et al.* 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* **40**(Database issue), D306–312. PMID: 22096229.
- Iimure, T., Kihara, M., Sato, K. 2014. Beer and wort proteomics. *Methods in Molecular Biology* **1072**, 737–754. PMID: 24136560.
- Imai, K., Nakai, K. 2010. Prediction of subcellular locations of proteins: where to proceed? *Proteomics* **10**(22), 3970–3983. PMID: 21080490.
- Ivanov, A.R., Colangelo, C.M., Dufresne, C.P. *et al.* 2013. Interlaboratory studies and initiatives developing standards for proteomics. *Proteomics* **13**(6), 904–909. PMID: 23319436.
- Jacq, B. 2001. Protein function from the perspective of molecular interactions and genetic networks. *Briefings in Bioinformatics* **2**, 38–50. PMID: 11465061.
- Jones, D.T. 2001. Protein structure prediction in genomics. *Briefings in Bioinformatics* **2**, 111–125.
- Käll, L., Krogh, A., Sonnhammer, E.L. 2007. Advantages of combined transmembrane topology and signal peptide prediction: the Phobius web server. *Nucleic Acids Research* **35**, W429–432.
- Kalume, D.E., Molina, H., Pandey, A. 2003. Tackling the phosphoproteome: tools and strategies. *Current Opinion in Chemical Biology* **7**, 64–69.
- Kristensen, D.M., Kannan, L., Coleman, M.K. *et al.* 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26**(12), 1481–1487. PMID: 20439257.
- Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**, 567–580.

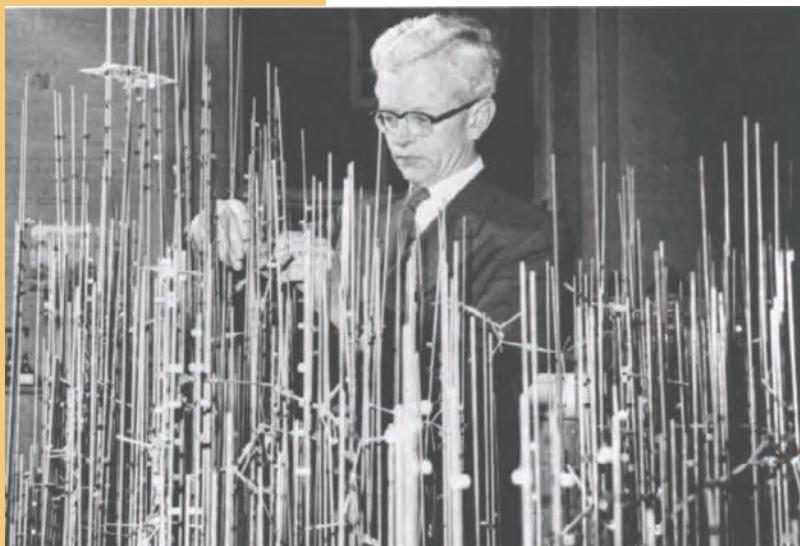
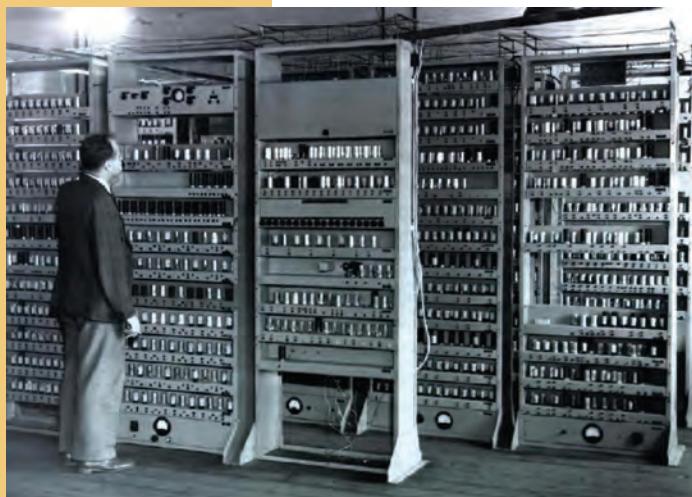
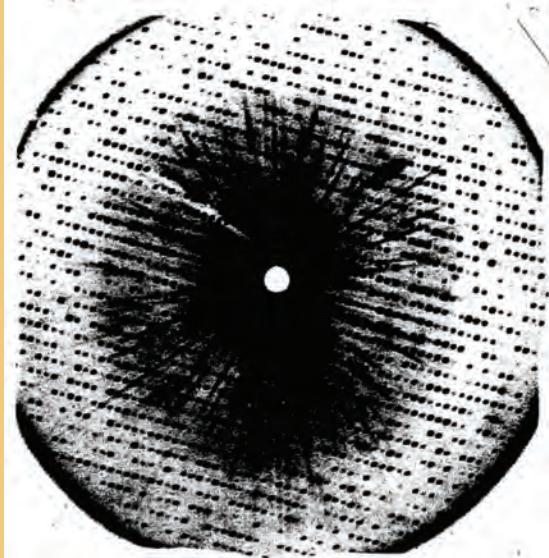
- Kumar, C., Mann, M. 2009. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Letters* **583**(11), 1703–1712. PMID: 19306877.
- Kyte, J., Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**(1), 105–132. PMID: 7108955.
- Langen, H., Berndt, P., Röder, D. *et al.* 1999. Two-dimensional map of human brain proteins. *Electrophoresis* **20**, 907–916. PMID: 10344266.
- Lee, D., Redfern, O., Orengo, C. 2007. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* **8**(12), 995–1005. PMID: 18037900.
- Lomize, A.L., Lomize, A.L., Pogozheva, I.D., Mosberg, H.I. 2011. Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *Journal of Chemical Information and Modeling* **51**(4), 930–946. PMID: 21438606.
- Lupas, A. 1997. Predicting coiled-coil regions in proteins. *Current Opinion in Structural Biology* **7**, 388–393.
- Lupas, A., Van Dyke, M., Stock, J. 1991. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164.
- Lüthy, R., Xenarios, I., Bucher, P. 1994. Improving the sensitivity of the sequence profile method. *Protein Science* **3**(1), 139–146. PMID: 7511453.
- Mann, G. 1906. *The Chemistry of the Proteids*. The Macmillan Company, New York.
- Marcotte, E.M. 2007. How do shotgun proteomics algorithms identify proteins? *Nature Biotechnology* **25**, 755–757.
- Malik, R., Dulla, K., Nigg, E.A., Körner, R. 2010. From proteome lists to biological impact—tools and strategies for the analysis of large MS data sets. *Proteomics* **10**(6), 1270–1283. PMID: 20077408.
- Martens, L., Orchard, S., Apweiler, R., Hermjakob, H. 2007. Human proteome organization proteomics standards initiative: data standardization, a view on developments and policy. *Molecular and Cellular Proteomics* **6**, 1666–1667.
- Martínez-Bartolomé, S., Deutsch, E.W., Binz, P.A. *et al.* 2013. Guidelines for reporting quantitative mass spectrometry based experiments in proteomics. *Journal of Proteomics* **95**, 84–88. PMID: 23500130.
- Martínez-Bartolomé, S., Binz, P.A., Albar, J.P. 2014. The Minimal Information About a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. *Methods in Molecular Biology* **1072**, 765–780. PMID: 24136562.
- Martins-de-Souza, D., Guest, P.C., Guest, F.L. *et al.* 2012. Characterization of the human primary visual cortex and cerebellum proteomes using shotgun mass spectrometry–data–independent analyses. *Proteomics* **12**(3), 500–504. PMID: 22162416.
- Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D. *et al.* 2013. The HUPO proteomics standards initiative: mass spectrometry controlled vocabulary. *Database (Oxford)* **2013**, bat009. PubMed PMID: 23482073.
- Mazumder, R., Vasudevan, S., Nikolskaya, A.N. 2008. Protein functional annotation by homology. *Methods in Molecular Biology* **484**, 465–490. PMID: 18592196.
- Miller, M.L., Blom, N. 2009. Kinase-specific prediction of protein phosphorylation sites. *Methods in Molecular Biology* **527**, 299–310. PMID: 19241022.
- Minden, J.S. 2012. DIGE: past and future. *Methods in Molecular Biology* **854**, 3–8. PMID: 22311749.
- Mishra, G.R., Suresh, M., Kumaran, K. *et al.* 2006. Human protein reference database: 2006 update. *Nucleic Acids Research* **34**(Database issue), D411–414.
- Muthusamy, B., Thomas, J.K., Prasad, T.S., Pandey, A. 2013. Access guide to human proteinpedia. *Current Protocols in Bioinformatics Chapter 1*, Unit 1.21. PMID: 23504933.
- Nakazawa, T., Yamaguchi, M., Okamura, T.A. *et al.* 2008. Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics* **8**(4), 673–685. PMID: 18214847.
- Nugent, T., Jones, D.T. 2012. Membrane protein structural bioinformatics. *Journal of Structural Biology* **179**(3), 327–337. PMID: 22075226.
- O'Farrell, P. H. 1975. High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry* **250**, 4007–4021.

- Orchard, S. 2012. Molecular interaction databases. *Proteomics* **12**(10), 1656–1662. PMID: 22611057.
- Orchard, S. 2014. Data standardization and sharing: the work of the HUPO-PSI. *Biochimica et Biophysica Acta* **1844**(1 Pt A), 82–87. PMID: 23524294.
- Orchard, S., Hermjakob, H. 2011. Data standardization by the HUPO-PSI: how has the community benefitted? *Methods in Molecular Biology* **696**, 149–160. PMID: 21063946.
- Orchard, S., Binz, P.A., Borchers, C. *et al.* 2012. Ten years of standardizing proteomic data: a report on the HUPO-PSI Spring Workshop: April 12–14th, 2012, San Diego, USA. *Proteomics* **12**(18), 2767–2772. PMID: 22969026.
- Panfoli, I., Calzia, D., Santucci, L. *et al.* 2012. A blue dive: from ‘blue fingers’ to ‘blue silver’. A comparative overview of staining methods for in-gel proteomics. *Expert Review of Proteomics* **9**(6), 627–634. PMID: 23256673.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**(18), 3551–3567. PMID: 10612281.
- Pevsner, J., Hou, V., Snowman, A. M., Snyder, S. H. 1990. Odorant-binding protein. Characterization of ligand binding. *Journal of Biological Chemistry* **265**, 6118–6125. PMID: 2318850
- Ponting, C. P. 2001. Issues in predicting protein function from sequence. *Briefings in Bioinformatics* **2**, 19–29. PMID: 11465059.
- Poole, R.K., Hughes, M.N. 2000. New functions for the ancient globin family: bacterial responses to nitric oxide and nitrosative stress. *Molecular Microbiology* **36**, 775–783.
- Powell, S., Szklarczyk, D., Trachana, K. *et al.* 2012. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Research* **40**(Database issue), D284–289. PMID: 22096231.
- Ptacek, J., Snyder, M. 2006. Charging it up: global analysis of protein phosphorylation. *Trends Genetics* **22**, 545–554.
- Ptacek, J., Devgan, G., Michaud, G. *et al.* 2005. Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679–684. PMID: 16319894.
- Punta, M., Forrest, L.R., Bigelow, H. *et al.* 2007. Membrane protein prediction methods. *Methods* **41**(4), 460–474. PMID: 17367718.
- Raes, J., Harrington, E.D., Singh, A.H., Bork, P. 2007. Protein function space: viewing the limits or limited by our view? *Current Opinion in Structural Biology* **17**, 362–369.
- Ratnam, M., Nguyen, D. L., Rivier, J., Sargent, P. B., Lindstrom, J. 1986. Transmembrane topography of nicotinic acetylcholine receptor: Immunochemical tests contradict theoretical predictions based on hydrophobicity profiles. *Biochemistry* **25**, 2633–2643.
- Righetti, P.G. 2013. Bioanalysis: Heri, hodie, cras. *Electrophoresis* **34**(11), 1442–1451. PMID: 23417314.
- Roepstorff, P. 2012. Mass spectrometry based proteomics, background, status and future needs. *Protein and Cell* **3**(9), 641–647. PMID: 22926765.
- Sabidó, E., Selevsek, N., Aebersold, R. 2012. Mass spectrometry-based proteomics for systems biology. *Current Opinion in Biotechnology* **23**(4), 591–597. PMID: 22169889.
- Sanger, F., Tuppy, H. 1951. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochemistry Journal* **49**(4), 463–481. PMID: 14886310.
- Shively, J.E. 2000. The chemistry of protein sequence analysis. *EXS* **88**, 99–117.
- Sigrist, C. J. *et al.* 2002. PROSITE: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics* **3**, 265–274.
- Sigrist, C.J., de Castro, E., Cerutti, L. *et al.* 2013. New and continuing developments at PROSITE. *Nucleic Acids Research* **41**(Database issue), D344–347. PMID: 23161676.
- Sonnhammer, E. L., Kahn, D. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Science* **3**, 482–492.

- Stroud, R. M., Walter, P. 1999. Signal sequence recognition and protein targeting. *Current Opinion in Structural Biology* **9**, 754–759.
- Takai, Y., Sasaki, T., Matozaki, T. 2001. Small GTP-binding proteins. *Physiology Review* **81**, 153–208.
- Tanford, C. 1980. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*. John Wiley & Sons, New York.
- Tatusov, R.L., Koonin, E.V., Lipman, D.J. 1997. A genomic perspective on protein families. *Science* **278**, 631–637.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D. et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41. PMID: 12969510.
- Taylor, C.F., Paton, N.W., Lilley, K.S. et al. 2007. The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology* **25**, 887–893. PMID: 17687369.
- Temporini, C., Calleri, E., Massolini, G., Caccialanza, G. 2008. Integrated analytical strategies for the study of phosphorylation and glycosylation in proteins. *Mass Spectrometry Review* **27**(3), 207–236. PMID: 18335498.
- Thelen, J.J., Miernyk, J.A. 2012. The proteomic future: where mass spectrometry should be taking us. *Biochemistry Journal* **444**(2), 169–181. PMID: 22574775.
- Thomas, P.D., Mi, H., Lewis, S. 2007. Ontology annotation: mapping genomic regions to biological function. *Current Opinion in Chemical Biology* **11**, 4–11.
- Thompson, J.D., Higgins, D.G., Gibson, T.J. 1994a. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Applications in the Biosciences* **10**(1), 19–29. PMID: 8193951.
- Thompson, J.D., Higgins, D.G., Gibson, T.J. 1994b. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**(22), 4673–4680. PMID: 7984417.
- Trost, B., Kusalik, A. 2011. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* **27**(21), 2927–2935. PMID: 21926126.
- Turewicz, M., Deutsch, E.W. 2011. Spectra, chromatograms, Metadata: mzML—the standard data format for mass spectrometer output. *Methods in Molecular Biology* **696**, 179–203. PMID: 21063948.
- Tusnády, G.E., Simon, I. 2010. Topology prediction of helical transmembrane proteins: how far have we reached? *Current Protein and Peptide Science* **11**(7), 550–561. PMID: 20887261.
- UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* **41**(Database issue), D43–47. PMID: 23161681.
- Venter, J.C. et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74.
- Viswanathan, S., Unlü, M., Minden, J.S. 2006. Two-dimensional difference gel electrophoresis. *Nature Protocols* **1**, 1351–1358. PMID: 17406422.
- Vizcaíno, J.A., Côté, R.G., Csordas, A. et al. 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research* **41**(Database issue), D1063–1069. PMID: 23203882.
- von Heijne, G. 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology* **225**(2), 487–494. PMID: 1593632.
- Walsh, C.T. 2006. *Posttranslational Modification of Proteins: Expanding Nature's Inventory*. Roberts and Company, Englewood, CO.
- Walsh, G.M., Rogalski, J.C., Klockenbusch, C., Kast, J. 2010. Mass spectrometry-based proteomics in biomedical research: emerging technologies and future strategies. *Expert Reviews in Molecular Medicine* **12**, e30. PMID: 20860882.
- Washburn, M.P. 2011. Driving biochemical discovery with quantitative proteomics. *Trends in Biochemical Science* **36**(3), 170–177. PMID: 20880711.
- Westerbrink, H.G.K. 1966. Biochemistry in Holland. *Clio Medica* **1**(2), 153.

- Whetzel, P.L., Parkinson, H., Stoeckert, C.J. Jr. 2006. Using ontologies to annotate microarray experiments. *Methods in Enzymology* **411**, 325–339.
- Wong, W.C., Maurer-Stroh, S., Eisenhaber, F. 2010. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput. Biol.* **6**(7), e1000867. PMID: 20686689.
- Yooseph, S. *et al.* 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology* **5**, e16.
- Zanzoni, A., Carbajo, D., Diella, F. *et al.* 2011. Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Research* **39**(Database issue), D268–271. PMID: 20965970.





Beginning in the 1940s, Max Perutz and John Kendrew realized the goal of determining the structure of globular proteins by solving the structure of myoglobin and hemoglobin. In recognition of this work, they shared the Nobel Prize in Chemistry in 1962. Top: X-ray precession photograph of a myoglobin crystal (from <http://www.nobel.se/chemistry/laureates/1962/kendrew-lecture.pdf>, WebLink 13.1). Kendrew studied myoglobin from the sperm whale (*Physeter catodon*), and incorporated a heavy metal by the method of isomorphous replacement. He could then bombard the crystals with X-rays in order to obtain an X-ray diffraction pattern (such as that shown here) with which to deduce the electron density throughout the crystal. This required the analysis of 25,000 reflections. Middle: Perutz and Kendrew used the EDSAC I computer (introduced in 1949, from <http://www.cl.cam.ac.uk/relics/jpegs/edsac99.36.jpg>, WebLink 13.2). This computer was essential to interpret the diffraction patterns. For a simulator that shows the capacity of the EDSAC machine, see <http://www.dcs.warwick.ac.uk/~edsac/> (WebLink 13.3). Bottom: photograph by Max Perutz of John Kendrew with his model of myoglobin in 1959.

Source: (Top) Kendrew, 1962. (Middle) Computer Laboratory, University of Cambridge. (Bottom) MRC Laboratory of Molecular Biology, 1959. Reproduced with permission from MRC Laboratory of Molecular Biology.

# Protein Structure

# CHAPTER 13

*A visitor to the Accademia in Florence can see magnificent images that emerged from blocks of marble at the hands of Michelangelo. By analogy, the noncrystallographer can capture the vision that a crystallographer has when admiring a rigorously shaped crystal before exploring the marvelous structure hidden within. So the Protein Data Bank is our museum, with models of molecules reflecting the wonders of nature and complex shapes that may be as old as life itself. With the aid of interactive graphics and networking, the PDB makes these images readily available. What wonders still remain hidden as we build, compare, and extend our database?*

*Edgar F. Meyer (1997)*

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

- understand the principles of protein primary, secondary, tertiary, and quaternary structure;
- use the NCBI tool CN3D to view a protein structure;
- use the NCBI tool VAST to align two structures;
- explain the role of PDB including its purpose, contents, and tools;
- explain the role of structure annotation databases such as SCOP and CATH; and
- describe approaches to modeling the three-dimensional structure of proteins.

## OVERVIEW OF PROTEIN STRUCTURE

Proteins adopt a spectacular range of conformations and interact with their cellular milieu in diverse ways. There are three major classes of proteins: structural proteins (such as tubulin and actin), membrane proteins (such as photoreceptors and ion channels), and globular proteins (such as globins).

The three-dimensional structure of a protein determines its capacity to function. This structure is determined from its primary (linear) amino acid sequence. In the 1950s Christian Anfinsen and others performed a remarkable set of experiments. They purified the enzyme ribonuclease from bovine pancreas, and denatured it with urea. This enzyme includes eight sulphydryl groups that form four disulfide bonds. After removing the urea, the ribonuclease refolded and adopted a conformation that was indistinguishable from native ribonuclease. Anfinsen stated the thermodynamic hypothesis that the three-dimensional structure of a native protein under physiological conditions is the one

Christian Anfinsen won part of the 1972 Nobel Prize in Chemistry "for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation" ([http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1972/](http://nobelprize.org/nobel_prizes/chemistry/laureates/1972/), WebLink 13.4).

An angstrom (abbreviated Å) is  $0.1\text{ nm}$  or  $10^{-10}\text{ m}$ ; a carbon–carbon bond has a distance of about  $1.5\text{ \AA}$ . John Kendrew and Max Perutz shared the 1962 Nobel Prize in Chemistry "for their studies of the structures of globular proteins;" see [http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1962/](http://nobelprize.org/nobel_prizes/chemistry/laureates/1962/) (WebLink 13.5).

It is difficult to make a pairwise alignment of rat retinol-binding protein (P04916) and rat odorant-binding protein (NP\_620258). If you use BLASTP no significant match is found, even using a large expect value and a scoring matrix appropriate for distantly related proteins (PAM250). If you perform a DELTA-BLAST search with rat OBP as a query, you will eventually detect retinol-binding protein after many iterations. We compare the three-dimensional structures of these two proteins in computer lab problem (13.4) in this chapter using the DaliLite server, and see evidence that they are homologous.

in which the Gibbs free energy of the system is lowest (Anfinsen, 1973). We can picture an energy landscape in which many conformations are possible, and proteins tend to adopt the structure(s) that minimize the free energy. Anfinsen's work helped to solidify the concept that the three-dimensional structure of a protein is inherently specified by the linear amino acid sequence.

In the 1950s researchers applying the techniques of X-ray crystallography to proteins focused on the structures of hemoglobin, myoglobin, ribonuclease, and insulin. By 1957 John Kendrew and colleagues reported the three-dimensional structure of myoglobin to  $6\text{ \AA}$  resolution, sufficient to reconstruct the main outline of the protein. Soon after, the resolution was improved to  $2\text{ \AA}$ . For the first time, all the atoms comprising a protein could be spatially described and the structural basis of the function of a protein – here, myoglobin as an oxygen carrier – was elucidated. Today the central repository of protein structures, the Protein Data Bank, contains over 100,000 structures (see "Protein Data Bank" below).

In this chapter we consider the structure of individual proteins from the principles of primary, secondary, tertiary, and quaternary structure. We also consider structural genomics initiatives in which a very broad range of high-resolution tertiary structures are determined for proteins, spanning organisms across the tree of life and also spanning the set of all possible conformations that protein structures can adopt. We introduce the main repository of protein structures, the Protein Data Bank (PDB), as well as three software tools to visualize structures: WebMol at PDB, Cn3D at NCBI, and DeepView at ExPASy. Many databases provide analyses of structural data and we describe three prominent databases: CATH, SCOP, and the Dali Domain Dictionary. Finally, we discuss protein structure prediction which underlies the newly emerging field of structural genomics.

## Protein Sequence and Structure

As described in Chapter 12, one of the most fundamental questions about a protein is its function. Function is often assigned based upon homology to another protein whose function is perhaps already known or inferred (Holm, 1998; Domingues *et al.*, 2000). Two proteins that share a similar structure are usually assumed to also share a similar function. For example, two receptor proteins may share a very similar structure; even if they differ in their ability to bind ligands or transduce signals, they still share the same basic function.

Various types of BLAST searching are employed to identify such relationships of homology (Chapters 4 and 5). However, for many proteins sequence identity is extremely limited. We may take retinol-binding protein and odorant-binding protein as examples: these are both lipocalins of about 20 kDa and are abundant, secreted carrier proteins. They share a GXW motif that is characteristic of lipocalins. However, it is difficult to detect homology based upon analysis of the primary amino acid sequences. By pairwise alignment, the two proteins share less than 20% identity. Both structure and function are preserved over evolutionary time more than sequence identity. The three-dimensional structures of these proteins are therefore extraordinarily similar. We have seen similar relationships for myoglobin relative to alpha globin and beta globin (Fig. 3.1).

Can we generalize the relationship between amino acid sequence identity and protein structures? It is clear that even a single amino acid substitution can in some instances cause a dramatic change in protein structure, as exemplified by disease-causing mutations (discussed at the end of this chapter in "Protein Structure and Disease"). Many other substitutions have no observable effects on protein structure (discussed in Anfinsen, 1973). It is common for amino acid sequence to change more rapidly than three-dimensional structure, as in the case of lipocalins.

## Biological Questions Addressed by Structural Biology: Globins

We can use the globins to illustrate some of the key questions in structural biology:

- What ligand does each protein transport? For many the answer is unknown. Can structural studies reveal the binding domain to suggest the identity of the ligand? How much structural information is required in order to predict the ligand from sequence information?
- Mutations in globin genes result in a variety of human diseases, including thalassemias and sickle cell anemia (Chapter 21). Can we predict the structural and functional consequences of a specific mutation?
- Globins have been divided into subgroups based upon phylogenetic analyses and based upon their localization. To what extent do those groupings reflect structural and functional similarities?
- When a genome is sequenced and a gene encoding a putative novel globin is discovered, can we use information about other globins of known structure in order to predict a new structure?

## PRINCIPLES OF PROTEIN STRUCTURE

Protein structure is defined at several levels. Primary structure refers to the linear sequence of amino acid residues in a polypeptide chain, such as human beta globin (**Fig. 13.1a**). Secondary structure refers to the arrangements of the primary amino acid sequence into motifs such as  $\alpha$  helices,  $\beta$  sheets, and coils (or loops; **Fig. 13.1b**). The tertiary structure is the three-dimensional arrangement formed by packing secondary structure elements into globular domains (**Fig. 13.1c**). Finally, quaternary structure involves this arrangement of several polypeptide chains. **Figure 13.1d** depicts two alpha globin chains and two beta globin chains joined to form mature hemoglobin, with four heme groups attached. Functionally important areas of a protein such as ligand-binding sites or enzymatic active sites are formed at the levels of tertiary and quaternary structure. We describe these levels of protein structure in the following sections, using myoglobin and hemoglobin as examples.

### Primary Structure

In nature, the primary amino acid sequence specifies a three-dimensional structure that forms for each protein. A protein folds to form its native structure(s), sometimes including the participation of chaperones. This process is rapid, typically taking from seconds to minutes; consider for example the bacterium *Escherichia coli* that can double every 20 minutes, requiring all its thousands of proteins to be functionally expressed within that time constraint. Formation of the native structure(s) may depend on some post-translational modifications, such as the addition of sugars or disulfide bridges. The central issue, called the protein-folding problem, is that each cell interprets the information in a primary amino acid sequence to form an appropriate structure. Challenges to structural biologists include: (1) how to understand the biological process of protein folding; and (2) how to predict a three-dimensional structure based on primary sequence data alone.

Proteins are synthesized from ribosomes where amino acids are joined by peptide bonds into a polypeptide chain. Each amino acid consists of an amino group, a central carbon atom  $C\alpha$  to which a side chain R is attached, and a carboxyl group (**Fig. 13.2a**). The peptide bond is a carbon-nitrogen amide linkage between the carboxyl group of one amino acid and the amino group of the next amino acid. One water molecule is eliminated during the formation of a peptide bond. The basic repeating unit of a polypeptide chain is therefore  $NH-C\alpha H-CO$  with a different R group extending from various  $C\alpha$  of various amino acids. In glycine, the R group is a hydrogen and that amino acid is therefore not chiral. For the other amino acids the R group is not a hydrogen; there are therefore four different moieties attached to  $C\alpha$ , allowing chiral (L- and D-) forms of most amino acids.

We discuss intrinsically disordered proteins towards the end of this chapter (“Intrinsically Disordered Proteins”); they do not adopt a unique native structure.

Remarkably, the discovery of the peptide bond was announced at a meeting on the same day (22 September 1902) by two researchers: Franz Hofmeister and Emil Fisher. Fisher won a 1902 Nobel Prize “in recognition of the extraordinary services he has rendered by his work on sugar and purine syntheses” ([http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1902/](http://nobelprize.org/nobel_prizes/chemistry/laureates/1902/), WebLink 13.6). In the area of protein research, he discovered proline and oxyproline, synthesized peptides up to eight amino acids in length, and devised new methods of compositional analysis of proteins such as casein.

(a) Primary structure

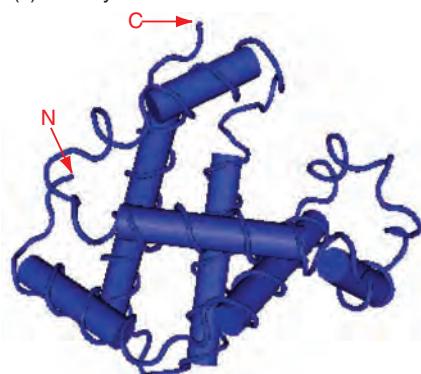
MVHLTPEEKSAVTALWGKVNVDVGGEALGRLLVVYPWTQRFFESFGDLSPTDAVMGNPKVKAHGKKVLGAFSD  
GLAHLNDLNKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHFGKEFTPVOAYAQOKVUVAGVANALAHKYH

(b) Secondary structure

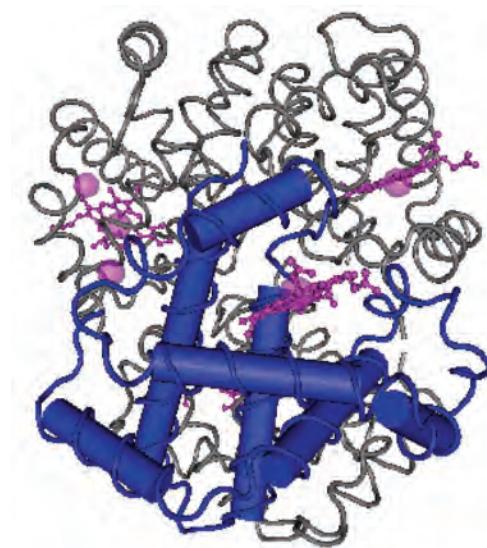
	10	20	30	40	50	60	70
UNK_257900	MVHL TPEEK SAVT ALWG KVNV D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S T P D A V M G N P K V K A H G K K V L G						
DSC	cccc hhhhhh hhhh ccccccc hhhhhh hhhh ccccccc hhhhhh hhhh hhhh hhhh hhhh						
MLRC	cccccc hhhhhh hhhh cccccccc hhhh hhhh eeeccc hhhh cccccccc hhhh hhhh hhhh						
PHD	cccccc hhhhhh hhhh cccc hhcc hhhh hhhh eeeccc hhhh hhhh cccc hhhh ec hhhh hhhh hhhh						
Sec. Cons.	cccccc hhhhhh hhhh ccccccc hcc hhhh hhhh eeeccc hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh						
	80	90	100	110	120	130	140
UNK_257900	A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D P E N F R L L G N V L V C V L A H H F G K E F T P V Q A A Y Q K V V A G V A N						
DSC	hhhhhhh hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh						
MLRC	hhhhhhh hhhh hhhh hhhh hhhh hhhh cccccccc hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh						
PHD	hhhhhhh hhhh hhhh hhhh hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh						
Sec. Cons.	hhhhhhh hhhh hhhh hhhh hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh ccccccc hhhh hhhh hhhh hhhh						

UNK\_257900 ALAHKYH  
DSC hhhhccc  
MLRC hhhhccc  
PHD hhhhhcc  
Sec.Cons. hhhhccc

(c) Tertiary structure



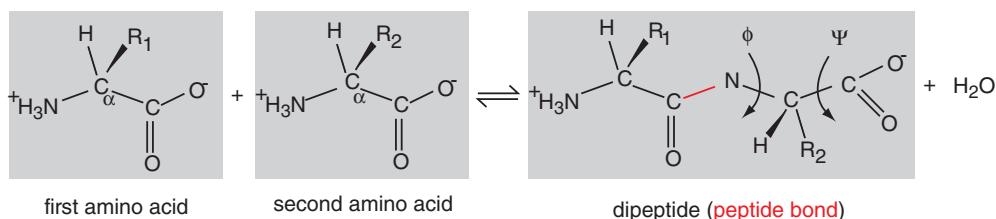
(d) Quaternary structure



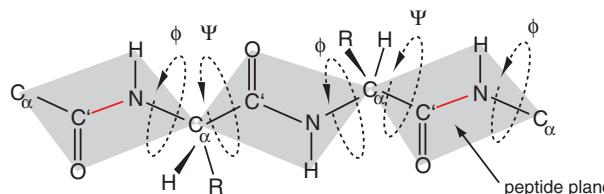
**FIGURE 13.1** A hierarchy of protein structure. (a) The primary structure of a protein refers to the linear polypeptide chain of amino acids. Here, human beta globin is shown (NP\_000539). (b) The secondary structure includes elements such as alpha helices and beta sheets. Here, beta globin protein sequence was input to the POLE server for secondary structure where three prediction algorithms were run and a consensus was produced. h: alpha helix; c: random coil; e: extended strand. (c) The tertiary structure is the three-dimensional structure of the protein chain. Alpha helices are represented as thickened cylinders. Arrows labeled N and C point to the amino and carboxy termini, respectively. (d) The quaternary structure includes the interactions of the protein with other subunits and heteroatoms. Here, the four subunits of hemoglobin are shown (with an  $\alpha 2\beta 2$  composition and one beta globin chain highlighted) as well as four noncovalently attached heme groups.

Source: (b) Produced using PBIL software, <http://pbil.univ-lyon1.fr/>. Courtesy of IBCP-FR 3302. (c, d) Produced using Cn3D software from NCBI.

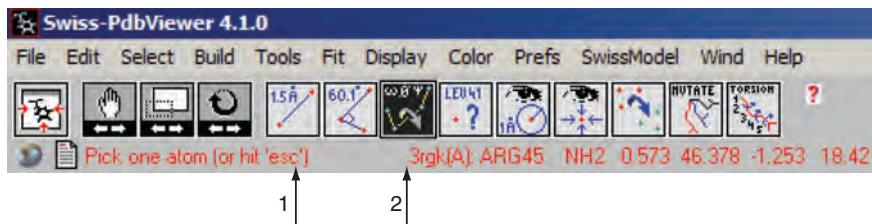
(a) peptide bond



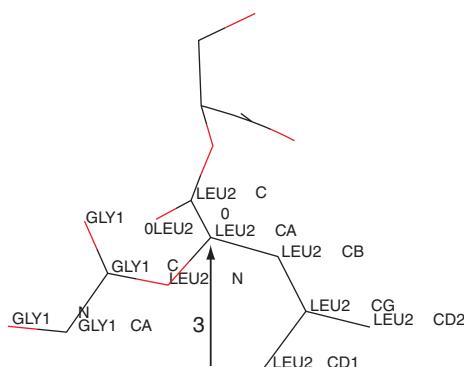
(b) phi and psi angles of polypeptide



(c) DeepView control bar



(d) DeepView viewer



**FIGURE 13.2** Peptide bonds and angles. (a) Each amino acid includes an amino group, an alpha carbon ( $C\alpha$ ) from which a side group R is attached, and a carboxyl group (having carbon  $C'$ ). Two amino acids condense to form a dipeptide with the elimination of water. The peptide bond (highlighted in red) is an amide linkage. (b) Polypeptide chains can be thought of as extending from one  $C\alpha$  atom to the next with the peptide bond constrained to lie along a plain. The  $N$ - $C\alpha$  bond is called phi ( $\phi$ ), and the  $C\alpha$ - $C'$  bond is called psi ( $\psi$ ). The angle of rotation around  $\phi$  and  $\psi$  for each peptide defines the entire main chain conformation. (c) The DeepView software from ExPASy (called Swiss-PdbViewer) includes a control bar with buttons for manipulating a molecule (translation, rotation, and zoom). There are additional tools that measure the following (from left to right): distance between two atoms (arrow 1); angle between three atoms; dihedral angles (arrow 2; here this tool has been selected and the  $\phi$ ,  $\psi$ , and  $\omega$  values are shown); select groups a certain distance from an atom; center the molecule on one atom; fit one molecule onto another; mutation tool; torsion tool. (d) Myoglobin (3RGK) was loaded into DeepView and, using the Control Panel, the first three amino acid residues (Gly-Leu-Ser) were selected. The nitrogens are indicated in bright red, the oxygens in pale red, and  $C\alpha$  carbons (CA) and  $C'$  carbons are indicated. By selecting the dihedral angle tool and clicking the leucine  $C\alpha$  carbon (arrow 1), the bond values in (b) were shown.

Source: (c, d) DeepView software from ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

The amino acid residues of the backbone of the polypeptide chain are constrained to the surface of a plane, and only have mobility around a restricted set of bond angles (**Fig. 13.2.b**; reviewed by Branden and Tooze, 1991; Shulz and Schirmer, 1979). Phi ( $\phi$ ) is the angle around the N-C $\alpha$  bond, and psi ( $\psi$ ) is the angle around the C $\alpha$ -C' bond. Glycine is an exceptional amino acid because it has the flexibility to occur at  $\phi\psi$  combinations that are not tolerated for other amino acids. For most amino acids the  $\phi$  and  $\psi$  angles are constrained to allowable regions in which there is a high propensity for particular secondary structures to form.

We describe how to obtain DeepView in computer lab exercise (13.3) at the end of this chapter. The PDB file for a human myoglobin, 3RGK, is available as Web Document 13.1 at <http://www.bioinfbook.org/chapter13>. SwissModel is available at <http://swissmodel.expasy.org/> (WebLink 13.7).

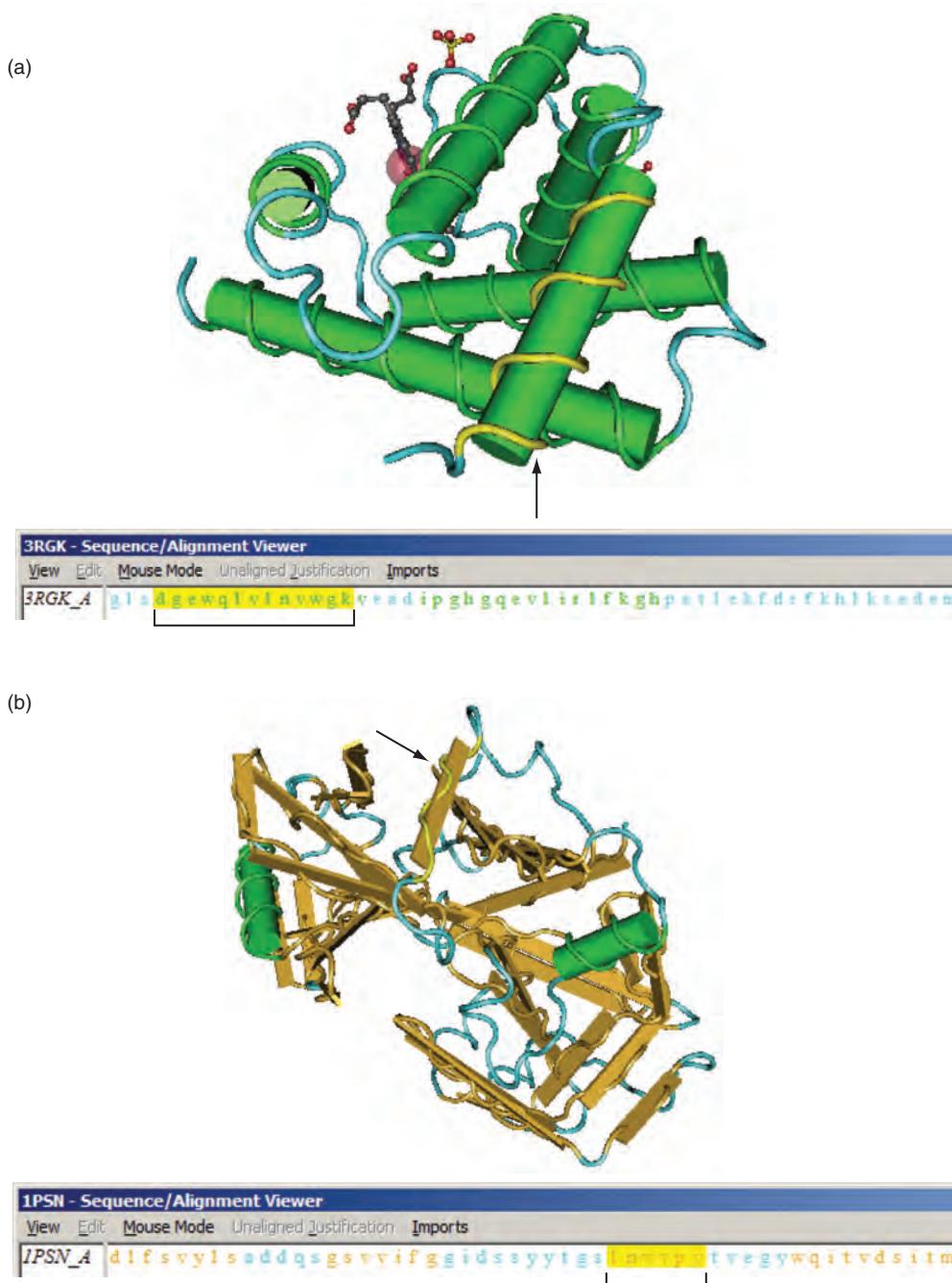
DeepView is a popular software program used to visualize protein structures and to analyze many features of one or more protein structures. It is also used in conjunction with SwissModel, an automated comparative modeling server. DeepView is available for download from the ExPASy website (Chapter 12). When we upload a file in the PDB (Protein Data Bank) format for myoglobin, we can view a control bar with assorted options for manipulating and analyzing the structure (**Fig. 13.2c**). Using the control panel of DeepView we can select just the first two amino acids of myoglobin (gly-leu) and obtain a description of the bond angles (**Fig. 13.2c, d**). One reason why it is useful to inspect these bond angles is that they provide information about the secondary structure of a protein, which we describe in the following section.

## Secondary Structure

Proteins tend to be arranged with hydrophobic amino acids in the interior and hydrophilic residues exposed to the surface. This hydrophobic core is produced in spite of the highly polar nature of the peptide backbone of a protein. The most common way that a protein solves this problem is to organize the interior amino residues into secondary structures consisting of  $\alpha$  helices and pleated  $\beta$  sheets. Linus Pauling and Robert Corey (1951) described these structures from studies of hemoglobin, keratins, and other peptides and proteins. Their models were later confirmed by X-ray crystallography. These secondary structures consist of patterns of interacting amino acid residues in which main chain amino (NH) and carboxy (C'O) groups form hydrogen bonds. There are three types of helices: (1)  $\alpha$  helices have 3.6 amino acids per turn and represent ~97% of all helices; (2) 3.10 helices have 3.0 amino acids per turn (and are therefore more tightly packed) and account for ~3% of all helices; and (3)  $\pi$  helices, which occur only rarely, have 4.4 amino acids per turn. Myoglobin is an example of a protein with  $\alpha$  helices (**Fig. 13.3a**); these helices are typically formed from contiguous stretches of 4–40 amino acid residues in length. The  $\beta$  sheets are formed from adjacent  $\beta$  strands composed of 2–15 residues (typically 5–10 residues). They are arranged in either parallel or antiparallel orientations which have distinct hydrogen-bonding patterns. Pepsin (1PSN) provides an example of a protein comprised largely of  $\beta$  sheets (**Fig. 13.3b**).  $\beta$  sheets have higher-order properties including the formation of barrels and sandwiches and “super secondary structure motifs” such as  $\beta$ - $\alpha$ - $\beta$  loops and  $\alpha$ / $\beta$  barrels. Proteins commonly contain combinations of both  $\alpha$  helices and  $\beta$  sheets.

A Ramachandran plot displays the  $\phi$  and  $\psi$  angles for essentially all amino acids in a protein (proline and glycine are not displayed). The Ramachandran plot for beta globin shows a preponderance of  $\phi\psi$  angle combinations in a region that is typical of proteins with a helical content (**Fig. 13.4a**). In contrast, for pepsin the majority of  $\phi\psi$  angles occur in a region that is characteristic of  $\beta$  sheets (**Fig. 13.4b**). Ramachandran plots can be created using a variety of software packages including DeepView from ExPASy.

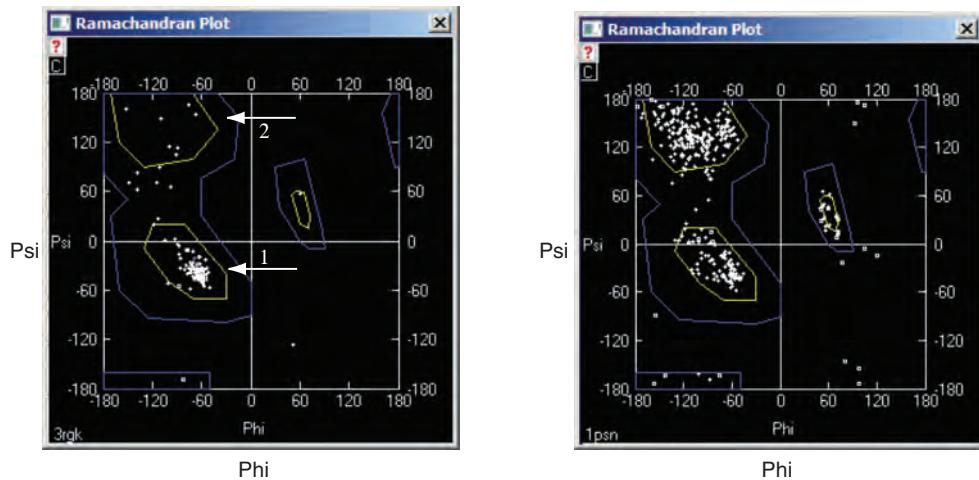
Dozens of methods have been developed to predict the secondary structure of a protein from its primary amino acid sequence (Pirovano and Heringa, 2010; Zhang *et al.*, 2011). Prediction began in the early 1970s. Chou and Fasman (1978) developed a method to predict secondary structure based on the frequencies of residues found in  $\alpha$  helices and



**FIGURE 13.3** Examples of secondary structure. (a) Myoglobin (Protein Data Bank accession 3RGK) is composed of large regions of  $\alpha$  helices, shown as strands wrapped around barrel-shaped objects. By entering the accession 3RGK into NCBI's structure site, the three-dimensional structure can be viewed using Cn3D software. The accompanying sequence viewer shows the primary amino acid sequence. By clicking on a colored region (bracket) corresponding to an alpha helix, that structure is highlighted in the structure viewer (arrow). (b) Human pepsin (PDB 1PSN) is an example of a protein primarily composed as  $\beta$  strands, drawn as large arrows. Selecting a region of the primary amino acid sequence (bracket) results in a highlighting of the corresponding  $\beta$  strand (arrow).

Source: Cn3D, NCBI.

(a) Ramachandran plot: myoglobin (3RGK) (b) Ramachandran plot: pepsin (1PSN)



**FIGURE 13.4** A Ramachandran plot displays the  $\varphi$  and  $\psi$  angles for each amino acid of a protein (except proline and in some cases glycine). Examples are shown for (a) myoglobin, a protein characterized by alpha helical secondary structure and (b) pepsin, a protein largely comprising beta sheets. The plots were generated using DeepView software from ExPASy. The arrows indicate the region of the Ramachandran plot in which  $\varphi\psi$  angles typical of alpha helices (arrow 1) and beta sheets (arrow 2) predominate.

Source: DeepView software from ExPASy. Reproduced with permission from SIB Swiss Institute of Bioinformatics.

$\beta$  sheets (together accounting for about half of all residues), as well as turns. Their algorithm calculates the propensity of each residue to form part of a helix, strand, or coil in the context of a sliding window of amino acids. For example, a proline is extremely unlikely to occur in an  $\alpha$  helix, and it is often positioned at a turn. The Chou–Fasman algorithm scans through a protein sequence and identifies regions where at least four out of six contiguous residues have a score for  $\alpha$  helices above some threshold value. The algorithm extends the search in either direction. Similarly, it searches for bends and turns. In a key study Williams *et al.* (1987) tabulated the conformational preferences of the amino acids (**Table 13.1**).

Subsequently, other approaches have been developed such as the GOR method (Garnier *et al.*, 1996). In most cases, these algorithms were used to analyze individual sequences (and they are still useful for this purpose). As multiply aligned sequences have become increasingly available, the accuracy of related secondary-structure prediction programs has increased. The PHD program (Rost and Sander, 1993a, b) is an example of an algorithm that uses multiple sequence alignment for this purpose.

In recent years the performance of secondary structure prediction software has improved. This is due to: the use of multiple alignments; the increased availability of large numbers of solved structures; and the application of machine-learning approaches such as neural networks. Neural network-based algorithms use layers of input signals (e.g., a multiple alignment of amino acid sequences of a protein, analyzed in sliding windows) and outputs (secondary structure predictions). A training protocol accepts sequences having known secondary structure elements. Zhang *et al.* (2011) compared 12 secondary structure predictors and found that those using neural networks tended to perform best.

The accuracy of the various algorithms has been assessed by evaluating their performance using databases of known structures. The standard measure for prediction accuracy, called Q3, is the proportion of all amino acids that have correct matches for the three states of helix, strand, and loop. Another measure is the segment overlap (Sov) which is relatively insensitive to small variations in secondary structure assignments, with less emphasis on assigning states to individual residues (Rost *et al.*, 1994). Current

**TABLE 13.1 Conformational preferences of the amino acids. Bold numbers indicate an increased propensity for a given amino acid to adopt a helix, strand, or turn position in a protein. Adapted from Williams *et al.* (1987) with permission from Elsevier.**

Amino acid	Preference			Properties
	Helix	Strand	Turn	
Glu	<b>1.59</b>	0.52	1.01	Helical preference; extended flexible side chain
Ala	<b>1.41</b>	0.72	0.82	
Leu	<b>1.34</b>	1.22	0.57	
Met	<b>1.30</b>	1.14	0.52	
Gln	<b>1.27</b>	0.98	0.84	
Lys	<b>1.23</b>	0.69	1.07	
Arg	<b>1.21</b>	0.84	0.90	
His	<b>1.05</b>	0.80	0.81	
Val	0.90	<b>1.87</b>	0.41	Strand preference; bulky side chains, beta-branched
Ile	1.09	<b>1.67</b>	0.47	
Tyr	0.74	<b>1.45</b>	0.76	
Cys	0.66	<b>1.40</b>	0.54	
Trp	1.02	<b>1.35</b>	0.65	
Phe	1.16	<b>1.33</b>	0.59	
Thr	0.76	<b>1.17</b>	0.90	
Gly	0.43	0.58	<b>1.77</b>	Turn preference; restricted conformations, side–main chain interactions
Asn	0.76	0.48	<b>1.34</b>	
Pro	0.34	0.31	<b>1.32</b>	
Ser	0.57	0.96	<b>1.22</b>	
Asp	0.99	0.39	<b>1.24</b>	

predictors reach about 82% accuracy based on Q3 and 81% accuracy based on Sov (Zhang *et al.*, 2011). Some of these leading software tools are listed in Table 13.2. This contrasts with 70–75% accuracy for methods around 2001 and 50–60% accuracy in the Chou–Fasman era.

In 1983 Wolfgang Kabsch and Christian Sander introduced a dictionary of secondary structure, including a standardized code for secondary structure assignment. These are applied in the DSSP database with eight states (Table 13.3). A variety of web servers allow you to input a primary amino acid sequence and delineate the secondary structure, often

**TABLE 13.2 Web servers for secondary-structure prediction based on neural networks. From Pirovano and Heringa (2010) and Zhang *et al.* (2011). Additional sites are listed at ExPASy (<http://www.expasy.org/tools/#secondary>, WebLink 13.38).**

Program	Source	URL
APSSP	Dr G. P. S. Raghava, Chandigarh, India	<a href="http://imtech.res.in/raghava/apssp/">http://imtech.res.in/raghava/apssp/</a>
Jpred	The Barton Group (Dundee)	<a href="http://www.compbio.dundee.ac.uk/~www-jpred/">http://www.compbio.dundee.ac.uk/~www-jpred/</a>
Porter	University College, Dublin	<a href="http://distill.ucd.ie/porter/">http://distill.ucd.ie/porter/</a>
PHD	Guy Yachdav & Burkhard Rost, Technical University of Munich	<a href="https://www.predictprotein.org/">https://www.predictprotein.org/</a>
Proteus	Wishart Research Group	<a href="http://wks80920.ccis.ualberta.ca/proteus/">http://wks80920.ccis.ualberta.ca/proteus/</a>
PSIPRED	Bloomsbury Centre for Bioinformatics	<a href="http://bioinf.cs.ucl.ac.uk/psipred/">http://bioinf.cs.ucl.ac.uk/psipred/</a>
SPINEX	Indiana University–Purdue University, Indianapolis	<a href="http://sparks.informatics.iupui.edu/SPINE-X/">http://sparks.informatics.iupui.edu/SPINE-X/</a>
SSpro	University of California, Irvine	<a href="http://scratch.proteomics.ics.uci.edu/">http://scratch.proteomics.ics.uci.edu/</a>

**TABLE 13.3 Secondary structure assignment from the DSSP database.**

DSSP code	Secondary structure assignment
H	Alpha helix
B	Residue in isolated beta-bridge
E	Extended strand, participates in beta ladder
G	3-helix (3/10 helix)
I	5 helix (pi helix)
T	Hydrogen bonded turn
S	Bend
Blank or C	Loop or irregular element, incorrectly called "random coil" or "coil."

Source: DSSP (<http://swift.cmbi.ru.nl/gv/dssp/>, WebLink 13.39), courtesy of G. Vriend.

DSSP software is available from  
 ↗ <http://swift.cmbi.ru.nl/gv/dssp/>  
 (WebLink 13.8). The PBIL website  
 is at ↗ <http://npsa-pbil.ibcp.fr>  
 (WebLink 13.9).

employing the DSSP codes. Some of the programs allow you to enter a single sequence, while others allow you to enter a multiple sequence alignment. As an example, the Pôle Bio-Informatique Lyonnais (PBIL) has a web server that offers secondary-structure predictions for a protein query. We used this server to generate the beta globin prediction in **Figure 13.1b**. This server also generates predictions using nine different algorithms and calculates a consensus. The various predictions differ somewhat in detail, but are generally consistent.

### Tertiary Protein Structure: Protein-Folding Problem

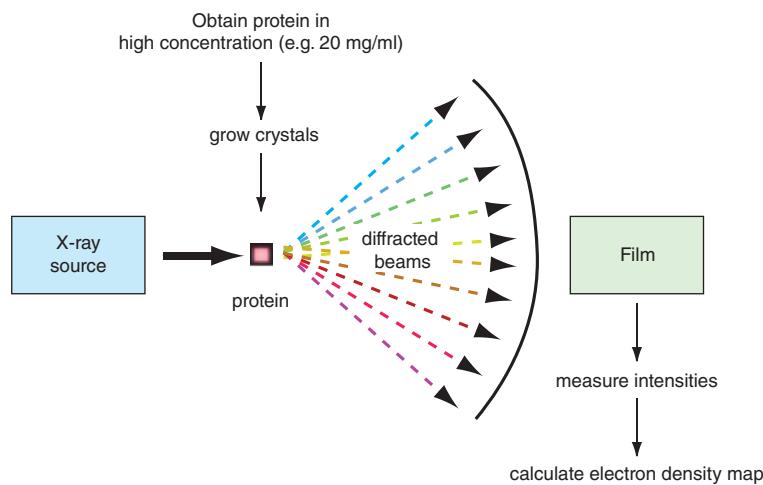
How does a protein fold into a three-dimensional structure? As mentioned above, this problem is solved very rapidly in nature. In 1969 Cyrus Levinthal introduced an argument (later called “Levinthal’s paradox”) that there are far too many possible conformations for a linear sequence of amino acids to adopt its native conformation through random samplings of the energy landscape. To find the most stable thermodynamic structure would require a period of time far greater than the age of the universe. Proteins must therefore adopt their three-dimensional conformations by following specific folding pathways. Progress in understanding protein folding has been reviewed by Dill *et al.* (2008), Hartl and Hayer-Hartl (2009), Travaglini-Allicatelli *et al.* (2009), and Dill and MacCallum (2012). Folding is thought to occur by incremental movements along a pathway toward favored low-energy native structures. This process is guided by factors such as hydrogen bonds (contributing to secondary structure), van der Waals interactions, backbone angle preferences, electrostatic interactions of amino acid side chains, hydrophobic interactions, and entropic forces. Additionally, chaperones stabilize nascent polypeptides on ribosomes and facilitate appropriate folding.

In structural biology, there are two main approaches to determining protein structure: X-ray crystallography; and nuclear magnetic resonance spectroscopy (NMR). Structures can also be predicted computationally using three approaches described near the end of this chapter (homology modeling, threading, and *ab initio* prediction; see “Protein Structure Prediction”).

X-ray crystallography is the most rigorous experimental technique used to determine the structure of a protein (Box 13.1), and about 80% of known structures were determined using this approach. The basic steps involved in this process are outlined in **Figure 13.5**. A protein must be obtained in high concentration and seeded in conditions that permit crystallization. The crystal scatters X-rays onto a detector, and the structure of the crystal is inferred from the diffraction pattern. The wavelength of X rays (about 0.5–1.5 Å) is useful to measure the distance between atoms, making this technique suitable to trace the amino acid side chains of a protein.

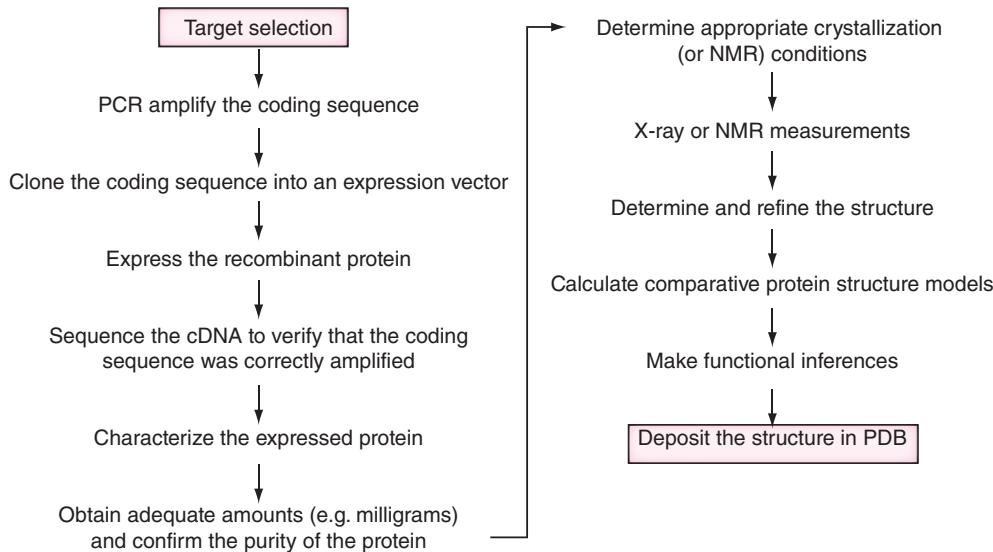
### BOX 13.1 X-RAY CRYSTALLOGRAPHY

A protein is obtained in high concentrations and crystallized in a solution such as ammonium sulfate. A beam of X-rays is aimed at the protein crystals. The protein is in a highly regular array that causes the X-rays to diffract (scatter) where they are detected on X-ray film. Spot intensities are measured, and an image is generated by Fourier transformation of the intensities. An electron density map is generated corresponding to the arrangements of the atoms that comprise the protein. Resolution of less than 2 Å is generally required for a detailed structure determination.



There have been continuous efforts to improve X-ray based technologies.

- X-ray free-electron lasers (XFELs) deliver ultrashort, high-intensity pulses of X-rays (Schlichting and Miao, 2012; Smith *et al.*, 2012). These pulses can be a billion times brighter than conventional sources, and this method has been called “diffraction-before-destruction.” Boutet *et al.* (2012) applied this approach to microcrystals ( $<1 \mu\text{M} \times 1 \mu\text{M} \times 3 \mu\text{M}$ ) of the enzyme lysozyme. The significance is that proteins and other molecules for which it is difficult to obtain large crystals may now have their structures solved. An important example is the structure of the  $\beta_2$  adrenergic receptor complexed with the alpha subunit of its corresponding heterotrimeric GTP-binding protein complex, solved by Brian Kobilka and colleagues (Rasmussen *et al.*, 2011). The  $\beta_2$  adrenergic receptor is an example of a G protein-coupled receptor (GPCR), the largest protein family encoded by the human genome. GPCRs respond to neurotransmitters, hormones, and signals such as light and odorants. XFELs have been used to advance the study of other membrane proteins (Kang *et al.*, 2013). In another application, Koopmann *et al.* (2012) solved the structures of a *Trypanosoma brucei* cathepsin enzyme



**FIGURE 13.5** The process involved in obtaining high-resolution structures. General procedure for obtaining a three-dimensional protein structure.

Brian Kobilka and Robert Lefkowitz were awarded the 2012 Nobel Prize in Chemistry "for studies of G-protein–coupled receptors;" see [http://www.nobelprize.org/nobel\\_prizes/chemistry/laureates/2012/](http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2012/) (WebLink 13.10). See the mention of GPCRs in "Protein Structure and Disease" below.

DARA, a DAtabase for RApid search of structural neighbors for proteins based on their X-ray small-angle scattering patterns, is available at <http://dara.embl-hamburg.de> (WebLink 13.11).

using crystals grown *in vivo*. Such crystals may preserve post-translational modifications, and may also be characterized in parallel by electron microscopy.

- Electron crystallography of two-dimensional crystals has been successfully employed to generate atomic models of many proteins (Kühlbrandt, 2013). The first of these was also the first membrane protein to have high-resolution structure determination, bacteriorhodopsin (Henderson and Unwin, 1975).
- Small-angle X-ray scattering allows low-resolution structure determination without the requirement for diffraction-quality crystals, and without size limitation (Perry and Tainer, 2013).

Nuclear magnetic resonance spectroscopy is an important alternative approach to crystallography. A magnetic field is applied to proteins in solution, and characteristic chemical shifts are observed. From these shifts, the structure is deduced. The largest structures that have been determined by NMR are about 350 amino acids (~40 kD), considerably smaller than the size of proteins routinely studied by crystallography. Other limitations are that the quality of NMR structures is less than those obtained by crystallography, and NMR yields multiple structure solutions rather than one. However, an advantage of NMR is that it does not require a protein to be crystallized, a notoriously difficult process.

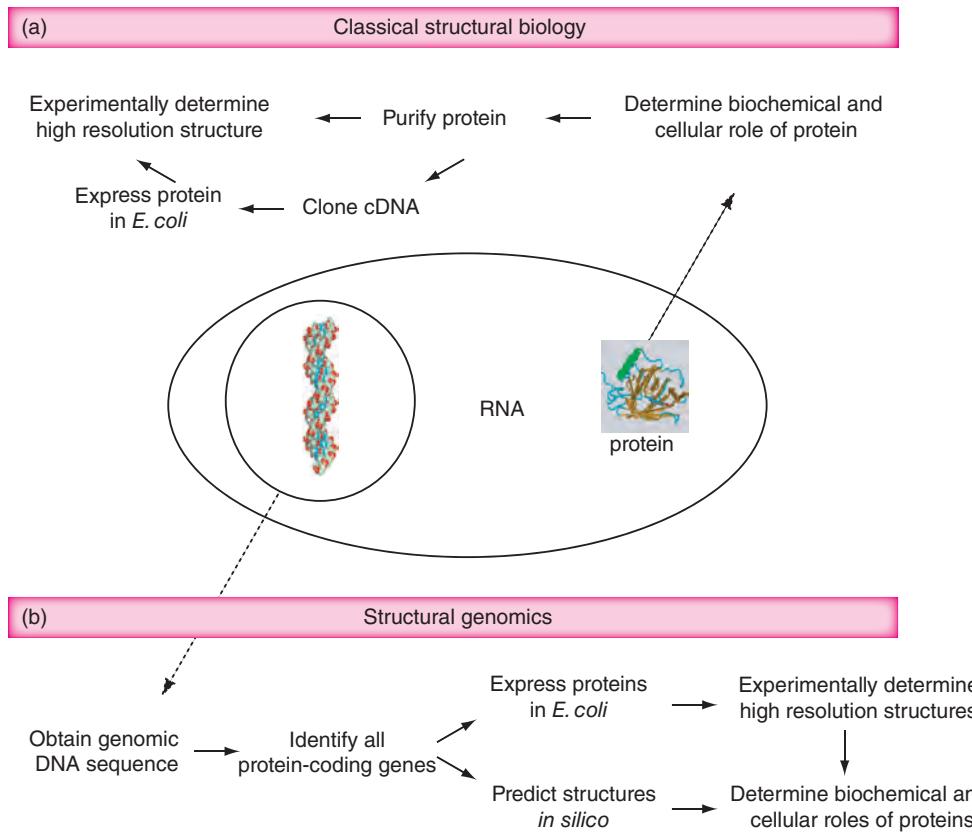
### Structural Genomics, the Protein Structure Initiative, and Target Selection

Structural genomics is an emerging field of research. Its goal is to determine the three-dimensional structure of all the major protein families throughout the tree of life, spanning fold space (Brenner, 2001; Koonin *et al.*, 2002; Andreeva and Murzin, 2010). Fold space refers to the total variety of three-dimensional protein structures that occur in nature. This mostly comprises proteins having  $\alpha$ ,  $\beta$ , or  $\alpha\beta$  secondary structure composition (Holm and Sander, 1997). This comprehensive approach will permit a deeper understanding of the relatedness of protein domains, and will also enable us to assign function to many proteins. Structure space (or fold space) may be cataloged in terms of protein sequence families, which generally are defined as containing members having greater than about 30% amino acid identity. Structural genomics ultimately aims to solve at least one high-resolution structure for every sequence family.

The relationship of structural genomics to traditional structural biology is outlined in **Figure 13.6**. Traditionally, researchers obtained the structure of individual proteins by starting with information about the known function of the protein. The new approach of structural genomics is based upon a reverse strategy: genome sequence projects generate predictions of protein-coding sequences. One fundamentally important aspect of each predicted protein is its structure. Predicted proteins may be expressed and their structures are solved to high resolution (**Fig. 13.6**). The recent identification of literally tens of millions of novel predicted proteins has enabled researchers to choose structures to solve (targets) based upon a variety of criteria. Once a target is selected and a cDNA encoding that protein is cloned, there are still many challenges in successfully expressing, purifying, and crystallizing the protein as well as obtaining its structure by either X-ray crystallography or NMR.

The general procedure for experimentally acquiring protein structural data, outlined in **Figure 13.5**, begins with target selection, the process of choosing which structure to solve (Brenner, 2000). Historically, proteins such as hemoglobin and cytochrome *c* were selected that were most amenable to experimental study: they are generally small, soluble, abundant, and known to have interesting biological functions. Today, additional criteria are considered in deciding priorities for which protein structures to solve (Carter *et al.*, 2008; Marsden and Orengo, 2008):

- All branches of life (eukaryotes, bacteria, archaea, and viruses) are studied.
- Should there be efforts to exhaustively solve all structures within an individual organism? This is being attempted for *Methanococcus jannaschii* and *Mycobacterium*



**FIGURE 13.6** Classical structural biology versus structural genomics. (a) In classical structural biology approaches, a protein is purified based upon some known function or activity. After biochemical purification of the protein, if there is sufficient yield, the protein may be crystallized and its structure determined. This in turn allows the biochemical function of the protein and its mechanism of action to be studied. Having obtained protein sequence, the corresponding complementary DNA (cDNA) may be cloned, allowing recombinant protein to be expressed and purified for structure analyses. (b) The field of structural genomics proceeds from genomic DNA sequence. Large numbers of protein-coding genes are predicted, often including all those encoded by a genome of interest. Selected proteins are either cloned and expressed for biochemical analysis or the structure is predicted computationally (“*in silico*”). The 3D structure of a protein may be determined experimentally using techniques such as X-ray crystallography or NMR spectroscopy. Finally, the biochemical role may be inferred based upon the nature of the structure. Additional insight into biochemical function is derived from database searches of the protein sequence (e.g., using DELTA-BLAST).

*tuberculosis*. The Bacterial Structural Genomics Initiative (Matte *et al.*, 2007) includes efforts to determine structures for a large number of *Escherichia coli* proteins.

- Should representatives from previously uncharacterized protein families be selected preferentially?
- Should medically important proteins such as drug discovery targets be chosen first?
- How can structures be solved for more proteins having transmembrane-spanning domains? These are among the most technically challenging proteins to study (Kang *et al.*, 2013). Chang and Roth (2001) successfully solved the structure of a multidrug-resistant ABC transporter from *E. coli*. They screened 96,000 crystallization conditions to find several that were adequate for X-ray structure determination.

The Protein Structure Initiative (PSI) has had a major impact in the direction of structural genomics, including target selection (Andreeva and Murzin, 2010; Montelione, 2012). The PSI was established in the United States in 2000, with similar structural

genomics projects conducted in other countries (Canada, Israel, Japan, and in Europe). The PSI is a coordinated effort by the academic, industry, and federal research communities to develop the technology needed to determine the three-dimensional structures of most proteins based on knowledge of the corresponding DNA sequences. There have been three phases. The pilot phase of the project (conducted during 2000–2005) involved nine structural genomics centers that solved more than 1100 structures at high resolution. A key feature of this project is that solving the structure of proteins that are closely related to those having known structures is relatively easy, but predicting structures without close structure neighbors can be extremely difficult. Of the 1100 solved structures over 700 were unique, that is, the structures shared less than 30% amino acid sequence identity with other known proteins.

The second phase of PSI (2005–2010) included the goal of depositing >4000 structures into the Protein Data Bank (we discuss the PDB in the following section). Most of these structures were of proteins, with additional protein-ligand complexes and paired X-ray and NMR structures. An analysis by Levitt (2007) emphasizes the general decline in the number of novel structures into the PDB beginning in 1995, and a reversal of this trend because of contributions from structural genomics initiatives. Chandonia and Brenner (2005) proposed that the Pfam5000 set be selected: these are the 5000 largest Pfam families for which no structure has yet been solved. As part of PSI phase 2, the structure of representative members of Pfam families having domains of unknown function were determined at high resolution. These include Pfam family PF06684 (Bakolitsa *et al.*, 2010), PF06938 (Han *et al.*, 2010), and PF04016 (Miller *et al.*, 2010). In each case these families contain hundreds of proteins, and solving the structures suggested possible functions in amino acid synthesis, signal transduction, and heavy metal chelation.

Phase 2 of PSI characterized the structures of only >100 human proteins. In the ongoing Phase 3, called PSI:Biology, the focus changed from solving as many structures as possible to solving structures that are biologically or medically relevant (such as GPCRs; Depietro *et al.*, 2013). The community can nominate proteins for structure determination. Current progress and project resources are available at the Structural Biology Knowledgebase (SBKB) website (Gifford *et al.*, 2012). As of February 2015, there have been ~332,000 protein targets deposited in the SBKB Target Track. Of these, a subset were successfully cloned, expressed, shown to be soluble, and crystallized. A subset of these yielded diffraction-quality crystals, and ~10,000 structures were solved and deposited in the Protein Data Bank repository.

In 1992, even before the first genome of a free-living organism had been fully sequenced, Cyrus Chothia estimated that there may be about 1500 distinct protein folds. Structural genomics initiatives such as PSI continue to bring us closer to identifying all of them.

## PROTEIN DATA BANK

The PDB was established at Brookhaven National Laboratories in Long Island in 1971. Initially, it contained seven structures. It moved to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1998. PDB is accessed at <http://www.rcsb.org/pdb/> or <http://www.pdb.org> (WebLink 13.14).

Once a protein sequence is determined, there is one principal repository in which the structure is deposited: the Protein Data Bank (PDB) (Rose *et al.*, 2013; reviewed in Berman, 2012; Berman *et al.*, 2013a–c; Goodsell *et al.*, 2013). A broad range of primary structural data is collected, such as atomic coordinates, chemical structures of cofactors, and descriptions of the crystal structure. The PDB then validates structures by assessing the quality of the deposited models and by how well they match experimental data.

The main page of the PDB website includes categories by which information may be accessed (Fig. 13.7). This database currently has over 100,000 structure entries (Table 13.4), with new structures being added at a rapid rate (Fig. 13.8). The database can be accessed directly by entering a PDB identifier into the query box on the main page, that is, by entering an accession number consisting of one number and three letters

The SBKB website is <http://sbkb.org/> (WebLink 13.12), which includes target selection data (<http://sbkb.org/tt/>). The main PSI website at the National Institutes of Health is <http://www.nigms.nih.gov/research/specificareas/PSI/Pages/default.aspx> (WebLink 13.13).



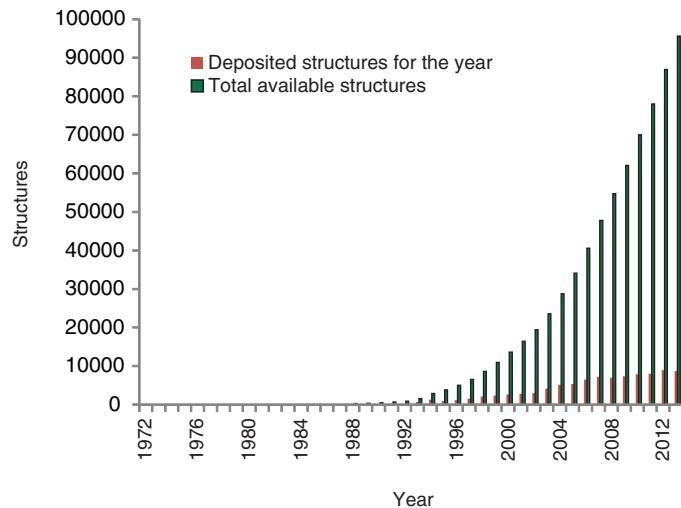
**FIGURE 13.7** The PDB is the main repository for three-dimensional structures of proteins and other macromolecules. Information about PDB holdings is organized into categories such as organism, taxonomy, experimental method (greater than 85% of which are derived from X-ray structure determination), and resolution (with less than 1.5 Å corresponding to the highest resolution structures). The home page of PDB allows queries such as a PDB identifier (e.g., 3RGK for a myoglobin structure) or a molecule name.

Source: RCSB PDB ([www.rcsb.org](http://www.rcsb.org)). Reproduced with permission from RCSB PDB.

**TABLE 13.4** Types of molecules, according to PDB Holdings.

Experimental technique	Proteins	Nucleic acids	Protein and nucleic acid complexes	Other	Total
X-ray diffraction	88,991	1,608	4,398	4	95,001
NMR	9,512	1,112	224	8	10,856
Electron microscopy	539	29	172	0	740
Hybrid	68	3	2	1	74
Other	164	4	6	13	187
Total	99,274	2,756	4,802	26	106,858

Source: RCSB PDB. ([www.rcsb.org](http://www.rcsb.org)). Reproduced with permission from RCSB PDB.



**FIGURE 13.8** Number of searchable structures per year in PDB. The PDB database has grown dramatically in the past decade. The yearly (red) and total (green) numbers of structures are shown.

Source: RCSB PDB ([www.rcsb.org](http://www.rcsb.org)). Reproduced with permission from RCSB PDB.

(e.g., 4HQB for hemoglobin). The PDB database can also be searched by keyword; the result of a keyword search for myoglobin is shown in **Figure 13.9**. In this case there are hundreds of results, and the list can be refined using options on the left sidebar. The result of searching for a specific hemoglobin identifier, 3RGK, links to a typical PDB entry (of which a portion is shown in **Fig. 13.10**). By clicking on an icon the 3RGK.pdb file can be downloaded locally for further analysis with a variety of tools such as DeepView. Information provided on the 3RGK page includes the resolution of the experimentally derived structure, the space group, and the unit cell dimensions of the crystals. There are links to a series of tools to visualize the three-dimensional structure, including Jmol (**Fig. 13.10**, arrow 2). **Table 13.5** lists some additional visualization software. Using Jmol does not require the installation of software (other than Java), and it is versatile (**Fig. 13.11**).

It is also possible to search within the PDB website using dozens of advanced search features (accessed via the top of the home page). This includes the use of BLAST or FASTA programs, allowing convenient access to PDB structures related to a query. Other advanced search features allow you to query based on properties of the molecule (e.g., its molecular weight), PubMed identifier, Medical Subject Heading (MeSH term; Chapter 2), deposit date, or experimental method.

377 Structure Hits | 10 Unreleased Structures | 124 Citations | 82 Ligand Hits | 48 Web Page Hits | Query Details | Save Query to MyPDB

**Query Parameters:**

Text Search for: myoglobin  
Other search suggestions:

Molecule Name	Structural Domains	Molecule of the Month	PDB Text	Ontology Terms
* Myoglobin (362) Find all	* Myoglobin [SCOP] (254) * Trematode hemoglobin/myoglobin [SCOP] (2)	* Myoglobin	* myoglobin * myoglobins Find all	* SS : MYOGLOBIN [Genome... (17) * D12.776....: Myoglobin [MeSH... (326) * PC : MYOGLOBIN [Genome... (246) * D12.776....: Myoglobin [MeSH... (311) * EC : MYOGLOBIN [Genome... (62) * HS : MYOGLOBIN [Genome... (1)

**Query Refinements: Select an item or pie chart** Hide

Organism	Taxonomy	Experimental Method	X-ray Resolution	Release Date	Polymer Type
* Physeter catodon (247) * Equus caballus (66) * Sus scrofa (17) * Homo sapiens (12) * Thunnus atlanticus (8) * Aplysia limacina (7) * Chironomus thummi thummi (4) * Other (16)	* Eukarya only (377)	* X-ray (372) * Neutron Diffraction (3) * Other (1) * Solution NMR (1)	* less than 1.5 Å (96) * 1.5 - 2.0 Å (203) * 2.0 - 2.5 Å (69) * 2.5 - 3.0 Å (4) * more choices...	* before 2000 (168) * 2000 - 2005 (48) * 2005 - 2010 (102) * 2010 - 2015 (55) * 2015 - today (4) * more choices...	* Protein (377)
Enzyme Classification	SCOP Classification	Protein Symmetry	Protein Stoichiometry		
* 3: Hydrolases (2) * 1: Oxidoreductases (1)	* Globin-like ( core: 6 helices; ... (269) * Cytochrome c ( core: 3 helices; ... (1)	* Asymmetric (353) * Cyclic (9) * more choices...	* Monomer (353) * Heteromer (5) * Homomer (4) * more choices...		

**FIGURE 13.9** Result of a PDB query for myoglobin. There are several hundred results organized into categories such as UniProt gene names, structural domains, and ontology terms. The search results further show how to explore myoglobin entries with the same categories shown in **Figure 13.7**.

Source: RCSB PDB (www.rcsb.org). Reproduced with permission from RCSB PDB.

Summary | 3D View | Sequence | Annotations | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

**Crystal Structure of Human Myoglobin Mutant K45R** 3RGK Display Files Download Files Share this Page DOI:10.2210/pdb3rgk/pdb

**ENTRY 3RGK SUPERSEDES 2MM1**

**Primary Citation**

X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 Å resolution.  
Hubbard, S.R., Lambright, S.G., Boxer, S.G., Hendrickson, W.A. ▾  
Journal: (1990) J.Mol.Biol. 20: 215-218  
PubMed: 2342104 ▾  
DOI: 10.1016/S0022-2836(05)80181-0 ▾  
Search Related Articles in PubMed ▾

**PubMed Abstract:**  
We have grown crystals in trigonal space group P3(2)21 of a mutant human myoglobin, aquomet form, in which lysine at position 45 has been replaced by arginine and cysteine at position 110 has been replaced by alanine. Suitable crystals of...  
[ Read More & Search PubMed Abstracts ]

**Molecular Description** Hide

**Classification:** Oxygen Transport  
**Structure Weight:** 17984.47 ▾

**Molecule:** Myoglobin  
**Polymer:** 1      **Type:** protein  
**Chains:** A      **Length:** 153

**Biological Assembly**

1 ↓ 3D View More Images...  
No symmetry. Stoichiometry: Monomer  
Biological assembly 1 assigned by authors and generated by PISA (software)  
Downloadable viewers:  
Simple Viewer Protein Workshop  
Kiosk Viewer

**FIGURE 13.10** Result of a search for a myoglobin structure, 3RGK. The summary information includes a description of the resolution (2.8 Å), the space group, unit cell dimensions, ligands, and external database annotation. Available links include a variety of visualization software (including Jmol, arrow 1).

Source: RCSB PDB (www.rcsb.org). Reproduced with permission from RCSB PDB.

**TABLE 13.5 Interactive visualization tools for protein structures.** The Protein Data Bank maintains a list of molecular graphics software links, accessible from the PDB home page via software tools/molecular viewers at [http://www.pdb.org/pdb/static.do?p=software/software\\_links/molecular\\_graphics.html](http://www.pdb.org/pdb/static.do?p=software/software_links/molecular_graphics.html) (WebLink 13.40).

Tool	Comment	URL
Cn3D	From NCBI	<a href="http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml">http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml</a>
JMol	Open-source Java viewer for chemical structures in 3D	<a href="http://jmol.sourceforge.net/">http://jmol.sourceforge.net/</a>
Kiosk Viewer	Uses Java Web Start	<a href="http://pdb.org/">http://pdb.org/</a>
Mage	Reads Kinemages	<a href="http://kinemage.biochem.duke.edu">http://kinemage.biochem.duke.edu</a>
Protein Workshop Viewer	Uses Java Web Start	<a href="http://pdb.org/">http://pdb.org/</a>
RasMol	Molecular graphics visualization tool	<a href="http://www.rasmol.org/">http://www.rasmol.org/</a>
RasTop	Molecular visualization software adapted from RasMol	<a href="http://www.geneinfinity.org/rastop/">http://www.geneinfinity.org/rastop/</a>
Simple Viewer	Uses Java Web Start	<a href="http://pdb.org/">http://pdb.org/</a>
SwissPDB viewer	At ExPASy	<a href="http://spdbv.vital-it.ch">http://spdbv.vital-it.ch</a>
VMD	Visual Molecular Dynamics; University of Illinois	<a href="http://www.ks.uiuc.edu/Research/vmd/">http://www.ks.uiuc.edu/Research/vmd/</a>

The PDB in Europe database (PDBe) is at <http://www.ebi.ac.uk/pdbe/> (WebLink 13.15). PDB Japan is at <http://pdbj.org> (WebLink 13.16).

The PDB-related WHAT IF servers are available at <http://swift.cmbi.ru.nl/servers/html/> (WebLink 13.17).

PDB is maintained by members of the WorldWide PDB. These include the RCSB PDB, the Protein Data Bank in Europe (operated by the European Bioinformatics Institute), and PDB Japan.

A series of databases are complementary to PDB and hold information corresponding directly to PDB entries. These include the following (Joosten *et al.*, 2011):

- DSSP includes secondary structure data;
- PDBREPORT includes data on structure quality and errors;
- PDBFINDER offers summaries of PDB content (information includes Enzyme Commission numbers for enzymes);
- PDB\_RED0 includes re-refined (often improved) copies of structures (e.g., the orientation of peptide planes can be optimized); and
- WHY\_NOT explains why particular files were not produced (e.g., a brand new PDB entry might not yet have entries in ancillary databases, or a PDB entry solved by NMR spectroscopy would not have PDB\_RED0 entries).

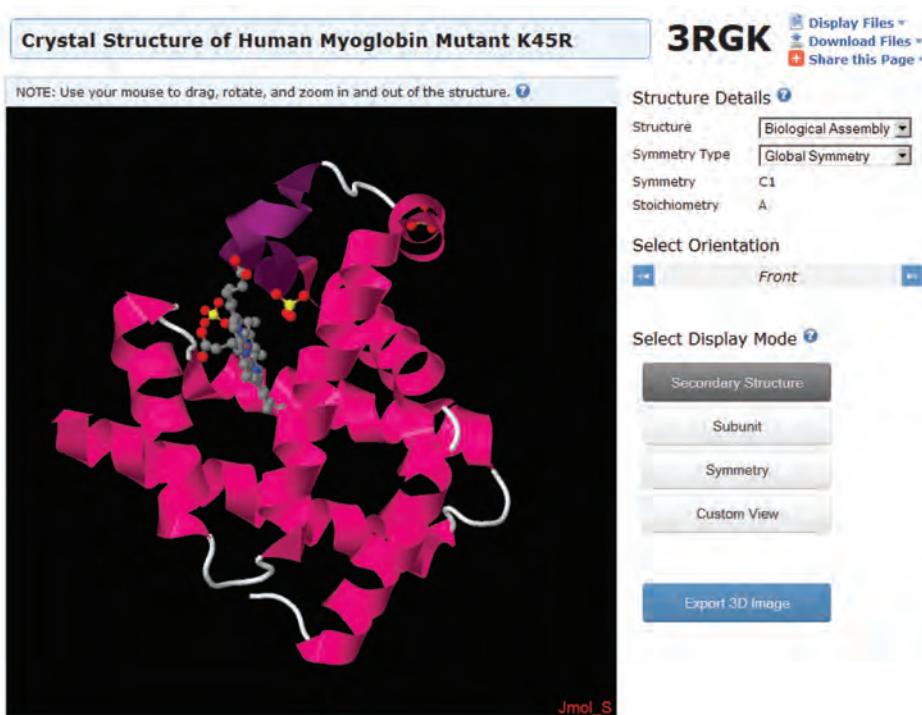
The PDB database occupies a central position in structural biology. Several dozen other databases and web servers link directly to it or incorporate its data into their local resources. We next explore NCBI and other sites that allow a single protein structure to be analyzed or several structures to be compared. We then explore databases that create comprehensive classification systems or taxonomies for all protein structures.

### Accessing PDB Entries at NCBI Website

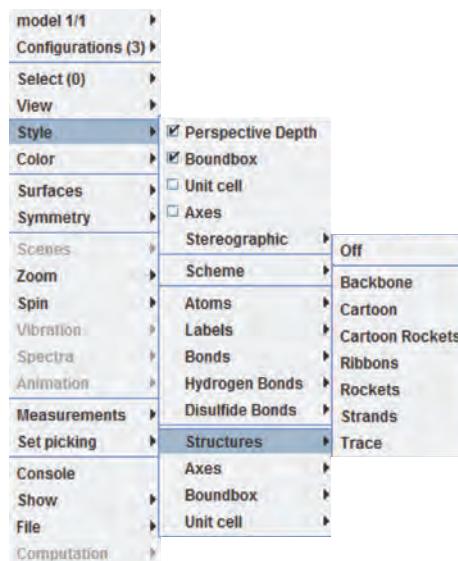
There are three main methods of finding a protein structure in the NCBI databases:

1. Text searches allow access to PDB structures. These searches can be performed on the structure page or through Entrez, and they can consist of keywords or PDB identifiers.

(a) Structure visualization in PDB using the Jmol applet



(b) Jmol options menus



**FIGURE 13.11** Jmol applet software permits the visualization and analysis of macromolecular structures. (a) View of a human myoglobin structure. This can be manipulated (e.g., zoomed or rotated), colored according to criteria such as secondary structure, visualized (e.g., to show van der Waal radii), and analyzed (e.g., by measuring interatomic distances). (b) Right-clicking (on a PC) opens a menu of Jmol viewing options.

Source: RCSB PDB ([www.rcsb.org](http://www.rcsb.org)). Reproduced with permission from RCSB PDB.

A keyword search of Entrez structures for hemoglobin yields a list of ~1300 proteins with four-character PDB identifiers. If you know a PDB identifier of interest, such as 3RGK for myoglobin, use it as a search term and to find an NCBI Structure entry with useful links, including to the Molecular Modeling Database (Fig. 13.12), the Cn3D viewer, the VAST comparison tool (see below), and the Conserved Domain

**FIGURE 13.12** The Molecular Modeling DataBase (MMDB) at NCBI offers tools to analyze protein (and other) structures. You can view the structure (lower right) using the Cn3D structure viewer (alternatively, you can view the PDB file corresponding to this entry, a human myoglobin with PDB accession 3RGK). A link to VAST (upper right) allows identification and visualization of related structures (see Fig. 13.14). The entry includes a literature citation (upper left) as well as further information on the myoglobin molecule and its interactions (not shown).

Source: Molecular Modeling DataBase (MMDB), NCBI.

The NCBI structure page is at  
 ↗ <http://www.ncbi.nlm.nih.gov/structure> (WebLink 13.18).

Database (Chapter 5). The Molecular Modeling Database is the main NCBI database entry for each protein structure (Madej *et al.*, 2012) from the group of Steve Bryant. It includes literature and taxonomy data, sequence neighbors (as defined by BLAST), structure neighbors (as defined by VAST; see following section), and visualization options.

2. It is possible to search by protein similarity. To do this, use the NCBI Protein database to select a protein of interest and look for a link to “Related Structures.” Alternatively, perform a BLASTP search and restrict the output to the PDB database. All database matches have entries in the NCBI Structure database (Fig. 13.13, right-hand column).
3. Searching using nucleotide queries is another option. It is possible to use a BLASTX search with a DNA sequence as input, restricting the output to the PDB database.

Cn3D is the NCBI software for structure visualization. We describe its use in computer lab exercise (13.1) and used it to generate Figure 13.3. Upon launching Cn3D two windows open: a Cn3D Viewer and a OneD-Viewer (Fig. 13.3.). The Cn3D Viewer shows the structure of the protein in seven available formats (such as ball-and-stick or space-filling models), and it can be rotated for exploration of the structure. The corresponding OneD-Viewer shows the amino acid sequence of the protein, including  $\alpha$  helices and

[Sequences producing significant alignments with E-value BETTER than threshold](#)

Select: [All](#) [None](#) Selected:0

	Description	Max score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">Chain B, Pigeon Hemoglobin (Oxy Form) &gt; pdb 2R80 D Chain D, Pigeon Hemoglobin (Ox)</a>	164	99%	1e-53	69%	<a href="#">2R80_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Crystal Structure Of Parrot Hemoglobin (Psittacula Kramerii) At Ph 7.5</a>	160	99%	3e-52	69%	<a href="#">2ZFB_B</a>
<input type="checkbox"/>	<a href="#">Chain B, R-State Form Of Chicken Hemoglobin D &gt; pdb 1HBR D Chain D, R-State Form O</a>	159	99%	1e-51	69%	<a href="#">1HBR_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Crystal Structure Determination Of Japanese Quail (Coturnix Coturnix Japonica)</a>	159	99%	2e-51	68%	<a href="#">3MJP_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Graylag Goose Hemoglobin (Oxy Form) &gt; pdb 1FAW D Chain D, Graylag Goose</a>	158	99%	2e-51	69%	<a href="#">1FAW_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Structure Determination Of Haemoglobin From Turkey(meleagris Gallopavo) At 2.1</a>	158	99%	3e-51	68%	<a href="#">2QMB_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Crystal Structure Determination Of Duck (Anas Platyrhynchos) Hemoglobin At 2.1</a>	157	99%	4e-51	69%	<a href="#">3EOK_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Bar-Headed Goose Hemoglobin (Oxy Form) &gt; pdb 1C40 B Chain B, Bar-Headed</a>	157	99%	4e-51	69%	<a href="#">1A4F_B</a>
<input type="checkbox"/>	<a href="#">Chain B, Crystal Structure Determination Of Ostrich Hemoglobin At 2.2 Angstrom Resolution</a>	155	99%	2e-50	68%	<a href="#">3FS4_B</a>
<input type="checkbox"/>	<a href="#">Chain A, R-State Form Of Chicken Hemoglobin D &gt; pdb 1HBR C Chain C, R-State Form O</a>	144	97%	4e-46	42%	<a href="#">1HBR_A</a>
<input type="checkbox"/>	<a href="#">Chain A, Crystal Structure Of Parrot Hemoglobin (Psittacula Kramerii) At Ph 7.5</a>	130	97%	2e-40	38%	<a href="#">2ZFB_A</a>
<input type="checkbox"/>	<a href="#">Chain A, Crystal Structure Determination Of Ostrich Hemoglobin At 2.2 Angstrom Resolution</a>	125	97%	9e-39	36%	<a href="#">3FS4_A</a>

**FIGURE 13.13** Structure entries can be retrieved from NCBI by performing a BLASTP search (with a protein query) or a BLASTX search (with DNA), restricting the output to the PDB database. Here, a DELTA-BLAST search with human beta globin (NP\_000509.1) restricted to birds (aves) produces matches against a variety of pigeon, parrot, duck, ostrich, and chicken globins. Since the database was set to PDB, all of these entries are of known structure and the PDB accession numbers are given (right-most column).

Source: BLASTP, BLASTX and PDB database, NCBI.

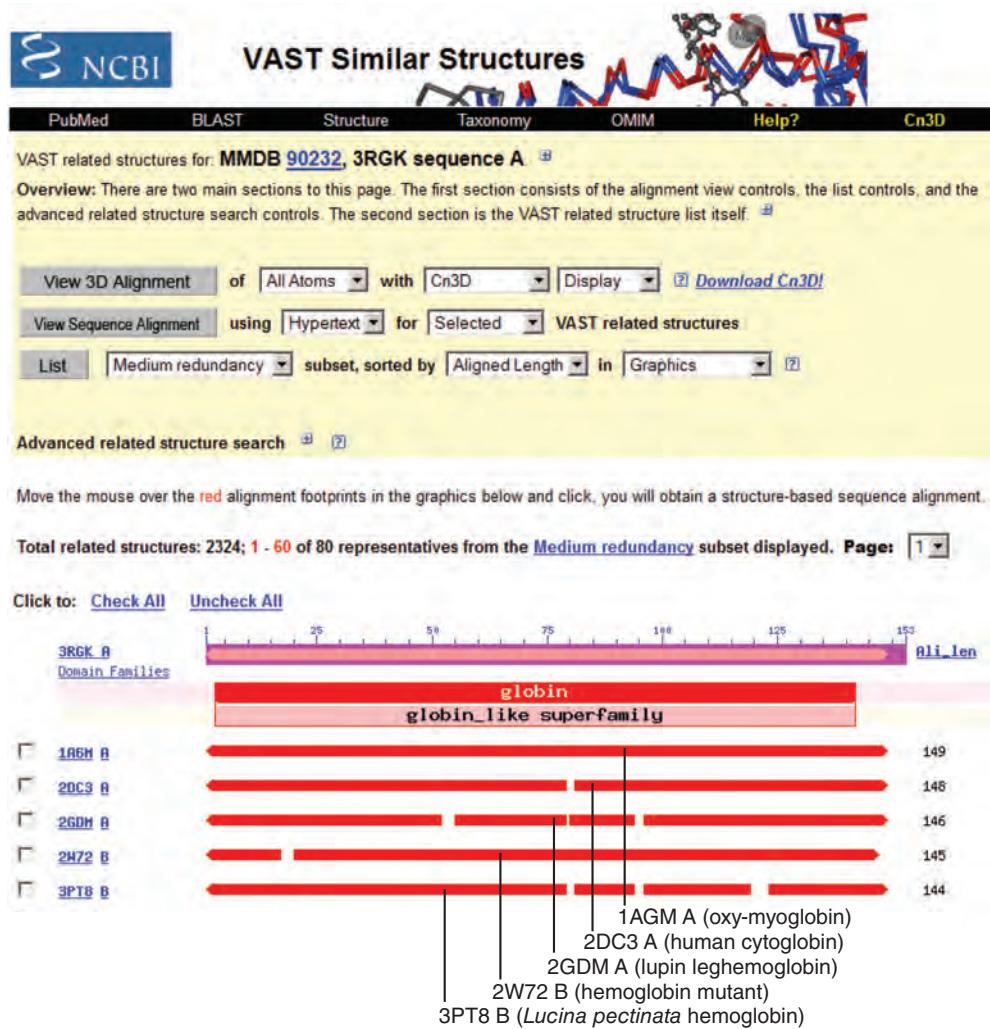
$\beta$  sheets. Highlighting any individual amino acid residue or group of residues in either the Cn3D Viewer or the OneD-Viewer causes the corresponding region of the protein to be highlighted in the other viewer.

Cn3D is short for “see in 3D.”

In addition to investigating the structure of an individual protein, multiple protein structures can be compared simultaneously. Beginning at the main MMDB structure summary for a protein such as myoglobin (Fig. 13.12), click “VAST” to obtain a list of related proteins for which PDB entries are available (Fig. 13.14). This list is part of the Vector Alignment Search Tool (VAST). Select the entries related to structures, or (using the advanced query feature) enter an accession such as 4HHB for hemoglobin. This results in a Cn3D image of both structures as well as a corresponding sequence alignment (Fig. 13.15). VAST provides many kinds of structural data (Box 13.2).

### Integrated Views of Universe of Protein Folds

We have examined how to view individual proteins and how to compare small numbers of structures. Chothia (1992) predicted about 1500 folds; how many different protein folds are now thought to exist? How many structural groups are there? Kolodny *et al.* (2013) note that these questions are complicated by the challenge of defining a domain and by the sometimes complex relationship between sequence, structure, and function. Several databases have been established to explore the broad question of the total protein



**FIGURE 13.14** The Vector Alignment Search Tool (VAST) at NCBI allows the comparison of two or more structures. These may be selected by checking boxes (lower left) or by entering a specific PDB accession number (under the advanced search options). This site also provides links to data on the structures being compared (see Box 13.2) and links to the Conserved Domain Database at NCBI. In this case, a myoglobin structure (3RGK) was accessed (Fig. 13.12) and the VAST link was clicked. The first 5 of >2300 related structures are shown.

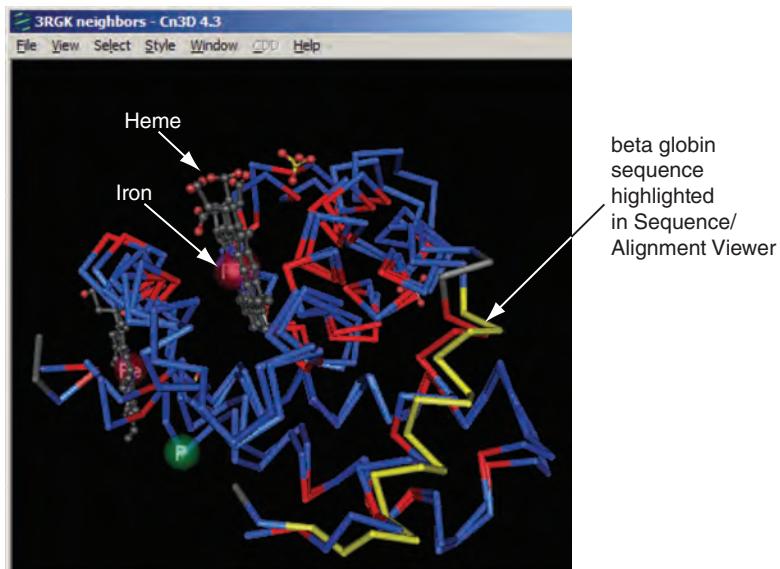
Source: Vector Alignment Search Tool (VAST), NCBI.

fold space (Andreeva and Murzin, 2010). We examine several of these databases: SCOP, CATH, and the Dali Domain Dictionary. These databases also permit searches for individual proteins. Christine Orengo and colleagues have suggested that, as the number of new folds is declining in their CATH database (with ~1300 fold groups), this represents the majority of fold groups that are easily accessible (Sillitoe *et al.*, 2013).

#### Taxonomic System for Protein Structures: SCOP Database

The Structural Classification of Proteins (SCOP) database provides a comprehensive description of protein structures and evolutionary relationships based upon a hierarchical classification scheme (Andreeva *et al.*, 2008). Recently, the SCOP-extended (SCOPe) database has maintained SCOP updates, and an entirely new SCOP2 is being introduced (see below). The SCOP database can be navigated by browsing the hierarchy, by a keyword query or PDB identifier query, or by a homology search with a protein sequence. A key feature of this database is that it has been manually curated by experts including

(a) Vector alignment search tool (VAST) alignment of myoglobin and beta globin



(b) Sequence/Alignment Viewer: corresponding amino acid sequences of myoglobin and beta globin

3RGK neighbors - Sequence/Alignment Viewer	
	View Edit Mouse Mode Unaligned Justification Imports
3RGK_A	~GLSDGEWQLVLNIVWGKV~eADIPGHGQEVLIRLFKGHPETLEKFDRFKHLKSEDEMKA
4HHB_B	vHLTPEEKSAVTLWGKV~NVDEVGGEALGRILVVYPWTQRFFESFGDLSTPDAVM

**FIGURE 13.15** Two structures that are selected in VAST (here myoglobin, 3RGK, and beta globin, 4HHB) are compared as overlaid structures in the Cn3D viewer and in the form of a sequence alignment. Despite the relatively low-sequence identity between these proteins, they adopt highly similar three-dimensional folds. The heme group and iron atoms are indicated. (b) A stretch of amino acids in the sequence viewer (yellow) corresponding to that highlighted on the corresponding structure(s) in (a).

Source: Vector Alignment Search Tool (VAST), NCBI.

## BOX 13.2 VAST INFORMATION

For each structural neighbor detected by VAST (such as Fig 13.15), the following information is listed:

- checkbox: allows for selection of individual neighbors;
- PDB: four-character PDB-identifier of the structural neighbor;
- PDB chain name;
- MMDB domain identifier;
- VAST structure similarity score based on the number of related secondary structure elements and the quality of the superposition;
- RMSD: root-mean-square superposition residual in angstroms (a descriptor of overall structural similarity);
- NRES: number of equivalent pairs of  $C\alpha$  atoms superimposed between the two structures (the alignment length, i.e., how many residues have been used to calculate the three-dimensional superposition);
- %Id: percent identical residues in the aligned sequence region;
- description: string parsed from PDB records;
- metric (Loop Hausdorff Metric): describes how well two structures match in loop regions; and
- gapped score: combines RMSD, the length of the alignment, and the number of gapped regions.

Data from VAST, NCBI (<http://www.ncbi.nlm.nih.gov/Structure/VAST/vasthelp.html#VASTTable>, WebLink 13.37).

**TABLE 13.6 Release notes from SCOPe database, release 2.03. For each fold, there are between one and dozens of superfamilies.**

Class	Number of folds	Number of proteins
All alpha proteins	284	46,456
All beta proteins	174	48,724
Alpha and beta proteins ( $\alpha/\beta$ )	147	51,349
Alpha and beta proteins ( $\alpha + \beta$ )	376	53,931
Multidomain proteins	66	56,572
Membrane and cell surface proteins	57	56,835
Small proteins	90	56,992
Coiled coil proteins	7	57,942
Low resolution protein structures	25	58,117
Peptides	120	58,231
Designed proteins	44	58,788
Total	1390	603,937

Source: SCOPe. Fox *et al.* (2014). Courtesy of SCOPe.

While SCOP was last updated in 2009, the Structural Classification of Proteins extended (SCOPe) database continued with Release 2.05 in 2015 (<http://scop.berkeley.edu/>, WebLink 13.19). The main classes in the latest release include ~1200 folds, ~2000 superfamilies, ~4500 families, and >200,000 domains. SCOPe is created and maintained by Naomi Fox, Steven Brenner, and John-Marc Chandonia.

Alexey Murzin, John-Marc Chandonia, Steven Brenner, Tim Hubbard, and Cyrus Chothia. Because of their expertise, SCOP has a reputation as being one of the most important and trusted databases for classifying protein structures. Automatic classification is now performed in SCOP, partly due to the increase in structures through structural genomics initiatives, with manual annotation for particularly difficult problems.

We explore SCOPe using a myoglobin query (3rgk) as an example. The hierarchy consists of classes; these are subsequently classified into folds, superfamilies, and families, until we reach protein domains and individual PDB protein structure entries. Beginning at the top of the hierarchy, eleven SCOPe classes are listed in Table 13.6. For myoglobin, the class is all alpha proteins (Fig. 13.16). The folds level of the hierarchy describes proteins sharing a particular secondary structure with the same arrangement and topology. Myoglobin is classified as having a globin-like fold (this is one of 284 different folds in this class). In SCOPe, different proteins with the same fold are not necessarily evolutionarily related.

As we continue down the SCOP hierarchy, we arrive at the level of the superfamily. Here proteins probably do share an evolutionary relationship, even if they share relatively low amino acid sequence identity in pairwise alignments. Myoglobin is in the globin-like superfamily, with a different, distantly related superfamily (alpha-helical ferredoxins) also categorized as having a globin-like fold. Beneath the globin-like superfamily there are five families, including one for globins including myoglobin. (The other four families include distantly related truncated hemoglobin and neural globins; Pfam links are sometimes provided.) While the superfamily level is defined by structural, functional, and sequence evidence for a common ancestor, the basis for classifying structures into SCOP families is less clear and has been examined by Pethica *et al.* (2012). SCOPe families include globins with 27 different domains such as myoglobin, beta globin, and alpha globin from various species. We can view myoglobin structures that have been determined from nine species (examples from an elephant and a seal are shown in Fig. 13.16).

SCOP will be replaced by a fully redesigned database, SCOP2, that is currently in development as a prototype (Andreeva *et al.*, 2014). Instead of a hierarchy, SCOP2 features a directed acyclic graph structure (such as that used by Gene Ontology; Chapter 12). The SCOP2 database will feature protein types (soluble, membrane, fibrous, and intrinsically disordered); evolutionary events (e.g., describing structural rearrangements);

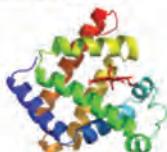
### Lineage for Protein: Myoglobin

1. Root: [SCOPe 2.03](#)
2. Class [a: All alpha proteins](#) [46456] (284 folds)
3. Fold [a.1: Globin-like](#) [46457] (2 superfamilies)
  - core: 6 helices; folded leaf, partly opened*
4. Superfamily [a.1.1: Globin-like](#) [46458] (5 families)
5. Family [a.1.1.2: Globins](#) [46463] (27 protein domains)
  - Heme-binding protein*
6. Protein Myoglobin [46469] (9 species)

### Species:

1. [Asian elephant \(Elephas maximus\)](#) [TaxId:9783] [46476] (1 PDB entry)

Domain for [1emy](#):



Domain [d1emya : 1emy A](#): [15204]  
complexed with cyn, hem

2. [Common seal \(Phoca vitulina\)](#) [TaxId:9720] [46472] (1 PDB entry)

Domain for [1mbs](#):



Domain [d1mbsa : 1mbs A](#): [15156]  
complexed with hem

**FIGURE 13.16** The Structural Classification of Proteins-extended (SCOPe) database includes a hierarchy of terms. The results of a search for myoglobin are shown, including its membership in a class (all alpha proteins), fold, superfamily, and family. Two of the myoglobin structures are shown, including their PDB accessions, species and taxonomy identifiers, domain assignment, and complexed ligands.

Source: SCOPe. Fox et al. (2014). Courtesy of SCOPe.

structural classes (based on secondary structural content); and protein relationships (e.g., structural or evolutionary). The motivation for these changes includes the need to classify proteins that are evolutionarily related but structurally distinct (and were therefore inappropriately classified as having the same fold in SCOP).

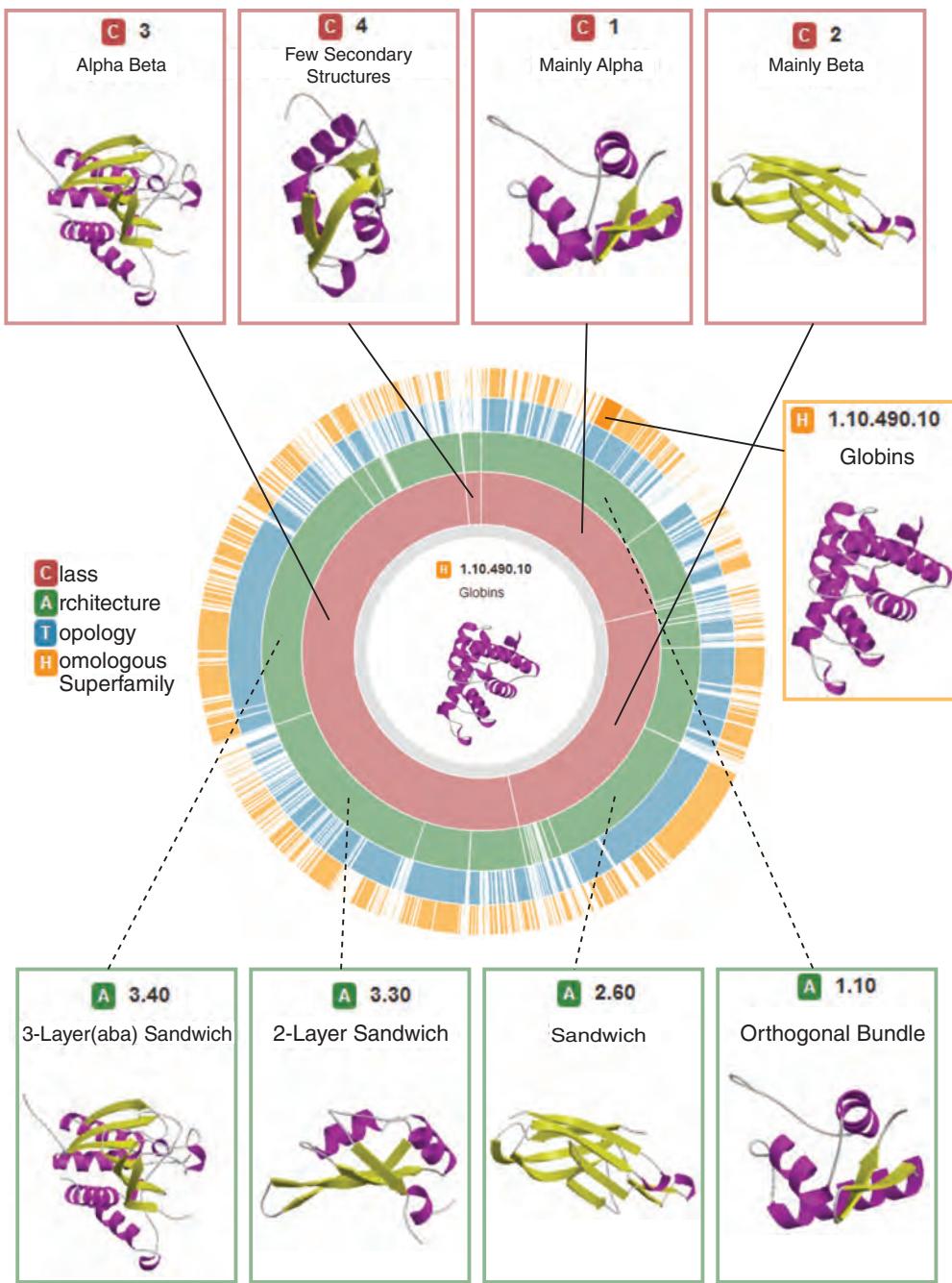
### CATH Database

CATH is a hierarchical classification system that describes all known protein domain structures (Cuff *et al.*, 2011). It has been developed by David Jones, Janet Thornton, Christine Orengo and colleagues with a particular emphasis on defining domain boundaries. While some parts of the classification system are automated, expert manual curation is also employed for tasks such as classifying remote folds and remote homologs. CATH clusters proteins at four major levels: class (C); architecture (A); topology (T); and homologous superfamily (H) (Fig. 13.17). A search of CATH with the term myoglobin (or hemoglobin) results in an output displaying the various hierarchy levels. The output for the globin superfamily includes a wealth of data on structures and their annotation (Fig. 13.18). This includes functional families (called FunFams) that rely on Gene Ontology and Enzyme Commission nomenclature to assign functions to structures (Sillitoe *et al.*, 2013).

At the highest level (class), the CATH database describes main folds based on secondary-structure prediction: mainly  $\alpha$ , mixed  $\alpha$  and  $\beta$ , and mainly  $\beta$  as well as a category of few secondary structures. Assignment at this level resembles the SCOP database

The SCOP2 website is <http://scop2.mrc-lmb.cam.ac.uk/> (WebLink 13.20). You can access data through its web browser or via SCOP2-graph, a graph-based web tool.

CATH is accessed at <http://www.cathdb.info> (WebLink 13.21). Version 4.0 includes about 235,000 domains and 2700 superfamilies from >69,000 annotated PDB structures (February 2015).



**FIGURE 13.17** The CATH resource organizes protein structures by a hierarchical scheme of class, architecture, topology (fold family), and homologous superfamily (globins are highlighted). The hierarchy can be browsed using an interactive wheel at the CATH website (<http://www.cathdb.info>).

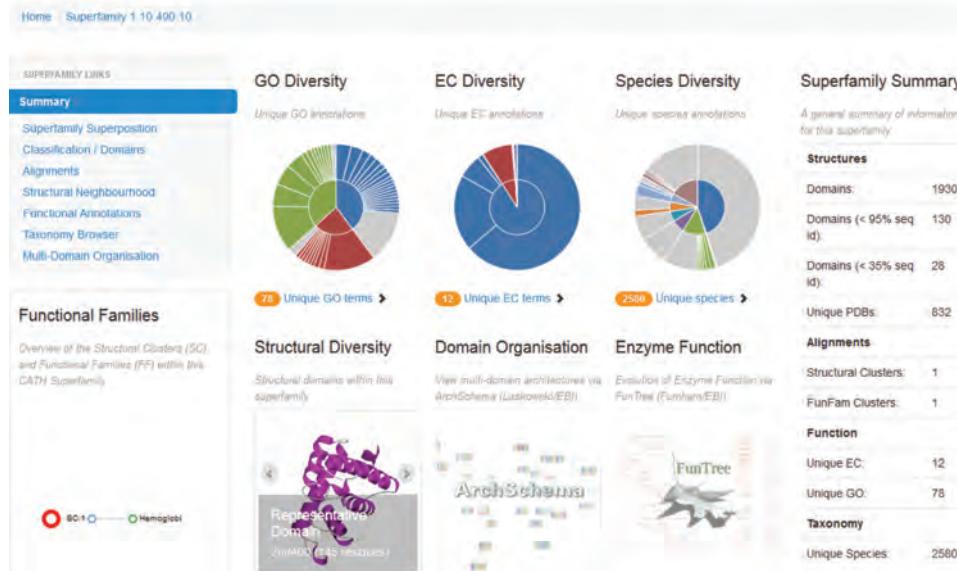
Source: CATH. Courtesy of Dr I. Sillitoe.

The SCOP classification system distinguishes alpha and beta proteins ( $\alpha/\beta$ , consisting of mainly parallel beta sheets with  $\beta-\alpha-\beta$  units) from  $\alpha+\beta$  (mainly antiparallel beta sheets, segregating  $\alpha$  and  $\beta$  regions). CATH does not make this distinction.

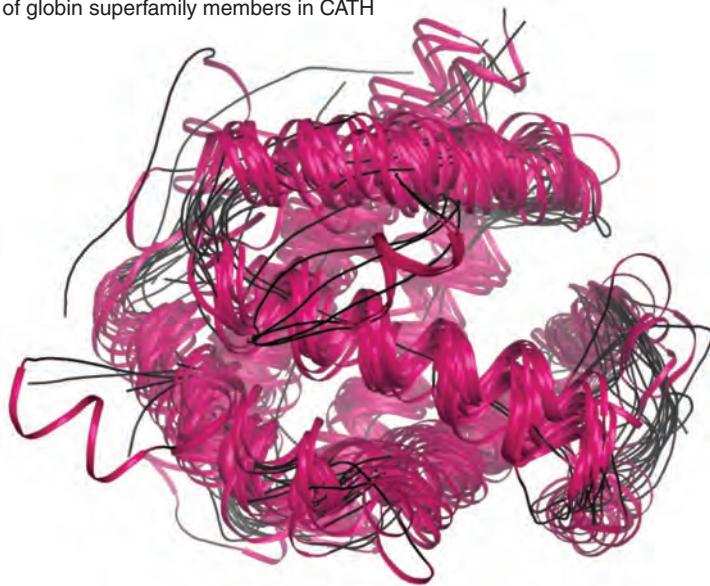
system (Table 13.6). The architecture (A) level of CATH describes the shape of the domain structure as determined by the orientations of the secondary structures. Examples are the TIM barrel (named for triose phosphate isomerase) and jelly roll. These assignments are made by expert judgment rather than by an automated process.

The topology (T) level of CATH describes fold families. Protein domains are clustered into families using several approaches including the SSAP algorithm of Taylor and Orengo (1989a, b). While at the Architecture level proteins share structural elements, they may differ in their connectivities; at the topology level structures are assembled into

## (a) CATH globin superfamily

**CATH Superfamily 1.10.490.10****Globins**

## (b) Superposition of globin superfamily members in CATH



**FIGURE 13.18** (a) A search of the CATH database with the term myoglobin (or hemoglobin) leads to a view of the globin superfamily. (b) This is richly annotated with Gene Ontology and Enzyme Commission functional assignments, and a variety of links (upper left of figure) such as superfamily superposition.

Source: CATH. Courtesy of Dr I. Sillitoe.

groups sharing both shape and connectivity. Proteins sharing topologies in common are not necessarily homologous. In contrast, the homologous superfamily (H) level clusters proteins that are likely to share homology (i.e., descent from a common ancestor).

**Dali Domain Dictionary**

Dali is an acronym for distance matrix alignment. The Dali database provides a classification of all structures in PDB and a description of families of protein sequences associated

The SSAP algorithm compares two protein structures. It can be accessed at <http://www.cathdb.info/cgi-bin/cath/GetSsapRasmol.pl> (WebLink 13.22). To compare two globins, try using 3rgk (a myoglobin structure) and 4hhbB (a beta globin structure). The output includes a PDB file, an option to launch Rasmol, and alignments in several formats.

(a) DaliLite query form  
(with myoglobin and alpha globin accessions)

**Upload first structure (mol1):**

No file selected.

**Or enter PDB identifier:**  **chain:**  (optional)

**Upload second structure (mol2):**

No file selected.

**Or enter PDB identifier:**  **chain:**  (optional)

(b) DaliLite structure comparison with Jmol



(c) DaliLite summary of results and pairwise structural alignments

## Summary

**FIGURE 13.19** The Dali server allows a comparison of two 3D structures based on analyses using distance matrices. (a) The PDB identifiers for myoglobin and beta globin are entered in the input form. (b) The output includes a pairwise structural alignment. (c) The output also includes a Z score (here a highly significant value of 21.4) based on quality measures such as: the resolution and amount of shared secondary structure; a root mean squared deviation (RMSD); percent identity; and a sequence alignment indicating secondary structure features.

Source: Holm and Rosenström (2010). Dali Server.

with representative proteins of known structure (Holm and Sander, 1993, 1996). For pairwise alignments, Dali uses a distance matrix that contains all pairwise distance scores between C $\alpha$  atoms in two structures. These scores from structural alignments are derived as a weighted sum of similarities of intramolecular distances. The Dali output reports Z scores which are useful to report biologically interesting matches of proteins, even if they are of different lengths.

Dali can be used to compare two structures with the DaliLite server (Holm *et al.*, 2006, 2008). An example is shown in **Figure 13.19** for myoglobin and beta globin. The result includes a Jmol interactive viewer that superimposes the two structures. You can also search the Dali database with a query, and browse a comprehensive classification of folds. For example, a search of the Dali fold index at the website hosted in Finland yields a classification of structural domains in PDB90 (a subset of the PDB in which no two chains share more than 90% sequence identity).

Dali is at  <http://ekhidna.biocenter.helsinki.fi/dali/start> (WebLink 13.23).

**TABLE 13.7** Partial list of protein structure databases.

Database	Comment	URL
3dee	Structural domain definitions	<a href="http://www.compbio.dundee.ac.uk/3Dee/">http://www.compbio.dundee.ac.uk/3Dee/</a>
Enzyme Structures Databases	Enzyme classifications and nomenclature	<a href="http://www.ebi.ac.uk/thornton-srv/databases/enzymes/">http://www.ebi.ac.uk/thornton-srv/databases/enzymes/</a>
FATCAT	Flexible structure alignment by chaining aligned fragment pairs allowing twists	<a href="http://fatcat.burnham.org/">http://fatcat.burnham.org/</a>
PDBeFold	Secondary-structure matching for fast protein structure alignment in three dimensions	<a href="http://www.ebi.ac.uk/msd-srv/ssm/">http://www.ebi.ac.uk/msd-srv/ssm/</a>
PDBePISA	Proteins, interfaces, structures and assemblies	<a href="http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html">http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html</a>
NDB	Database of three-dimensional nucleic acid structures	<a href="http://ndbserver.rutgers.edu/">http://ndbserver.rutgers.edu/</a>
PDBSum	Summary information about protein structures	<a href="http://www.ebi.ac.uk/pdbsum/">http://www.ebi.ac.uk/pdbsum/</a>
SWISS-MODEL Repository	Database of annotated three-dimensional comparative protein structure models	<a href="http://swissmodel.expasy.org/repository/">http://swissmodel.expasy.org/repository/</a>

## Comparison of Resources

We have described SCOP, CATH, and the Dali Domain Dictionary. Many other databases are available that classify and analyze protein structures, and some of these are listed in **Table 13.7**. It is notable that for some proteins, such as the four listed in **Table 13.8**, authoritative resources such as SCOP, CATH, and Dali-based databases provide different estimates of the number of domains in a protein (Sillitoe *et al.*, 2013 review the extent of overlap between CATH and SCOP). The field of structural biology provides rigorous measurements of the three-dimensional structure of proteins, and yet classifying domains can be a complex problem requiring expert human judgments (Kolodny *et al.*, 2013). There may be differing interpretations as to whether a particular segment of a protein exists as an independent folding unit, or whether the main principle of domain decomposition involves compactness or the density of residue-residue contacts within a putative domain (as is the case for DomainParser). SCOP is especially oriented towards classifying whole proteins, while CATH is oriented towards classifying domains.

The Genome3D project was established to facilitate comparisons between three-dimensional model predictions and structural annotation from leading resources (Lewis *et al.*, 2013). Genome3D also involves a collaboration to map common entries in the SCOP and CATH databases.

You can access DomainParser at <http://compbio.ornl.gov/structure/domainparser/> (WebLink 13.24).

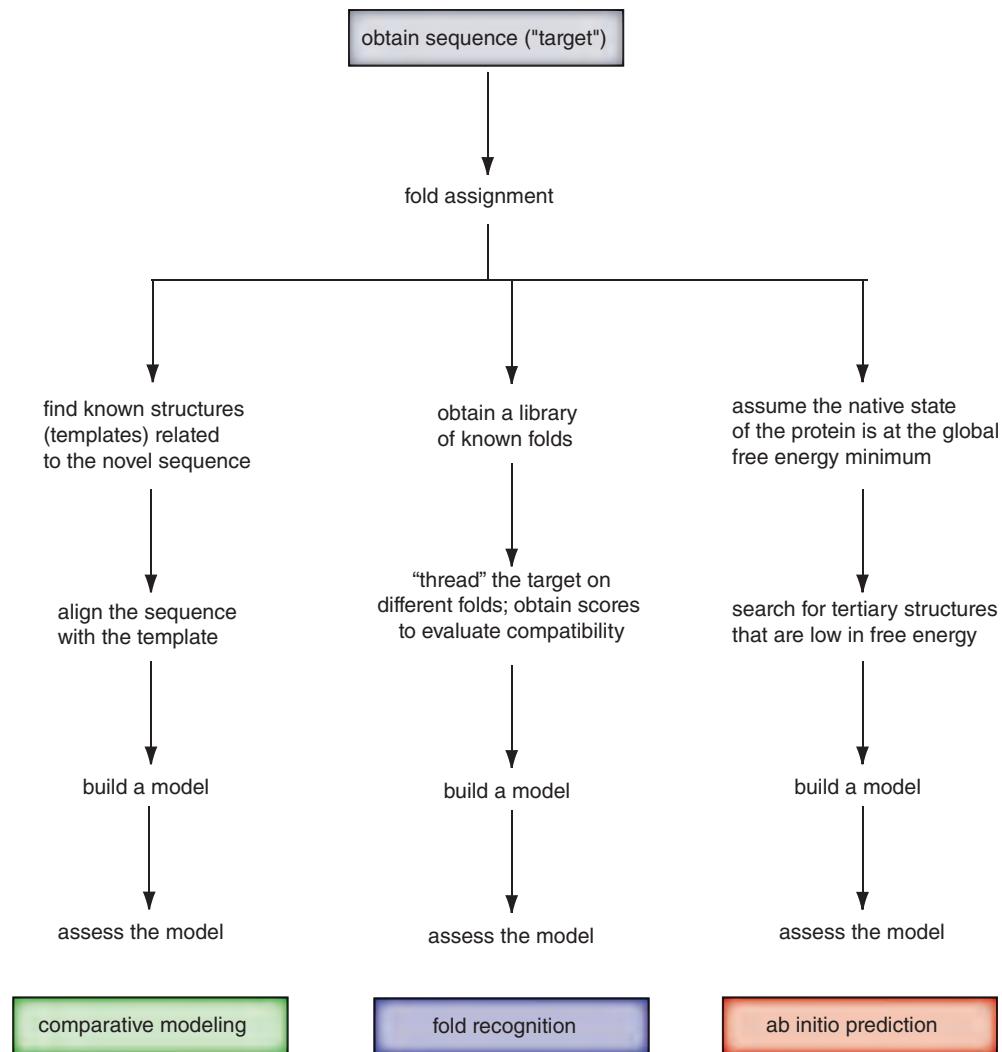
Genome3D is available at <http://genome3d.eu/> (WebLink 13.25).

## PROTEIN STRUCTURE PREDICTION

Structure prediction is a major goal of proteomics. There are three principal ways to predict the structure of a protein (**Fig. 13.20**; Cozzetto and Tramontano, 2008; Pavlopoulou and Michalopoulos, 2011). First, for a protein target that shares substantial

**TABLE 13.8** Proteins having different numbers of domains assigned by SCOP, CATH, and DALI. Values are the number of domains assigned by each database. Data from CATH, SCOP, and DALI from the Protein Data Bank (<http://www.pdb.org>).

Name	PDB accession	SCOP	CATH	DALI
Glycogen phosphorylase	1gpb	1	2	3
Annexin V	1avh_A	1	4	4
Submaxillary renin	1smr_A	1	2	1
Fructose-1,6-bisphosphatase	5fbp_A	1	2	2



**FIGURE 13.20** Approaches to predicting protein structures (adapted from Baker and Sali, 2001). Comparative modeling is the most powerful approach when a target sequence has any indications of homology with a known structure. Threading is used to compare segments of a protein to a library of known folds. In the absence of homologous structures, *ab initio* prediction is used to model protein structure. Adapted from Baker and Sali (2001).

similarity to other proteins of known structure, homology modeling (also called comparative modeling) is applied. Second, for proteins that share folds but are not necessarily homologous, threading is a major approach. Proteins that are analogous (related by convergent evolution rather than homology) can be studied this way. Third, for targets lacking identifiable homology (or analogy) to proteins of known structure, *ab initio* approaches are applied.

### Homology Modeling (Comparative Modeling)

While over 100,000 protein structures have been deposited in PDB, over half a million protein sequences have been deposited in the SwissProt database and 84 million more in TrEMBL (Chapter 12). For the vast majority of proteins, the assignment of structural models relies on computational biology approaches rather than experimental determination. As protein structures continue to be solved by X-ray crystallography and NMR

spectroscopy, the most reliable method of modeling and evaluating new structures is by comparison to previously known structures (Baker and Sali, 2001; Jones, 2001). This is the method of comparative modeling of protein structure, also called homology modeling. This method is fundamental to the field of structural genomics.

Comparative modeling consists of four sequential steps (Marti-Renom *et al.*, 2000).

1. Template selection and fold assignment are performed. This can be accomplished by searching for homologous protein sequences and/or structures with tools such as BLAST and DELTA-BLAST. The target can be queried against databases described in this chapter, such as PDB, CATH, and SCOP. As part of this analysis, structurally conserved regions and structurally variable regions are identified. It is common for structurally variable regions to correspond to loops and turns, often at the exterior of a protein.
2. The target is aligned with the template. As for any alignment problem, it is especially difficult to determine accurate alignments for distantly related proteins. For 30% sequence identity between a target and a template protein, the two proteins are likely to have a similar structure if the length of the aligned region is sufficient (e.g., more than 60 amino acids). The use of multiple sequence alignments (Chapter 6) can be especially useful.
3. A model is built. A variety of approaches are employed, such as rigid-body assembly and segment matching.
4. The model must be evaluated (see below).

There are several principal types of errors that occur in comparative modeling (see Marti-Renom *et al.*, 2000):

- errors in side-chain packing;
- distortions within correctly aligned regions;
- errors in regions of a target that lack a match to a template;
- errors in sequence alignment; and
- use of incorrect templates.

The accuracy of protein structure prediction is closely related to the percent sequence identity between a target protein and its template (Fig. 13.21). When the two proteins share 50% amino acid identity or more, the quality of the model is usually excellent. For example, the root-mean-square deviation (RMSD) for the main-chain atoms tends to be 1 Å in such cases. Model accuracy declines when comparative models rely on 30–50% identity, and the error rate rises rapidly below 30% identity. *De novo* models are able to generate low-resolution structure models.

In Chapter 3, we discussed the importance of the length of the alignment in considering percent identity between two proteins.

Many web servers offer comparative modeling including quality assessment, such as SWISSMODEL at ExPASy, MODELLER, and the PredictProtein server (Table 13.9). After a model is generated it is necessary to assess its quality. The goal is to assess whether a particular structure is likely, based on a general knowledge of protein structure principles. Criteria for quality assessment may include whether the bond lengths and angles are appropriate; whether peptide bonds are planar; whether the carbon backbone conformations are allowable (e.g., following a Ramachandran plot); whether there are appropriate local environments for hydrophobic and hydrophilic residues; and solvent accessibility. Quality assessment programs include VERIFY3D, PROCHECK, and WHATIF at CMBl (Netherlands; Table 13.9).

## Fold Recognition (Threading)

While there are currently >100,000 entries in the Protein Data Bank, there may be only 1000–2000 distinct folds in nature. Fold recognition, also called threading, is useful when

sequence identity	model accuracy	resolution	technique	applications
100%	100%	1.0 Å	X-ray crystallography, NMR	Studying catalytic mechanisms
				Designing and improving ligands
				Prediction of protein partners
				Defining antibody epitopes
50%	95%	1.5 Å	comparative protein structural modeling	Supporting site-directed mutagenesis
				Refining NMR structures
				Fitting into low-resolution electron density
30%	80%	3.5 Å	threading	Identifying regions of conserved surface residues
<<20%	80 aa	4-8 Å	de novo structure prediction	

**FIGURE 13.21** Protein structure prediction and accuracy as a function of the relatedness of a novel structure to a known template. Modified from Baker and Sali (2001). aa: amino acids. Used with permission.

Websites for fold recognition include 3D-PSSM (<http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html>, WebLink 13.26) and its successor PHYRE (<http://www.sbg.bio.ic.ac.uk/~phyre/>, WebLink 13.27), FUGUE (<http://tardis.nibio.go.jp/fugue/>, WebLink 13.28).

a target sequence of interest lacks identifiable sequence matches and yet may have folds in common with proteins of known structure. The target might assume a fold that occurs in a characterized protein because of convergent evolution, or because the two proteins are homologous but extremely distantly related. An input sequence is parsed into subfragments and “threaded” onto a library of known folds. Scoring functions allow an assessment of how compatible the sequence is with known structures. A variety of web servers provide automatic threading.

**TABLE 13.9 Websites for structure prediction by comparative modeling, and for quality assessment.**

Website	Comment	URL
3D-JIGSAW	Laboratory of Paul Bates	<a href="http://bbmm.cancerresearchuk.org/~3djigsaw/">http://bbmm.cancerresearchuk.org/~3djigsaw/</a>
Geno3D	POLE	<a href="http://pbil.ibcp.fr/htm/index.php">http://pbil.ibcp.fr/htm/index.php</a>
MODELLER	From Andrej Sali's group	<a href="http://www.salilab.org/modeller/">http://www.salilab.org/modeller/</a>
PredictProtein	Laboratory of Burkhard Rost	<a href="http://www.predictprotein.org/">http://www.predictprotein.org/</a>
SWISS-MODEL	ExPASy	<a href="http://swissmodel.expasy.org/">http://swissmodel.expasy.org/</a>
PROCHECK	Quality assessment	<a href="http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/">http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/</a>
VERIFY3D	Quality assessment	<a href="http://nihserver.mbi.ucla.edu/Verify_3D/">http://nihserver.mbi.ucla.edu/Verify_3D/</a>
WHATIF	Quality assessment	<a href="http://swift.cmbi.ru.nl/whatif/">http://swift.cmbi.ru.nl/whatif/</a>

## Ab Initio Prediction (Template-Free Modeling)

In the absence of detectable homologs, protein structure may be assessed by *ab initio* (or *de novo*) structure prediction. “*Ab initio*,” meaning “from the beginning,” is the most difficult approach to structure prediction (Osguthorpe, 2000; Simons *et al.*, 2001; Jothi, 2012). It is based on two assumptions: (1) all the information about the structure of a protein is contained in its amino acid sequence; and (2) a globular protein folds into the structure with the lowest free energy. Finding such a structure requires both a scoring function and a search strategy. While the resolution of *ab initio* methods is generally low, this approach is useful to provide structural models.

The Rosetta method is one of the most successful *ab initio* strategies (Simons *et al.*, 2001; Rohl *et al.*, 2004; Adams *et al.*, 2013). The target protein is evaluated in fragments of nine amino acids. These fragments are compared to known structures in PDB. From this analysis, structures can be inferred for the entire peptide chain. Typically, models generated with Rosetta have accuracies of 3–6 Å root mean square deviation from known structures for aligned segments of 60 or more amino acids (Rohl *et al.*, 2004). Bonneau *et al.* (2002) used the Rosetta method to model the structure of all Pfam-A sequence families (Chapter 6) for which three-dimensional structures are unknown. By calibrating their method on known structures, they estimated that for 60% of the proteins studied (80 of 131), one of the top five ranked models successfully predicted the structure within 6.0 Å RMSD.

The Robetta server from David Baker’s lab, located at <http://robbetta.bakerlab.org/> (WebLink 13.29), applies the Rosetta method (Kim *et al.*, 2004).

## A Competition to Assess Progress in Structure Prediction

How well can the community predict the structures of proteins, particularly those with novel folds? The state-of-art protein prediction is assessed by the structural genomics community at Critical Assessment of Techniques for Protein Structure Prediction (CASP; Kryshtafovych *et al.*, 2014a). This structure prediction experiment (or competition) has occurred every two years since the first competition in 1996. While 35 groups participated in CASP1, over 200 prediction servers and manual groups joined CASP10 in 2012, coming from dozens of countries. Approximately 100 experimentally determined targets were evaluated, and tens of thousands of models were deposited with a team of assessors. The structures of the targets were known but withheld from publication so that the community could perform predictions in a blind fashion (Kryshtafovych *et al.*, 2014b). Predictors consisted of either scientists who performed modeling of each target, or automatic servers that produced predictions in a short time period (48 hours) without human intervention. By 2014, CASP11 generated nearly 60,000 predictions.

The CASP targets include those that require:(1) comparative modeling with close evolutionary relationships (e.g., those identifiable by BLAST); (2) comparative modeling to distantly related targets (e.g., those requiring PSI- or DELTA-BLAST or hidden Markov models to detect relationships of a template to proteins having known structure); (3) threading; (4) template-free modeling; (5) refinement of protein models; or (6) assessment of intramolecular residue-residue contacts (Monastyrskyy *et al.*, 2014a; Nugent *et al.*, 2014; Taylor *et al.*, 2014). Kryshtafovych *et al.* (2014a) reviewed the overall progress of CASP. In its first 10 years (CASP1 through CASP5) there was substantial improvement in model quality. In the second decade, improvements through CASP10 have been more modest, with overall model accuracy being comparable to that in CASP5. There are several reasons for this. Each target undergoes comparative modeling using an existing experimental structure as a guide that may be superimposed on the target. There has been progress in the ability to identify best templates (with 10% improvement in the past decade), partly through the development of methods involving multiple templates. The increased availability of known structures has however (surprisingly) made it more difficult to identify best templates in some cases. Major challenges include: the need for

improved alignments; the need for models of close evolutionary relationships to approach the accuracy obtained by experimental structure determination; the need to better refine models of remote evolutionary relationships; and the need to discriminate among the best template-free models (Moult, 2005; Tai *et al.*, 2005; Moult *et al.*, 2007).

The CASP website provides detailed results of the competition. One criterion for the accuracy of a prediction is the GDT\_TS metric which compares the difference in position of the main chain C $\alpha$  atoms in a model relative to the position in the experimentally determined structure. **Figure 13.22** shows examples of an easy protein target from CASP10 that was solved by most groups (**Fig. 13.22a**) and a difficult target that no group solved (**Fig. 13.22b**). **Figure 13.22c** depicts an example of a target that was aligned either very well or very poorly by many groups; those with poor results misaligned the sequence of the target, highlighting the difficulty of correctly aligning a target sequence onto available template structures for template-based models.

The Protein Structure Prediction Center organizes CASP information (<http://predictioncenter.org/>, WebLink 13.30) including results from each CASP competition.

## INTRINSICALLY DISORDERED PROTEINS

Jane Dyson and Peter Wright (2006) wrote an article entitled “According to current textbooks, a well-defined three-dimensional structure is a prerequisite for the function of a protein. Is this correct?” Many proteins do not adopt stable three-dimensional structures, and this may be an essential aspect of their ability to function properly. Intrinsically disordered proteins are defined as having unstructured regions of significant size such as at least 30 or 50 amino acids (Dyson and Wright, 2005; Le Gall *et al.*, 2007; Radivojac *et al.*, 2007; Babu *et al.*, 2012; Bellay *et al.*, 2012). Such regions do not adopt a fixed three-dimensional structure under physiological conditions, but instead exist as dynamic ensembles in which the backbone amino acid positions vary over time without adopting stable equilibrium values.

Keith Dunker and colleagues have estimated that about 10% of the PDB proteins have disordered regions longer than 30 amino acids (Le Gall *et al.*, 2007). Only ~7% of the protein structures in PDB correspond to the full-length sequence in Swiss-Prot (and only ~25% of the proteins correspond to the structures that match >95% of the length of the protein in Swiss-Prot). The lack of full-length sequences among proteins with solved structures may reflect the common occurrence of intrinsic disorder. Furthermore, these authors suggest that >25% of the proteins in SwissProt have disordered regions. DisProt, the Database of Disordered Proteins, centralizes information on this class of proteins (Sickmeier *et al.*, 2007). The Protein Structure Initiative has encountered great difficulty in obtaining many crystal structures, and analyses by Johnson *et al.* (2012) indicate that many of those challenging proteins contain long intrinsically disordered regions. Similarly, such disordered regions present challenges in the CASP experiments (Monastyrskyy *et al.*, 2014b).

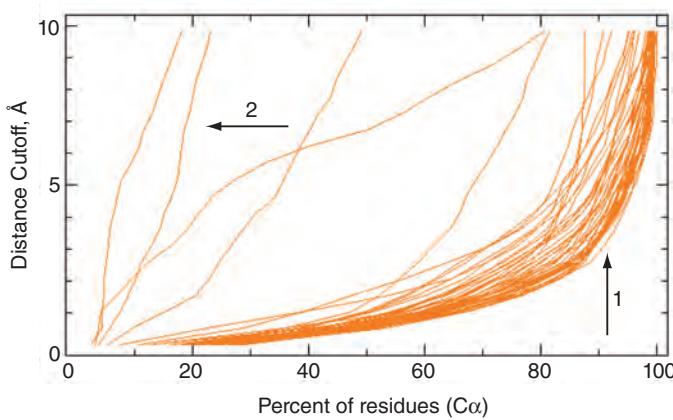
The Database of Intrinsic Disorder is available at <http://www.disprot.org/> (WebLink 13.31). As of March 2015 it includes ~700 proteins and >1500 disordered regions.

Intrinsically disordered regions may have important cellular functions (Babu *et al.*, 2012). They may change conformation upon binding to a biological target (a ligand) in a process in which folding and binding are coupled. Many disordered regions of proteins are highly conserved, consistent with their having functionally important roles. Dunker *et al.* (2005) discuss the role of intrinsic disorder in protein–protein interaction networks, in which it is thought that the average protein has few connections but “hub” proteins serve central roles with many (tens to hundreds) of links. Intrinsic disorder in hub proteins could facilitate their ability to bind to structurally diverse protein partners.

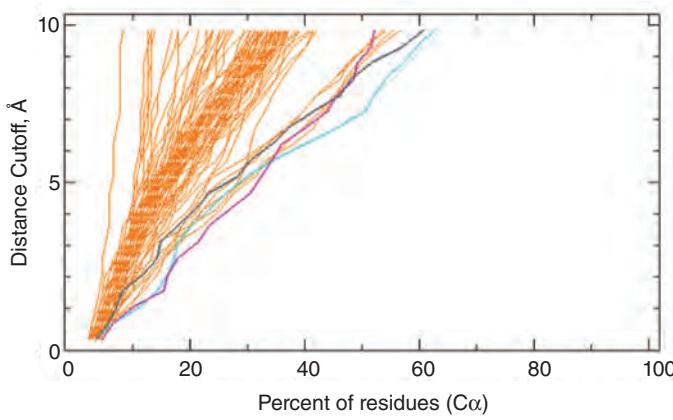
## PROTEIN STRUCTURE AND DISEASE

The linear sequence of amino acids specifies the three-dimensional structure of a protein. A change in even a single amino acid can cause a profound disruption in structure. For example, cystic fibrosis is caused by mutations in the gene-encoding

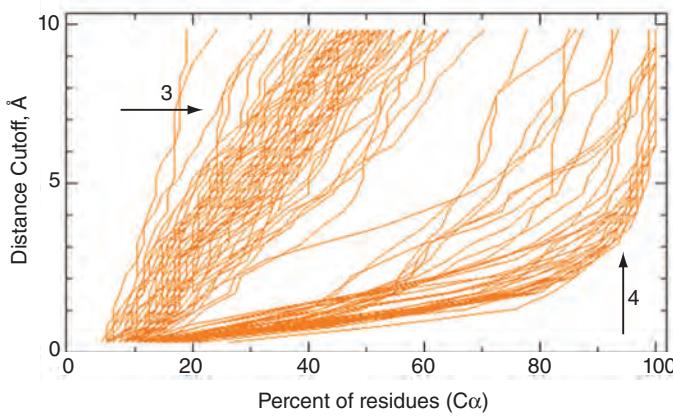
(a) CASP10 target T0645-D1: solved by most teams



(b) CASP10 target T0658-D1: not solved by any team



(c) CASP10 target T0651-D1: solved by many teams, misaligned by many teams



**FIGURE 13.22** Examples of results from the CASP10 competition. Each plot (called a GDT plot or “Hubbard plot”) shows the percent of CA or C $\alpha$  residues (i.e., the percent of the modeled structure; x axis) versus the distance cutoff in Ångstroms (from 0 Å to 10 Å; y axis). Each line represents a summary of a single prediction of that protein’s structure; multiple lines are from the many groups that submitted predictions. (a) Example of a protein target (T0645) whose structure was modeled extremely well by many teams participating in the CASP competition. Note that a very high percentage of the residues in the predictions that could be overlaid on the correct structure (x axis values approaching 100%) with only a very small RMSD (distance cutoff, y axis) as indicated by arrow 1. A small number of predictions were wrong (arrow 2) because they correctly matched the true structure over only a small percent of residues even at large distance cutoffs. (b) Example of a protein target (T0658) whose true structure was not predicted by any group in the CASP competition. Several groups’ predictions (colored lines from the Seok, Jiang, and Zhang groups) were better than all others. (c) Example of a target (T0651) that was predicted incorrectly by many teams (arrow 3) but correctly by others (arrow 4). Such a broad discrepancy in prediction accuracy is often attributable to incorrect sequence alignments in homology modelling.

Source: CASP10 results at <http://www.predictioncenter.org>. Reproduced with permission from University of California, Davis.

cystic fibrosis transmembrane regulator (CFTR; Ratjen and Döring, 2003). The most common mutation is  $\Delta F508$ , a deletion of a phenylalanine at position 508. The consequence of removing this residue is to alter the alpha helical content of the protein (Massiah *et al.*, 1999). This in some way impairs the ability of the CFTR protein to traffic through the secretory pathway to its normal location on the plasma membrane of lung epithelial cells.

Changes in protein sequence that are associated with disease do not necessarily cause large changes in protein structure. An example is provided by sickle cell anemia (Online Mendelian Inheritance in Man or OMIM #603903), the most common inherited blood disorder. It is caused by mutations in the gene encoding beta globin on chromosome 11p15.4. Adult hemoglobin is a tetramer consisting of two alpha chains and two beta chains. The protein carries oxygen in blood from the lungs to various parts of the body. A substitution of a valine for a normally occurring glutamic acid residue forms a hydrophobic patch on the surface of the beta globin, leading to clumping of many hemoglobin molecules.

Many human diseases are associated with defective protein folding. This may lead to a toxic gain of function as is thought to occur in Alzheimer's disease (OMIM #104300), Parkinson's disease, Huntington's disease, and prion diseases (Hartl and Hayer-Hartl, 2009). Several examples of proteins associated with human disease are listed in **Table 13.10**, including CFTR and beta globin.

Earlier in this chapter we described the high-resolution structure of a G protein-coupled receptor (GPCR). Of all 21,000 drugs listed by the Food and Drug Administration, there are >1300 unique drugs which affect just 324 drug targets (Pitt *et al.*, 2009). Half of all drugs act upon four protein families: GPCRs; nuclear receptors; ligand-gated ion channels; and voltage-gated ion channels. PDB contains structures for >100 of these targets.

David Baker and 70 colleagues performed a community experiment to assess the ability to predict the effects of mutations on protein–protein interactions (Moretti *et al.*, 2013). They designed two proteins capable of binding to hemagglutinin from influenza virus (Chapter 16), then created single point mutant variants corresponding to all 20 amino acids at all positions of these short proteins. Computational predictions of the effects of mutations on protein–protein binding were compared to experimentally derived measurements. About a third of the mutations associated with increased binding were identified (at a 10% false discovery rate). The factors that led to the most accurate predictions included consideration of protein stability, packing, electrostatics, and solvation. As more such datasets become available, prediction methods can be expected to improve.

**TABLE 13.10 Examples of proteins associated with diseases for which subtle change in protein sequence leads to change in structure. CFTR: cystic fibrosis transmembrane regulator. Note that OMIM refers to the disease entry (rather than the protein entry) and PDB refers to the accession of an example of the protein structure, cited on the NCBI Protein site.**

Disease	OMIM	Gene/Protein	RefSeq	PDB
Alzheimer disease	#104300	Amyloid precursor protein	NP_000475.1	2M4J
Cystic fibrosis	#219700	CFTR	NP_000483.3	2LOB
Huntington disease	#143100	Huntingtin	NP_002102.4	4FED
Creutzfeldt-Jakob disease	#123400	Prion protein	NP_000302.1	2M8T
Parkinson disease	#168600	alpha-synuclein isoform NACP140	NP_000336.1	2M55
Sickle cell anemia	#603903	Hemoglobin beta	NP_000509.1	2M6Z

These studies are relevant to the interpretation of the clinical significance of single-nucleotide variations in human genomes, as many such variants are expected to have deleterious effects on both binding interactions and hence on fitness.

## PERSPECTIVE

The aim of structural genomics is to define structures that span the entire space of protein folds. This project has many parallels to the Human Genome Project. Both are ambitious endeavors that require the international cooperation of many laboratories. Both involve central repositories for the deposit of raw data, and in each the growth of the databases is exponential.

It is realistic to expect that the great majority of protein folds will be defined in the near future. Each year, the proportion of novel folds declines rapidly. A number of lessons are emerging:

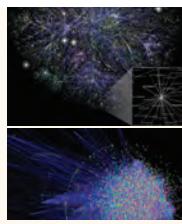
- proteins assume a limited number of folds;
- a single three-dimensional fold may be used by proteins to perform entirely distinct functions; and
- the same function may be performed by proteins using entirely different folds.

## PITFALLS

One of the great mysteries of biology is how the linear amino acid sequence of a protein folds quickly into the correct three-dimensional conformation. One set of challenges concerns the experimental solution of three-dimensional structures that span the extent of sequence space. At the present time, no representative structures have been solved for thousands of protein families. Another set of challenges concerns protein structure prediction. While structures can be predicted with high confidence when a closely related template of known structure is available, it is still difficult to predict entirely novel protein structures. *Ab initio* methods are continually improving, particularly for predicting the structures of small proteins.

## ADVICE FOR STUDENTS

Choose one protein of interest that has a known structure and has been described in the literature, and analyze it in depth. An example is a neurotransmitter receptor that functions as a chloride channel (PDB structure 3rhw; Hibbs and Gouaux, 2011; see computer lab problem (13.5) below). Follow the principles of primary, secondary, tertiary, and quaternary structure. Try to reproduce the figures in the paper(s) you choose. Compare its structure to other known structures in databases (e.g., via BLAST) and perform direct structural comparisons. Explore its folds, domains, and other features from SCOP and CATH.



## Discussion Questions

**[13-1]** The Protein Data Bank (PDB) is the central repository of protein structure data. What do databases such as SCOP and CATH offer that PDB lacks?

**[13-2]** A general rule is that protein structure evolves more slowly than primary amino sequence. Two proteins can therefore have only limited amino acid sequence iden-

tity, while sharing highly similar structures. (A good example of this is the lipocalins, where retinol-binding protein, odorant-binding protein, and  $\beta$ -lactoglobulin share highly related structures with low sequence identity.) Are there likely to be exceptions to this general rule?

### PROBLEMS/COMPUTER LAB

**[13-1]** View the structure of a protein using Cn3D at NCBI. (1) Download Cn3D from the NCBI Structure site (<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>, WebLink 13.32). (2) Go to NCBI Entrez Structures and select a lipocalin. You can access this from the main NCBI page by navigating to “structure.” Alternatively, in Entrez you can type a query, select “limits,” and restrict the output to PDB. If you select “odorant-binding protein,” there are entries for odorant-binding proteins from several different species. From cow, there are entries deposited independently from different research groups (e.g., PDB identifiers 1OBP, 1PBO). (3) Select “View 3D Structure” in the MMDB web page. Explore the links on the page. Click “View/Save Structure.” (4) Two windows open: the Cn3D viewer and the 1D-viewer. Click on each of these, and notice how they are interconnected. Change the “style” of the Cn3D viewer. Identify the  $\alpha$  helices and  $\beta$  sheets of the protein.

**[13-2]** View the structure of a protein using Jmol at PDB. (1) Go to <http://www.pdb.org> (WebLink 13.33) and enter the term 4HHB (for hemoglobin) in the search box. Note that the title of this page is “the crystal structure of human deoxyhaemoglobin at 1.74 angstroms resolution.” An icon at the top includes the option to download the PDB file to your desktop; by doing this you can easily load the 4HHB file into other programs later. Next, under the heading “display options” click Jmol. (2) The Jmol program opens (running Java) without the need to install software locally. There are pull-down menus and command tabs along the side and bottom of the image of hemoglobin, including a help document. Explore its dozens of features including viewing options.

**[13-3]** View the structure of a protein using DeepView at ExPASy. (1) Visit the website for DeepView, the Swiss PDB Viewer, at <http://expasy.org/spdbv/> (WebLink 13.34). Select download and install the software locally. (2) Open the file 3RGK (a myoglobin PDB file). You can find this by visiting PDB (<http://www.pdb.org>), querying 3RGK, and downloading the PDB file to your desktop. There is a main toolbar (see Fig. 13.2b); use its File → Open command. (3) Under the Window pull-down menu, open the control panel. Click the column header “show” to deselect all the amino acid residues, then click the first two to view just them. On the main toolbar, click the  $\omega$ ,  $\phi$ ,  $\psi$  button (see Fig. 13.2b) to view the bond angles. (4) Compare the structures of two lipocalins using VAST at NCBI:

- Go back to the MMDB page for 1PBO and select “Structure neighbors.” (This can be accessed by mousing over the protein graphic.) You are now looking at the

NCBI VAST (Vector Alignment Search Tool) site. There is a list of proteins related to OBP. Select one or two other proteins, such as  $\beta$ -lactoglobulin or retinol-binding protein, by clicking on the box(es) to the left. Now view/save the alignments.

- Notice that two windows open up: Cn3D and DDV (the two-dimensional viewer). Again, explore the relationship between these two visualization tools. What are the similarities between the proteins you are comparing? What are their differences? Highlight the regions of conserved amino acids both in the alignment viewer and the graphical viewer. Where are the invariant GXW residues located?

**[13-4]** Compare the structures of two homologous proteins using Dali at <http://ekhidna.biocenter.helsinki.fi/dali/start> (WebLink 13.35). Try structures such as 1PBO (for an odorant-binding protein) and 1RBP (for retinol-binding protein). Are the structures significantly related? By what criteria? Are the sequences significantly related, and by what criteria?

**[13-5]** This problem involves identifying a known structure related to your sequence of interest, and then modeling the structure of your protein. A child has seizures and intellectual disability. To try to find the genetic cause you sequence the exome of the child and his parents, and find a *de novo* mutation (M79nn) in GABRB encoding the beta subunit of the GABA receptor. (1) Find a related structure by identifying the GABRB protein sequence at NCBI, and performing a BLASTP search restricted to the PDB. Alternatively, you can go to PDB and perform a database search. (2) Visit the SWISS-MODELLER (via ExPASy), register, and go to the SwissModel Automatic Modelling Mode. There, paste in the FASTA-formatted GABRB protein, and specify the PDB accession of the closest related known structure (e.g., 3rhw chain A).

**[13-6]** Titin is the largest human protein (>34,000 amino acids). What is known about its structure, including domains? Try the following:

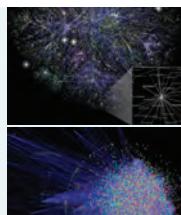
- the NCBI structure page;
- the PDB (see <http://www.rcsb.org/pdb/101/motm.do?momID=185>);
- CATH or SCOP;
- a BLASTP search against the PDB at the NCBI website.

**[13-7]** Sickle-cell anemia is caused by a specific mutation in HBB, E7V (i.e., a glutamic acid residue at amino acid position 7 is substituted with a valine). As a consequence of this mutation, hemoglobin tetramers can clump together. This causes the entire red blood cell to deform, adopting a sickled shape. Use PDB identifier 4HHB for wildtype

hemoglobin and 2HBS for a mutant form. Compare the structures using the VAST tool at NCBI. Is the glutamate at position 7 on the surface of the protein or is it buried inside? Does the mutation to a valine cause a change in the predicted secondary or tertiary structure of the protein?

**[13-8]** One way to obtain a list of identifiers for protein structures is by using a Perl script to query NCBI databases. For those without experience of writing Perl scripts, NCBI offers an interactive web tool called EBot. This constructs an E-utility pipeline. (1) Visit the EBot web page (<http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi>, WebLink 13.36). (2) Enter your email address (note that you could have selected from dozens of databases to

begin). (3) Enter a PubMed text query, Perutz M[Au]. This should limit results to articles by Max Perutz. Click “Add Step to Pipeline.” (4) Choose “Link the entire dataset to one set of related records (elink)” and “Build Step.” (5) Scroll down to select “Structure Links” then “Add Step to Pipeline.” (6) Choose “Stop here and download the UIDs” then Build Step. (7) Provide an output file name (ebot\_globins) and “End Pipeline.” (8) Choose a file name (e.g., ebot1.pl) and generate the Perl script. (9) Save the script to your computer, and run it by typing perl in Windows (use the command prompt), Mac OS/X (use the terminal), or Linux (use the shell). A copy of this Perl script is available as a text file (Web Document 13.2).



## Self-Test Quiz

**[13-1]** In comparing two homologous but distantly related proteins, which of the following is true?

- (a) They tend to share more three-dimensional structure features in common than percent amino acid identity.
- (b) They tend to share more percent amino acid identity in common than three-dimensional structure features.
- (c) They tend to share three-dimensional structure features and percent amino acid identity to a comparable extent.
- (d) It is not reasonable to generalize about the extent to which they share three-dimensional structure features and percent amino acid identity.

**[13-2]** Protein secondary structure prediction algorithms typically calculate the likelihood that a protein forms:

- (a)  $\alpha$  helices;
- (b)  $\alpha$  helices and  $\beta$  sheets;
- (c)  $\alpha$  helices,  $\beta$  sheets, and coils; or
- (d)  $\alpha$  helices,  $\beta$  sheets, coils, and multimers.

**[13-3]** An advantage of X-ray crystallography relative to NMR for structure determination is that when using X-ray crystallography it is easier to:

- (a) solve the structure of transmembrane domain-containing proteins;
- (b) grow crystals than prepare samples for NMR;
- (c) interpret diffraction data; or
- (d) determine the structures of large proteins.

**[13-4]** The Protein Data Bank (PDB):

- (a) functions primarily as the major worldwide repository of macromolecular secondary structures;
- (b) contains approximately as many structures as there are protein sequences in SwissProt/TrEMBL;
- (c) includes data on proteins, DNA–protein complexes, as well as carbohydrates; or
- (d) is operated jointly by the NCBI and EBI.

**[13-5]** The NCBI VAST algorithm:

- (a) is a web browser tool for the visualization of related protein structures by threading;
- (b) is a visualization tool that allows the simultaneous comparison of as many as two structures;
- (c) allows searches of all the NCBI structure database with queries that have known structures (i.e., having PDB accession numbers), but this tool is not useful for the analysis of uncharacterized structures; or
- (d) allows searches of all the NCBI structure database entries against each other and provides a list of “structure neighbors” for a given query.

**[13-6]** Cn3D is a molecular structure viewer at NCBI. Its features:

- (a) a menu-driven program linked to automated homology modeling;
- (b) a command line interface useful for a variety of structure analyses;
- (c) a structure viewer that is accompanied by a sequence viewer; or
- (d) a structure viewer that allows stereoscopic viewing of structure images.

**[13-7]** The CATH database offers a hierarchical classification of protein structures. The first three levels, class (C), architecture (A), and topology (T), all describe:

- (a) protein tertiary structure (e.g., tertiary structure composition, packing, shape, orientation, and connectivity);
- (b) protein secondary structure (e.g., secondary structure composition, packing, shape, orientation, and connectivity);
- (c) protein domain structure; or
- (d) protein superfamilies grouped according to homologous domains.

**[13-8]** Homology modeling may be distinguished from *ab initio* prediction because:

- (a) homology modeling requires a model to be built;
- (b) homology modeling requires alignment of a target to a template;

(c) homology modeling is usefully applied to any protein sequence; or

(d) the accuracy of homology modeling is independent of the percent identity between the target and the template.

**[13-9]** You have a protein sequence and you want to quickly predict its structure. After performing BLAST and DELTA-BLAST searches, you identify the most closely related proteins with a known structure as having 15% amino acid identity to your protein with a nonsignificant expect value. Which of these options is best?

- (a) X-ray crystallography;
- (b) NMR;
- (c) submitting your sequence to a protein structure prediction server that performs homology modeling; or
- (d) submitting your sequence to a protein structure prediction server that performs *ab initio* modeling.

## SUGGESTED READING

There are many superb overviews of structural genomics and protein structure prediction. For overviews on protein folding, see Dill *et al.* (2008), Fersht (2008) and Hartl and Hayer-Hartl (2009). The central structure repository PDB is described by Berman *et al.* (2013a). Structural genomics initiatives are reviewed and analyzed by Andreeva and Murzin (2010), Marsden *et al.* (2007), Chandonia and Brenner (2006), Levitt (2007), and others cited above. Michael Levitt and colleagues (Kolodny *et al.*, 2013) have written a clear, thought-provoking review on the universe of protein folds including the problem of classifying structures according to domains. Very early but useful reviews are by Holm and Sander (1996, 1997).

Jenny Gu and Philip Bourne have edited an excellent textbook, *Structural Bioinformatics* (2009).

## REFERENCES

- Adams, P.D., Baker, D., Brunger, A.T. *et al.* 2013. Advances, interactions, and future developments in the CNS, Phenix, and Rosetta structural biology software systems. *Annual Review of Biophysics* **42**, 265–287. PMID: 23451892.
- Andreeva A., Murzin, A.G. 2010. Structural classification of proteins and structural genomics: new insights into protein folding and evolution. *Acta Crystallographica Section F Structural Biology and Crystallization Communications* **66**(Pt 10), 1190–1197. PMID: 20944210.
- Andreeva, A., Howorth, D., Chandonia, J.M. *et al.* 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research* **36**(Database issue), D419–425.
- Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., Murzin, A.G. 2014. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* **42**(Database issue), D310. PMID: 24293656.
- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**, 223–230.
- Babu, M.M., Kriwacki, R.W., Pappu, R.V. 2012. Structural biology. Versatility from protein disorder. *Science* **337**(6101), 1460–1461. PMID: 22997313.

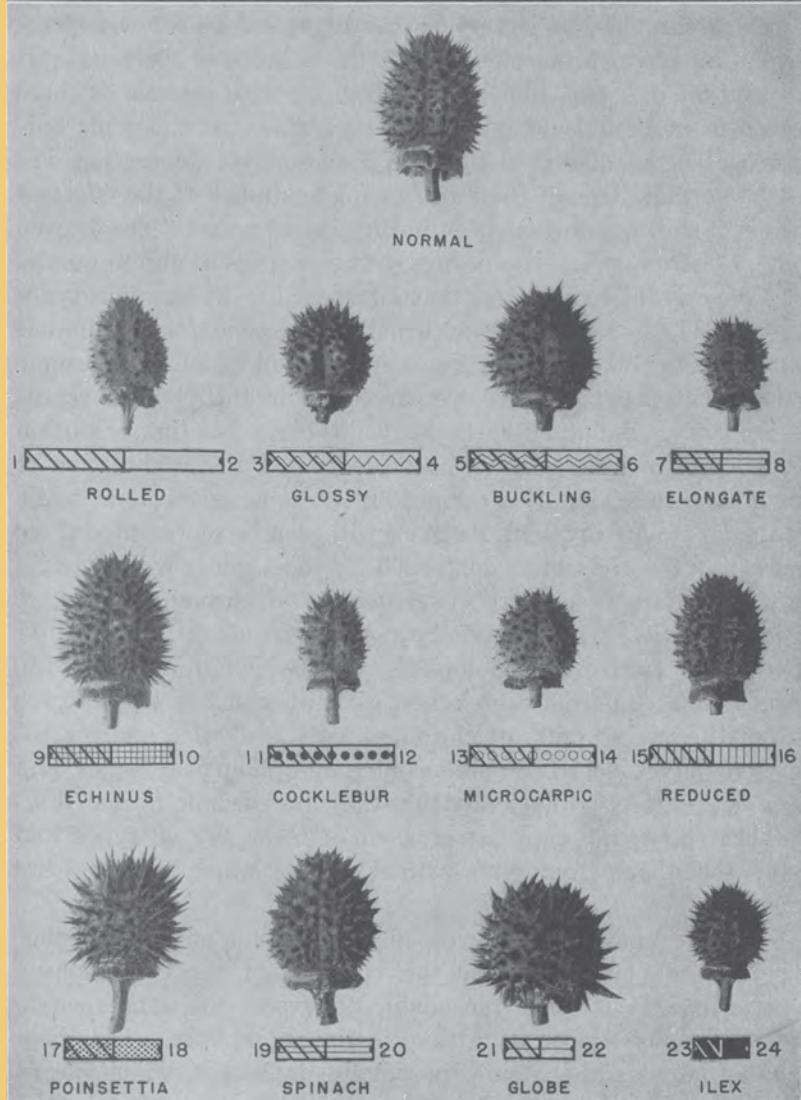
- Baker, D., Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294**, 93–96.
- Bakolitsa, C., Kumar, A., Jin, K.K. et al. 2010. Structures of the first representatives of Pfam family PF06684 (DUF1185) reveal a novel variant of the Bacillus chorismate mutase fold and suggest a role in amino-acid metabolism. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **66**(Pt 10), 1182–1189. PMID: 20944209.
- Bellay, J., Michaut, M., Kim, T. et al. 2012. An omics perspective of protein disorder. *Molecular Biosystems* **8**(1), 185–193. PMID: 22101230.
- Berman, H.M. 2012. Creating a community resource for protein science. *Protein Science* **21**(11), 1587–1596. PMID: 22969036.
- Berman, H.M., Coimbatore Narayanan, B., Di Costanzo, L. et al. 2013a. Trendspotting in the Protein Data Bank. *FEBS Letters* **587**(8), 1036–1045. PMID: 23337870.
- Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L. 2013b. How community has shaped the Protein Data Bank. *Structure* **21**(9), 1485–1491. PMID: 24010707.
- Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L. 2013c. The future of the protein data bank. *Biopolymers* **99**(3), 218–222. PMID: 23023942.
- Bonneau R., Strauss C. E., Rohl C. A. et al. 2002. De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology* **322**, 65–78.
- Boutet, S., Lomb, L., Williams, G.J. et al. 2012. High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337**(6092), 362–364. PMID: 22653729.
- Branden, C., Tooze, J. 1991. *Introduction to Protein Structure*. Garland Publishing, New York.
- Brenner, S. E. 2000. Target selection for structural genomics. *Nature Structural Biology* **7** (Suppl.), 967–969.
- Brenner, S.E. 2001. A tour of structural genomics. *Nature Reviews Genetics* **2**, 801–809.
- Carter, P., Lee, D., Orengo, C. 2008. Target selection in structural genomics projects to increase knowledge of protein structure and function space. *Advances in Protein Chemistry and Structural Biology* **75**, 1–52. PMID: 20731988.
- Chandonia, J.M., Brenner, S.E. 2005. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **58**, 166–179.
- Chandonia, J.M., Brenner, S.E. 2006. The impact of structural genomics: expectations and outcomes. *Science* **311**, 347–351.
- Chang, G., Roth, C. B. 2001. Structure of MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**, 1793–800.
- Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543–544.
- Chou, P. Y., Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology* **47**, 45–148 (1978).
- Cozzetto, D., Tramontano, A. 2008. Advances and pitfalls in protein structure prediction. *Current Protein and Peptide Science* **9**(6), 567–577 (2008). PMID: 19075747.
- Cuff, A.L., Sillitoe, I., Lewis, T. et al. 2011. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research* **39**(Database issue), D420–426. PMID: 21097779.
- Depietro, P.J., Julfayev, E.S., McLaughlin, W.A. 2013. Quantification of the impact of PSI: Biology according to the annotations of the determined structures. *BMC Structural Biology* **13**(1), 24. PMID: 24139526.
- Dill, K.A., MacCallum, J.L. 2012. The protein-folding problem, 50 years on. *Science* **338**(6110), 1042–1046. PMID: 23180855.
- Dill, K.A., Ozkan, S.B., Shell, M.S., Weikl, T.R. 2008. The protein folding problem. *Annual Review of Biophysics* **37**, 289–316. PMID: 18573083.
- Domingues, F. S., Koppensteiner, W. A., Sippl, M. J. 2000. The role of protein structure in genomics. *FEBS Letters* **476**, 98–102.
- Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., Uversky, V.N. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS Journal* **272**, 5129–5148.

- Dyson, H.J., Wright, P.E. 2005. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology* **6**, 197–208.
- Dyson, H.J., Wright, P.E. 2006. According to current textbooks, a well-defined three-dimensional structure is a prerequisite for the function of a protein. Is this correct? *IUBMB Life* **58**, 107–109.
- Fersht, A.R. 2008. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. *Nature Reviews Molecular Cell Biology* **9**(8), 650–654. PMID: 18578032.
- Fox, N.K., Brenner, S.E., Chandonia, J.M. 2014. SCOPe: Structural Classification of Proteins: extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* **42**(Database issue), D304–309. PMID: 24304899.
- Garnier, J., Gibrat, J.F., Robson, B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology* **266**, 540–553.
- Gifford, L.K., Carter, L.G., Gabanyi, M.J., Berman, H.M., Adams, P.D. 2012. The Protein Structure Initiative Structural Biology Knowledgebase Technology Portal: a structural biology web resource. *Journal of Structural and Functional Genomics* **13**(2), 57–62. PMID: 22527514.
- Goodsell, D.S., Burley, S.K., Berman, H.M. 2013. Revealing structural views of biology. *Biopolymers* **99**(11), 817–824. PMID: 23821527.
- Gu, J., Bourne, P.E. (eds). 2009. *Structural Bioinformatics*. Second edition. Hoboken, NJ, Wiley-Blackwell.
- Han, G.W., Bakolitsa, C., Miller, M.D. et al. 2010. Structures of the first representatives of Pfam family PF06938 (DUF1285) reveal a new fold with repeated structural motifs and possible involvement in signal transduction. *Acta Crystallography Section F: Structural Biology and Crystallization Communications* **66**(Pt 10), 1218–1225. PMID: 20944214.
- Hartl, F.U., Hayer-Hartl, M. 2009. Converging concepts of protein folding in vitro and in vivo. *Nature Structural and Molecular Biology* **16**(6), 574–581. PMID: 19491934.
- Henderson, R., Unwin, P.N. 1975. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**(5521), 28–32. PMID: 1161000.
- Hibbs, R.E., Gouaux, E. 2011. Principles of activation and permeation in an anion-selective Cys-loop receptor. *Nature* **474**(7349), 54–60. PMID: 21572436.
- Holm, L. 1998. Unification of protein families. *Current Opinion in Structural Biology* **8**, 372–379.
- Holm, L., Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* **233**, 123–138.
- Holm, L., Sander, C. 1996. Mapping the protein universe. *Science* **273**, 595–603.
- Holm, L., Sander, C. 1997. New structure: novel fold? *Structure* **5**, 165–171.
- Holm, L., Rosenström, P. 2010. Dali server: conservation mapping in 3D. *Nucleic Acids Research* **38**(Web Server issue), W545–549. PMID: 20457744.
- Holm, L., Kääriäinen, S., Wilton, C., Plewczynski, D., Wilton, C. 2006. Using Dali for structural comparison of proteins. *Current Protocol in Bioinformatics Chapter 5*, Unit 5.5. PMID: 18428766.
- Holm, L., Kääriäinen, S., Rosenström, P., Schenkel, A. 2008. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **24**(23), 2780–2781. PMID: 18818215.
- Johnson, D.E., Xue, B., Sickmeier, M.D. et al. 2012. High-throughput characterization of intrinsic disorder in proteins from the Protein Structure Initiative. *Journal of Structural Biology* **180**(1), 201–215. PMID: 22651963.
- Jones, D.T. 2001. Protein structure prediction in genomics. *Briefings in Bioinformatics* **2**, 111–125.
- Joosten, R.P., te Beek, T.A., Krieger, E. et al. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**(Database issue), D411–419. PMID: 21071423.
- Jothi, A. 2012. Principles, challenges and advances in ab initio protein structure prediction. *Protein and Peptide Letters* **19**(11), 1194–1204. PMID: 22587787.
- Kabsch, W., Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637.
- Kang, H.J., Lee, C., Drew, D. 2013. Breaking the barriers in membrane protein crystallography. *International Journal of Biochemistry and Cell Biology* **45**(3), 636–644. PMID: 23291355.

- Kim, D.E., Chivian, D., Baker, D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* **32**(Web Server issue), W526–W531.
- Kolodny, R., Pereyaslavets, L., Samson, A.O., Levitt, M. 2013. On the universe of protein folds. *Annual Review of Biophysics* **42**, 559–582. PMID: 23527781.
- Koonin, E. V., Wolf, Y. I., Karev, G. P. 2002. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223.
- Koopmann, R., Cupelli, K., Redecke, L. et al. 2012. In vivo protein crystallization opens new routes in structural biology. *Nature Methods* **9**(3), 259–262. PMID: 22286384.
- Kryshtafovych, A., Fidelis, K., Moult, J. 2014a. CASP10 results compared to those of previous CASP experiments. *Proteins* **82**(2), 164–174. PMID: 24150928.
- Kryshtafovych, A., Monastyrskyy, B., Fidelis, K. 2014b. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins* **82**(2), 7–13. PMID: 24038551.
- Kühlbrandt, W. 2013. Introduction to electron crystallography. *Methods in Molecular Biology* **955**, 1–16. PMID: 23132052.
- Le Gall, T., Romero, P.R., Cortese, M.S., Uversky, V.N., Dunker, A.K. 2007. Intrinsic disorder in the Protein Data Bank. *Journal of Biomolecular Structure and Dynamics* **24**, 325–342.
- Levinthal, C. 1969. How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems* (eds P.Debrunner, J.C.M.Tsibris, E.Munck), pp. 22–24. University of Illinois, Urbana IL.
- Levitt, M. 2007. Growth of novel protein structural data. *Proceedings of the National Academy of Science, USA* **104**(9), 3183–3188. PMID: 17360626.
- Lewis, T.E., Sillitoe, I., Andreeva, A. et al. 2013. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. *Nucleic Acids Research* **41**(Database issue), D499–507. PMID: 23203986.
- Madej, T., Addess, K.J., Fong, J.H. et al. 2012. MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Research* **40**(Database issue), D461–464. PMID: 22135289.
- Marsden, R.L., Orengo, C.A. 2008. Target selection for structural genomics: an overview. *Methods in Molecular Biology* **426**, 3–25. PMID: 18542854.
- Marsden, R.L., Lewis, T.A., Orengo, C.A. 2007. Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics* **8**, 86. PMID: 17349043.
- Marti-Renom, M. A., Stuart, A. C., Fiser, A. et al. 2000. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure* **29**, 291–325. PMID: 10940251.
- Massiah, M. A., Ko, Y. H., Pedersen, P. L., Mildvan, A. S. 1999. Cystic fibrosis transmembrane conductance regulator: Solution structures of peptides based on the Phe508 region, the most common site of disease-causing DeltaF508 mutation. *Biochemistry* **38**, 7453–7461.
- Matte, A., Jia, Z., Sunita, S., Sivaraman, J., Cygler, M. 2007. Insights into the biology of Escherichia coli through structural proteomics. *Journal of Structural and Functional Genomics* **8**(2–3), 45–55. PMID: 17668295.
- Meyer, E.F. 1997. The first years of the Protein Data Bank. *Protein Science* **6**, 1591–1597.
- Miller, M.D., Aravind, L., Bakolitsa, C. et al. 2010. Structure of the first representative of Pfam family PF04016 (DUF364) reveals enolase and Rossmann-like folds that combine to form a unique active site with a possible role in heavy-metal chelation. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **66**(Pt 10), 1167–1173. PMID: 20944207.
- Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., Kryshtafovych, A. 2014a. Evaluation of residue–residue contact prediction in CASP10. *Proteins* **82**(2), 138–153. PMID: 23760879.
- Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., Fidelis, K. 2014b. Assessment of protein disorder region predictions in CASP10. *Proteins* **82**(2), 127–137. PMID: 23946100.
- Montelione, G.T. 2012. The Protein Structure Initiative: achievements and visions for the future. *F1000 Biology Reports* **4**, 7. PMID: 22500193.
- Moretti, R., Fleishman, S.J., Agius, R. et al. 2013. Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions. *Proteins* **81**(11), 1980–1987. PMID: 23843247.

- Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15**, 285–289.
- Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., Tramontano, A. 2007. Critical assessment of methods of protein structure prediction–Round VII. *Proteins* **69**, 3–9.
- Nugent, T., Cozzetto, D., Jones, D.T. 2014. Evaluation of predictions in the CASP10 model refinement category. *Proteins* **82**(2), 98–111. PMID: 23900810.
- Osguthorpe, D. J. 2000. Ab initio protein folding. *Current Opinion in Structural Biology* **10**, 146–152.
- Pauling, L., Corey, R.B. 1951. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proceedings of the National Academy of Science, USA* **37**, 729–740.
- Pavlopoulou, A., Michalopoulos, I. 2011. State-of-the-art bioinformatics protein structure prediction tools (Review). *International Journal of Molecular Medicine* **28**(3), 295–310. PMID: 21617841.
- Perry, J.J., Tainer, J.A. 2013. Developing advanced X-ray scattering methods combined with crystallography and computation. *Methods* **59**(3), 363–371. PMID: 23376408.
- Pethica, R.B., Levitt, M., Gough, J. 2012. Evolutionarily consistent families in SCOP: sequence, structure and function. *BMC Structural Biology* **12**, 27. PMID: 23078280.
- Pirovano, W., Heringa, J. 2010. Protein secondary structure prediction. *Methods in Molecular Biology* **609**, 327–348. PMID: 20221928.
- Pitt, W.R., Higuero, A.P., Groom, C.R. 2009. Structural bioinformatics in drug discovery. In: *Structural Bioinformatics*, second edition (eds Gu, J., Bourne, P.E.), pp. 809–845. Hoboken, NJ, Wiley-Blackwell.
- Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N., Dunker, A.K. 2007. Intrinsic disorder and functional proteomics. *Biophysics Journal* **92**, 1439–1456.
- Rasmussen, S.G., DeVree, B.T., Zou, Y. et al. 2011. Crystal structure of the  $\beta 2$  adrenergic receptor-Gs protein complex. *Nature* **477**(7366), 549–555. PMID: 21772288.
- Ratjen, F., Döring, G. 2003. Cystic fibrosis. *Lancet* **361**, 681–689.
- Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D. 2004. Protein structure prediction using Rosetta. *Methods in Enzymology* **383**, 66–93.
- Rose, P.W., Bi, C., Bluhm, W.F. et al. 2013. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Research* **41**(Database issue), D475–482. PMID: 23193259.
- Rost, B., Sander, C. 1993a. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* **232**, 584–599.
- Rost, B., Sander, C. 1993b. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Science USA* **90**, 7558–7562.
- Rost, B., Sander, C., Schneider, R. 1994. Redefining the goals of protein secondary structure prediction. *Journal of Molecular Biology* **235**, 13–26.
- Schlichting, I., Miao, J. 2012. Emerging opportunities in structural biology with X-ray free-electron lasers. *Current Opinion in Structural Biology* **22**(5), 613–626. PMID: 22922042.
- Shulz, G.E., Schirmer, R.H. 1979. *Principles of Protein Structure*. Springer-Verlag, New York.
- Sickmeier, M., Hamilton, J.A., LeGall, T. et al. 2007. DisProt: the Database of Disordered Proteins. *Nucleic Acids Research* **35**(Database issue), D786–D793.
- Sillitoe, I., Cuff, A.L., Dessimoz, B.H. et al. 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research* **41**(Database issue), D490–498. PMID: 23203873.
- Simons, K. T., Strauss, C., Baker, D. 2001. Prospects for ab initio protein structural genomics. *Journal of Molecular Biology* **306**, 1191–1199.
- Smith, J.L., Fischetti, R.F., Yamamoto, M. 2012. Micro-crystallography comes of age. *Current Opinion in Structural Biology* **22**(5), 602–612. PMID: 23021872.
- Tai, C.H., Lee, W.J., Vincent, J.J., Lee, B. 2005. Evaluation of domain prediction in CASP6. *Proteins* **61** Suppl 7, 183–192.
- Taylor, T.J., Tai, C.H., Huang, Y.J. et al. 2014. Definition and classification of evaluation units for CASP10. *Proteins* **82**(2), 14–25. (2013). PMID: 24123179.

- Taylor, W. R., Orengo, C. A. 1989a. Protein structure alignment. *Journal of Molecular Biology* **208**, 1–22.
- Taylor, W. R., Orengo, C. A. 1989b. A holistic approach to protein structure alignment. *Protein Engineering* **2**, 505–519.
- Travaglini-Alcocatelli, C., Ivarsson, Y., Jemth, P., Gianni, S. 2009. Folding and stability of globular proteins and implications for function. *Current Opinion in Structural Biology* **19**(1), 3–7. PMID: 19157852.
- Williams, R.W., Chang, A., Juretic, D., Loughran, S. 1987. Secondary structure predictions and medium range interactions. *Biochimica et Biophysica Acta* **916**, 200–204.
- Zhang, H., Zhang, T., Chen, K. *et al.* 2011. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in Bioinformatics* **12**(6), 672–688. PMID: 21252072.



It is of great interest to understand the relationship between the genotype (e.g., having an altered chromosome number) and the phenotype (the appearance of the organism including its fitness). When an organism has an extra copy of a chromosome it is trisomic. The mechanisms by which trisomy occurs were understood in detail by the 1940s. The jimson-weed (*Datura stramonium L.*), a flowering plant of the potato family (*Solanaceae*), normally has 12 pairs of chromosomes. Albert Blakeslee (1874–1954) investigated the seed capsule from wildtype *Datura* (top) and 12 distinct trisomic types. For each trisomic, a diagram of the extra chromosome is shown including a numbering system for the chromosome ends (telomeres). Blakeslee noted that since each chromosome has a distinctive set of genes, each trisomic plant has a distinctive phenotype.

Source: Riley (1948, p. 420). Used with permission.

# Functional Genomics

# CHAPTER 14

Nil adeo quoniam natum'st in corpore, ut uti possemus, sed quod natum'st, id procreat usum. *[In fact, nothing in our bodies was born in order that we might be able to use it, but rather, having been born, it begets a use.]*

—Lucretius (c. 100–c. 55 bc), *De Rerum Natura*, IV, 834–835 (1772, p. 160)

*A world of made is not a world of born.*

—E. E. Cummings (1954, p. 397)

## LEARNING OBJECTIVES

Upon reading this chapter, you should be able to:

- define functional genomics;
- describe the key features of eight model organisms;
- explain techniques of forward and reverse genetics;
- discuss the relation between the central dogma and functional genomics; and
- describe proteomics-based approaches to functional genomics.

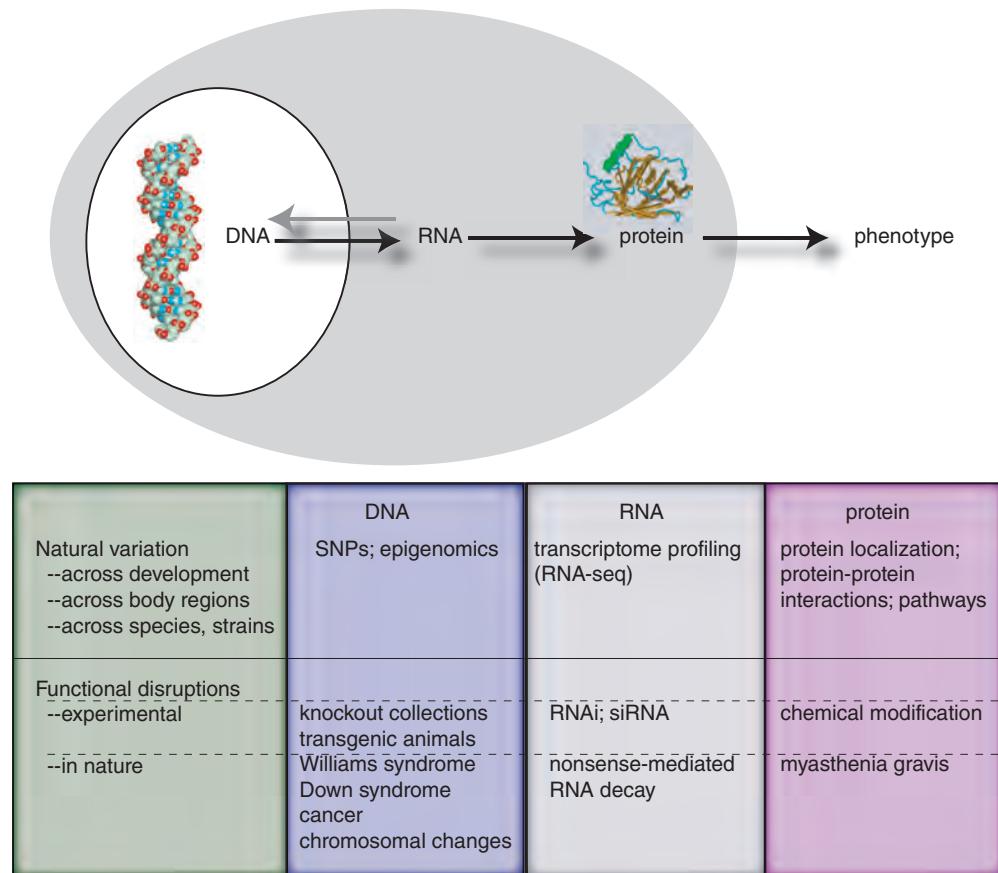
## INTRODUCTION TO FUNCTIONAL GENOMICS

A genome is the collection of DNA that comprises an organism. Functional genomics is the genome-wide study of the function of DNA (including genes and nongenic elements) as well as the nucleic acid and protein products encoded by DNA. We may further consider the meaning of the term functional genomics by considering some examples of the ways in which it has been characterized in recent years.

- Functional genomics may be applied to the complete collection of DNA (the genome), RNA (the transcriptome), or protein (the proteome) of an organism. The assessment of RNA transcripts that are expressed at various times of development or various body regions constitutes an example of functional genomics.
- Functional genomics implies the use of high-throughput screens, in contrast to traditional methods of biology in which one gene or protein has been characterized experimentally in depth. Such traditional methods commonly complement high-throughput approaches. For example, after performing a yeast two-hybrid screen to identify thousands of interacting protein partners in some model organism, further validation of selected binding partners is subsequently performed.

- Functional genomics often involves the perturbation of gene function to investigate the consequence on the function of other genes in a genome. For example, in the yeast *Saccharomyces cerevisiae*, each gene has been individually knocked out and simultaneously “bar-coded” as discussed below.
- One of the most challenging and fundamental problems in modern biology is to understand the relationship between genotype and phenotype (discussed in the following section). Connecting the two is a fundamental part of functional genomics.

We provide an overview of functional genomics in **Figure 14.1** with a schematic of a cell. We can consider the three cellular constituents of genomic DNA (including genes); RNA (including coding and noncoding RNA; Chapter 10); and proteins (Chapters 12 and 13). Other constituents, such as lipids and various metabolites, are also worthy of



**FIGURE 14.1** Functional genomics approaches to high-throughput protein analysis. From left to right, we can consider several aspects of a cell: the functions associated with DNA, RNA, and protein as well as higher-order aspects such as protein interactions, biochemical pathways, cell metabolism, and ultimately the phenotype of the cell and of the organism. We can also consider functional genomics approaches in the two broad categories of natural variation and of functional disruptions. Natural variation includes comparisons of the state of DNA, RNA, protein, or other cellular constituents as changes occur over time, under different physiological conditions, or (in the case of multicellular organisms) across different cell types and body regions. Functional disruptions occur in nature (such as chromosomal abnormalities); Williams syndrome is an example of a microdeletion syndrome causing the hemizygous (single-copy) loss of dozens of genes on chromosome 7, and Down syndrome is caused by the gain of an extra copy of chromosome 21. In this chapter we discuss high-throughput experimental approaches to disrupting gene function. Such studies elucidate the normal function of genes.

consideration but are not “informational” in the same sense as the polymers above. The scope of functional genomics includes two levels.

1. *Natural variation.* How do genes, RNA transcripts, and proteins change across body regions, or across developmental stages? In terms of genomic DNA we see in Chapter 18 that the genomes of many closely related yeast species have been sequenced, and in Chapter 19 we describe the recent sequencing of 12 *Drosophila* species and 15 mouse strains. In Chapter 20, we also discuss the variation in individual human genome sequences. Variation encompasses other aspects such as epigenetics (the study of heritable changes in gene function that occur without a change in DNA sequence, as when DNA is reversibly methylated). In terms of RNA transcripts, techniques such as RNA-seq (Chapter 10) are used to define region- and time-specific features of RNA transcripts.
2. *Functional disruptions* occur in nature and are experimentally studied. These include deletions, insertions, inversions, and translocations. The scale includes entire genomes (we discuss fish, plant, and *Paramecium* genome duplications in Chapter 19), entire chromosomes (which may become aneuploid, i.e., having an abnormal copy number), segments of chromosomes, or single nucleotides. Examples of naturally occurring deletions include the many microdeletion syndromes in which there is a hemizygous loss of chromosomal material, often spanning several million base pairs and including the loss of one copy of dozens of genes. We can find many examples of RNA loss (such as nonsense-mediated decay) and protein loss (e.g., in one form of myasthenia gravis, muscle weakness results from an autoimmune reaction that destroys copies of the nicotinic acetylcholine receptor at the neuromuscular junction; Drachman, 1994).

In this chapter we will describe many experimental approaches to deleting genes as well as intentionally reducing protein levels as a way to probe function. Amplifications also commonly occur in nature; Down syndrome is a well-known example in which the presence of three copies of chromosome 21 (instead of the usual two) is associated with increased levels of mRNA and possibly of protein derived from chromosome 21, leading to a panoply of phenotypes. Experimentally, transgenic or other models can be used to overexpress DNA, RNA, or protein.

We can summarize our focus in this chapter as the consideration of both natural variation and also disrupted cellular function. We explore how to disrupt gene, gene expression, or protein function, and what the consequences are of such disruptions.

## The Relationship Between Genotype and Phenotype

The genotype of an individual consists of the DNA that comprises the organism. The phenotype is the outward manifestation in terms of properties such as size, shape, movement, and physiology. We can consider the phenotype of a cell (e.g., a precursor cell may develop into a brain cell or liver cell) or the phenotype of an organism (e.g., a person may have a disease phenotype such as sickle-cell anemia). We can trace the history of how genotype and phenotype are defined back to August Weismann in the late nineteenth century (Web Document 14.1).

A great challenge of biology is to understand the relationship between genotype and phenotype (Ryan *et al.*, 2013). We can gather information about either one alone. Considering the genotype, we have now sequenced thousands of genomes (including viral and organellar genomes), and defined many of the coding and noncoding genes. It is possible to further describe the transcription of DNA into both coding and noncoding RNA. Protein products are also characterized in depth, both alone and in the context of interaction partners, pathways, and networks.

Considering the phenotype, we can describe many categories of phenotype from variation in the natural state (such as hair color or other quantitative traits) to disease. Model organisms undergo extensive phenotypic analyses. Diseases affecting humans, other animals, plants, or other organisms represent another aspect of phenotype. We discuss the major database for human disease (OMIM) in Chapter 21. As an example of a disease phenotype, Rett syndrome primarily affects girls, leading to hand-wringing, the loss of purposeful hand movements, and autism-like features. The syndrome was recognized (and named) in the 1980s when a group of patients with a similar phenotype were gathered at a meeting in Austria. Eventually, Huda Zoghbi and colleagues identified mutations in the X-linked gene *MECP2* as causing Rett syndrome (Amir *et al.*, 1999). *MECP2* encodes a protein that functions as a transcriptional repressor, regulating gene expression. This case typifies the challenge in understanding the relationship between the genotype (a mutation in a specific gene encoding a transcriptional repressor) and a phenotype (a syndrome having unique features). We have thousands of patients with diagnoses from intellectual disability to learning disorders; beginning with a phenotype, how do we find the corresponding genotype for disorders that have a genetic basis? In the case of diseases such as Rett syndrome, for which both genotype and phenotype are known, how do we connect them? Through understanding the cellular phenotype we may rationally devise therapeutic strategies aimed at correcting abnormalities that are introduced by a mutant gene product.

The field of functional genomics involves experimental and computational strategies to elucidate the function of DNA and chromosomes in relation to phenotype at the levels of the cell, the tissue, and the organism. There is a large gap in our understanding of how genotype and phenotype are related. For many diseases, understanding a primary genetic mutation (or insult) has not led to effective treatment or to a cure because of this gap in our understanding. We know that Down syndrome is caused by the occurrence of an extra copy of chromosome 21, but we do not understand why Down syndrome individuals have characteristic symptoms ranging from intellectual disability to abnormal facial features to common heart problems, and we do not know why the phenotype ranges from mild to extremely severe (e.g., profound intellectual disability and self-injurious behavior).

The remainder of this chapter is organized into three parts. First, we introduce eight model organisms that are prominent in functional genomics studies. We then describe two basic approaches to genetic studies of gene function: reverse and forward genetics. Finally, we explore functional genomics as related to proteomics, networks, and pathways as molecular biology intersects with systems biology.

## EIGHT MODEL ORGANISMS FOR FUNCTIONAL GENOMICS

Leonelli and Ankeny (2012) consider the origin of the concept of model organisms and the impact on the research community of focusing resources on these organisms. For estimates of species' divergence times see <http://www.timetree.org> (WebLink 14.1; Hedges *et al.*, 2006).

The tree of life has three great domains: the bacteria, archaea, and eukaryotes, as well as the separate group of viruses. Thousands of organisms across the tree of life are studied intensively. We can describe eight of them that have particularly important roles in the field of functional genomics. This is not a comprehensive list of model organisms, but helps to define the strengths and limitations of different experimental systems as well as the types of questions that can be addressed. We discuss the properties of their genomes in more detail in Chapters 15 (providing an overview of genomes), 17 (*Escherichia coli*), 18 (for *S. cerevisiae*), 19 (various eukaryotic genomes), and 20 and 21 (the human genome). The time when these organisms last shared a common ancestor with humans is approximately 2.5 billion years ago (BYA) for *E. coli*, 1.5 BYA for *Arabidopsis* and *S. cerevisiae*, 900 million years ago (MYA) for *C. elegans* and *Drosophila*, 450 MYA for zebrafish, and 90 MYA for mouse.

Leading bioinformatics and genomics organizations have initiated a broad range of functional genomics projects related to model organisms. These include efforts by the

Wellcome Trust Sanger Institute, the National Institutes of Health (NIH), and the National Human Genome Research Institute (NHGRI) at NIH. The Encyclopedia of DNA Elements (ENCODE) project (Chapter 8), which focused on characterizing functional elements of the human genome in great depth, also includes efforts to assess function in model organisms.

## 1. The Bacterium *Escherichia coli*

The bacterium *Escherichia coli* serves as the best-characterized bacterial organism, if not the best-characterized living organism. For decades it served as a leading model organism for bacterial genetics and molecular biology studies. Its 4.6 megabase (million base pairs) genome was sequenced by Blattner *et al.* (1997); we further describe the genome in Chapter 17. At the time of the initial genome sequencing, some function could be assigned to 62% of its genes. The principal website for *E. coli* is EcoCyc, the Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism (Keseler *et al.*, 2013). Today EcoCyc assigns a function to >75% of the 4501 annotated genes. Its content includes enzymes, transporters, transcription factors, and a range of regulatory interactions. EcoCyc is complemented by PortEco which includes high-throughput data from techniques such as RNA expression studies, single-gene knockouts, and chromatin immunoprecipitation (Hu *et al.*, 2014). EcoCyc also links to the BioCyc database, a collection of ~3000 databases of pathways and organisms (Latendresse *et al.*, 2012).

As an introduction to the use of the EcoCyc database, a query with the term globin links to nitric oxide dioxygenase, a flavohemoglobin. The result includes links to the protein sequence, and functional annotation from the Gene Ontology project (Chapter 12) and Multifun (a classification scheme similar to that of Clusters of Orthologous Groups (COGs) described in Chapter 12). There is extensive annotation of thousands of *E. coli* genes at EcoCyc.

Genome databases are available for all prominent organisms. Lourenço *et al.* (2011) emphasize the challenges of trying to integrate information across different levels (genes, proteins, and compounds) of *E. coli*. For example, there is a lack of standard nomenclature (even water has different designations across chemical databases). Many genes share synonyms, so EcoCyc and KEGG (see “Pathways, Networks, and Integration” below) list two different genes (*argA* and *argD*) that are each associated with a variant named Arg1.

Reed *et al.* (2006) described four dimensions of genome annotation, encompassing both experimental and computational (in silico) approaches.

1. One-dimensional annotation refers to identifying genes and assigning predicted functions. For *E. coli* this has been achieved to a high degree. For a variety of eukaryotes (Chapters 18–20), obtaining a trusted, precise catalog of genes has been extremely challenging because of the difficulty of identifying genes in genomic DNA. The task is becoming easier as more genomes are sequenced and comparative genomics approaches facilitate gene discovery.
2. Two-dimensional annotation refers to specifying the cellular components and their interactions, a topic we discuss “Proteomics Approaches to Functional Genomics” below. For *E. coli* this has to a great extent been achieved through the description of transcriptional regulatory networks in the RegulonDB database (Gama-Castro *et al.* 2008) and protein interactions in Bacteriome.org (Su *et al.*, 2008), for example. The MetaCyc database (Caspi *et al.*, 2008) includes over 2000 metabolic pathways from ~2500 organisms as of 2014.
3. Three-dimensional annotation is a description of the intracellular arrangement of chromosomes and of cellular components.
4. Four-dimensional annotation refers to characterizing genome changes that occur during evolution. This is a major theme of our study of bacterial, archaeal, viral, and

A Wellcome Trust Sanger Institute model organism site is available at <http://www.sanger.ac.uk/research/areas/mouseandzebrafish/> (WebLink 14.2). The NIH offers a website on model organisms for biomedical research (<http://www.nih.gov/science/models/>, WebLink 14.3). The NHGRI Functional Analysis Program is available at <http://www.genome.gov/10000612> (WebLink 14.4). The website for the ENCODE project at UCSC is <http://genome.ucsc.edu/ENCODE/> (WebLink 14.5).

EcoCyc is online at <http://ecocyc.org/> (WebLink 14.6), PortEco is at <http://porteco.org/> (WebLink 14.7), and BioCyc is at <http://biocyc.org/> (WebLink 14.8). Additional related resources are Regulon at <http://regulondb.ccg.unam.mx/> (WebLink 14.9) and EcoGene at <http://ecogene.org/> (WebLink 14.10).

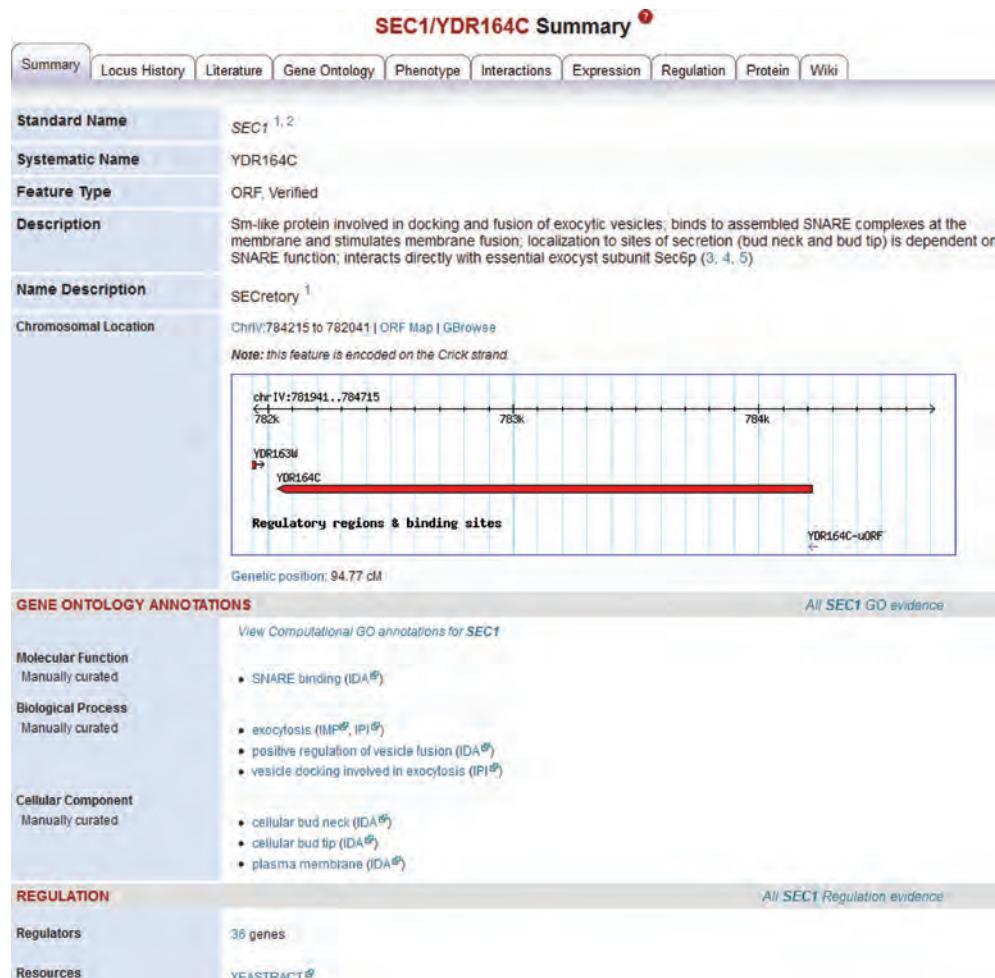
MetaCyc is available at <http://metacyc.org/> (WebLink 14.11).

eukaryotic chromosomes, where comparative genomics approaches have allowed the delineation of evolution from the level of whole genomes and chromosomes to individual DNA segments that are under positive or negative selection (Chapter 7).

## 2. The Yeast *Saccharomyces cerevisiae*

SGD is online at <http://www.yeastgenome.org/> (WebLink 14.12). Genome statistics are available from Genome Snapshot (Hirschman *et al.*, 2006) at <http://www.yeastgenome.org/cache/genomeSnapshot.html> (WebLink 14.13).

The budding yeast *S. cerevisiae* is the best-characterized organism among the eukaryotes. This single-celled fungus was the first eukaryote to have its genome sequenced (see Chapters 15 and 18). Its 13 megabase genome encodes about 6000 proteins. The *Saccharomyces* Genome Database (SGD) offers a remarkably deep insight into many aspects of the genome, including access to the results of hundreds of functional genomics experiments (Cherry *et al.*, 2012; Engel and Cherry, 2013). There are currently ~6600 annotated open reading frames (ORFs, corresponding to genes), including ~5000 that are verified, 750 that are uncharacterized (likely to be functional based on conservation across species but not experimentally validated), and <800 dubious (ORFs that are neither well conserved nor validated). Approximately 4200 gene products have been annotated to the root gene ontology terms (molecular function, biological process, cellular component; see Chapter 12).



**FIGURE 14.2** The *Saccharomyces* Genome Database (SGD) offers a wealth of functional genomics information. The top portion of a search for a typical gene, *SEC1*, is shown. This includes chromosomal location, gene ontology annotations, and regulators.

Source: *Saccharomyces* Genome Database (SGD). Reproduced with permission from Stanford University.

**MUTANT PHENOTYPES**

All *SEC1* Phenotype evidence

Classical genetics	<ul style="list-style-type: none"> <li>conditional</li> <li>endomembrane system morphology: abnormal</li> <li>heat sensitivity: increased</li> <li>Invertase secretion: decreased</li> <li>Invertase accumulation: increased</li> <li>Bgl2p distribution: abnormal</li> <li>sporulation: absent</li> </ul>
reduction of function	<ul style="list-style-type: none"> <li>sporulation: decreased</li> </ul>
Large-scale survey	
null	<ul style="list-style-type: none"> <li>inviable</li> <li>resistance to mefloquine: decreased</li> <li>resistance to cloquind: decreased</li> </ul>
overexpression	<ul style="list-style-type: none"> <li>vegetative growth: decreased rate</li> </ul>
reduction of function	<ul style="list-style-type: none"> <li>competitive fitness: decreased</li> </ul>
Resources	<a href="#">PROPHET</a>   <a href="#">PhenoM</a>   <a href="#">SCMD</a>   <a href="#">ScreenTroll</a>   <a href="#">Yeast Fitness Database</a>

**INTERACTIONS**

All *SEC1* Interaction evidence

Physical Interactions	<p>190 total interaction(s) for 86 unique genes/features.</p> <ul style="list-style-type: none"> <li>Affinity Capture-MS: 19</li> <li>Affinity Capture-RNA: 1</li> <li>Affinity Capture-Western: 25</li> <li>Co-fractionation: 1</li> <li>Co-purification: 1</li> <li>PCA: 4</li> <li>Reconstituted Complex: 23</li> <li>Two-hybrid: 6</li> </ul>
Genetic Interactions	<ul style="list-style-type: none"> <li>Dosage Growth Defect: 2</li> <li>Dosage Lethality: 5</li> <li>Dosage Rescue: 26</li> <li>Negative Genetic: 20</li> <li>Phenotypic Enhancement: 2</li> <li>Phenotypic Suppression: 4</li> <li>Positive Genetic: 5</li> <li>Synthetic Growth Defect: 6</li> <li>Synthetic Lethality: 39</li> <li>Synthetic Rescue: 1</li> </ul>

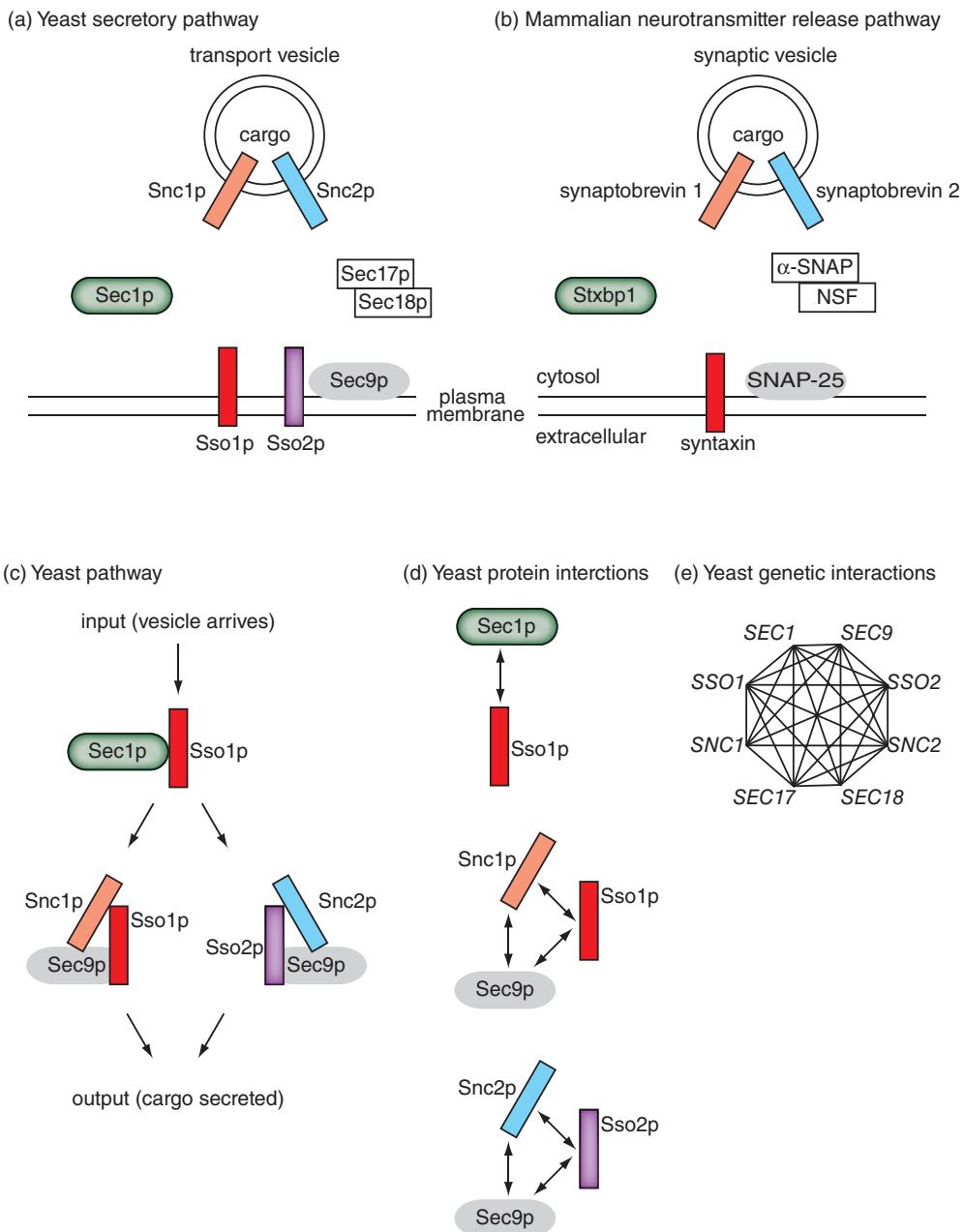
**Number of Experiments vs. Log<sub>2</sub> Ratios**

SGD      Dec 7, 2013

**FIGURE 14.3** Further portion of a search for *SEC1* in SGD (see Fig. 14.2). This provides information on mutant phenotypes from classical genetics and high-throughput technologies, as well as physical and genetic interactions. The inset shows a clickable summary of expression experiments (based on the Serial Pattern of Expression Levels Locator or SPELL tool), including those in which the *SEC1* transcript was up- or down-regulated.

Source: *Saccharomyces* Genome Database (SGD). Reproduced with permission from Stanford University.

To introduce SGD, we perform a search with a typical query, *SEC1* (Figs. 14.2 and 14.3). *SEC1* is a gene that encodes a protein (Sec1p) involved in vesicle trafficking (Fig. 14.4a). *SEC1* was discovered in a genetic screen (described in “Functional Genomics Using Reverse and Forward Genetics” below) for mutants that fail to secrete the enzyme invertase properly. Later experiments showed that Sec1p is related to the *SSO1* gene (named as “suppressor of *SEC1*”) and that the Sec1p and Sso1p proteins bind to each other to facilitate vesicle-mediated secretion in yeast. Sso1p, localized to the plasma membrane, is called a SNARE protein ( $\alpha$ -soluble NSF attachment protein receptor) that also interacts with the vesicular SNARE protein Snc1p. Sec1p, Sso1p, and Snc1p are therefore proteins that function in the process of delivering a vesicle and its contents to an appropriate compartment in a eukaryotic cell; in this case, the vesicles deliver proteins to the plasma membrane which are then secreted outside the cell. All of these yeast trafficking proteins have mammalian counterparts (indicated in Fig. 14.4b). The SGD entry for *SEC1* includes a wealth of information, including a description of its role in vesicle trafficking, and an explanation that the null (or knockout) phenotype is inviable and accumulates secretory vesicles (consistent with its required role in



**FIGURE 14.4** Diagram of *S. cerevisiae* and mammalian proteins involved in secretion as an illustration of functional genomics principles and approaches. (a) A constitutive trafficking pathway exists in yeast with a set of proteins, including several of the sec (secretory pathway) mutants. The cytosolic protein Sec1p interacts with Sso1p, a plasma membrane protein and ortholog of mammalian syntaxin. Sso1p also interacts with a protein complex that includes the vesicle-associated proteins Snc1p and Snc2p (mammalian synaptobrevin/VAMP) and the membrane-associated protein Sec9p (mammalian SNAP-25). Sec17p and Sec18p are required for this step and for other intracellular trafficking pathways such as from the Golgi apparatus to the vacuole. In yeast, the paralogous *SNC1/SNC2* and *SSO1/SSO2* genes arose after an ancient whole-genome duplication event (see Chapter 18). The presence of two copies of each molecule could allow functional redundancy, so that if one copy is lost (e.g., through mutation) the organism could be viable. Alternatively, the duplicated genes could acquire distinct functions, such as conferring the specificity of the docking and fusion events of transport vesicles with the appropriate intracellular target membrane. (b) Simplified diagram of proteins in the mammalian nerve terminal. Syntaxin binding protein 1 (Stxbp1, also called Munc18-1/N-sec1) binds tightly to the plasma membrane protein syntaxin. Separately, syntaxin binds to the synaptic vesicle protein synaptobrevin as well as SNAP-25 to form a protein complex, and subsequently the proteins NSF and α-SNAP further bind. Through this pathway, synaptic vesicles fuse with the plasma membrane and release their neurotransmitter contents by exocytosis. (c) Hypothetical pathway diagram showing two sets of proteins that could accomplish the task of secretion in yeast using parallel pathways. (d) Biochemical studies can reveal pairwise protein interactions and can also reveal complexes of multiple proteins. However, physical interactions would not reveal the relationship of proteins that do not interact directly but are part of the same pathway (such as Sec1p and Sec9p). (e) Genetic interaction maps reveal functionally related genes, including those involved in parallel pathways and those that do not physically interact. Adapted from Ooi *et al.* (2006), with permission from Elsevier.

trafficking; **Fig. 14.2**). The SGD page also provides dozens of resources including links to a genome browser (GBrowse), literature, interaction databases, and information on physical and genetic interactions (**Fig. 14.3**).

As we introduce functional genomics approaches we will return to *SEC1* as an example. Over 100 million years ago the entire *S. cerevisiae* genome duplicated, followed by a massive loss of duplicated genes. We discuss this in Chapter 18, and use *SSO1* and its paralog *SSO2* as examples to discuss the evidence for whole-genome duplication and the possible fates of duplicated genes.

In this chapter we introduce a variety of functional genomics assays in yeast. One reason that yeast offer an appealing experimental system is that virtually any desired genomic change can be introduced at the native locus using very efficient homologous recombination-based methods. In addition, they grow rapidly, they can make colored colonies, and it is easy to construct yeast strains with “reporters” that allow for selection of mutants with interesting traits, even if very rare. A variety of selectable colony color markers are available such as *MET15* or *ADE2* in which mutants can be selected for color upon growth in a particular medium. In this way the phenotypic consequence of genetic manipulations can be readily determined. SGD provides access to this wealth of mutant phenotype data (Engel *et al.*, 2010).

The 2013 Nobel Prize in Physiology or Medicine was awarded to James E. Rothman, Randy W. Schekman and Thomas C. Südhof “for their discoveries of machinery regulating vesicle traffic, a major transport system in our cells.” Schekman’s work included the identification of yeast secretory (SEC) mutants; Rothman focused on vesicular transport between Golgi stacks; and Südhof studied vesicle function in the mammalian nerve terminal. See [http://www.nobelprize.org/nobel\\_prizes/medicine/laureates/2013/](http://www.nobelprize.org/nobel_prizes/medicine/laureates/2013/) (WebLink 14.14).

### 3. The Plant *Arabidopsis thaliana*

The thale cress *Arabidopsis thaliana* was the first plant to have its genome sequenced (and the third finished eukaryotic genome sequence). It has served as a model for eukaryotic functional genomics projects (reviewed in Borevitz and Ecker, 2004; Koornneef and Meinke, 2010). The principal web site, The *Arabidopsis* Information Resource (TAIR), centralizes a vast amount of information about its genome (Lamesch *et al.*, 2012). **Figure 14.5** shows some of the diversity of information that is accessible from the home page pull-down menus. Under the browse menu, a link to “2010 projects” describes dozens of projects designed to reach a National Science Foundation goal to functionally annotate all *Arabidopsis* genes by 2010. As an example of a gene search at TAIR, a query for *Arabidopsis SEC1A* (RefSeq accession NP\_563643; locus tag At1g02010) reveals information about its chromosomal location and available mutants.

The TAIR website is <http://www.arabidopsis.org/> (WebLink 14.15).

*Arabidopsis* offers many appealing features as a model plant, including its short generation time, prolific seed production, compact genome size, and opportunities for genetic manipulation. As an example Atwell *et al.* (2010) performed a genome-wide association study (GWAS) to examine over 100 phenotypes (broadly involving flowering, plant defense, element concentrations, and developmental traits) in nearly 200 inbred lines. GWAS is a technique in which thousands to millions of single-nucleotide polymorphisms (SNPs) are identified as a proxy for genotype and associated with phenotypes. Most GWAS have been applied to human disease (see Chapter 21). Atwell *et al.* successfully identified many common alleles of major effect, and in some cases found single genes that are associated with particular phenotypes.

### 4. The Nematode *Caenorhabditis elegans*

Among the metazoans (animals), the soil-dwelling nematode *Caenorhabditis elegans* is a key model organism. This was the first multicellular animal to have its genome sequenced. This roundworm, like fruit flies and humans, is capable of complex behaviors, but its body is simple and all the 959 somatic cells in its body have been mapped including their lineages throughout development. Wormbase is the main online information repository (Harris *et al.*, 2014).

The *C. elegans* genome encodes ~20,400 protein-coding genes, almost exactly the same number as in humans. Almost 7000 genes have been deleted (*C. elegans* Deletion Mutant Consortium, 2012), parallel to similar efforts for other model organisms. In the

WormBase is available at <http://www.wormbase.org> (WebLink 14.16). The trans-NIH *C. elegans* initiative website is [http://www.nih.gov/science/models/c\\_elegans/](http://www.nih.gov/science/models/c_elegans/) (WebLink 14.17).

Search	Tools	ABRC Stocks
Search Overview	Tools Overview	Stocks Overview
DNA/Clones	GBrowse	ABRC Home
Ecotypes	Synteny Viewer	Browse ABRC Catalog
Genes	Seqviewer	Supplement to ABRC Catalog
Gene Ontology Annotations	Mapviewer	Search ABRC DNA/Clone Stocks
Plant Ontology Annotations	AraCyc Metabolic Pathways	Search ABRC Seed/Germplasm Stocks
Keywords	N-Browse	ABRC Stock Order History
Locus History	Integrated Genome Browser	ABRC Fee Structure
Markers	BLAST	Place ABRC Order
Microarray Element	WU-BLAST	Search My ABRC Orders
Microarray Experiment	FASTA	Search ABRC Invoices
Microarray Expression	Patmatch	How to Make Payments to ABRC
People/Labs	Motif Analysis	ABRC Stock Donation
Polymorphisms/Alleles	VxInsight	
Proteins	Java Tree View	
Protocols	Bulk Data Retrieval	
Publication	Chromosome Map Tool	
Seed/Germplasm	Gene Symbol Registry	
Textpresso Full Text	Textpresso Full Text	
Browse	Portals	Download
Browse Overview	Portals Overview	Download Overview
ABRC Catalog	Clones/DNA Resources	ABRC Documents
2010 Projects	Education and Outreach	Genes
Monsanto SNP and Ler Collections	Gene Expression Resources	GO and PO Annotations
Gene Families	Genome Annotation	Maps
Transposon Families	MASC/Functional Genomics	Metabolic Pathways
Gene Class Symbols	Mutant and Mapping Resources	Polymorphisms
Ontologies/Keywords	Nomenclature	Proteins
Archived e-Journals	Proteomics Resources	Protocols
The Arabidopsis Book (TAB)	Metabolomics Resources	Microarray Data

**FIGURE 14.5** The *Arabidopsis* Information Resource (TAIR) is the principal genome database for *Arabidopsis*. The screen capture shows some of the menu options including search strategies, analysis tools, available stocks, functional classification, and access to functional genomics initiatives.

Source: The *Arabidopsis* Information Resource (TAIR), courtesy of Phoenix Bioinformatics.

million mutation project, ~2000 mutagenized *C. elegans* strains were mutagenized then sequenced, allowing >800,000 unique single-nucleotide variants to be identified (eight nonsynonymous changes per gene) as well as 16,000 indels (Thompson *et al.*, 2013). That project also involved sequencing 40 wild isolate strains, producing a comparable number of SNVs and indels.

The modENCODE project, in parallel to the human ENCODE project, characterized the *C. elegans* transcriptome as well as transcription factor-binding sites and chromatin organization (Gerstein *et al.*, 2010). This resulted in more complete and accurate gene models as well as models of transcription factor-binding sites associated with microRNAs.

## 5. The Fruit Fly *Drosophila melanogaster*

The fruit fly *Drosophila melanogaster*, another metazoan invertebrate, has long served as a model for genetics. Early studies of *Drosophila* resulted in the descriptions of the nature of the gene as well as linkage and recombination, producing gene maps a century ago. The recent sequencing of 12 *Drosophila* genomes (*Drosophila* 12 Genomes Consortium *et al.*, 2007) and 192 inbred lines of the *Drosophila* genus (*Drosophila* Genetic Reference Panel; Chapter 19) are already providing unprecedented insight into mechanisms of genome evolution (Russell, 2012). The central *Drosophila* database, FlyBase, combines molecular and genetic data on the *Drosophilidae* (McQuilton *et al.*, 2012; St Pierre *et al.*, 2014).

A strength of *Drosophila* as a model organism is that genomic changes can be induced with extreme precision, from single-nucleotide changes to introducing large-scale chromosomal deletions, duplications, inversions, or other modifications. At the same time, it is a multicellular animal that features a complex body plan. Loss of function mutations have been introduced into all of its ~14,000 protein-coding genes, and over half of these have an identifiable phenotype. Many human disease gene mutations have been modeled in *Drosophila* to further understand pathogenesis (Chen and Crowther, 2012). As for *C. elegans*, the modENCODE consortium has comprehensively mapped transcripts, histone modifications, and other biochemical signatures. This has tripled the portion of the *Drosophila* genome that has been annotated (modENCODE Consortium *et al.*, 2010).

## 6. The Zebrafish *Danio rerio*

Although the lineages leading to modern fish and humans diverged approximately 450 million years ago, both are vertebrate species and orthologs are identifiable for the great majority of their protein-coding genes (with an average of about 80% amino acid identity between orthologs). The first four fish genomes to be sequenced were the pufferfish *Takifugu rubripes* and *Tetraodon nigroviridis*, the medaka *Oryzias latipes*, and the zebrafish *Danio rerio* (Chapter 19). Of these, the zebrafish has emerged as an important model organism for functional genomics (Henken *et al.*, 2004). It is a small tropical freshwater fish having a genome size of 1.8 billion base pairs (Gb) organized into 25 chromosomes. There are >26,000 protein-coding genes (Howe *et al.*, 2013b), modestly more than in humans. For functional genomics studies, the zebrafish has served as a model for understanding both normal and abnormal development. Mutations in large numbers of human disease gene orthologs have been generated and characterized, and both forward and reverse genetic screens (introduced in “Functional Genomics Using Reverse and Forward Genetics” below) have been applied (e.g., Kettleborough *et al.*, 2013; Varshney *et al.*, 2013). Some of the advantages of zebrafish as a model organism include the following:

- Its generation time is short, especially for a vertebrate.
- It produces large numbers of progeny.
- The developing embryo is transparent. For example, if a transgene is inserted into the genome with a promoter that drives the expression of green fluorescent protein (GFP), it is possible to see this expression from the outside of each animal’s body.
- It is a vertebrate and therefore a close model for human disease.
- Its genome is well annotated. The vertebrate genome annotation (Vega) database at the Sanger Institute focuses on high-quality manual annotation with a particular focus on genomes such as human, mouse, and zebrafish (Wilming *et al.*, 2008).

The principal zebrafish website is the Zebrafish Information Network (ZFIN ; Howe *et al.*, 2011, 2013a).

Flybase is at <http://www.flybase.org> (WebLink 14.18). The *Drosophila* Genetic Reference Panel website is <http://dgrp.gnets.ncsu.edu> (WebLink 14.19).

Two of the giants of genetics research focused their studies on *Drosophila*: Thomas Hunt Morgan and Hermann J. Muller. Morgan was awarded a Nobel Prize in 1933 “for his discoveries concerning the role played by the chromosome in heredity” ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1933/](http://nobelprize.org/nobel_prizes/medicine/laureates/1933/), WebLink 14.20). He and his contemporaries A.H. Sturtevant, C.B. Bridges, and Muller discovered a broad array of properties of genes and chromosomes. They described chromosomal deficiencies including nondisjunction, balanced lethals, chromosomal duplication (trisomy) and monosomy, and translocations. Muller was awarded a 1946 Nobel Prize “for the discovery of the production of mutations by means of X-ray irradiation.” His finding of position effect variegation laid the foundation for modern epigenetics research. The 1995 Nobel Prize in Physiology or Medicine was awarded to Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus “for their discoveries concerning the genetic control of early embryonic development.” These studies were also performed in *Drosophila* ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1995/](http://nobelprize.org/nobel_prizes/medicine/laureates/1995/), WebLink 14.21).

The Vega database is available at <http://vega.sanger.ac.uk/> (WebLink 14.22).

ZFIN is online at <http://www.zfin.org> (WebLink 14.23). The trans-NIH zebrafish initiative website is <http://www.nih.gov/science/models/zebrafish/> (WebLink 14.24).

The Trans-NIH Mouse Initiatives homepage is <http://www.nih.gov/science/models/mouse/> (WebLink 14.25).

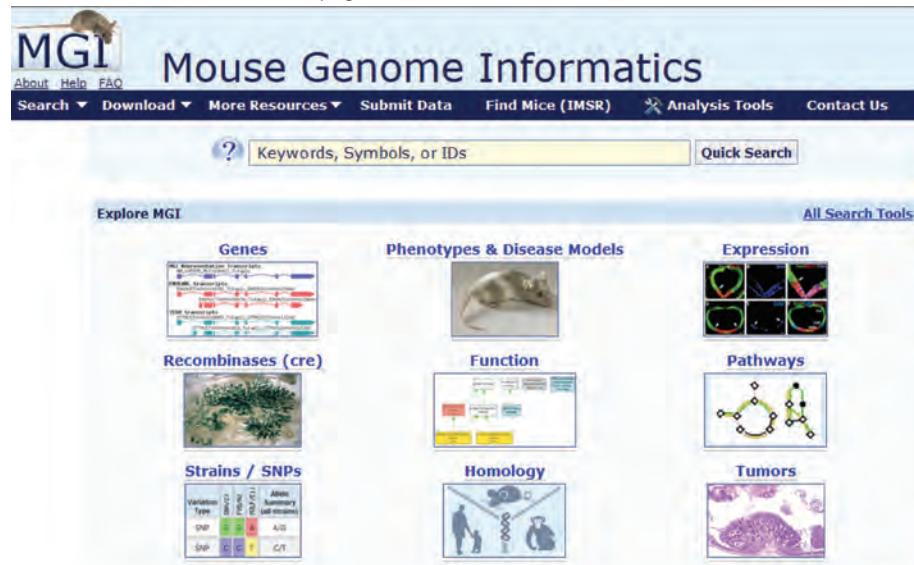
The MGI website is <http://www.informatics.jax.org/> (WebLink 14.26).

## 7. The Mouse *Mus musculus*

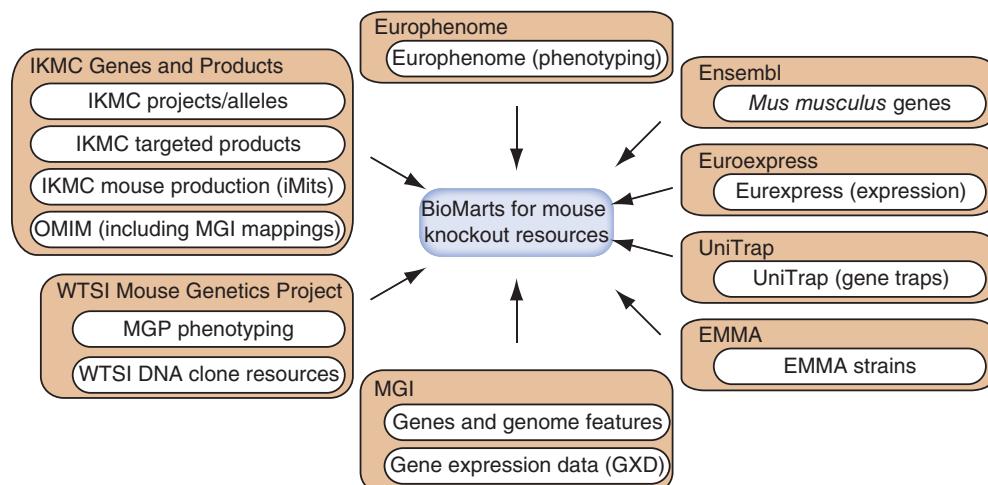
The rodents diverged from the primate lineage relatively recently (90 million years ago) and share almost all of their genes with humans. The mouse *Mus musculus* is one of the most important model organisms for the study of human gene function because of the close structural and functional relationship between the two genomes combined with a relatively short generational span, and powerful tools have been developed to manipulate its genome.

The main mouse genome website is the Mouse Genome Informatics (MGI) site (Blake *et al.*, 2014). In common with other leading organism-specific website resources, MGI provides a portal to mouse-specific resources including sequence data, a web browser, available mutant strains, gene expression studies, and literature (Fig. 14.6a).

(a) Mouse Genome Informatics home page



(b) Customized BioMarts for mouse functional genomics



**FIGURE 14.6** The Mouse Genome Informatics (MGI) Database is the principal website for mouse genomics information. (a) The home page provides a portal to a vast number of resources. (b) There are many specialized BioMarts focused on mouse functional genomics, including: MGI; Ensembl, providing a genomic context; UniTrap for gene trapping; International Mouse Knockout Consortium (IMKC) resources; Wellcome Trust Sanger Institute (WTSI) mouse genetics; and the European Mutant Mouse Archive (EMMA).

Source: Redrawn from MGD, Blake *et al.* (2014). Reproduced with permission from MGI.

About 10,000 mouse genes have been knocked out (Koscielny *et al.*, 2014). The International Mouse Phenotyping Consortium (IMPC) provides access to mutant mice and associated data, including coordination with MGI and Ensembl. IMPC follows earlier efforts such as the Knockout Mouse Project (KOMP), the European Conditional Mouse Mutagenesis Program (EUCOMM), and the North American Conditional Mouse Mutagenesis Project (NorCOMM) (International Mouse Knockout Consortium *et al.*, 2007). We discuss their strategies for mutating all protein-coding genes in mouse, including the two main knockout approaches of gene targeting and gene trapping (Guan *et al.*, 2010; White *et al.*, 2013). IMPC provides links to a variety of mouse-specific BioMarts (**Fig. 14.6b**). All major genotyping projects are associated with detailed phenotyping efforts such as the pipelines described by Fuchs *et al.* (2011).

In a project called the Collaborative Cross, 1000 recombinant inbred strains of mouse are being bred (Complex Trait Consortium, 2004; Collaborative Cross Consortium, 2012; Welsh *et al.*, 2012). This project is producing large numbers of genetically related mice that have nonlethal phenotypic diversity, and also that can be exposed to manipulations such as phenotypic screens (see “Forward Genetics: Chemical Mutagenesis” below). The 1000 strains are derived from eight inbred founder strains that were systematically crossed. These strains will be fully genotyped and used to model human populations and diseases. If we denote the eight inbred founder strains A–H, then the  $G_1$  generation will consist of AB, CD, EF, and GH genotypes (from mating of AA  $\times$  BB mice, CC  $\times$  DD, etc.), the  $G_2$  generation will consist of AB  $\times$  CD mice yielding ABCD genotypes and EF  $\times$  GH yielding EFGH. After 23 generations there will be 99% inbreeding with unique recombination events. The 1000 mouse strains are expected to provide an important resource for modeling human populations and diseases.

The Diversity Outbred (DO) population offers a complementary approach to mouse genetics, offering the benefit of wildtype levels of heterozygosity (thus buffering against the consequences of mutation). The allelic diversity of these mice facilitates their usefulness in mapping phenotypes (Churchill *et al.*, 2012; Logan *et al.*, 2013). The founders of the DO are from breeding lines of the Collaborative Cross.

## 8. *Homo sapiens*: Variation in Humans

Humans are not considered to be a model organism by some, and we do not consider ourselves to be an experimental system *per se*. We are however motivated to investigate the range of phenotypic expression to understand how we acquire our characteristic features, how we have evolved, and how we fit into the ecosystem. One of the strongest motivations for studying humans is to understand the causes of disease in order to search for more effective diagnoses, preventions, treatments, and ultimately cures if possible. While in most contexts we do not experiment on ourselves invasively, nature does perform functional genomics experiments on us. For example, human fecundity is extraordinarily low relative to other mammalian and vertebrate species (see Chapter 21). Of all conceptuses that appear normal after one week of development as a zygote, perhaps over 80% are not viable. This is due to massive aneuploidy that commonly occurs, causing trisomy, monosomy, and even tetrasomy (four copies) or nullisomy (zero copies) of many chromosomes. Functional genomics is an experimental science in which gene function is often assessed by perturbing a system. Genes may be selectively deleted or duplicated, and then the functional consequence is measured to infer the function of the gene. Nature produces the equivalent of functional genomics experiments through the many forms of variation that organisms experience. Experimentally, the emergence of next-generation sequencing is having a profound impact on studies of genetic variation in humans (Kilpinen and Barrett, 2013).

**The International Mouse Phenotyping Consortium (IMPC)**  
website is <https://www.mousephenotype.org/> (WebLink 14.27). Mouse-centered BioMarts are available from MGI (<http://biomart.informatics.jax.org/>, WebLink 14.28); IKMC (<http://www.i-dcc.org/>, WebLink 14.29); UniTrap (<http://biomart.helmholtz-muenchen.de/>, WebLink 14.30); the Europhenome Mouse Phenotyping Resource (<http://www.europhenome.org/biomart/martview/>, WebLink 14.31); and WTSI (<http://www.sanger.ac.uk/htgt/biomart/martview/>, WebLink 14.32). Additionally, the prominent Ensembl BioMart includes mouse genes under Ensembl builds (<http://www.ensembl.org/biomart/martview/>, WebLink 14.33).

**Collaborative Cross websites**  
include <http://churchilljax.org/research/cc.shtml> (WebLink 14.34).

Aneuploidy refers to a change in chromosomal copy number. A euploid individual has the normal two copies of a set of chromosomes.

## FUNCTIONAL GENOMICS USING REVERSE AND FORWARD GENETICS

There are many different basic approaches to identifying the function of a gene. Biochemical strategies can be employed, which typically involves studying one gene or gene product at a time. This is often the most rigorous way to study gene function, and has been the main approach for the past century. For example, in order to understand the function of a globin gene, its protein product can be purified to homogeneity and its physical properties (such as molecular mass, isoelectric point, oxygen- and heme-binding properties, and post-translational modifications; Chapter 12), its interactions with other proteins, its role in cellular pathways, and the consequence of mutating the gene characterized. We described eight different aspects of protein function in **Figure 12.18**. While invaluable, analysis of a single gene and its products is almost always laborious and time-consuming; a variety of complementary high-throughput strategies have therefore been introduced. These strategies can produce thousands of mutant alleles that are then available to facilitate the research of scientists who focus on the study of any particular genes.

One high-throughput method of assessing gene function is to examine messenger RNA levels in various conditions or states using RNA-seq (described in Chapters 10 and 11) or to measure protein levels (Chapter 12). These studies usually give only indirect rather than direct information about gene function. For example, if red blood cells are treated with a drug that inhibits heme biosynthesis in mitochondria, the cell may respond with a complex program of responses that serve to regulate the expression of heme-binding proteins such as the globins. Globin messenger RNA and protein levels might be reduced dramatically, but it would be incorrect to infer that the drug acted directly on the globin gene, messenger RNA, or protein. Similarly, when RNA transcript levels are measured in tissues or cell lines derived from individuals with a disease, significantly regulated transcripts might reflect adaptive changes made in response to a primary insult such as a genetic mutation. Changes might also occur because of downstream effects: a gene defect could disrupt a pathway, leading to degeneration of a brain region, and other cells such as glia could proliferate as a downstream response. Such experiments are not likely to directly reveal the gene-causing mutation, although they may reveal information about its secondary consequences and are essentially a molecular phenotype for the mutant.

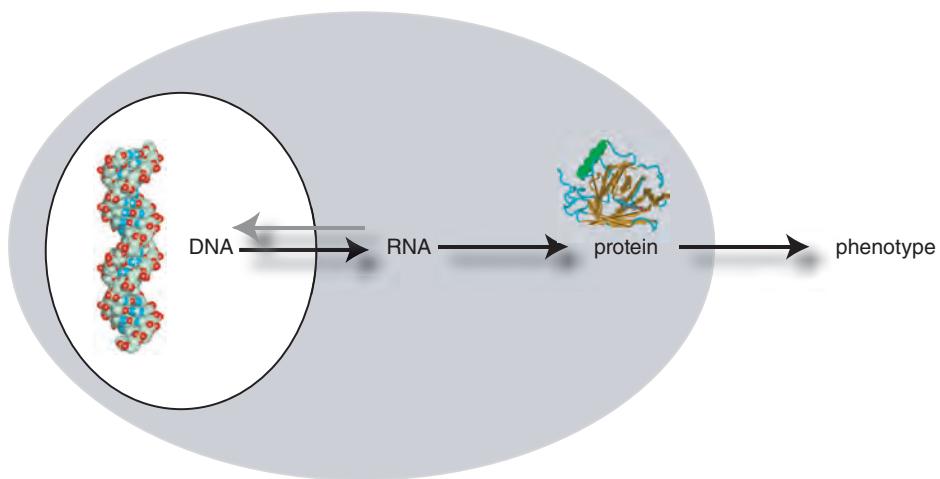
There are two main kinds of genetic screens that are used to identify gene function in a high-throughput fashion: reverse and forward genetics (reviewed in Schulze and McMahon, 2004; Ross-Macdonald, 2005; Alonso and Ecker, 2006; Caspary and Anderson, 2006). These two approaches are illustrated in **Figure 14.7**. In reverse genetic screens, a large number of genes (or gene products) is systematically inhibited one by one. This can be accomplished in many ways, for example by deleting genes using homologous recombination, gene trapping, or by selectively reducing messenger RNA abundance. One or more phenotypes of interest are then measured.

The main challenge of this approach is that for some organisms it is difficult to disrupt large numbers of genes (such as tens of thousands) in a systematic fashion. It can also be challenging to discern the phenotypic consequences for a gene that is disrupted. As an example of reverse genetics, Thomas Südhoff and colleagues targeted the deletion of mouse *syntaxin binding protein 1* (*Stxb1*; also called *Munc18-1* or *N-sec1*), a gene encoding a nerve terminal protein (Verhage *et al.*, 2000). The phenotype was lethality at the time of birth, with neurons unable to secrete neurotransmitter. Remarkably, brain development appeared normal up to the time of death. This targeted deletion allowed the dissection of the functional role of this gene; **Figure 14.4b** depicts its function.

In forward genetic screens the starting point is a defined phenotype of interest, such as the ability of plants to grow in the presence of a drug, neurons to extend axons to appropriate targets in the mammalian nervous system, or an eukaryotic cell to transport

### Reverse genetics (mutate genes then examine phenotypes)

- Strategy: Systematically inhibit the function of every gene in a genome  
 Approach 1: gene targeting by homologous recombination  
 Approach 2: gene trap mutagenesis  
 Approach 3: inhibit gene expression using RNA interference  
 Measure the effect of gene disruption on a phenotype



- Strategy: Identify a phenotype (e.g. growth in the presence of a drug)  
 Mutate genomic DNA (e.g. by chemical mutagenesis)  
 Identify individuals having an altered phenotype  
 Identify the gene(s) that were mutated  
 Confirm those genes have causal roles in influencing the genotype

### Forward genetics ("phenotype-driven" screen)

**FIGURE 14.7** Reverse and forward genetics. In reverse genetics, genes are targeted for deletion through approaches such as homologous recombination. After a knockout animal is produced, the phenotype is investigated to discern the function of the gene. This is called a “gene-driven” approach because it begins with targeted deletion or disruption of a gene. In forward genetics, the starting point is a phenotype of interest. The genome is subjected to a process of mutagenesis (typically with a chemical such as ENU or an exogenous DNA transposon). Mutants are collected and screened for those that display an altered phenotype. Next, the genes underlying the altered phenotype are mapped and identified. This is called a “phenotype-driven” approach, since the starting point is an altered phenotype and not particular disrupted genes.

cargo. An experimental intervention is made, such as administering a chemical mutagen or radiation to cells (or to an organism). This results in the creation of mutants. The phenotype of interest is observed in rare representatives among a large collection of mutants. If individuals need to be assayed for the phenotype one at a time (as part of a screen), this can be extremely laborious. If a specific selective condition can be defined in which only the desired mutant grows (a selection), the process is greatly facilitated. A second challenge of forward genetics approaches is to then identify the responsible gene(s) using mapping and sequencing strategies. As an example of this approach, Peter Novick, Randy Schekman and colleagues characterized temperature-sensitive yeast mutants that accumulate secretory vesicles (Novick and Schekman, 1979; Novick *et al.*, 1980). These secretion (*sec*) mutants occurred in a series of dozens of complementation groups (yeast strains harboring different mutant alleles of the same gene). All the *sec* mutant genes

The accession number of *S. cerevisiae* Sec1p is NP\_010448.  
 The accession of an ortholog, human syntaxin-binding protein 1a is NP\_003156.

were subsequently identified. For example, the *SEC1* gene encodes the Sec1p protein that functions in vesicle docking at the cell surface. Sec1p is a yeast ortholog of mammalian syntaxin-binding protein 1. A schematic showing the role of Sec1p and three other sec proteins in vesicle trafficking is shown in **Figure 14.4b**.

### Reverse Genetics: Mouse Knockouts and the $\beta$ -Globin Gene

Knocking out a gene refers to creating an animal model in which a homozygous deletion is created, that is, there are zero copies (denoted  $(-/-)$ ) and referred to as a null allele) instead of the wildtype situation of two copies in a diploid organism  $(+/+)$ . In a hemizygous deletion, one copy is deleted and one copy remains  $(+/-)$ .

We can illustrate the use of knockouts with the example of the  $\beta$ -globin gene. In normal adult humans, hemoglobin is a tetramer that consists of two  $\alpha$ -globin subunits and two  $\beta$ -globin subunits ( $\alpha_2\beta_2$ ), with a minor amount (~2–3%) consisting of  $\alpha_2\delta_2$  tetramers. The  $\beta$  and  $\delta$  genes are part of a cluster of  $\beta$ -like genes on chromosome 11 (**Fig. 14.8a**). There is a similar arrangement on mouse chromosome 7 (**Fig. 14.8b**). The globin genes are expressed at different developmental stages and cell types in a manner that is exquisitely choreographed. Within the  $\beta$ -globin cluster,  $\epsilon$ -globin is expressed in the blood island of the yolk sac until 6–8 weeks of gestation when it is silenced and  $\gamma$ -globin genes are activated. At birth,  $\delta$ -globin and  $\beta$ -globin gene expression increase, while  $\gamma$ -globin expression declines until it is silenced at about age 1. This process is called hemoglobin switching, and it is thought to occur because of interactions between the globin genes and the upstream locus control region (reviewed in Li *et al.*, 2006). Various protein complexes interact with the locus control region (Mahajan and Wissman, 2006). As indicated in **Figure 14.8a**, specific regulatory sites have been identified by techniques such as DNase I hypersensitivity assays that reveal regions of exposed chromatin.

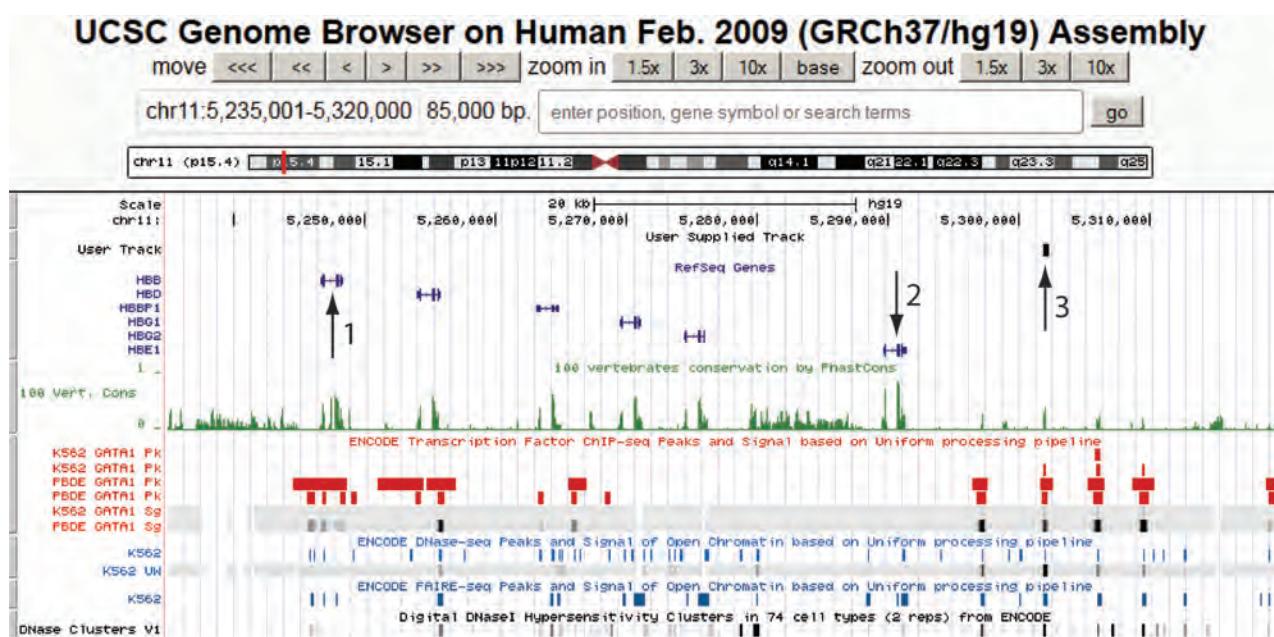
Several diseases are associated with perturbations of globin function (discussed in Chapter 21 on human disease). Sickle-cell anemia is caused by mutations in a copy of the  $\beta$ -globin gene. Thalassemias are hereditary anemias that result from an imbalance in the usual one-to-one proportion of  $\alpha$  and  $\beta$  chains. In an effort to create an animal model of thalassemias, and to further understand the function of the  $\beta$  globin gene, Oliver Smithies and colleagues used homologous recombination in embryonic stem cells to disrupt the mouse major adult  $\beta$ -globin gene *b1* (Shehee *et al.*, 1993). In homologous recombination, recombinant DNA introduced into the cell recombines with the endogenous, homologous sequence (Capecci, 1989).

The 2007 Nobel Prize in Physiology or Medicine was awarded to Mario Capecci, Sir Martin Evans, and Oliver Smithies "for their discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells;" see [http://nobelprize.org/nobel\\_prizes/medicine/laureates/2007/](http://nobelprize.org/nobel_prizes/medicine/laureates/2007/) (WebLink 14.35).

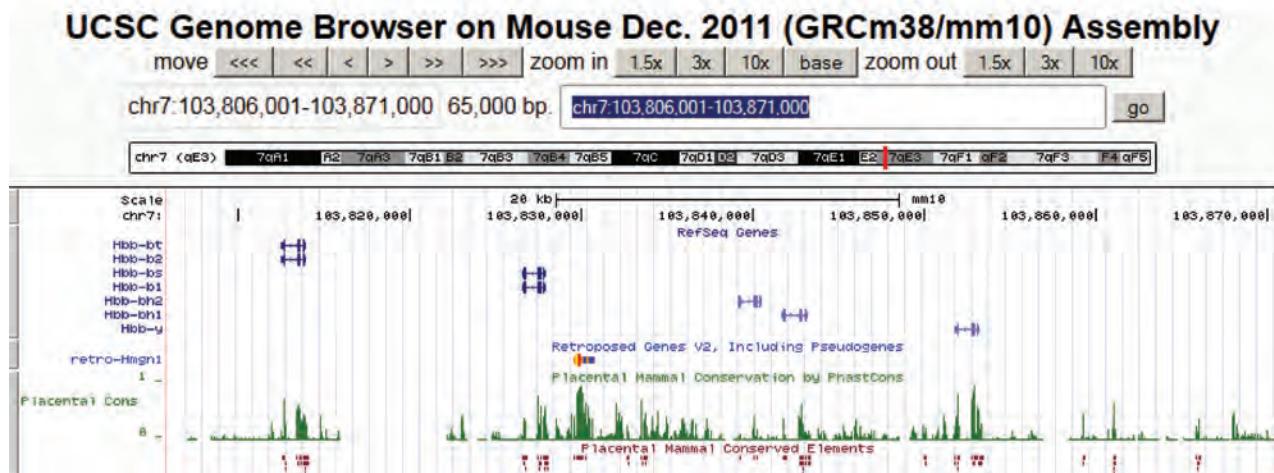
The approach, depicted in **Figure 14.9**, requires a targeting vector that includes the  $\beta$ -globin gene having a portion modified by insertion of the *neo* gene into exon 2. This targeting vector is introduced into embryonic stem cells by electroporation. When the cells are cultured in the presence of the drug G418, wildtype cells die whereas cells having the *neo* cassette survive. The successful introduction of an interrupted form of the  $\beta$ -globin gene into stem cells can be confirmed by using the polymerase chain reaction and/or Southern blots (in which a radiolabeled fragment of the insert is hybridized to membranes containing extracts of genomic DNA from wildtype and targeted cells). Targeted embryonic cell lines are injected into mouse blastocysts and implanted into the uterus of a foster mother to generate chimeric offspring. The mice that were heterozygous for the disrupted gene  $(+/-)$  appeared normal, while homozygous mutants  $(-/-)$  died *in utero* or near the time of birth. The knockout therefore caused a lethal thalassemia with abnormal red blood cells and lack of protein produced from the deleted *b1* gene.

In nature, the same *b1* gene is sometimes deleted in mice. Surprisingly, this naturally occurring deletion results in only a mild thalassemia, rather than the lethal phenotype that results from the knockout. Shehee *et al.* (1993) hypothesized that the locus control region normally regulates the *b1* and *b2* genes, but there is a rate-limiting amount of promoter

(a) Human beta globin cluster region

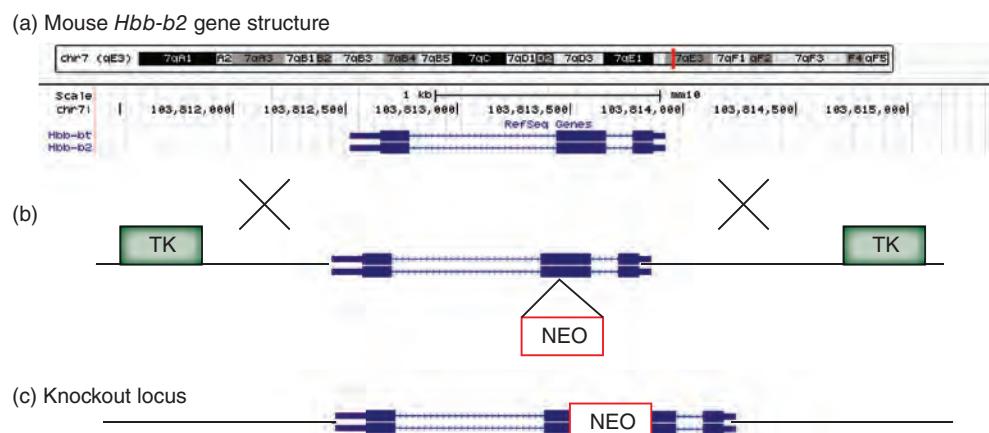


(b) Mouse beta globin cluster region



**FIGURE 14.8** The  $\beta$  globin locus (a) on human 11 (85 kilobases on chr11:5,235,001–5,320,000, GRCh37/hg19 assembly) and (b) on mouse chromosome 7 (65 kilobases on chr7:103,806,001–103,871,000, GRCm38/mm10 assembly). In (a), this region includes six globin RefSeq genes (including a pseudogene), ranging from HBB (arrow 1) to HBE (arrow 2). A locus control region is indicated (arrow 3) in the intergenic region upstream of HBE. Annotation tracks (“hub tracks”) from the UCSC Genome browser, derived from the ENCODE project are shown including ChIP-seq peaks for GATA1. Other tracks show DNAase I hypersensitivity, indicating genomic loci that are likely to have regulatory functions because they are in a conformation that is susceptible to DNase cleavage. Other annotation tracks show comparable patterns. A BED file was created corresponding to the beta globin locus control region sequence (given in accession AY195961). Note the prominent GATA1 peaks in the upstream regulatory region, some of which are conserved based on the PhastCons alignments. The properties of gene regulatory regions vary across cell types (e.g., erythrocytes and hematopoietic precursor K562 cells prominently display hypersensitivity sites) as well as at different developmental stages (e.g., fetal versus adult erythrocytes). In (b) mouse globin genes are displayed with a conservation track, showing multi-species conservation corresponding to exons as well as some conservation in noncoding regions, corresponding to cis-regulatory elements.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.



**FIGURE 14.9** Method of gene knockout by homologous recombination. (a) Structure of the  $\beta$  globin gene locus (from the UCSC Genome Browser, mouse GRCm38/mm10 assembly chr7:103,811,401–103,815,400), showing three exons that are transcribed from right to left. (b) Schematic of the linearized targeting vector used by Shehee *et al.* (1993). It includes the  $\beta$  globin gene with a neo gene inserted into exon 2 to allow for selection based on conferring resistance to the drug G418. Copies of the thymidine kinase (TK) gene from herpes simplex virus 1 flank the homologous segments and are also used for selection. The large X symbols indicate regions where crossing-over can occur between homologous segments. (c) The successfully targeted locus includes a  $\beta$  globin gene that is interrupted by the *neo* gene. Note that this gene is labeled *Hbb-b1* in the UCSC Genes track and *Hbb-b2* in the RefSeq and Ensembl tracks.

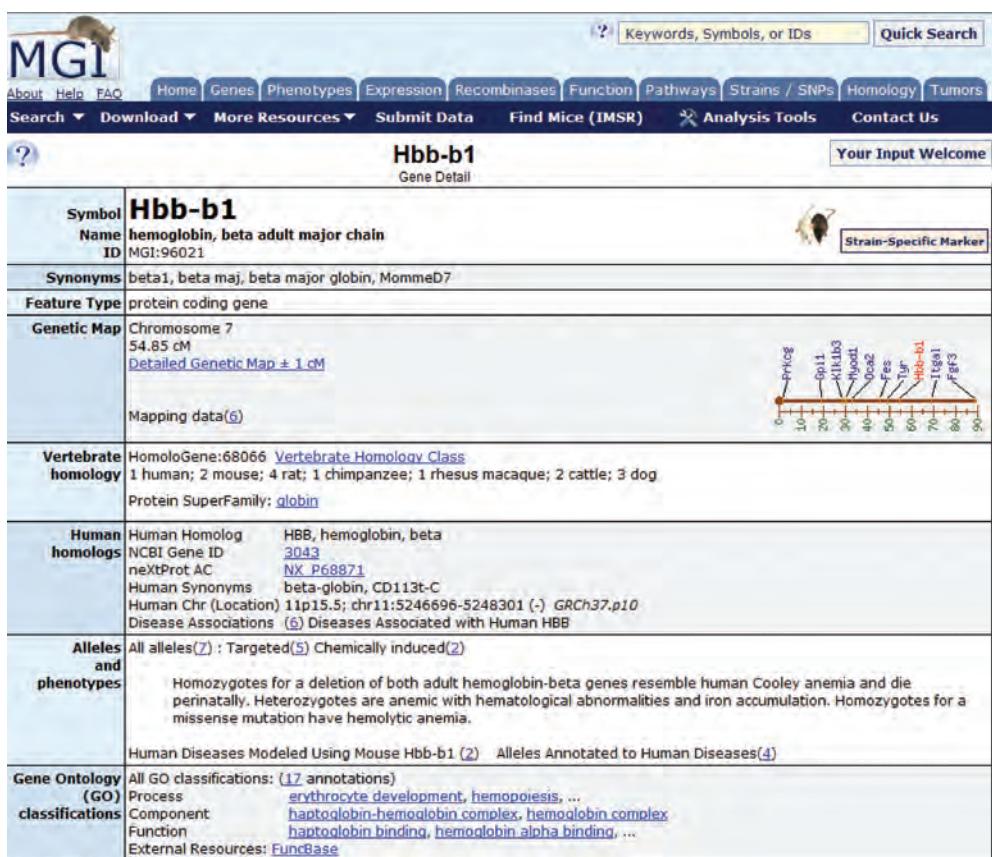
Source: (a) <http://genome.ucsc.edu>, courtesy of UCSC. (b) Adapted from Shehee *et al.* (1993), with permission from the National Academy of Sciences.

sequence neighboring each gene that the locus control region can regulate. In the naturally occurring deletion associated with nonlethal thalassemia, the locus control region interacts with just the *b2* gene (and mediates a compensatory increase in *b2*-derived globin protein). However, in the targeted mutant the locus control region regulates *b2* and also interacts with two more promoters: the inserted *tk* promoter driving the *neo* gene; and the promoter of the deleted *b1* gene. The three promoters compete for factors associated with the locus control region, and so relatively functional *b2* mRNA is produced and the phenotype is lethal instead of mild.

This example highlights the complexity of creating a null allele with an insertion vector. Many other strategies have been introduced (reviewed in van der Weyden *et al.*, 2002) including a variety of positive and negative selection markers and the use of replacement vectors instead of insertion vectors that leave behind no selectable markers (and are therefore less likely to interfere with endogenous processes). Conditional knockouts permit activation (for “gain-of-function”) or inactivation (for “loss-of-function”) *in vivo*, and can be invoked at any time of development or, through the use of tissue-specific promoters, in any region of the body. Conditional knockouts can be used to study the effects of disrupting a gene while avoiding embryonic lethality.

Major, coordinated efforts are underway to collect knockout mice via initiatives at the Sanger Institute, the National Institutes of Health, and elsewhere (Austin *et al.*, 2004; White *et al.*, 2013). An ultimate goal is to systematically knock out all mouse genes using several approaches. It is proposed to generate null alleles, including a null-reporter allele for each gene (such as  $\beta$ -galactosidase or green fluorescent protein). The reporter allows the determination of the cell types that normally express that gene. It is further proposed to make mutated alleles using gene targeting, gene trapping, and RNA interference (discussed in “Reverse Genetics: Gene Silencing by Disrupting RNA” below). Mouse strain C57BL/6 is widely used and was the first

The NIH Knockout Mouse Project (KOMP) has websites at <http://www.nih.gov/science/models/mouse/knockout/> (WebLink 14.36) and <http://www.genome.gov/17515708> (WebLink 14.37). Data coordination is via the IMPC website <https://www.mousephenotype.org/> (WebLink 14.38). Currently, ~4500 mice have been produced, >700 are assigned for production and phenotyping, and ~10,000 embryonic stem cells have been produced (February 2015).



**FIGURE 14.10** The Mouse Genome Informatics (MGI) website entry for the major beta globin gene (*Hbb-b1*) summarizes molecular data on that gene and includes a phenotype category, indicating that seven mutant alleles are indexed (five targeted and two chemically induced). Additional alleles are reported for the broader query of the hemoglobin beta chain complex (not shown).

Source: MGD, Blake *et al.* (2014). Reproduced with permission from MGI.

strain to have its genome sequenced. Efforts include the Knockout Mouse Project (KOMP), the European Conditional Mouse Mutagenesis Program (EUCOMM), and the North American Conditional Mouse Mutagenesis Project (NorCOMM) as well as a series of European initiatives (Ayadi *et al.*, 2012; International Mouse Knockout Consortium *et al.*, 2007).

The MGI website (**Fig. 14.6**) provides portals for browsing available knockout resources. This includes Deltagen and Lexicon Knockout Mice, and KOMP genes. As an example of a search for a specific gene, enter “globin” into the main search box at the MGI website and follow the link to *Hbb-b1* (the beta globin adult major chain on chromosome 7; **Fig. 14.10**). This page includes information about the gene as well as a link to phenotypic alleles (**Fig. 14.11**). Detailed phenotypic data are provided, such as the body weight and effects on the hematopoietic system.

## Reverse Genetics: Knocking Out Genes in Yeast Using Molecular Barcodes

Knockout studies in the yeast *S. cerevisiae* are far more straightforward and also much more sophisticated than in the mouse for several reasons. The yeast genome is extremely compact, having very short noncoding regions and introns in fewer than 7% of its ~6000 genes. Also, homologous recombination can be performed with high efficiency. A consortium of researchers achieved the remarkable goal of creating yeast strains representing

The *Saccharomyces* Genome Deletion Project website is [http://www-sequence.stanford.edu/group/yeast\\_deletion\\_project/deletions3.html](http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html) (WebLink 14.39). It includes lists of available deletion strains, protocols, and datasets.

**Phenotypic Alleles**  
Query Results -- Summary

Symbol	Hbb-b1				
Name	hemoglobin, beta adult major chain				
ID	MGI:96021				
7 matching Alleles (1 Gene/Marker represented)					
Allele Symbol Gene; Allele Name	Chr	Synonyms	Category	Abnormal Phenotypes Reported in these Systems	Human Disease Models
<a href="#">Hbb-b1<sup>MormmeD7</sup></a> hemoglobin, beta adult major chain; modifier of murine metastable epialleles, D7	7	RBC14	Chemically induced (ENU)	hematopoietic, immune, integument, liver/biliary, mortality/aging	<a href="#">Beta-Thalassemia</a> 613985
<a href="#">Hbb-b1<sup>Rbc13</sup></a> hemoglobin, beta adult major chain; red blood cell mutant 13	7	RBC13	Chemically induced (ENU)	hematopoietic, immune, integument, liver/biliary, mortality/aging	<a href="#">Beta-Thalassemia</a> 613985
<a href="#">Hbb-b1<sup>tm1Ley</sup></a> hemoglobin, beta adult major chain; targeted mutation 1, Timothy J Ley	7		Targeted (knock-in)	no abnormal phenotype observed	
<a href="#">Hbb-b1<sup>tm1Shs</sup></a> hemoglobin, beta adult major chain; targeted mutation 1, Takaji Shirasawa	7	Hbb <sup>Pres</sup>	Targeted (knock-in)	cellular, hematopoietic, homeostasis, immune, muscle, respiratory	<a href="#">Hemoglobin--Beta Locus; HBB</a> 141900
<a href="#">Hbb-b1<sup>tm1Unc</sup></a> hemoglobin, beta adult major chain; targeted mutation 1, University of North Carolina	7	Hbb <sup>th-3</sup> , Hbb <sup>th3</sup>	Targeted (knock-out)		
<a href="#">Hbb-b1<sup>tm1(KOMP)Mbp</sup></a> hemoglobin, beta adult major chain; targeted mutation 1, Mouse Biology Program, UCDavis	7		Targeted (knock-out) <i>(Cell Line)</i>		
<a href="#">Hbb-b1<sup>tm1(KOMP)Wtsi</sup></a> hemoglobin, beta adult major chain; targeted mutation 1, Wellcome Trust Sanger Institute	7		Targeted (knock-out) <i>(Cell Line)</i>		

**FIGURE 14.11** The MGI description of beta globin mutants includes phenotypic data such as the type of mutation (e.g., targeted knockout or conditional knock-in), the observed phenotypes, the human disease relevance, and the allelic composition (genetic background).

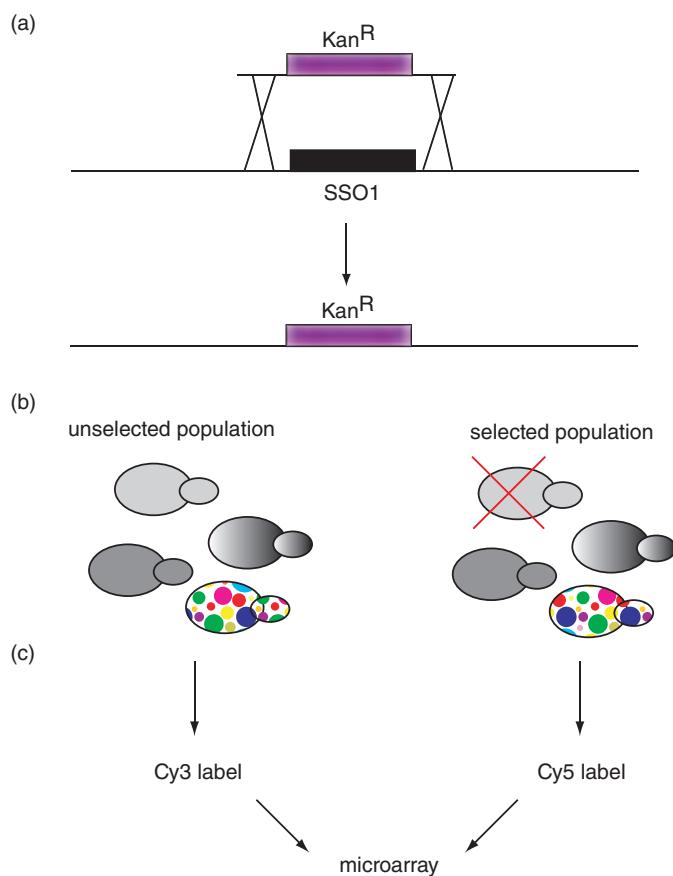
Source: MGD, Blake *et al.* (2014). Reproduced with permission from MGI.

the targeted deletion of virtually every known gene (Giaever *et al.*, 2002). The goals of this project were as follows:

- to create a yeast knockout collection in which all of the ~6000 ORFs in the *S. cerevisiae* genome are disrupted;
- to provide all nonessential genes (85% of the total) in four useful forms: (1) diploids heterozygous for each yeast knockout (*MATa* and *MATα* strains); (2) diploids homozygous for each yeast knockout; (3) a mating type (*MATa* haploid); and (4) α-mating-type (*MATα* haploid); knockouts of essential genes are only viable in the heterozygous diploids; and
- to provide all essential genes (15% of the total) as diploids heterozygous for each yeast knockout.

Within five years of the creation of the knockout strains, more than 5000 genes were associated with a phenotype based on three dozen publications (reviewed in Scherens and Goffeau, 2004). The strategy employed for this project is gene replacement by PCR, relying on the high rate of homologous recombination that occurs in yeast (Fig. 14.12a). A short region of DNA (about 50 bp), corresponding to the upstream and downstream portions of each open reading frame, is placed on the end of a selectable marker gene. Additionally, two “molecular barcodes” (an UPTAG and a DOWNTAG) are unique 20-base-pair oligonucleotide sequences included in each such deletion/substitution strain.

Budding yeasts have two mating types: *MATa* and *MATα*. Haploid *MATa* and *MATα* cells can mate with each other to form diploid *MATa/α* cells. Both haploid and diploid phases of the life cycle grow mitotically.



**FIGURE 14.12** Targeted deletion of virtually all *S. cerevisiae* genes. (a) The strategy is to use gene replacement by homologous recombination. Each gene (e.g., *SSO1*) is deleted and replaced by a *KanR* gene, with unique UPTAG and DOWNTAG primer sequences located at either end. (b) A variety of selection conditions can be used. (c) Genomic DNA is isolated from each condition, labeled with Cy3 or Cy5, and hybridized to a microarray. In this way, genes functionally involved in each growth condition can be identified.

This feature allows thousands of deletion strains to be pooled and assayed in parallel in a variety of growth conditions. The molecular barcode approach is extremely powerful. A collection of thousands of yeast knockouts can be grown in routine medium (Fig. 14.12b, unselected population) or in the presence of drug, temperature change, or other experimental condition (selected population). Some of the strains in the selected population might grow slowly (or die), and others might grow favorably. Genomic DNA is isolated, the TAGs (or molecular barcodes) are PCR amplified, labeled with Cy3 or Cy5 dyes, and hybridized to a microarray which contains all 12,000 molecular barcodes (20-mers) on its surface (Fig. 14.12c). Strains that are represented at high or low levels relative to the unselected population are identified based on unequal Cy3/Cy5 ratios on the microarray.

Giaever *et al.* (2002) used the yeast knockout collection to describe genes that are necessary for optimal growth under six conditions: high salt, sorbitol, galactose, pH 8, minimal medium, and treatment with the antifungal drug nystatin. Their findings include the following:

- About 19% of the yeast genes (1105) were essential for growth on rich glucose medium. Only about half of these genes were previously known to be essential. Beyond these 1105 genes, additional genes could be essential in other growth conditions.

- Nonessential ORFs are more likely to encode yeast-specific proteins.
- Essential genes are more likely to have homologs in other organisms.
- Few of the essential genes are duplicated within the yeast genome (8.5% of the non-essential genes have paralogs, while only 1% of the essential genes have paralogs). This supports the hypothesis that duplicated genes have important redundant functions (see Chapter 18).

The systematic deletion method offers a number of important advantages:

- All known genes in the *S. cerevisiae* genome are assayed.
- Each mutation is of a defined, uniform structure.
- Mutations are guaranteed to be null.
- Mutant knockout strains are recovered, banked, and made available to the scientific community.
- Studies of multigene families are facilitated.
- Parallel phenotypic analyses are possible, and many different phenotypes can be assayed.
- Once the strains have been generated, the labor requirement is low when a new phenotype is assessed.

This method also has limitations:

- The labor investment to generate these knockouts was very large.
- For each gene, only null alleles were generated for study. (Additional alleles may be available from other studies.)
- No new genes are discovered with this approach, in contrast to random transposon insertion approaches (described in the following section).
- All nonannotated ORFs are missed. In particular, short ORFs may not be annotated.
- Deletions in overlapping genes may be difficult to interpret.

Since over 80% of the yeast genes are nonessential, this implies that yeast can compensate for their loss through functional redundancy, perhaps by the presence of paralogs (such as *SSO1* and *SSO2*) in which the loss of one is compensated by the presence of the other. A similar scenario explains why deletion of the *b1* beta globin gene in mouse results in a mild disease due to upregulation of the activity of the paralogous *b2* gene. Another possibility is that parallel pathways exist such that if one is compromised the other can compensate. In this scenario, depicted in **Figure 14.4c**, the genes encoding members of each pathway need not be homologous. Another idea is that nonessential genes do not have redundancy or compensatory pathways, but are functionally required only under highly specific circumstances; under some experimental condition, they would therefore be found to be essential or at least to confer improved fitness.

How can we determine the functions of nonessential genes in yeast? One approach is to study synthetic lethality, in which a combination of two separate nonlethal mutations causes inviability (reviewed in Ooi *et al.*, 2006). A related concept is synthetic fitness in which two nonlethal mutations combine to confer a growth defect or other disruption that is more severe than that of either single mutation. Tong *et al.* (2001) devised a high-throughput strategy called synthetic genetic array (SGA) analysis to generate haploid double mutants (reviewed in Tong and Boone, 2006). A “query” mutation is crossed to an array of ~4700 “target” deletion mutants, and double mutant meiotic progeny that are inviable indicate that the two mutants are functionally related. Using 132 different query genes, Tong *et al.* (2004) identified a genetic interaction network having ~1000 genes and ~4000 interactions. The queries included nonessential genes as well as conditional alleles of essential genes. The results were consistent with

the behavior of a “small world network” in which immediate neighbors of a gene tend to interact together. In a related TAG array-based approach, Jef Boeke and colleagues defined functionally related networks of genes that are responsible for maintaining DNA integrity, the processes by which cells protect themselves from chromosomal damage (Pan *et al.*, 2006). They identified ~5000 interactions involving 74 query genes. This illustrates how functional pathways can be inferred using a genetic screen to identify modules of interacting proteins.

Another approach to gene function based on the yeast knockout collection is heterozygous diploid-based synthetic lethality by microarray analysis (dSLAM) (Ooi *et al.*, 2006; Pan *et al.*, 2007). In dSLAM, a “query” mutation is introduced into a population of ~6000 heterozygous diploid yeast “target” mutants. The pool of double heterozygotes is then haploidized by sporulation and the haploids are analyzed. A control pool consists of single target mutants, while the experimental pool consists of double (query plus target) mutants. TAGs from these two pools are labeled and analyzed on microarrays to define differential growth properties. Advantages of dSLAM are its use of molecular barcodes to quantify synthetic lethal relationships on microarrays, and its use of heterozygous diploid cells which accumulate fewer suppressor mutations that can confound analysis. A concern for all genetic interaction methods is that the false positive and false negative error rates may vary according to many factors, including the nature of the particular query.

A practical approach to finding genetic relationships between yeast genes is to use the SGD database. As shown for *SEC1* in **Figure 14.3**, five different types of genetic interaction were observed using a variety of genetic screens. (1) There were five dosage lethality interactions. These involved *SEC1*, *SEC4*, *SEC8*, and *SEC15* genes, and the identification of additional *SEC* genes suggest that these genes all function in a common pathway. In a dosage lethality experiment, overexpression of one gene causes lethality in a strain that is mutated or deleted for another gene. (2) There were 13 dosage rescue interactions in which overexpression of one gene rescues the deleterious phenotype (lethality or growth defect) caused by deletion of another gene. These interactions included *SEC3*, *SEC5*, *SEC10*, and *SEC15*. (3) There were five phenotypic suppression interactions in which mutation (or overexpression) of one gene suppresses the phenotype (other than a lethality or growth defect) caused by mutation or overexpression of another gene. These interactors included both *SEC* genes (*SEC6*, *SEC14*, *SEC18*) and *SNC1* (**Fig. 14.4**). (4) There was one synthetic growth defect interaction, in which the expression of two mutant genes in a strain, each of which causes a mild phenotype under some experimental condition, results in the phenotype of slow growth. This occurred between *SEC1* and *SRO7*. (5) There were 38 synthetic lethality interactions that resulted in the phenotype of inviability. These synthetic lethals included a range of genes, both in the *SEC* family and others.

### **Reverse Genetics: Random Insertional Mutagenesis (Gene Trapping)**

We have discussed targeted gene knockouts in mouse and yeast. Many other reverse genetics techniques have been developed (summarized in **Table 14.1**). Another high-throughput approach to disrupting gene function is called gene trapping. When this technique is applied to mouse, insertional mutations are introduced across the genome in embryonic stem cells (reviewed in Stanford *et al.*, 2006; Abuin *et al.*, 2007; Lee *et al.*, 2007; Ullrich and Schuh, 2009). Gene trapping is performed using vectors that insert into genomic DNA leaving sequence tags that often include a reporter gene. In this way, mutagenesis of a gene can be accomplished and the gene expression pattern of the mutated gene can be visualized. When the random insertional mutagenesis technique is applied to *Arabidopsis*, DNA is often introduced

**TABLE 14.1 Reverse genetics techniques. Adapted from Alonso and Ecker (2006), with permission from Macmillan Publishers Ltd.**

Method	Advantages	Disadvantages
Homologous recombination (e.g., gene knockouts)	A targeted gene can be replaced, deleted, or modified precisely; stable mutations are produced; specific (no off-target effects)	Low throughput; low efficiency
Gene silencing (e.g., RNAi)	Can be high-throughput; can be used to generate an allelic series; can restrict application to specific tissues or developmental stages	Unpredictable degree of gene silencing; phenotypes not stable; off-target effects are possible
Insertional mutagenesis	High-throughput; used for loss-of-function and gain-of-function studies; results in stable mutations	Random or transposon-mediated insertions target only a subset of the genome; limited effectiveness on tandemly repeated genes; limited usefulness for essential genes
Ectopic expression	Similar to gene silencing	Similar to gene silencing

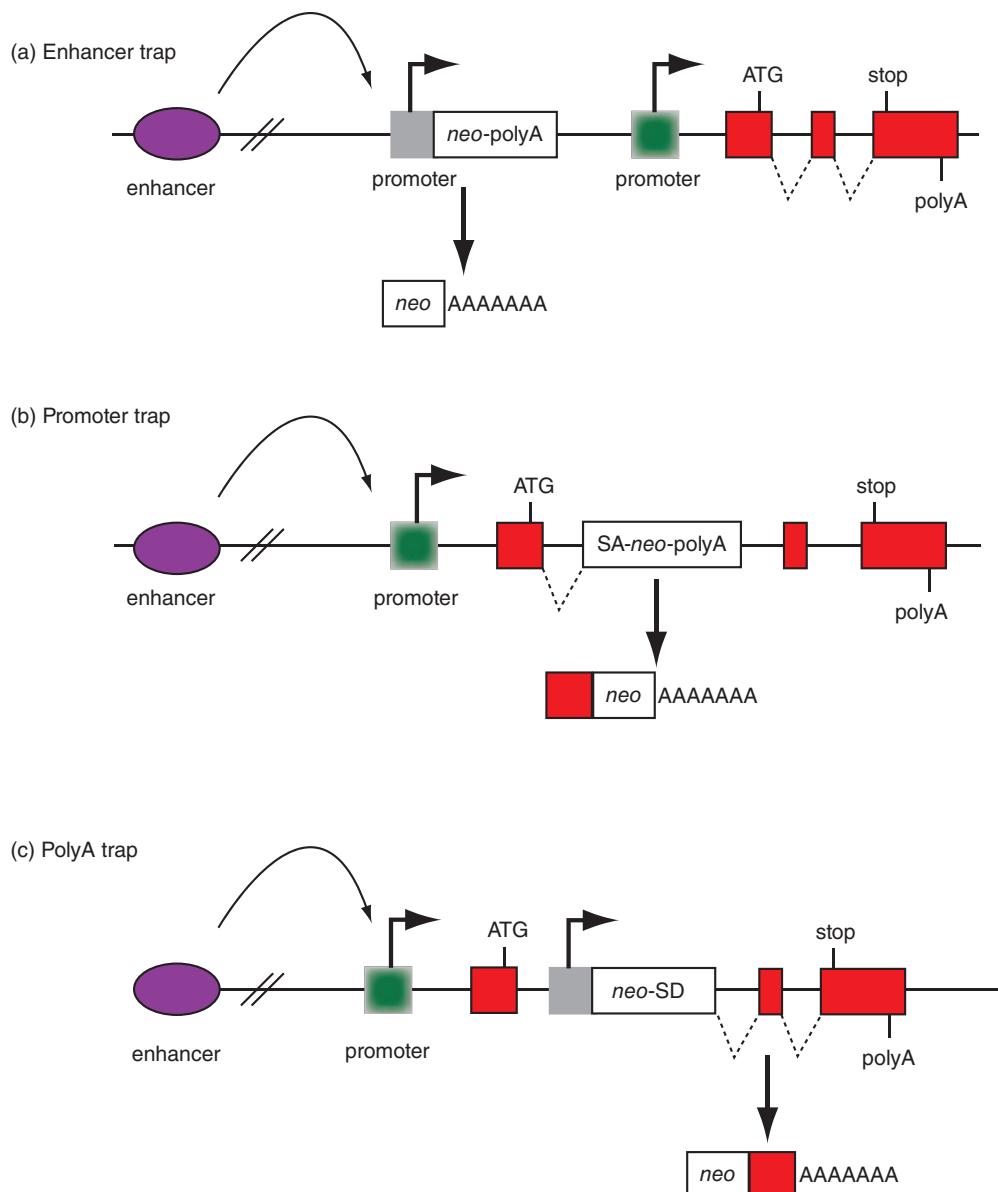
using the bacterium *Agrobacterium tumefaciens* as a vector (reviewed in Alonso and Ecker, 2006).

Gene trap vectors are typically transfected into mouse embryonic stem cells with subsequent expression of a selectable marker and resistance to antibiotics. Figure 14.13 shows three strategies for using gene traps in mouse. Each gene trap vector lacks an essential transcriptional component. An enhancer trap includes a promoter, neomycin resistance (neo) gene, and polyadenylation signal (Fig. 14.13a). It requires an endogenous enhancer to drive expression of the neo mRNA. A promoter trap lacks a promoter (but includes a splice acceptor and a selectable marker), and its expression is driven by the function of an endogenous promoter (Fig. 14.13b). PolyA traps have their own promoter that drives expression of neo, but they depend on external polyadenylation signals to successfully confer drug resistance (Fig. 14.13c). These traps are useful to trap untranscribed genes since they do not depend on activity of an endogenous promoter.

Gene trapping is a method of random mutagenesis and is not used to target a specific gene or locus. One strength of the method is that a single vector can be used to both mutate and identify thousands of genes. The technique also has the potential to trap genes that were not previously mapped; this contrasts with targeted approaches that require prior knowledge of the gene sequence. A limitation is that specific genes of interest cannot be targeted. Even a large-scale random mutagenesis experiment may fail to trap genes because of the nonrandom nature of the insertion sites in the genome (Hansen *et al.*, 2003).

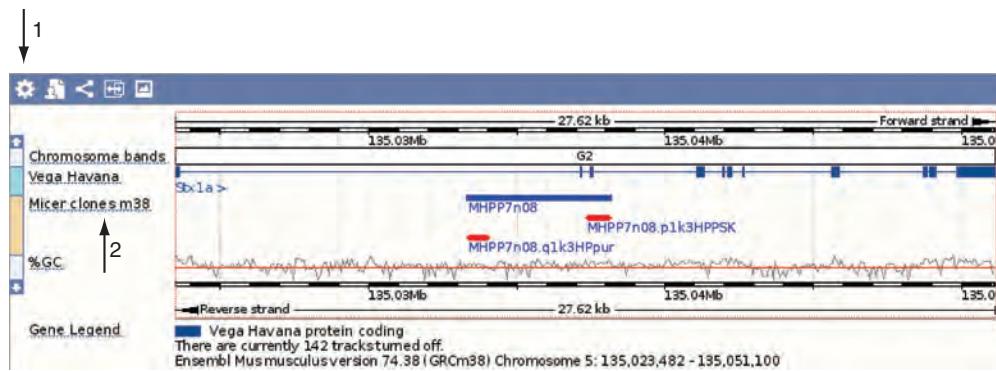
There are several large-scale insertional mutagenesis projects. The International Gene Trap Consortium (IGTC) manages a collection of ~45,000 mouse embryonic stem cell lines that represent ~45% of known mouse genes (Skarnes *et al.*, 2004; Nord *et al.*, 2006). The Mutagenic Insertion and Chromosome Engineering Resource (MICER) includes ~120,000 insertional targeting constructs that can be used to inactivate genes with a high targeting efficiency (28%; Adams *et al.*, 2004). You can view IGTC gene trap constructs at the UCSC Genome Browser, and both MICER and IGTC resources are available as annotation tracks at the Ensembl mouse genome browser (Fig. 14.14).

The International Gene Trap Consortium website is <http://www.genetrap.org> (WebLink 14.40).



**FIGURE 14.13** Strategies for gene trap mutagenesis. (a) An enhancer trap consists of a vector containing a promoter, a *neo* gene that confers antibiotic resistance (and therefore allows for selection of successfully integrated sequences), and a polyadenylation signal (polyA). This construct is activated by an endogenous enhancer, and disrupts the function of the endogenous gene. The endogenous gene is depicted with its own promoter, start codon (ATG), three exons in this schematic example, a stop codon, and a polyadenylation signal. (b) A promoter trap lacks an exogenous promoter and instead depends on an endogenous enhancer and promoter. It includes a splice acceptor (SA), *neo* cassette, and polyadenylation site. Integration of this vector disrupts the expression of an endogenous gene. (c) A poly(A) trap vector includes its own promoter and *neo* cassette but depends on an endogenous polyadenylation signal for successful expression.

Source: Abuin et al. (2007). Reproduced with permission from Springer Science and Business Media.



**FIGURE 14.14** Access to information on gene trapped genes at the Ensembl mouse genome browser. From the home page of Ensembl, select mouse syntaxin 1a (*Stx1a*) then use the configuration menu (arrow 1) to select MICER data (arrow 2) and International GeneTrap Consortium (not shown). Several MICER constructs are shown; these are vectors that are useful for generating knockout mice and for chromosome engineering.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

### Reverse Genetics: Insertional Mutagenesis in Yeast

We will describe two powerful approaches to gene disruption in yeast, in addition to homologous recombination: (1) genetic footprinting using transposons; and (2) harnessing exogenous transposons.

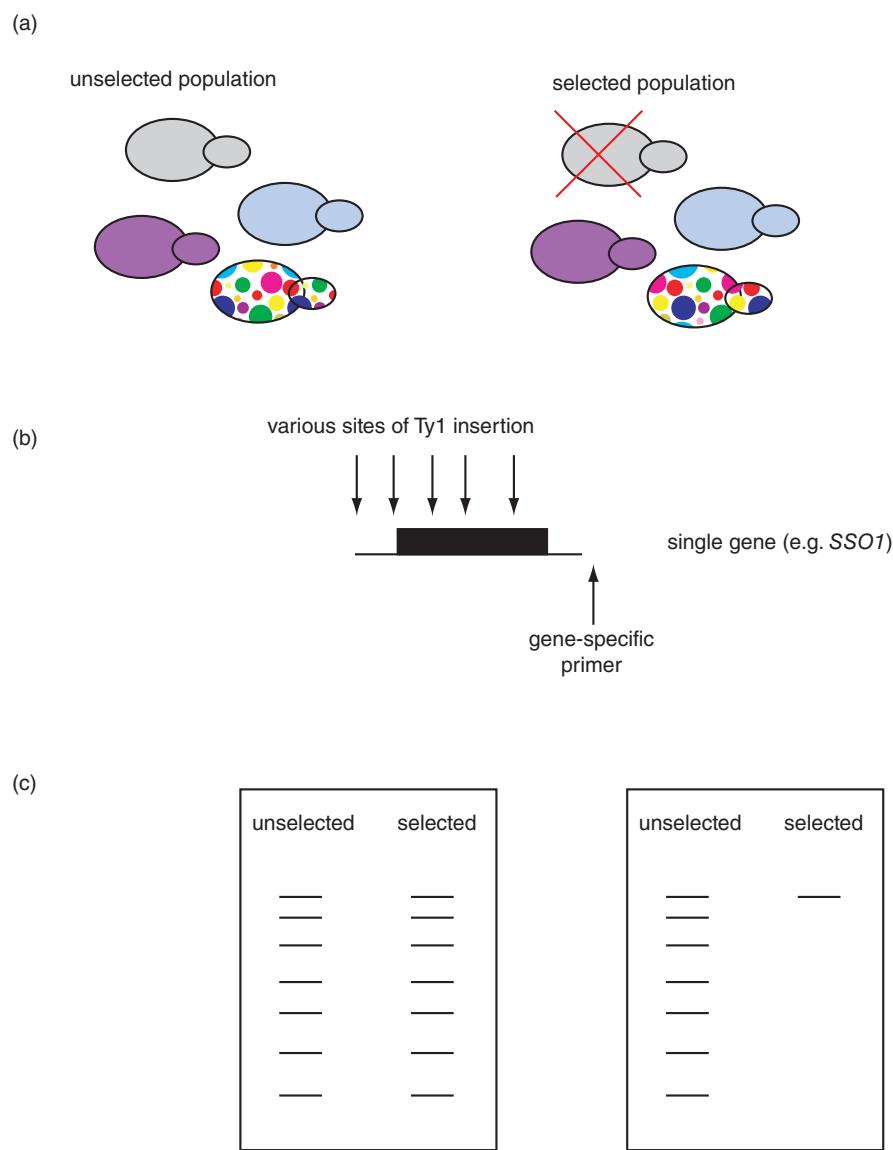
Transposons are DNA elements that physically move from one location to another in the genome (Chapter 8). They accomplish either with an RNA intermediate (retrotransposons) or without (DNA transposons). The Ty1 element is a yeast retrotransposon that inserts randomly into the genome. Patrick Brown, David Botstein, and colleagues developed a strategy in which populations of yeast are grown under several different conditions (e.g., rich medium versus minimal medium) and subjected to Ty1 transposon-mediated mutagenesis (Smith *et al.*, 1995, 1996; Fig. 14.15). Following the insertion, the polymerase chain reaction (PCR) is performed using primers that are specific to the gene and to the Ty1 element. This results in a series of DNA products of various molecular weights. The premise of the approach is that an individual gene (e.g., *SSO1*) might be important for growth under certain conditions. There will be a loss of PCR products (a “genetic footprint”) that indicates the importance of that gene for a particular condition.

This approach has several advantages:

- Any gene of interest can be assayed or genes can be selected randomly.
- Multiple mutations can be assayed for any given gene.
- It is possible to perform phenotypic analyses in parallel in a population.
- Many different phenotypes can be selected for analysis.
- The approach can succeed even for overlapping genes.

There are also several disadvantages:

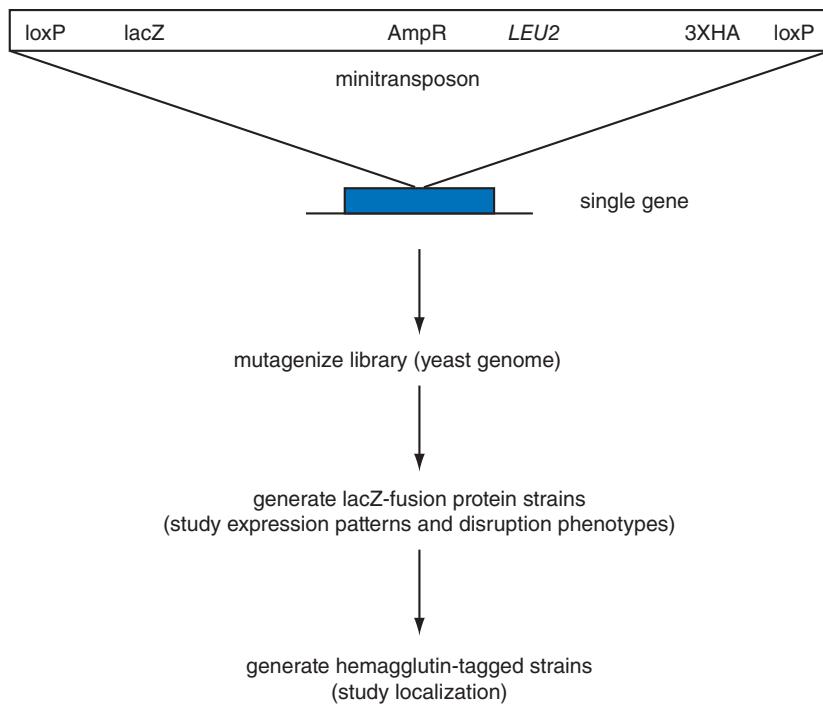
- Mutant strains are not recovered.
- Multiple mutations (alleles) are generated, but they are all insertions (rather than knockouts or other types of mutation).
- The approach is labor-intensive and entails a gene-by-gene analysis.
- The role of duplicated genes with overlapping functions may be missed.



**FIGURE 14.15** Genetic footprinting. (a) A population of yeast is selected (e.g., by changing the medium or adding a drug); some genes will be unaffected by the selection process. (b) Random insertion of a transposon allows gene-specific PCR to be performed and (c) subsequent visualization of DNA products electrophoresed on a gel. Some genes will be unaffected by the selection process (panel at left). Other genes, tagged by the transposition, will be associated with a reduction in fitness. Less PCR product will be observed (in (c)), therefore identifying this gene as necessary for survival of yeast in that selection condition.

Another mutagenesis approach involves the random insertion of reporter genes and insertional tags into genes using bacterial or yeast transposons (Ross-Macdonald *et al.*, 1999; Fig. 14.16). A minitransposon derived from a bacterial transposon Tn3 contains a lacZ reporter gene lacking an initiator methionine or upstream promoter sequence. When randomly inserted into a protein-coding gene, it is expected to be translated in-frame in one out of six cases. When this happens, the yeast will produce  $\beta$ -galactosidase, allowing the insertion event to be detected. The construct includes loxP sites that allow a recombination event in which the lacZ is removed and the target gene is tagged with only a short amount of DNA encoding three copies of a hemagglutinin (HA) epitope tag.

An HA-tagged protein can be localized within a cell using an antibody specific to HA.



**FIGURE 14.16** Transposon tagging and gene disruption to assess gene function in yeast. Adapted from Ross-Macdonald *et al.* (1999), with permission from Macmillan Publishers.

This minitransposon construct allows a genome-wide analysis of disruption phenotypes, gene expression studies, and protein localization. Ross-Macdonald *et al.* (1999) generated 11,000 yeast strains in which they characterized disruption phenotypes under 20 different growth conditions. These studies resulted in the identification of 300 previously nonannotated ORFs.

### Reverse Genetics: Gene Silencing by Disrupting RNA

We have discussed reverse genetics approaches in which a gene is deleted by homologous recombination. Another approach to identifying gene function is to disrupt the messenger RNA rather than the genomic DNA. RNA interference (RNAi) is a powerful, versatile, and relatively novel technique that allows genes to be silenced by double-stranded RNA (reviewed in Lehner *et al.*, 2004; Sachidanandam, 2004; Martin and Caplen, 2007). In plants and animals, small RNAs (21–23 nucleotides) regulate the expression of target genes. The extent of inhibition of gene function may be variable, in contrast to null alleles created by gene knockouts. Mechanistically, RNAi is a form of post-transcriptional gene silencing that is mediated by double-stranded RNA. It may function as a host defense system to protect against viruses, and RNAi may also serve to regulate endogenous gene expression. When double-stranded RNAs are introduced into *Drosophila*, nematode, plant, or human cells they are processed by the endoribonuclease Dicer into small interfering RNAs (siRNAs). These siRNAs cleave target messenger RNAs through the actions of an RNA-induced silencing complex (RISC) composed of proteins (such as Argonaute proteins) and RNA. The endogenous RNAi process seems to involve microRNAs (described in Chapter 10) rather than double-stranded RNAs.

RNAi has been used in genome-wide screens to systematically survey the phenotypic consequence of disrupting almost every gene. In *Drosophila*, Boutros *et al.* (2004) ascribed functions to 91% of all genes and reported 438 double-stranded RNAs that inhibited the function of essential genes. A further extension of the RNAi approach was provided by creating a transgenic RNAi library in *Drosophila* that permits targeted, conditional gene inactivation in virtually any cell type at any developmental stage. Dietzl *et al.* (2007) created an RNAi library that targets over 13,000 genes (97% of the predicted protein-coding genes in *Drosophila*). There are many false negative results, based on comparisons to a positive control set consisting of known phenotypes that are expected to occur based on previous classical genetics studies. This may occur because the library was constructed by randomly inserting transgenes into the fly genome, and not all transgenes express at sufficiently high levels. (The false negative rate for the library was ~40% and for the genes was ~35%.) There were also false positive results; some could occur because of off-target effects such as changes in the expression levels of genes flanking the target. As an example of the usefulness of this approach, Dietzl *et al.* described the use of a neuronal promoter to screen neuronal genes, and reported a lethal phenotype for many including *n-syb* (a homolog of *SNC1*/synaptobrevin, Fig. 14.4), *Snap* (a homolog of *SEC17*/ $\alpha$ SNAP), and *Syx5* (a homolog of *SSO1*/syntaxin).

While it is known that false positive results can occur, Ma *et al.* (2006) emphasized how extensive this problem can be. Off-target effects consist of RNAi constructs that inhibit the expression of endogenous genes other than those that are targeted. It is expected that sequences sharing a high degree of conservation to the small RNA regulator over a span of 19 or more nucleotides will also be targeted. In RNAi studies of *Drosophila*, Ma *et al.* noted off-target effects mediated by short stretches of double-stranded RNA. These false positives often contain tandem trinucleotide repeats (CAN where N represents any of the four nucleotides, with especially strong effects observed with CAA and CAG repeats). Such genes are overrepresented in the results of published RNAi screens. Ma *et al.* propose that libraries should be designed to avoid even short sequences present in multiple genes and, further, that identified phenotypic effects should be independently confirmed using more than one non-overlapping double-stranded RNA for each candidate.

RNAi screens have been performed in other organisms such as *C. elegans* (e.g., Kamath *et al.*, 2003; Kim *et al.*, 2005; Sönnichsen *et al.*, 2005). Remarkably, *C. elegans* can be fed bacteria that express double-stranded RNA to inhibit gene function (Fraser *et al.*, 2000). Kamath *et al.* performed a genome-wide RNAi screen and described mutant phenotypes for ~1500 genes, about two-thirds of which did not previously have an assigned phenotype. The most common RNAi phenotype they observed is embryonic lethality, observed in over 900 strains. In human, Berns *et al.* (2004) targeted ~7900 genes using retroviral vectors that encode over 23,000 short hairpin RNAs, and identified novel modulators of proliferation arrest dependent on p53, a key tumor suppressor and regulator of the cell cycle. Brass *et al.* (2008) used RNAi to systematically inhibit the function of human genes in a HeLa cell line transfected with short interfering RNAs. They identified 273 messenger RNAs that are required for human immunodeficiency virus (HIV) infection and replication in human cells. These human genes and gene products are potential targets for antiviral drugs. Unlike other antiretroviral drugs, potential drugs targeting these key human host proteins would not be affected by the extraordinary diversity of HIV genotypes (even within a single infected individual, there may be one million variant HIV genomes).

An HIV interaction database is available at <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/> (WebLink 14.41). Currently it lists ~7000 HIV-1 protein interactions involving 3500 different proteins. We discuss HIV in Chapter 16.

The GenomeRNAi database (Schmidt *et al.*, 2013) is available at <http://www.genomernai.de/> (WebLink 14.42). Currently it includes information on >140,000 *Drosophila* RNAi constructs and >320,000 RNAi reagents for human studies (February 2015). FLIGHT is available online at <http://flight.icr.ac.uk/> (WebLink 14.43). The RNAi Database (RNAiDB) is available at <http://rnai.org/> (WebLink 14.44) and focuses on *C. elegans* RNAi resources.

The Morpholino Database (MODB) is available at <http://www.morpholinodatabase.org/> (WebLink 14.45). It currently contains over ~1000 morpholinos.

ZiFiT Targeter is available online at <http://zifit.partners.org/> ZiFiT/ (WebLink 14.46). You can also perform CRISPR search and analysis at <http://crispr.mit.edu> (WebLink 14.47).

There are several prominent database resources for RNAi data. (1) The GenomeRNAi database integrates sequence data for RNAi reagents with phenotypic data from RNAi screens, primarily in cultured *Drosophila* cells (Horn *et al.*, 2007). A search of the GenomeRNAi database with the query *rop* (a *Drosophila* homolog of yeast *SEC1*) shows several RNAi probes (including the phenotype, the specificity, the occurrence of off-target effects, and the efficiency) as well as a link to the FlyBase gene entry. (2) FLIGHT also provides data on high-throughput RNAi screens (Sims *et al.*, 2006). Its scope and mission are comparable to the GenomeRNAi database. Both include a BLAST server, and FLIGHT contains additional analysis tools. (3) The RNAi Database is a similar database dedicated to *C. elegans* (Gunsalus *et al.*, 2004). A search for *UNC-18*, the *C. elegans* homolog of *SEC1/rop*/syntaxin-binding protein 1, shows a list of phenotypes observed in RNAi screens. For example, RNAi of *UNC-18* leads to resistance to the acetylcholinesterase inhibitor aldicarb, a drug that induces paralysis by preventing the normal breakdown of the neurotransmitter acetylcholine. This result is consistent with a functional role for *unc-18* in modulating the release of acetylcholine from vesicles in the presynaptic terminal at the neuromuscular junction.

A second approach to disrupting RNA is to knock down gene expression using morpholinos (Angerer and Angerer, 2004; Pickart *et al.*, 2004). Morpholinos are a form of antisense oligonucleotide consisting of a nucleic acid base with a morpholine ring and a phosphorodiamidate linkage between residues. They specifically bind to messenger RNAs (and microRNAs) and have been used to downregulate transcripts. They have been used extensively in zebrafish, and the ZFIN database describes the results of experiments using morpholinos. The MOpholino DataBase lists morpholinos and their targets, and associated phenotypic data (Knowlton *et al.*, 2008).

Several newly developed approaches offer the exciting possibility of engineering a genome by selectively modifying nucleotides of interest. (1) Zinc finger nucleases are engineered proteins that target genomic DNA of interest. Target specificity is gained through the amino acid sequences of DNA-binding zinc fingers, the number of fingers, and interaction of a nuclease with the target DNA. Zinc fingers have been used in model organisms including the rat (Geurts and Moreno, 2010) and zebrafish. (2) The transcription activator-like effector nucleases (TALENs) have been widely used to target sequences across a variety of organisms (Joung and Sander, 2013). They combine a nuclease that cleaves genomic DNA with a DNA-binding domain that can be directed to any target sequence of interest. (3) *Streptococcus pyogenes* and other bacteria and archaea use a system called clustered regularly interspaced short palindromic repeats (CRISPR)/Cas to defend against viruses and other foreign nucleic acids. RNA molecules guide a nuclease (Cas9) to a specific DNA site where cleavage occurs (Barrangou, 2013). The CRISPR/Cas system has been adapted to target and disrupt one or many genes in human and other cells (Le Cong *et al.*, 2013; Mali *et al.*, 2013) and to activate transcription (Perez-Pinera *et al.*, 2013).

The Zinc Finger Consortium produced a software package (ZiFiT Targeter) to help design zinc finger target sites as well as TALENs (Sander *et al.*, 2010). Its website recently expanded to include CRISPR/Cas resources as well. The group of George Church (Mali *et al.*, 2013) provided a resource of ~190,000 unique guide RNAs targeting ~41% of the human genome.

Zinc finger nuclease, TALEN, and CRISPR/Cas technologies do not have perfect specificity, and off-target or incidental cleavages are a potential concern. Cradick *et al.* (2013) targeted the beta globin (*HBB*) gene and identified incidental cleavage of the closely related delta globin (*HBD*) gene, sometimes using guide strands having just a one-base mismatch. They reported a series of insertion, deletions, and point mutations. As for any genome editing technology it is essential to control for such effects to correctly interpret research findings, especially as these technologies begin to have clinical applications.

## Forward Genetics: Chemical Mutagenesis

Forward genetics approaches are sometimes referred to as phenotype-driven screens. They are commonly performed using *N*-ethyl-*N*-nitrosurea (ENU), a powerful chemical mutagen used to alter the male germline (O'Brien and Frankel, 2004; Clark *et al.*, 2004; Probst and Justice, 2010; Stottmann and Beier, 2010; Horner and Caspary, 2011). ENU is more effective than X-ray irradiation,  $\gamma$ irradiation, or other chemical mutagens at inducing point mutations in organisms from mice to *Drosophila* to plants (Russell *et al.*, 1979). While the spontaneous mutation rate is about  $5\text{--}10 \times 10^{-6}$  for the average locus, ENU treatment typically yields a mutation frequency of about  $1 \times 10^{-3}$  per locus. These mutations tend to consist of single base substitutions, sometimes resulting in missense, splicing, or nonsense mutations. After ENU is administered to mice or other organisms, a phenotype of interest is observed (such as failure of neurons to migrate to an appropriate position in the spinal cord). Recombinant animals are created by inbreeding and the phenotype can then be demonstrated to be heritable. The mutagenized gene is mapped by positional cloning and identified by sequencing the genes in the mapped interval. In mice, ENU is used to mutagenize either spermatogonia or embryonic stem cells. O'Brien and Frankel (2004) reviewed the use of chemical mutagenesis in the mouse and emphasized the need for phenotyping that is both expert and high capacity.

Arnold *et al.* (2012) summarized their findings of 185 phenotypes associated with 129 genes, as well as 402 incidental mutations predicted to affect 390 genes (reviewed in Gunn, 2012). These findings are archived in the Mutagenetix database.

A major limitation of the ENU approach is that the gene(s) whose point mutations are responsible for the observed phenotypic change must be identified without the benefit of tags introduced into the genomic DNA. While positional cloning used to be a laborious process, the availability of complete genome sequences and dense maps of polymorphic markers has permitted relatively rapid identification of genes of interest. Michael Zwick and colleagues have applied next-generation sequencing to rapidly identify causal variants (Sun *et al.*, 2012). They apply multiplex chromosome-specific exome capture to simultaneously assess variants in mutant, parental, and background strains.

The use of balancer chromosomes has also facilitated the ENU approach (Hentges and Justice, 2004). In a balancer chromosome, a phenotypically marked chromosomal segment is inverted; this facilitates mapping as well as maintenance of mutations in the heterozygous state. This effect was first described by Hermann Muller (1918). Monica Justice and colleagues used the strategy of a balancer chromosome to characterize dozens of novel recessive lethal mutations on mouse chromosome 11 (Kile *et al.*, 2003). The balancer chromosome consists of mouse chromosome 11 harboring a large inversion (34 megabases). Male mice are treated with ENU, mated to females with the balancer chromosome and, through a strategy of successive intercrosses, mice that have a homozygous lethal mutation can be identified and the gene can be easily mapped.

Mutagenetix is online at <http://mutagenetix.utsouthwestern.edu/> (WebLink 14.48). It currently includes >300 phenotypes linked to genes, and ~200,000 incidental mutations identified in >20,000 genes.

## Comparison of Reverse and Forward Genetics

Both reverse and forward genetics approaches are powerful. We can contrast and compare several of their features.

- These approaches ask different questions. Reverse genetics asks “What is the phenotype of this mutant?” Forward genetics asks “What mutants have this particular phenotype?”
- Reverse genetics approaches attempt to generate null alleles as a primary strategy (and conditional alleles in many cases). Forward genetics strategies such as chemical mutagenesis are “blind” in that multiple mutant alleles are generated that affect a phenotype (Guénet, 2005). These alleles include hypomorphs (having reduced

function), hypermorphs (having enhanced function), and neomorphs (having novel function) as well as null alleles.

- We introduced techniques such as insertional mutagenesis (see above) as a form of reverse genetics. However, insertional mutagenesis has also been used in the context of forward genetics screens. In each case an attempt is made to infer the function of a set of genes based on the phenotypic consequence of disrupting the expression of a gene.

## FUNCTIONAL GENOMICS AND THE CENTRAL DOGMA

See <http://www.genome.gov/10000612> (WebLink 14.49) for a description of the NHGRI functional analysis program.

We have discussed reverse and forward genetics approaches to gene function. Another way to describe the scope of the field of functional genomics is to consider the central dogma that DNA is transcribed to RNA and translated to protein. These levels of analysis are reflected in the organization of functional genomics projects at the National Human Genome Research Institute (NHGRI) and elsewhere.

### Approaches to Function and Definitions of Function

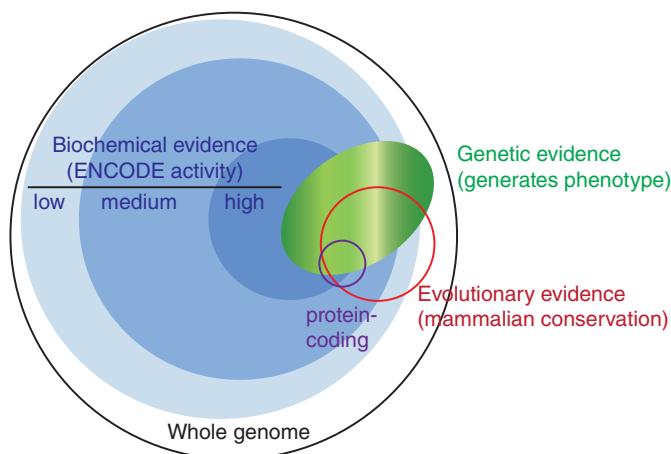
Function is a purpose or activity. In the context of bioinformatics and genomics, there is no single definition of function. Instead, function is considered in the context of a biological process such as development of the heart or metabolism of amino acids. We have already encountered function in our recent chapters. In Chapter 8 we discussed the assertion of the ENCODE Project Consortium that over 80% of the human genome is functional because it transcribes RNA and/or binds proteins that are involved in gene regulation.

It is important to distinguish three different *approaches* to function from three *definitions* of function (Fig. 14.17). In the context of interpreting the function of genomic DNA, genetic approaches can be adopted (e.g., establishing the consequence of sequence alterations by knocking out a gene in a mouse, or studying the consequence of a microdeletion syndrome in a patient). A second approach is evolutionary: we can align homologous DNA and/or proteins. In Chapter 13 we introduced structural genomics initiatives, many of which define inferred protein function according to the inference of homologous superfamilies. A third approach is biochemical: we can measure an activity in a given cell type and physiological condition. The ENCODE biochemical map describes many biochemical events that will facilitate hypothesis testing such as examining the consequence of knocking out long noncoding RNA genes.

The *approaches* are closely related to the *definitions* of function. The first definition is that functional processes are evolutionarily selected. Since the emergence of the neutral theory of evolution (Kimura 1968, 1983; Chapter 7) it has been posited that the majority of DNA changes are neutral or nearly neutral. However, functional elements are under positive natural selection according to this definition of function and can be determined by identifying genomic loci under constraint and by characterizing the consequences of naturally occurring mutations or targeted mutations such as described in this chapter.

A second definition of function is that of a causal role: in genetics a gene knockout results in a phenotype, allowing us to infer the normal function of the gene. From the evolutionary perspective, this definition of function implies that conserved loci, identified by comparative genomics, are functional. There are many caveats: some ultraconserved loci seem to be dispensable (highlighting the need to perform appropriate phenotyping to identify the particular conditions in which some change has a deleterious consequence). A criticism of this second definition of function is that causal roles may be identified that are irrelevant to biology. A heart causes the sound of a beat, but this does not mean that the key function of the heart is to make sound. A segment of DNA may be transcribed into RNA, but it remains to be established that this is biologically relevant.

		Approach to function		
		Genetic	Evolutionary	Biochemical
Definition of function	Establish consequence of sequence alterations	Comparative genomics: align DNA, proteins	Measure an activity in a given cell type	
Evolutionary selected effect	<ul style="list-style-type: none"> <li>Naturally occurring or targeted mutations can be a “gold standard”</li> <li>Possible to infer function based on selection</li> </ul>	<ul style="list-style-type: none"> <li>&lt;15% of genome under constraint</li> <li>Noncoding regions often hard to align</li> </ul>		
Causal role	<ul style="list-style-type: none"> <li>Example: knockout generates a phenotype</li> <li>Caveat: some phenotypes depend on a particular condition to be identified</li> </ul>	<ul style="list-style-type: none"> <li>Many conserved loci functionally important</li> <li>Caveat: some ultra-conserved loci dispensable</li> <li>Caveat: some poorly conserved loci are functionally equivalent</li> </ul>	<ul style="list-style-type: none"> <li>There are increasing numbers of examples of mutations in enhancer regions that cause disease</li> </ul>	
Inferred selected effect	<ul style="list-style-type: none"> <li>Question inspired by ENCODE biochemical map: do most biochemical signatures correspond to functional sites that impact fitness?</li> </ul>	<ul style="list-style-type: none"> <li>Creation of ENCODE biochemical map may inspire new discoveries of sequence conservation in biochemically functional noncoding regions</li> </ul>	<ul style="list-style-type: none"> <li>Majority of genome functional</li> <li>An uncertain % drift, noise</li> <li>ENCODE biochemical map will facilitate hypothesis testing</li> </ul>	



**FIGURE 14.17** Distinguishing different approaches to function (columns) from definitions of function (rows). Considering these definitions and approaches clarifies the conclusions that can be drawn from projects such as ENCODE that ascribe function to the great majority of genomic DNA. The bottom figure shows three circles corresponding to the magnitude of functional findings in ENCODE. The bottom portion of this figure was redrawn from Kellis *et al.* (2014), with permission from PNAS.

A third definition of function is that of inferred selected effect. Every receptor that has been identified must bind some endogenous ligand, it is believed, because that is the inherent function of a receptor. However, for some receptors no endogenous ligand has yet been identified. This definition of function places faith in biological function that rises above the background noise of biological processes.

## Functional Genomics and DNA: Integrating Information

BioSystems and FLink are available at <http://www.ncbi.nlm.nih.gov/biosystems/> (WebLink 14.50).

One goal of functional genomics is to provide integrated views of DNA, RNA, protein, and pathways. Many resources (such as those at Ensembl, EBI, and NCBI) offer this integrated view. As an example, the NCBI BioSystems database describes groups of molecules that interact in some biological system (Geer *et al.*, 2010). BioSystems serves as a repository of pathways and other data, and it is an interface to the Entrez system (Chapter 2). The Frequency weighted links (FLink) tool allows you to input a list of genes (or proteins or small molecules) and obtain a ranked list of biosystems. Begin by choosing a database (we select BioSystems; Fig. 14.18a) and input data of interest (we select globin in Fig. 14.18b via an Entrez search, but you can upload a list of identifiers). The output can include entries from KEGG, REACTOME, and GO from various species (Fig. 14.18c). Each entry has a BioSystems identifier (BSID), typically linking to a pathway map. The LinkTo option further provides links to relevant data in other NCBI databases, in this example to thousands of globin structure links (Fig. 14.18d).

It is routine for bioinformatics tools to link fields of information using relational databases. BioMart at Ensembl is a prominent example. Christopher Bouton introduced DRAGON in 2000, one of the earliest relational databases (Bouton and Pevsner, 2000). DRAGON automatically downloaded and interconnected databases such as UniGene, SwissProt, KEGG (see “Pathways, Networks, and Integration” below), and Pfam. Bouton recently introduced Cortellis™ Data Fusion, an analytic platform that integrates multiple data sources and is often used in the pharmaceutical and biotechnology industries.

The vision of the ENCODE project is inherently integrative. Its main goal has been to catalog a “parts list” of functional elements in the human genome at the level of genomic DNA elements that act at the RNA and protein levels, including regulatory elements that control gene activity.

What are the phenotypic effects of genomic variation? The Critical Assessment of Genome Interpretation (CAGI) is a community-based project in which research teams are provided genetic variants and must predict molecular, cellular, or organismal phenotypes. CAGI is modeled after Critical Assessment of Structure Prediction (CASP; Chapter 13). The 2013 CAGI experiment included <200 predictions from 33 different research groups. Examples of challenges are identifying asthma or other disease-associated variants in personal genomes, or predicting which *BRCA1* mutations confer increased risk of breast cancer.

Cortellis Data Fusion is available at <http://thomsonreuters.com/cortellis-data-fusion/> (WebLink 14.51). While DRAGON is no longer contemporary because it has been superceded by tools such as BioMart, it continues to operate at <http://pevsnerlab.kennedykrieger.org/> (WebLink 14.52).

The ENCODE website at the UCSC Genome Bioinformatics site is <http://genome.ucsc.edu/ENCODE/> (WebLink 14.53), and the ENCODE homepage at NHGRI is <http://www.genome.gov/10005107> (WebLink 14.54).

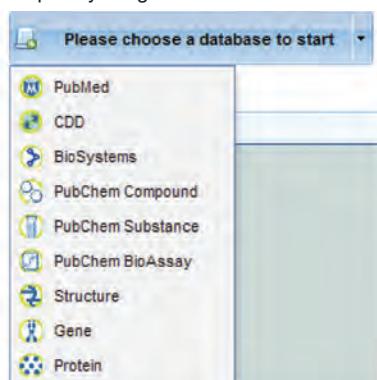
The CAGI website is <https://genomeinterpretation.org/> (WebLink 14.55).

## Functional Genomics and RNA

Surveys of RNA transcript levels across different regions (for multicellular organisms) and times of development provide fundamental information about an organism’s program of gene expression. (The expression “gene expression profiling” is commonly used, although more precisely it is steady-state mRNA levels that are measured rather than the process of gene expression.) Many studies have surveyed changes in RNA transcripts levels across developmental stages of organisms, or across body regions. Microarrays have been used to measure gene expression patterns for thousands of *Drosophila* genes across many developmental stages (Arbeitman *et al.*, 2002). Similar studies have been performed for the mosquito (Koutsos *et al.*, 2007), *C. elegans* (Kim *et al.*, 2001), and other species. These experiments have gradually been complemented by RNAseq with its extended dynamic range and improved coverage of the transcriptome.

The *Saccharomyces* Genome Database (SGD) offers many resources to describe gene expression in yeast. For each gene, an expression summary plots the log<sub>2</sub> ratio of gene expression (x axis) versus the number of experiments (y axis; Fig. 14.3, lower right). That plot is clickable, so experiments in which SEC1 RNA is dramatically up- or down-regulated can be quickly identified.

(a) Frequency-weighted link: select database



(b) FLink: input identifiers or search terms



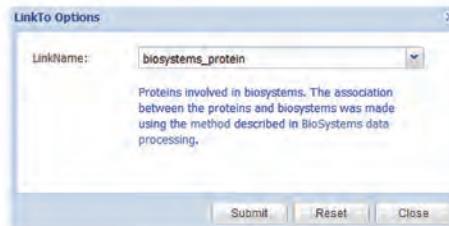
(c) FLink: table of globin results

**BioSystems**

BSID	Source	Name	Type	Organism
437	KEGG	Two-component system	conserved biosystem	
438	KEGG	Bacterial chemotaxis	conserved biosystem	
451	KEGG	Base excision repair	conserved biosystem	
83043	KEGG	Base excision repair	organism-specific biosystem	Homo sapiens
83240	KEGG	Base excision repair	organism-specific biosystem	Mus musculus
105837	REACTOME	DNA Repair	organism-specific biosystem	Homo sapiens
105838	REACTOME	Base Excision Repair	organism-specific biosystem	Homo sapiens
105839	REACTOME	Base-Excision Repair, AP Site Formation	organism-specific biosystem	Homo sapiens
105840	REACTOME	Depurination	organism-specific biosystem	Homo sapiens
105841	REACTOME	Recognition and association of DNA glycosylase with site containing an affected purine	organism-specific biosystem	Homo sapiens

(d) FLink LinkTo options

biosystems\_biosystems\_similar  
biosystems\_biosystems\_specific  
biosystems\_biosystems\_sub  
biosystems\_biosystems\_super  
biosystems\_cdd\_specific  
biosystems\_gene  
biosystems\_pcassay\_active  
biosystems\_pcassay\_target  
biosystems\_pccompound  
biosystems\_pcsubstance  
biosystems\_protein  
biosystems\_pubmed  
biosystems\_structure  
biosystems\_taxonomy



**FIGURE 14.18** NCBI offers the FLink resource to identify connections between an input list of proteins, genes, or other molecules and associated database entries. (a) Users first select a database, (b) enter search terms, and (c) obtain a table of results from assorted databases. Note the “LinkTo” option; (d) shows available links, each of which further connects the results to further database entries.

Source: FLink, NCBI.

RNA studies can also be integrated with DNA- and protein-level perspectives. As one example, Low *et al.* (2013) studied two rat strains for which the genome sequence was known (one of which is hypertensive). They performed RNA-seq from the liver of individuals of both strains (finding expression of >18,000 known genes) and mass spectrometry (finding evidence for ~26,000 peptides). This rich dataset allowed Low *et al.* to identify nonsynonymous variants, to find evidence for RNA editing (in which the genomic

DNA specifies a given codon but editing at the RNA level directs synthesis of a different protein sequence), and to characterize post-translational modifications. They could characterize the correlation between RNA and protein ( $r \sim 0.42$ ). Finally, they could identify a variant in the *Cyp17a1* gene that is a candidate for causing hypertension in one of the strains.

### Functional Genomics and Protein

Classical biochemical approaches to protein function involve an assay for the function of a protein (such as its enzymatic activity or a bioassay for its influence on a cellular process). This assay may be used as the basis of a purification scheme in which the protein is purified to homogeneity. Thousands of proteins have been studied individually with this approach. Each protein has its own personality in terms of biochemical properties and its propensity to interact with a variety of resins that separate proteins on the basis of size, charge, or hydrophobicity. We described several techniques to study proteins in Chapter 12, including two-dimensional gel electrophoresis and mass spectrometry.

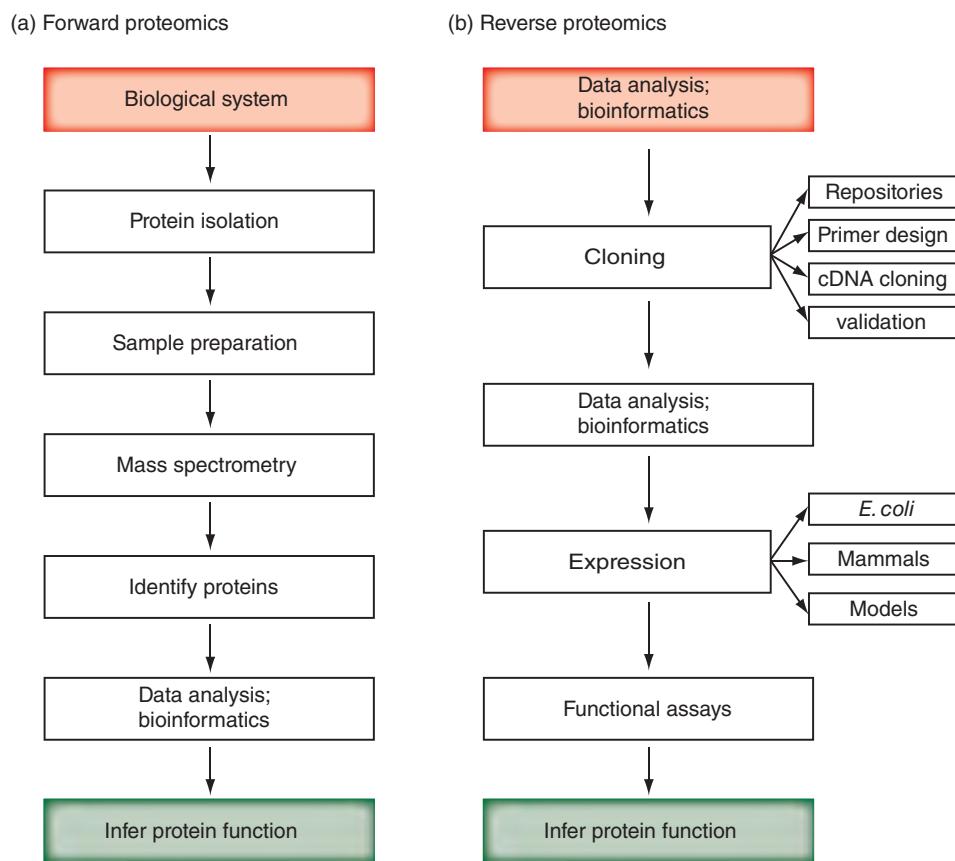
## PROTEOMICS APPROACHES TO FUNCTIONAL GENOMICS

In the remainder of this chapter we introduce proteomics approaches to functional genomics. We describe a protein function prediction experiment, protein–protein interactions, and then conclude with a study of protein pathways. The functions of most proteins are unknown. Even for relatively well-studied model organisms such as *Escherichia coli* and *S. cerevisiae*, functions have been assigned to only perhaps two-thirds of all proteins, and the function of the great majority of mouse or human proteins is unknown. The high-throughput proteomics projects attempt to assign function on a large scale, identifying the presence of proteins (in particular, physiological conditions) or identifying protein–protein interaction partners.

Basic features of proteins include their sequence, structure, homology relationships, post-translational modifications, localization, and function. In addition to the study of individual proteins, high-throughput analyses of thousands of proteins are possible (Molloy and Witzmann, 2002). We describe three such approaches: (1) identifying pairwise interactions between protein using the yeast two-hybrid system; (2) identifying protein complexes involving two or more proteins using affinity chromatography with mass spectrometry; and (3) analyzing protein pathways. While protein studies have been studied in-depth in a variety of model organisms, studies in *S. cerevisiae* are particularly advanced.

We have discussed forward genetics and reverse genetics approaches to gene function. A similar framework can be applied to proteomics (Palcy and Chevet, 2006). Forward proteomics approaches correspond to the classical approach to protein characterization (Fig. 14.19a). A biological system is selected, such as human cells from individuals with or without a disease. Proteins are compared by techniques such as mass spectrometry, differentially regulated proteins are identified, and from this the function of these proteins and their possible roles in the disease state may be inferred and further studied. In reverse proteomics, the starting point is genomic sequence from which genes, RNA transcripts, and protein products can be inferred (Fig. 14.19b). Complementary DNA (cDNA) clones can be obtained and expressed in a variety of systems so that their function may be assessed in assays for protein–protein interactions or other behaviors (cellular phenotypes).

Both forward and reverse proteomics approaches may be applied to discover protein function. Both of these may involve high-throughput techniques in which large numbers of samples and/or proteins are assayed. For example, in the forward proteomics approach of “isobaric tags for relative and absolute quantitation” (iTRAQ; Aggarwal



**FIGURE 14.19** Forward and reverse proteomics. (a) Forward proteomics. An experimental system is selected (such as a comparison of two developmental stages or normal versus diseased tissue). Proteins are extracted in a manner depending on the biological question that is addressed (e.g., selecting for membrane proteins or a subcellular organelle). Sample preparation may include steps such as polyacrylamide gel electrophoresis or chromatography columns to separate complex protein mixtures and reduce the complexity of the sample fractions being compared. Proteins may be labeled with fluorescent dyes or a variety of other tags, then they are separated and analyzed by techniques such as mass spectrometry (Chapter 12). Spectra are analyzed and differentially regulated proteins are identified. These regulated proteins may reflect functional differences in the comparison of the original samples. (b) Reverse proteomics. A genome sequence of interest is analyzed and genes, transcripts, and proteins are predicted based on a combination of computational and experimental evidence (discussed in Chapter 8 for eukaryotes). Complementary DNAs (cDNAs) are cloned based on information about open reading frames available in repositories and based on appropriate primer design. cDNAs are validated by sequence analysis and are then expressed in systems such as *E. coli* (for the production of recombinant proteins), mammalian cells, or other model organism systems. Functional assays are performed in order to assess function; assays include the yeast two-hybrid system or other protein interaction assays. Adapted from Palcy and Chevet (2006), with permission from John Wiley & Sons.

*et al.*, 2006), for eight or more protein samples of interest, the identity and relative quantity of 1000 proteins in each of these samples may be determined with high accuracy. Protein microarrays, analogous to DNA microarrays, consist of affinity reagents (such as specific antibodies) that are attached to a solid support (Sutandy *et al.*, 2013). Such technology can be challenging because of the difficulty in maintaining the structure (and function) of immobilized proteins. Nonetheless, it has been applied to diverse problems from characterizing enzyme activity to the detection of post-translational modifications, assessing antibody specificity, and measuring protein–protein interactions.

## Functional Genomics and Protein: Critical Assessment of Protein Function Annotation

The critical assessment of protein function annotation (CAFA) experiment is also modeled on CASP. More than 48,000 protein sequences were released to 30 participating teams who predicted Gene Ontology (GO; Chapter 12) annotations. The performances of various algorithms were assessed on a subset of 866 proteins for which “gold standard” GO annotation was available to the organizers, then used to assess the performance of prediction algorithms. CAGI involved many challenges inherent in the nature of protein function (Radivojac *et al.*, 2013):

- Protein function is defined at multiple levels, involving the role of a protein on its own and in pathways, cells, tissues, and organisms.
- Protein function is context dependent (e.g., many proteins change function in the presence of a signal such as calcium or a binding partner).
- Proteins are often multifunctional.
- Functional annotations are often incomplete and may be incorrect.
- Curation efforts map protein function to gene names, but multiple isoforms of a gene may have different functions.

Radivojac *et al.* concluded that the 2013 CAFA results included algorithms that outperform BLAST. In an independent analysis, Gillis and Pavlidis (2013) noted the fundamental reliance of all algorithms on sequence alignment with BLAST. Burkardt Rost and colleagues (Hamp *et al.*, 2013) used several homology-based predictors of protein function and noted that they performed well. Also, small changes to their algorithms could produce dramatically different results.

Gillis and Pavlidis (2013) also suggested that a major bottleneck is the inclusion of experimentally defined protein functions into annotation databases. Dessimov *et al.* (2013) also emphasize the occurrence of false positive errors. For example, a protein may be annotated as “receptor binding” by GO while other databases such as InterProScan may define it as “carbohydrate binding,” reflecting its authentic biological activity. If the InterProScan annotation is not transferred to other databases (such as SwissProt) and given GO functional annotation, then “carbohydrate binding” would be classified as a false positive result.

Future CAFA experiments are likely to continue evolving in design and in performance metrics. As with other competition experiments, they will move the community towards developing and benchmarking better methods for function prediction.

## Protein–Protein Interactions

Proteins are responsible for a dazzling variety of functions, from serving as enzymes to having structural roles. A consistent theme is that most proteins perform their functions in networks associated with other proteins and other biomolecules. As a basic approach to discerning protein function, pairwise interactions between proteins can be characterized (Williamson and Sutcliffe, 2010; Velasco-García and Vargas-Martínez, 2012). Proteins often interact with partners with high affinity. (The two main parameters of any binding interaction are the affinity, measured by the dissociation constant  $K_D$ , and the maximal number of binding sites  $B_{max}$ .) The interactions of two purified proteins can be measured with dozens of techniques such as the following:

- *Co-immunoprecipitation*, in which specific antibodies directed against a protein of interest are used to precipitate the protein to the bottom of a test tube along with any associated binding partners.

- *Affinity chromatography*, in which a cDNA construct is engineered that encodes a protein of interest in frame with glutathione S-transferase (GST) or some other tag such as polyhistidine. A resin to which glutathione is covalently attached is incubated with a GST fusion protein, and it binds to the resin along with any binding partners. Irrelevant proteins are eluted and then the specific binding complex is eluted and its protein content is identified.
- *Cross-linking with chemicals or ultraviolet radiation*, in which a protein is allowed to bind to its partners and then cross-linking is applied and the interactors are identified.
- *Surface plasmon resonance* (with the BIACore technology of GE Healthcare), in which a protein is immobilized to a surface and kinetic binding properties of interacting proteins are measured.
- *Equilibrium dialysis and filter binding assays*, in which bound free ligands (that is, a protein with and without its interacting partner) are separated and quantitated.
- *Fluorescent resonance energy transfer* (FRET), in which two labeled proteins yield a characteristic change in resonance energy upon sharing a close physical interaction.

We can approach the general issues associated with protein–protein interactions by considering the trafficking proteins shown in **Figure 14.4**. Some interactions occur in a pairwise fashion; for example, mammalian syntaxin binds to syntaxin-binding protein 1 in a binary complex (**Fig. 14.4d**). Syntaxin is also a member of several other complexes to the exclusion of syntaxin-binding protein; for example, syntaxin 1a, synaptobrevin-2/VAMP-2, and SNAP-25 (**Fig. 14.4b**) bind in a complex so tightly that they are able to migrate together as a trimer even in the harsh conditions of polyacrylamide gel electrophoresis that denatures most proteins. If purified syntaxin is immobilized on a column and mixed with an extract of rat brain, it is likely that two or more separate complexes will form as depicted in **Figure 14.4d** for Sso1p and other yeast orthologs. It would be incorrect to infer a direct binding interaction between syntaxin binding protein and synaptobrevin or SNAP-25. At the same time, it would be reasonable to conclude that all these proteins function as part of a common pathway. Finding genetic interactions can provide even more information about genes whose products function in a pathway or in parallel, related pathways (**Fig. 14.4c, e**). Genetic interaction data give less information about which particular proteins directly interact or which form protein complexes, but they may provide more information than studies of protein partners and protein complexes in terms of the members of protein pathways.

#### ***Yeast Two-Hybrid System***

The yeast two-hybrid system is a high-throughput method used to identify protein–protein interactions (Fields and Song, 1989; Fields, 2009). The assay is extremely versatile and has been used to identify protein-binding partners in many species. It is based upon the fact that the yeast *GAL4* transcriptional activator is composed of two independent activation and binding domains (Box 14.1). The cDNA encoding a protein of interest (the “bait”) is fused to the *GAL4* DNA binding domain. A large collection of cDNAs (a library consisting of various “prey”) is cloned into a vector containing the *GAL4* activation domain. Alone, the *GAL4* DNA binding domain does not activate transcription. However, when the bait binds to another fusion protein expressed from the cDNA library, the proximity of the two proteins enables transcription of a *GAL4* reporter gene. The name “two-hybrid” system refers to the use of two recombinant proteins that must interact.

In addition to the strategy of using a bait protein to screen a library, the yeast two-hybrid system has been used to measure the interaction of a known bait protein with individual, cloned prey proteins. In this way a set of many protein–protein interactions can be assayed. Compared to screening libraries, this approach has the advantage of systematically testing a matrix of possible protein–protein interactions; it has the disadvantage of

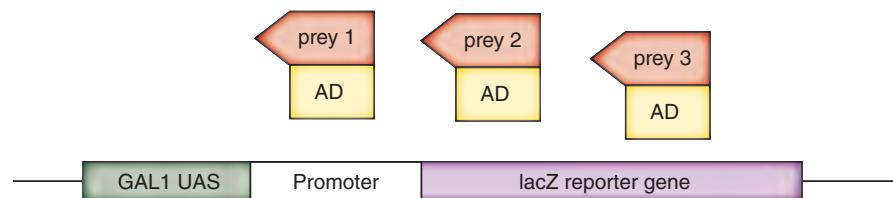
### BOX 14.1 YEAST TWO-HYBRID SYSTEM

The yeast two-hybrid system allows the identification of the binding partners of a protein. A cDNA encoding a protein of interest (such as huntingtin, the protein that is mutated in Huntington's disease) is used as a "bait" to identify interacting proteins in a library of cDNAs encoding human proteins expressed in brain ("prey"). A construct containing huntingtin cDNA, fused to a DNA binding domain (BD), is introduced into yeast cells. The BD interacts with a yeast *GAL1* upstream activating sequence (UAS) but, in the absence of an appropriate activator domain (AD), a *lacZ* reporter gene is not activated (see part (a) in figure below). A library of thousands of cDNAs is created, each fused to an activation sequence, but these alone are also unable to activate a reporter gene (see part (b)). When a clone from the library (AD fused to prey 1) binds to the bait/DNA BD construct, the activator domain is able to activate transcription of the *lacZ* reporter gene. This reporter allows identification of plasmid DNA from these yeast cells, and the prey 1 cDNA is sequenced. There may be many different binding partners identified from a yeast two-hybrid library. In one application of this technology, Li *et al.* (1995) identified huntingtin-associated protein (HAP-1), a protein enriched in brain that may affect the selective neuropathology of expanded polyglutamine repeats in Huntington's disease.

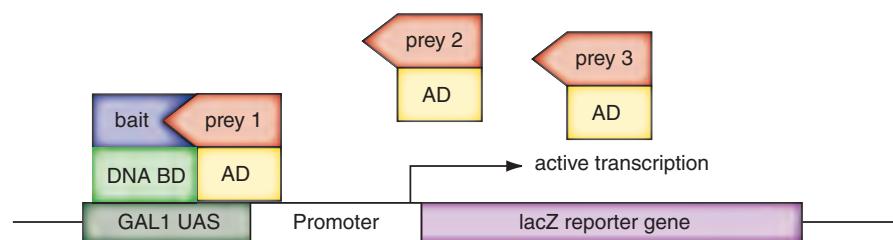
(a) DNA binding without activation



(b) Prey bound to activation domain



(c) Transcription activation upon prey binding to bait



not allowing the discovery of novel interacting partners that might be found in a complex cDNA library.

Yeast two-hybrid system technology has been applied to the analyses of essentially all possible pairwise protein–protein interactions in the yeast *S. cerevisiae*. Uetz *et al.* (2000) described 957 interactions involving 1004 yeast proteins, while Ito *et al.* (2001) identified 4549 interactions among 3278 proteins. These datasets are useful to define possible pathways of interacting proteins. Surprisingly, only about 20% of these two datasets overlap. The lack of concordance between these datasets may be due to differences in the physiological conditions in the studies, or to different sources of false positive and

false negative errors (discussed in the following). Other high-throughput yeast two-hybrid assays have been applied to *Drosophila* and other organisms (Giot *et al.*, 2003).

This experimental strategy entails a number of assumptions, including reasons for false positive results (biologically nonsignificant interactions) and false negative results (missed biological interactions; Schächter, 2002). False negative results may occur for the following reasons:

- The bait that is introduced into yeast cells must be localized to the nucleus. If the bait targets its native location, this could explain why some previously known interactions were not observed.
- The fusion protein construct must not interfere with the function of the bait protein.
- Transient protein interactions may be missed.
- Some protein complexes require highly specific physiological conditions in which to form, and may therefore be missed. Some interactions may fail in the specialized environment of the yeast nucleus.
- There may be a bias against hydrophobic proteins and low-molecular-weight proteins.

False positive results may also occur for a variety of reasons. Some proteins may be inherently susceptible to nonspecific binding interactions (i.e., they are “sticky” and activate many bait proteins). Proteins that are denatured may bind nonspecifically. A bait protein may autoactivate a reporter gene. Careful analysis of two-hybrid results allows these sources of false positive and false negative results to be reduced, for example by identifying promiscuous binding proteins.

Information about yeast two-hybrid data is available in several databases. The *Saccharomyces* Genome Database includes a link to physical interaction data including interactions from two-hybrid screens (Fig. 14.3, upper left). A search for Sec1p reveals several interaction partners including Sso2p and Mso1p. (When Mso1p was used as a bait in a reciprocal fashion, it was again found to bind to Sec1p.)

The yeast two-hybrid system has been extended to many other applications including RNA-protein interactions (Martin, 2012) and small molecule screening (Rezwan and Auerbach, 2012). Stynen *et al.* (2012) provide an extensive review of many related applications.

#### **Protein Complexes: Affinity Chromatography and Mass Spectrometry**

Affinity chromatography is a technique in which a ligand such as a protein is chemically immobilized to a matrix on a column. A major difference between the yeast two-hybrid strategy and the affinity chromatography approach is that the yeast two-hybrid system is only used to detect pairwise interactions between proteins. In contrast, an affinity chromatography approach allows subunits consisting of many proteins to be isolated and identified.

Many groups have employed a strategy of identifying thousands of multiprotein complexes in the yeast *S. cerevisiae* and other organisms (e.g., Gavin *et al.*, 2002, 2006; Ho *et al.*, 2002; Krogan *et al.*, 2006). Each group selected large numbers of “bait” proteins containing a tag that allowed each bait to be introduced into yeast, where they could form native protein complexes. After complexes were allowed to form under physiologically relevant conditions, the bait was extracted, copurifying associated proteins. These protein complexes were resolved by one-dimensional SDS-PAGE. Thousands of individual protein gel bands (from experiments with many different bait proteins) were excised from the gel with a razor, digested with trypsin to form relatively small protein fragments, and identified by MALDI-TOF mass spectrometry (Chapter 12).

Employing this strategy, Gavin *et al.* (2002) obtained 1167 yeast strains expressing tagged proteins, from which they purified 589 tagged proteins and identified 232 protein

complexes. Ho *et al.* (2002) selected 725 bait proteins and also detected thousands of protein–protein associations. In each case, a large number of the protein complexes that were identified included proteins of previously unknown function, highlighting the strength of these large-scale approaches. Gavin *et al.* (2006) performed a more comprehensive screen using tandem affinity purification coupled to mass spectrometry (TAP-MS) to create ~2000 TAP-fusion proteins. A total of 88% of these interacted with at least one partner, and the abundance of the identified binding partners ranged from 32 to 500,000 copies per cell. Gavin *et al.* developed a “socio-affinity” index measuring the log-odds of the number of times two proteins are observed interacting divided by the expected occurrence based on the frequency in the dataset. Krogan *et al.* (2006) also used TAP-MS and reported over 7000 protein–protein interactions involving ~2700 proteins. Employing a clustering algorithm they defined ~550 protein complexes averaging 4.9 subunits per complex. There was a large number of complexes with few members (two to four proteins), and few complexes with many members. Each of these various studies reported many complexes that were absent from the MIPS database, and they also revealed new members of previously characterized complexes. Krogan *et al.* reported enhanced coverage and accuracy because of technical improvements such as: (1) avoiding artifacts associated with protein overproduction; (2) systematically tagging and purifying both interacting partners; (3) using two methods of sample preparation and two methods of mass spectrometry; and (4) assigning confidence values to protein interaction predictions.

IntAct is available at the European Bioinformatics Institute (<http://www.ebi.ac.uk/intact>) (WebLink 14.56). Currently it contains ~87,000 proteins, >520,000 interactions, and >13,000 publications (February 2015). The main species covered in IntAct are *S. cerevisiae*, human, *Drosophila*, *E. coli* strain K12, *C. elegans*, mouse, and *Arabidopsis*.

Data from Gavin *et al.* (2006) and many other interaction experiments are available in the IntAct database (Kerrien *et al.*, 2012). A search for sec1 shows 16 interactors (although not Sso1p homologs, as recorded for yeast two-hybrid screens).

Basic questions about complexes include the stoichiometry (the number of various subunits), the subunit interactions, and the organization. Conventional biochemical techniques can be used to approach all these questions, and in some cases electron microscopy can reveal structural organization. Hernández *et al.* (2006) applied TAP-MS to several well-characterized complexes: the scavenger decapping and nuclear cap-binding complexes, as well as the exosome which contains ten different subunits. They could distinguish dimers from trimers and reveal subunit interactions that were not apparent using the yeast two-hybrid approach.

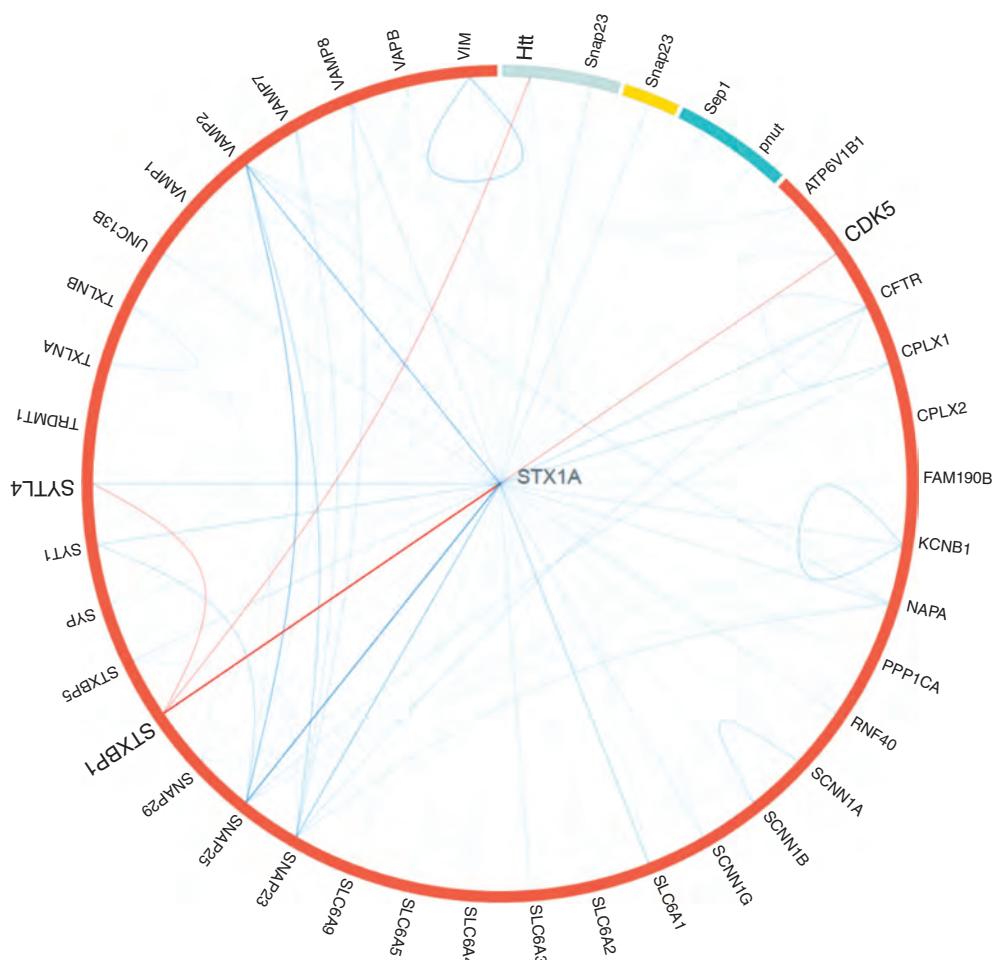
This approach yields false positive and false negative results for reasons similar to those presented above for yeast two-hybrid screens. While many complexes are identified repeatedly within a given experiment, indicating that saturation has been reached, this does not mean that those complexes are biologically real. Also, when a protein is identified by mass spectrometry it is usually accompanied by a confidence score. Peptides that are identified multiple times are associated with high confidence identifications, while “one-hit wonders” that are identified by one peptide observed in a single run are by definition present in low abundance and are more likely to be spurious or misidentified.

#### **Protein–Protein Interaction Databases**

Many prominent databases store information on protein–protein interactions as well as protein complexes; several of these are listed in Table 14.2. For example, the Biological General Repository for Interation Datasets (BioGRID) includes over 500,000 manually annotated interactions (Chatr-aryamontri *et al.*, 2013). An entry for human syntaxin (STX1A) shows connections such as to the syntaxin-binding protein STXBP1 (Fig. 14.20). Mathivanan *et al.* (2006) compared the content of eight major databases that include information on human protein–protein interactions. They emphasized the dramatic differences in their content including the number of reported interactions, the total number of proteins, the curation methodology, and the methods of detecting protein–protein interactions. Ooi *et al.* (2010) also reviewed major interaction databases,

**TABLE 14.2** Protein–protein interaction databases.

Database	Comment	URL
BioGrid	Repository for interaction datasets	<a href="http://www.thebiogrid.org/">http://www.thebiogrid.org/</a>
Biomolecular Object Network Databank (BOND)	Requires log-in; formerly BIND	<a href="http://bond.unleashedinformatics.com/">http://bond.unleashedinformatics.com/</a>
Comprehensive Yeast Genome Database (CYGD)	From the Munich Information Center for Protein Sequences (MIPS)	<a href="http://mips.helmholtz-muenchen.de/genre/proj/yeast/">http://mips.helmholtz-muenchen.de/genre/proj/yeast/</a>
Database of Interacting Proteins (DIP)	From UCLA	<a href="http://dip.doe-mbi.ucla.edu/">http://dip.doe-mbi.ucla.edu/</a>
Human Protein Reference Database (HPRD)	From Akhilesh Pandey's group at Johns Hopkins	<a href="http://www.hprd.org/">http://www.hprd.org/</a>
IntAct	At the European Bioinformatics Institute	<a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a>
Molecular Interactions (MINT) Database	Rome	<a href="http://mint.bio.uniroma2.it/mint/">http://mint.bio.uniroma2.it/mint/</a>
PDZBase	Database of PDZ domains	<a href="http://abc.med.cornell.edu/pdzbase">http://abc.med.cornell.edu/pdzbase</a>
Reactome	Curated resource of core human pathways and reactions	<a href="http://reactome.org/">http://reactome.org/</a>
Search Tool for the Retrieveal of Interacting Genes/Proteins (STRING)	Database of known and predicted protein–protein interactions	<a href="http://string.embl.de/">http://string.embl.de/</a>



**FIGURE 14.20** A BioGrid network map for human syntaxin and its binding partners.

Source: BioGrid, Courtesy of M. Tyers, TyersLab.

**TABLE 14.3 Overlap among interaction databases. Numbers of binary interactions are rounded to the nearest thousand, and percentages are rounded.**

	INTACT	MINT	BIOGRID	DIP	HPRD	MPACT	GNP	MPPI
INTACT <sup>a</sup>	83,000							
MINT <sup>b</sup>	54%	68,000						
BIOGRID <sup>c</sup>	16%	23%	138,000					
DIP <sup>d</sup>	46%	61%	61%	50,000				
HPRD <sup>e</sup>	22%	19%	15%	2%	37,000			
MPACT <sup>f</sup>	42%	46%	57%	49%	0%	12,000		
GNP <sup>g</sup>	1%	2%	1%	1%	5%	0%	1000	
MPPI <sup>h</sup>	10%	13%	8%	4%	36%	0%	0%	1000

<sup>a</sup><http://www.ebi.ac.uk/intact/>; <sup>b</sup><http://mint.bio.uniroma2.it/mint/Welcome.do>; <sup>c</sup><http://thebiogrid.org/>; <sup>d</sup><http://dip.doe-mbi.ucla.edu/dip/Main.cgi>; <sup>e</sup><http://www.hprd.org/>; <sup>f</sup>Currently unavailable; <sup>g</sup><http://genomenetwork.nig.ac.jp/public/sys/gnppub/portal.do>; <sup>h</sup><http://mips.helmholtz-muenchen.de/proj/ppi/>

Source: Ooi *et al.* (2010). Reproduced with permission from Springer Science + Business Media.

describing great differences in their scope and the limited overlap of their reported interactions (**Table 14.3**).

You can access PSICQUIC at the European Bioinformatics Institute (<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml>, WebLink 14.57).

Cytoscape can be downloaded from <http://www.cytoscape.org> (WebLink 14.58). Nearly 200 apps (formerly called plug-ins) are available.

We described some of the efforts of the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) in Chapter 12. Their work includes the introduction of a standard format to describe molecular interactions (Kerrien *et al.*, 2007) and efforts to unify search results from multiple databases through a PSI Common QUery InterfaCe (PSICQUIC; Orchard, 2012). Currently this includes 150 million binary interactions, searchable with free text or the Molecular Interaction Query Language. A search for syntaxin shows binding partners as defined by dozens of databases (**Fig. 14.21a**).

Protein network relationships can be viewed at the PSICQUIC website with a Cytoscape display at a broad scale (**Fig. 14.21b**) or zoomed in. Cytoscape is a popular software package for visualizing networks of genes, proteins, or other molecules (Cline *et al.*, 2007; Saito *et al.*, 2012). The input is a list of nodes (e.g., proteins) and inclusion of lists of edges and attributes is possible. By downloading Cytoscape software you can access pre-selected networks or import your own (**Fig. 14.21c, d**).

## From Pairwise Interactions to Protein Networks

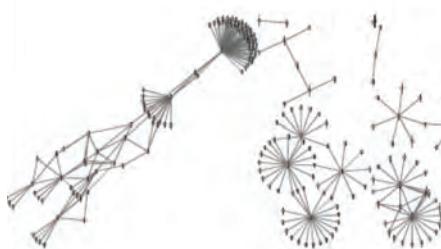
A typical mammalian genome has ~20,000 to 25,000 protein-coding genes, a subset of which (perhaps 10,000 to 15,000) are expressed in any given cell type. These proteins are localized to particular compartments (or are secreted) where many of them interact as part of their function. Some, such as the carrier proteins hemoglobin, myoglobin, retinol-binding protein, and odorant-binding protein, do not rely on protein–protein interactions but instead bind to a ligand (such as oxygen, vitamin A, or odorants) and transport it across a compartment by facilitated diffusion. Other proteins function through binary interactions; the majority function via protein complexes. In some cases, these complexes are spatially arranged in what Robinson *et al.* (2007) call the “molecular sociology of the cell.” These authors describe some of the techniques used to determine the structures of complexes, and they further describe the architecture of multisubunit structures such as the nuclear pore complex and the 26S proteasome.

Information about the roles of many proteins in a cell can be integrated in databases and visualized with protein network maps (Schächter, 2002; Bader *et al.*, 2003; Shafrazi *et al.*, 2004). A pathway is a linked set of biochemical reactions (Karp, 2001). The

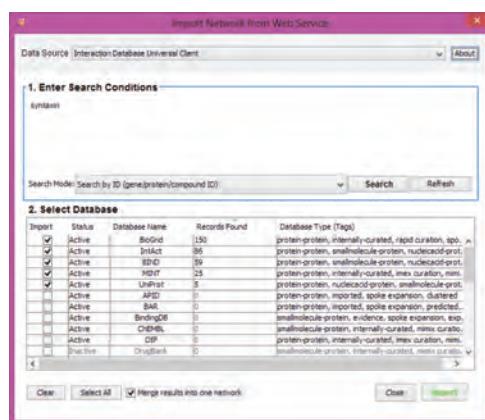
(a) PSICQUIC databases of protein interactions



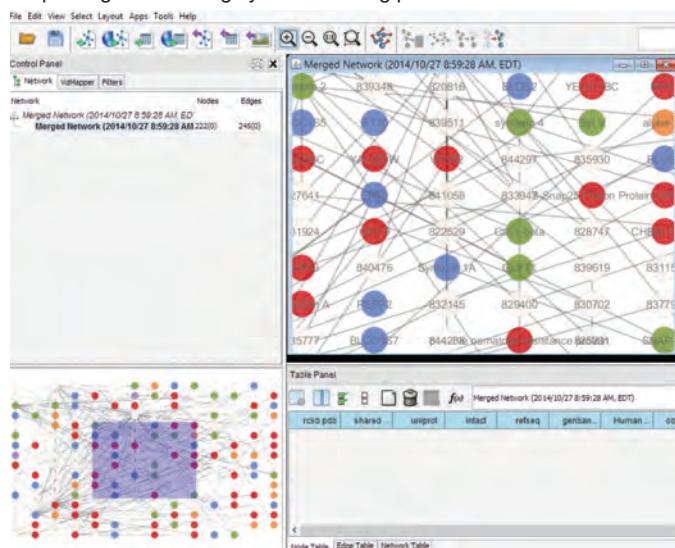
(b) PSICQUIC display of Cytoscape network for syntaxin



(c) Cytoscape data import



(d) Zoom of Cytoscape diagram showing syntaxin binding partners



**FIGURE 14.21** Protein interaction networks. (a) PSICQUIC is useful to retrieve data on protein-protein interactions from multiple sources; more than 150 million interactions may be queried. (b) A search for syntaxin shows >1,600 clustered binary interactions that are presented in a table and may be visualized on the PSICQUIC EBI website in a Cytoscape network. (c) Cytoscape software can be locally installed and, in a new session, data are imported from large, preset networks of human proteins. (d) An example is shown for syntaxin including a full network in a grid layout (bottom left) and a zoomed region (upper right, including several syntaxin paralogs and binding partners such as SNAP-25). PSICQUIC version 1.4.5 was queried from the EBI website. Cytoscape version 3.0.2 was used on a PC.

Source: PSIQUIC.

motivation behind making pathway maps is to visualize complex biological processes, that is, to use high-throughput data on protein interactions to generate a model of all functional pathways that is as complete as possible. There are unusual challenges associated with defining protein networks, as described in the following sections.

#### *Assessment of Accuracy*

One of the basic issues associated with a prediction is the assessment of its accuracy. How likely is it that a false positive or false negative error has occurred? To assess this, benchmark (“gold standard”) datasets are required that consist of trustworthy pathways. A particular approach to predicting or reconstructing pathways can then be tested to determine whether it is specific and sensitive. Unfortunately, relatively few interaction networks have been characterized in great detail; there are few accepted benchmark datasets comparable to those available for fields such as sequence alignment and structural biology. There is little concordance between major benchmark sets such as MIPS, Gene Ontology designations, and KEGG (see “Pathways, Networks, and Integration” below; Bork *et al.*, 2004). Improved validation methods have emerged in recent years (Braun, 2012). These include the use of reference sets as positive controls, random datasets as negative controls, assessing technical false positives (identified interactions that give a positive signal although the proteins do not physically interact), and identifying biological false positives (interactions that occur *in vitro* but not *in vivo*).

#### *Choice of Data*

A related issue is that the choice of data is critical. Many researchers integrate data from genomic sequences, expression of RNA transcripts, and protein measurements. It can be challenging to perform this integration since RNA and protein levels are often shown to be poorly correlated. Considering just protein–protein interaction data, for all high-throughput techniques the false positive and false negative error rates can be extremely high as we have seen (e.g., with yeast two-hybrid system data). Nonetheless many projects have proceeded to integrate the largest available datasets, including those with millions of predicted protein interactions, and also interactions as reported in thousands of literature references. For any study, it is essential to carefully evaluate the sources of error and the sensitivity and specificity of the assigned pathways.

#### *Experimental Organism*

The choice of experimental organism is important. Among the eukaryotes, *S. cerevisiae* is the best characterized: its genome encodes a relatively small number of genes; a tremendous amount of information is known about genes and gene products; and as a unicellular fungus it is simple compared to multicellular metazoans. In considering the use of different organisms to model pathways, a caveat is that even when orthologs of members of a particular pathway are identified, the function of homologs is not necessarily conserved across species. (When a protein has an established function in one species, an ortholog in a different species is often assigned the same function as a transitive property. When these orthologs do not actually share the same function, this situation has been called “transitive catastrophe.”) Mika and Rost (2006) analyzed high-throughput datasets from human, *Drosophila*, *C. elegans*, and *S. cerevisiae*. They introduced two metrics: an identity-based overlap measure that describes the overlap between two different datasets in the IntAct database within a single organism, and a homology-based measure that can be used to compare results from datasets in two different organisms. Their unexpected finding was that, for all organisms analyzed and at almost all levels of sequence similarity, inference of protein–protein interactions based on homology was dramatically more accurate for pairs of homologs from the same organism than for

homologs between different organisms. One significant aspect of this result is that, if two proteins are shown to interact in yeast, they do not necessarily interact in animals. Mika and Rost provide examples of protein sequences in *Drosophila* that have different binding partners than in yeast.

*C. elegans* represents another well-characterized organism, having 959 somatic cells and ~20,500 protein-coding genes that have been mapped into interaction networks (Walhout, 2011). A total of 940 of these nematode genes encode transcription factors that can regulate gene function.

### Variation in Pathways

In attempting to reconstruct networks on a global scale, another consideration is the great variation in the composition and behavior of different pathways. Some, such as the tricarboxylic acid (Krebs) cycle or urea cycle have been examined in depth for many decades; for example, extremely detailed maps of metabolic pathways are available at the ExPASy and KEGG websites. Other pathways are hypothetical or very poorly characterized. Some are constitutive, while others form transiently under particular physiological conditions or developmental stages. Some complexes are highly abundant, while others (such as the exocyst complex) appear to exist in vanishingly small quantities. For others, such as the vault complex (van Zon *et al.*, 2003), the function remains entirely obscure even after extensive studies.

The ExPASy website (Chapter 12) includes detailed maps for metabolic pathways and for cellular and molecular processes (<http://www.expasy.org/cgi-bin/search-biochem-index>, WebLink 14.59).

### Categories of Maps

There are different categories of network or pathway maps. These include maps based on metabolic pathways, physical and/or genetic interaction data, summaries of the scientific literature, or signalling pathways. For some screens (including the yeast two-hybrid system), information may be gained about the particular domain(s) within a protein that are responsible for interactions. Some maps are based on experimental data, while others mix computationally derived results (such as transfers of information from orthologous networks) with experimental data.

We can describe the properties of protein networks. In graphical representations of such complexes, nodes typically represent proteins while edges represent interactions. Most nodes are sparsely connected, while a few nodes are highly connected. Barabási and Albert (1999) suggested that most networks (including biological networks, social networks, and the World Wide Web) follow a scale-free power law distribution:

$$P(k) \sim k^{-\gamma} \quad (14.1)$$

where  $P(k)$  is the probability that a node in the network interacts with  $k$  other nodes, and  $P(k)$  decays following the constant  $\gamma$ . As a consequence, large networks self-organize into a scale-free state. According to this model, this power law distribution is a consequence of the continuous growth of networks and of the propensity of new nodes to attach preferentially to sites (nodes; here, proteins) that are already well connected. Two basic models that have emerged to describe protein complexes are a “spoke” model, in which a protein bait interacts with multiple partners like the spokes on a wheel, and a “matrix” model in which all proteins are connected (Bader and Hogue, 2002). Either of these models can encompass scale-free properties, although an analysis by Bader and Hogue indicates that a spoke model is more accurate. In reviewing eight databases of human protein interactions, Mathivanan *et al.* (2006) noted that the Human Protein Reference Database (HPRD) and Reactome databases include a large number of hub proteins that have many binary (direct) protein interactions. (The Reactome database assumes a matrix model with all proteins interconnected within a complex.) A similar finding applies to yeast; as discussed in “Protein Complexes” above, Krogan *et al.* (2006) described ~550 protein complexes of which about two dozen complexes had  $\geq 10$  members, while the majority

had 2–4 members. A property of networks having hub proteins is that random disruption of individual nodes (e.g., through mutation) is likely to be well tolerated, although the entire system is vulnerable to some failures at highly connected nodes (Albert *et al.*, 2000).

Many aspects of network properties have been further studied, such as the performance of different ways of creating and assessing confidence scores assigned to particular edges (interactions; Suthram *et al.*, 2006). Assigning confidence scores requires a benchmark (for example, STRING relies on KEGG, described in the following section), although it is challenging to define adequate benchmarks. Another aspect of protein networks is the nature of hub proteins. Haynes *et al.* (2006) showed that hub proteins (defined as having  $\geq 10$  interacting partners) have more intrinsic disorder than end proteins (those with one interacting partner) in worm, fly, and human. We described intrinsic disorder in Chapter 13. Yet another feature of networks is their modularity (Sharom *et al.*, 2004). One example of modularity is vesicle-mediated exocytosis of neurotransmitter in the mammalian nerve terminal (Fig. 14.4b). The components required for neurotransmitter release function autonomously at a great distance from the cell body, and respond to the arrival of an action potential (an electrical signal) in a local fashion by releasing neurotransmitters. This signaling system has a modular nature. Li *et al.* (2006) estimated the modularity as well as the clustering exponent  $\gamma$  for protein interaction networks in yeast, *C. elegans*, and *Drosophila*, reporting that all three have a scale-free nature and varying degrees of modularity.

PathGuide is at <http://www.pathguide.org/> (WebLink 14.60). BioGRID is available at <http://www.thebiogrid.org> (WebLink 14.61). SGD is at <http://www.yeastgenome.org> (WebLink 14.62). BioPAX is online at <http://www.biopax.org/> (WebLink 14.63).

MetaCyc is available at <http://metacyc.org/> (WebLink 14.64). There are currently over 2200 pathways, 5500 organism databases, and 49,000 citations in the database (February 2015).

KEGG is available at <http://www.genome.ad.jp/kegg/> (WebLink 14.65). The current release (February 2015) includes about 16 million genes from high-quality genomes (~300 eukaryotes, >3100 bacteria, and ~180 archaea), >130 million genes from metagenomes, and >350,000 pathways.

## Pathways, Networks, and Integration: Bioinformatics Resources

There are many database resources for global interaction networks. PathGuide is a website that lists 240 biological pathway resources (Bader *et al.*, 2006). These are organized into categories such as protein–protein interactions, metabolic pathways, signaling pathways, pathway diagrams, and genetic interaction networks. For *S. cerevisiae*, the BioGRID database (Reguly *et al.*, 2006) provides manual curation of ~32,000 publications describing physical and genetic interactions. It is available online at its own site and through the SGD (see Fig. 14.2, lower right side). In an effort to standardize the way various database projects present information, the Biological Pathway Exchange (BioPAX) consortium provides a data exchange ontology for biological pathway integration.

Several web servers provide pathway maps. MetaCyc is a database of metabolic pathways (Caspi *et al.*, 2008). It includes experimentally verified enzyme and pathway information, with links from pathways to genes, proteins, reactions, and metabolites. The SGD offers similar metabolic pathway maps for yeast, including data derived from MetaCyc.

A major pathway database is offered by the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa *et al.*, 2008; Fig. 14.22). The KEGG atlas contains a detailed map of metabolism based on 120 metabolic pathways, with links to various organisms. KEGG pathways are a collection of manually drawn maps in six areas: metabolism; genetic information processing; environmental information processing; cellular processes; human diseases; and drug development. An example of a pathway map is shown for vesicular transport (Fig. 14.23); by choosing *S. cerevisiae* from a menu of organisms, clicking on a box such as syntaxin links to an entry on yeast Sso1p. For all these pathway maps, the information obtained from biochemical studies is far richer and more accurate in terms of the identities of genes and gene products, their correct subcellular distributions, and the details of their interactions with partner proteins.

As another example of a KEGG pathway, by selecting human neurodegenerative disorders a pathway description of amyotrophic lateral sclerosis (ALS; Lou Gehrig's disease) can be found (Fig. 14.24). Mutations in the superoxide dismutase gene *SOD1* are a

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See [Release notes](#) for new and updated features).

### Main entry point to the KEGG web service

[KEGG2](#)      [KEGG Table of Contents](#)      [Update notes](#)

### Data-oriented entry points

<a href="#">KEGG PATHWAY</a>	KEGG pathway maps	[Pathway list]
<a href="#">KEGG BRITE</a>	BRITE functional hierarchies	[Brite list]
<a href="#">KEGG MODULE</a>	KEGG modules	[Module list]
<a href="#">KEGG DISEASE</a>	Human diseases	[Cancer   Infectious disease]
<a href="#">KEGG DRUG</a>	Drugs	[ATC drug classification]
<a href="#">KEGG ORTHOLOGY</a>	Ortholog groups	[KO system]
<a href="#">KEGG GENOME</a>	Genomes	[KEGG organisms]
<a href="#">KEGG GENES</a>	Genes and proteins	Release history
<a href="#">KEGG COMPOUND</a>	Small molecules	[Compound classification]
<a href="#">KEGG REACTION</a>	Biochemical reactions	[Reaction modules]

### Entry point for wider society

[KEGG MEDICUS](#)      Health-related information resource

### Organism-specific entry points

[KEGG Organisms](#)      Enter org code(s)   hsa hsa eco

### Analysis tools

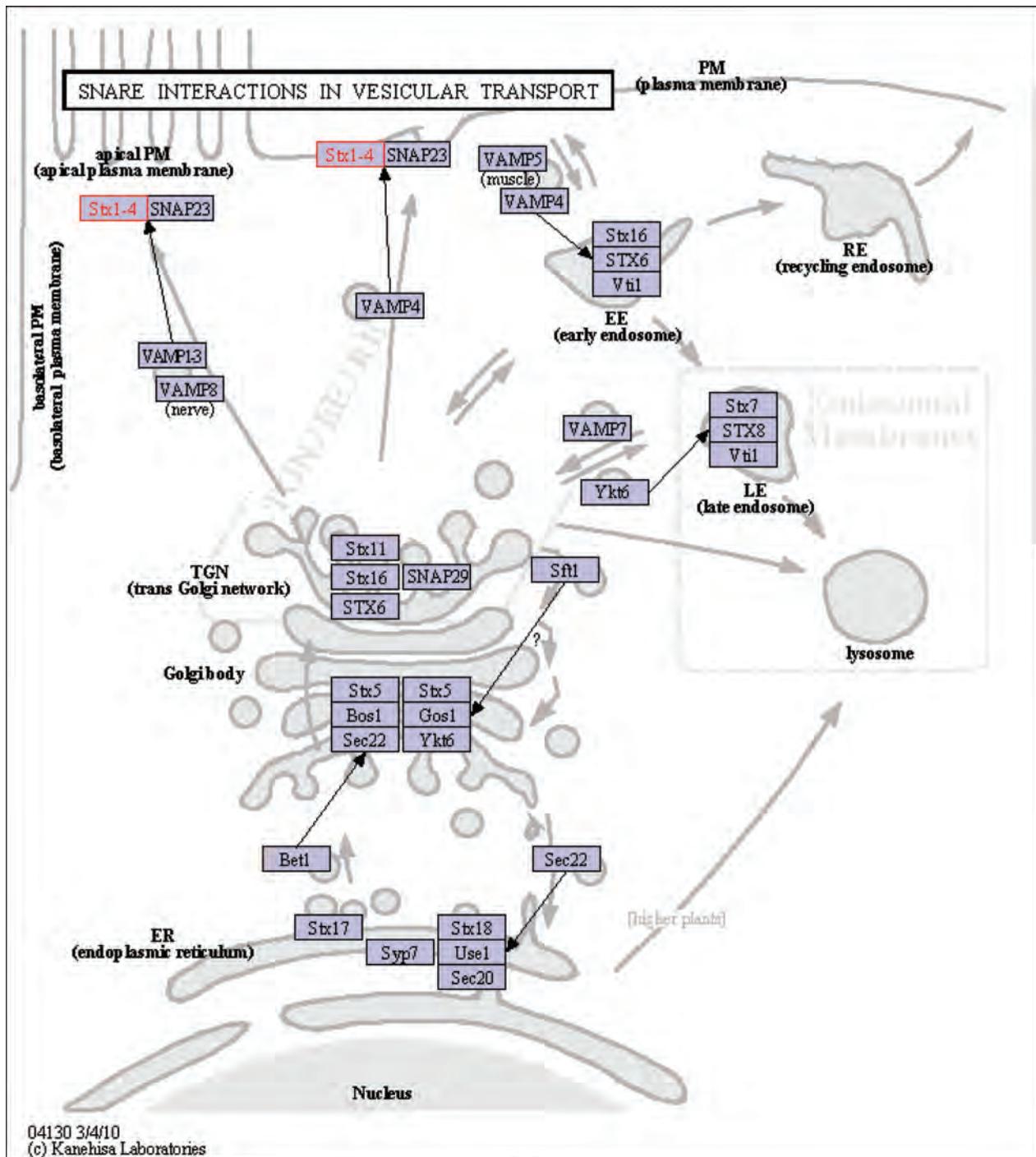
<a href="#">KEGG Mapper</a>	KEGG PATHWAY/BRITE/MODULE mapping tools
<a href="#">KEGG Atlas</a>	Navigation tool to explore KEGG global maps
<a href="#">KAAS</a>	KEGG automatic annotation server
<a href="#">BLAST/FASTA</a>	Sequence similarity search
<a href="#">SIMCOMP</a>	Chemical structure similarity search
<a href="#">PathPred</a>	Biodegradation/biosynthesis pathway prediction

**FIGURE 14.22** The KEGG database includes pathway maps, data for a broad range of organisms, and a variety of analysis tools.

Source: KEGG, Courtesy of Kanehisa Laboratories.

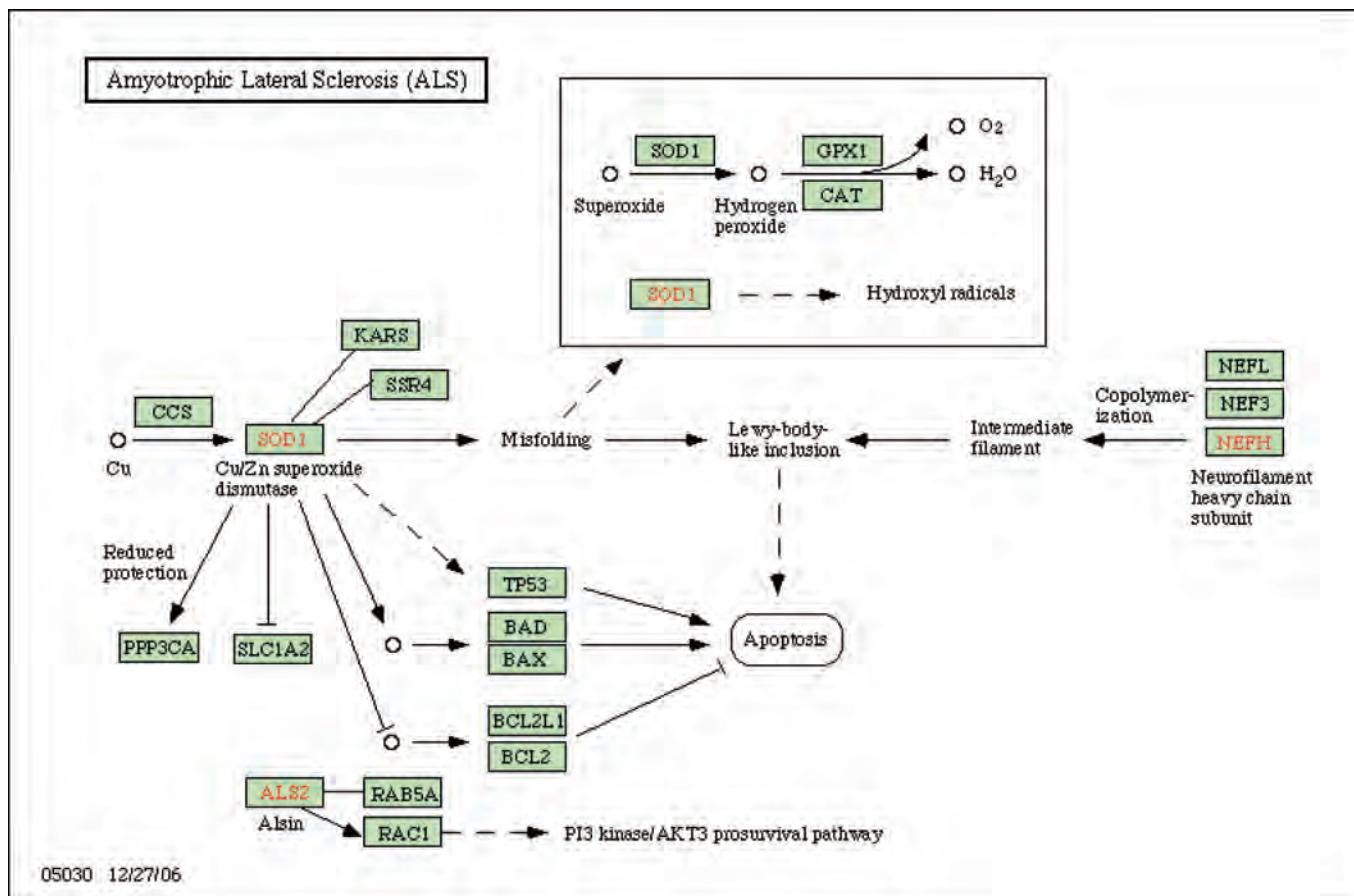
common cause of this debilitating disease. *SOD1* is an enzyme that normally converts the toxic oxygen metabolite superoxide ( $O_2^-$ ) into hydrogen peroxide and water. As shown in the KEGG pathway map, *SOD1* has been shown to interact directly and indirectly with a variety of other proteins, such as those involved in apoptosis (programmed cell death). Clicking on *SOD1*, an entry describing the protein and nucleotide sequence can be found as well as several external links such as the Enzyme Commission number and protein structure links, Pfam, Prosite, the Human Protein Reference Database, and Online Mendelian Inheritance in Man (OMIM; Chapter 21).

This example of *SOD1* highlights a strength of KEGG: its coverage of a broad range of proteins and cellular processes is comprehensive. The example also serves to show that some processes described in KEGG are likely to be organism specific. KEGG is based primarily on data generated from bacterial genomes, and pathways described in bacteria are not always applicable to eukaryotic organisms.



**FIGURE 14.23** The KEGG database includes pathway maps and data for a broad range of organisms. This pathway shows SNARE function (soluble N-ethylmaleimide-sensitive factor receptors, including syntaxin and other proteins described in Fig. 14.4). Syntaxin is represented in additional KEGG pathway maps.

Source: KEGG, Courtesy of Kanehisa Laboratories.



**FIGURE 14.24** KEGG includes pathways for diseases. A pathway for amyotrophic lateral sclerosis (ALS; Lou Gehrig's disease) is shown. Proteins in boxes link to detailed entries.

Source: KEGG, Courtesy of Kanehisa Laboratories.

## PERSPECTIVE

Many thousands of genomes have now been sequenced (including viral and organellar genomes). For the genomes of prominent organisms such as human, worms, flies, plants, and yeast, we are acquiring catalogs of the genes and gene products encoded by each genome. Defining the genes and the complete structure of the genome are challenging problems that we address in Part III of this book. We are already beginning to confront a problem that is perhaps even harder than identifying genes: identifying their function. Function has many definitions, as discussed for proteins in Chapter 12. In this chapter we have described many innovative, high-throughput functional genomics approaches to defining gene function. The field of functional genomics is broad, and can be considered using many different categories. (1) What type of organism do we wish to study? We highlighted eight model organisms, although many other models are commonly used. (2) What type of questions do we want to address: natural variation or experimental manipulations used to elucidate gene function? (3) What type of experimental approach do we wish to apply (e.g., forward versus reverse genetics)? (4) What type of molecules do we wish to study (i.e., from genomic DNA to RNA to protein or metabolites)? (5) What types of biological questions are we trying to address? For many investigators interested in human diseases or the function of human genes, there are yeast orthologs (see Chapters 18 and 21). If a yeast ortholog is identified then genetic screens can suggest many potential interacting partners that may elucidate the function of the human gene.

## PITFALLS

We have described a range of approaches to assessing gene function, including analyses at the levels of genes (e.g., creating null alleles or otherwise interfering with gene function), RNA, and proteins. The following caveats should be noted.

- Every method produces false negatives and false positives. It is important to estimate these rates, although it can be difficult to acquire trusted (“gold standard”) datasets with which to measure sensitivity and specificity.
- Many methods seem to work well with “knowns” but work much less well with unknown genes. Reasons may include functional redundancy, complex, multiple functions, or functions not evident under lab conditions.
- Combinatorial informatic approaches need weighting to help evaluate strength of “links” between genes. Also, any single set of gene “links” is incomplete.
- What is needed to have a better success rate at functional prediction is fewer links of low quality and more links of high quality.

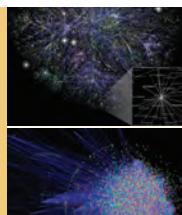
## ADVICE FOR STUDENTS

The field of functional genomics is growing at an extraordinary pace. Large numbers of genome-wide functional assays have been developed and applied to different organisms. If you have a favorite protein or gene it is a good idea to survey existing functional genomics datasets. If the gene is present in humans, have knockouts or other functional assays been performed in model organisms?

## WEB RESOURCES

NCBI offers a Probe Database which includes reagents for functional genomics; it serves as a repository for nucleic acid reagents.

Visit the Probe database at <http://www.ncbi.nlm.nih.gov/probe> (WebLink 14.66). See also its glossary at <http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/Glossary.shtml> (WebLink 14.67).



## Discussion Questions

**[14-1]** Define a functional genomics question. For example, how can we predict the functions of genes that currently lack functional annotation? How does the choice of experimental organism affect the approaches you might take to answer the question? How can a critical assessment competition help determine the accuracy of the predictions?

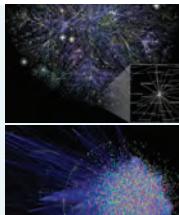
**[14-2]** Consider a human disease for which a gene has been implicated (such as  $\beta$ -globin in sickle cell anemia) and an animal model is available. How can forward genetics approaches be used to study this disease? How can reverse genetics approaches be used? What are some of the differences in the kinds of information these two approaches can provide?

## PROBLEMS/COMPUTER LAB

**[14-1]** Suppose you did not know anything about the function of hemoglobin but wanted to use bioinformatics resources to learn about its role in mouse and zebrafish. What information can you find?

**[14-2]** Select a yeast gene such as *SEC1*. Is it an essential gene? What proteins does it interact with based on physical (biochemical) or genetic assays? Are the interactions observed in yeast also found in mammalian systems?

**[14-3]** List all human genes for which there is a targeted knockout allele in mouse. To do this, use MGI BioMart (choose the allele type filter).



## Self-Test Quiz

**[14-1]** While there are many definitions of “functional genomics,” the best of these is:

- (a) the assignment of function to genes based primarily on genome-wide gene expression data using techniques such as microarrays or SAGE;
- (b) the assignment of function to genes based primarily on comprehensive surveys of protein–protein interactions and protein networks;
- (c) the combined use of genetic, biochemical, and cell biological approaches to study the function of a particular gene, its mRNA product, and its corresponding protein product; or
- (d) the assignment of function to genes and proteins using genome-wide screens and analyses.

**[14-2]** Reverse genetics approaches involve:

- (a) systematically inhibiting the functions of one or many genes (or gene products), and measuring the phenotypic consequences correctly;
- (b) measuring a phenotype of interest (such as cell growth), applying an intervention (such as radiation exposure) to generate a large collection of mutants, and identifying changes to the phenotype of interest;
- (c) treating an organism with a chemical mutagen or other agent to induce mutations, observing a phenotype of interest, and mapping the gene(s) responsible for the phenotype; or
- (d) all of the above.

**[14-3]** The “YKO” project is an effort to systematically knock out all yeast ORFs. A potential limitation of this approach is:

- (a) molecular barcodes may sometimes be toxic for yeast genes;
- (b) this approach is not suited to finding new genes, but instead focuses on already-known genes;
- (c) mutant knockout strains cannot be banked for later study by other investigators; or
- (d) mutations may not be null.

**[14-4]** A major advantage of genetic footprinting using transposons is:

- (a) the approach is technically easy and can be scaled up to study the function of many genes;
- (b) both insertion alleles and knockout alleles can be studied;

- (c) any known gene of interest can be studied with this approach; or
- (d) mutant strains can be banked for later study by other researchers.

**[14-5]** Forward genetics screens have become increasingly powerful. However, a major limitation is that:

- (a) mutations that are introduced through the use of mutagens or radiation do not leave molecular “tags” or barcodes in the genomic DNA, thus adding to the challenge of identifying DNA changes that are responsible for particular phenotypes correctly;
- (b) mutant alleles tend to be null rather than having a broad range of phenotypes;
- (c) these screens often involve morpholinos, but these compounds are effective in only a limited number of organisms; or
- (d) there is no universally preferred method to systematically inhibit the function of each gene in a genome.

**[14-6]** High-throughput screens such as the yeast two-hybrid system and affinity purification experiments can have false positive results because:

- (a) some proteins are inherently sticky;
- (b) some bait proteins that are introduced into cells become mislocalized;
- (c) some protein complexes form only very transiently;
- (d) affinity tags or epitope tags can interfere with protein–protein interactions; or
- (e) all of the above.

**[14-7]** Problems in determining protein networks include all of the following except for:

- (a) few benchmark datasets are available with which to assess false positive and false negative results;
- (b) false positive and negative error rates tend to be very high;
- (c) there is tremendous heterogeneity in the types of protein complexes that form; or
- (d) experimental data have been generated for bacteria and single-celled eukaryotes such as the yeast *S. cerevisiae*, but it has not yet been possible to obtain high-throughput data for organisms such as *Drosophila* and human.

**[14-8]** Hub proteins are proteins that occur at:

- (a) nodes that are highly connected within a protein network;
- (b) edges that are highly connected within a protein network;
- (c) nodes that are sparsely connected within a protein network; or
- (d) edges that are sparsely connected within a protein network.

**[14-9]** Which of the following best describes a major problem in evaluating large-scale cellular pathway diagrams?

- (a) the direction of the biochemical pathways is not usually known;
- (b) the pathway maps do not employ Gene Ontology nomenclature;
- (c) the pathway maps often depend on the correct identification of orthologs, but this can be problematic; or
- (d) the pathway maps tend to be derived from bacteria and archaea, but only limited information is available on eukaryotes.

## SUGGESTED READING

Excellent reviews of functional genomics approaches are available for the mouse (van der Weyden *et al.*, 2002; Guénet, 2005), plants (Borevitz and Ecker, 2004; Alonso and Ecker, 2006), and yeast. Abuin *et al.* (2007) thoroughly review gene trap mutagenesis.

## REFERENCES

- Abuin, A., Hansen, G. M., Zambrowicz, B. 2007. Gene trap mutagenesis. *Handbook of Experimental Pharmacology* **178**, 129–147.
- Adams, D. J., Biggs, P. J., Cox, T. *et al.* 2004. Mutagenic insertion and chromosome engineering resource (MICER). *Nature Genetics* **36**, 867–871. PMID: 15235602.
- Aggarwal, K., Choe, L. H., Lee, K. H. 2006. Shotgun proteomics using the iTRAQ isobaric tags. *Briefings in Functional Genomics and Proteomics* **5**, 112–120.
- Albert, R., Jeong, H., Barabasi, A. L. 2000. Error and attack tolerance of complex networks. *Nature* **406**, 378–382.
- Alonso, J. M., Ecker, J. R. 2006. Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. *Nature Reviews Genetics* **7**, 524–536.
- Amir, R.E., Van den Veyver, I.B., Wan, M. *et al.* 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics* **23**(2), 185–188. PMID: 10508514.
- Angerer, L. M., Angerer, R. C. 2004. Disruption of gene function using antisense morpholinos. *Methods in Cellular Biology* **74**, 699–711.
- Arbeitman, M. N., Furlong, E. E., Imam, F. *et al.* 2002. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275.
- Arnold, C.N., Barnes, M.J., Berger, M. *et al.* 2012. ENU-induced phenovariance in mice: inferences from 587 mutations. *BMC Research Notes* **5**, 577. PMID: 23095377.
- Atwell, S., Huang, Y.S., Vilhjálmsson, B.J. *et al.* 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**(7298), 627–631. PMID: 20336072.
- Austin, C. P., Battey, J. F., Bradley, A. *et al.* 2004. The knockout mouse project. *Nature Genetics* **36**, 921–924. PMID: 15340423.
- Ayadi, A., Birling, M.C., Bottomley, J. *et al.* 2012. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian Genome* **23**(9–10), 600–610. PMID: 22961258.
- Bader, G. D., Hogue, C. W. 2002. Analyzing yeast protein–protein interaction data obtained from different sources. *Nature Biotechnology* **20**, 991–997.

- Bader, G. D., Heilbut, A., Andrews, B. *et al.* 2003. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends in Cell Biology* **13**, 344–356.
- Bader, G. D., Cary, M. P., Sander, C. 2006. Pathguide: a pathway resource list. *Nucleic Acids Research* **34**, D504–506.
- Barabási, A. L., Albert, R. 1999. Emergence of scaling in random networks. *Science* **286**, 509–512.
- Barrangou, R. 2013. CRISPR-Cas systems and RNA-guided interference. *Wiley Interdisciplinary Reviews RNA* **4**(3), 267–278. PMID: 23520078.
- Berns, K., Hijmans, E. M., Mullenders, J. *et al.* 2004. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437.
- Blake, J.A., Bult, C.J., Eppig, J.T. *et al.* 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Research* **42**, D810–817. PMID: 24285300.
- Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A. *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**(5331), 1453–1462. PMID: 9278503.
- Borevitz, J. O., Ecker, J. R. 2004. Plant genomics: the third wave. *Annual Review of Genomics and Human Genetics* **5**, 443–477 (2004). PMID: 15485356.
- Bork, P., Jensen, L. J., von Mering, C. *et al.* 2004. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology* **14**, 292–299.
- Bouton, C.M., Pevsner, J. 2000. DRAGON: Database Referencing of Array Genes Online. *Bioinformatics* **16**(11), 1038–1039. PMID: 11159315.
- Boutros, M., Kiger, A. A., Armknecht, S. *et al.* 2004. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835.
- Brass, A. L., Dykxhoorn, D. M., Benita, Y. *et al.* 2008. Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**(5865), 921–926. PMID: 18187620.
- Braun, P. 2012. Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. *Proteomics* **12**(10), 1499–1518. PMID: 22589225.
- C. elegans* Deletion Mutant Consortium. 2012. Large-scale screening for targeted knockouts in the *Caenorhabditis elegans* genome. *G3 (Bethesda)* **2**(11), 1415–1425. PMID: 23173093.
- Capecci, M. R. 1989. Altering the genome by homologous recombination. *Science* **244**, 1288–1292.
- Caspari, T., Anderson, K.V. 2006. Uncovering the uncharacterized and unexpected: unbiased phenotype-driven screens in the mouse. *Developmental Dynamics* **235**, 2412–2423.
- Caspi, R., Foerster, H., Fulcher, C.A. *et al.* 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **36**(Database issue), D623–631. PMID: 17965431.
- Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S. *et al.* 2013. The BioGRID interaction database: 2013 update. *Nucleic Acids Research* **41**(Database issue), D816–823. PMID: 23203989.
- Chen, K.F., Crowther, D.C. 2012. Functional genomics in *Drosophila* models of human disease. *Briefings in Functional Genomics* **11**(5), 405–415. PMID: 22914042.
- Cherry, J.M., Hong, E.L., Amundsen, C. *et al.* 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research* **40**(Database issue), D700–705. PMID: 22110037.
- Churchill, G.A., Gatti, D.M., Munger, S.C., Svenson, K.L. 2012. The Diversity Outbred mouse population. *Mammalian Genome* **23**(9–10), 713–718. PMID: 22892839.
- Clark, A. T., Goldowitz, D., Takahashi, J. S. *et al.* 2004. Implementing large-scale ENU mutagenesis screens in North America. *Genetica* **122**, 51–64.
- Cline, M.S., Smoot, M., Cerami, E. *et al.* 2007. Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**(10), 2366–2382. PMID: 17947979.
- Collaborative Cross Consortium. 2012. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190**(2), 389–401. PMID: 22345608.
- Complex Trait Consortium. 2004. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics* **36**, 1133–1137.

- Cradick, T.J., Fine, E.J., Antico, C.J., Bao, G. 2013. CRISPR/Cas9 systems targeting  $\beta$ -globin and *CCR5* genes have substantial off-target activity. *Nucleic Acids Research* **41**(20), 9584–9592. PMID: 23939622.
- Cummings, E. E. 1954. Pity this busy monster manunkind. In *Poems*, 1923–1954. Harcourt, Brace, New York.
- Dessimoz, C., Škunca, N., Thomas, P.D. 2013. CAFA and the open world of protein function predictions. *Trends in Genetics* **29**(11), 609–610. PMID: 24138813.
- Dietzl, G., Chen, D., Schnorrer, F. et al. 2007. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448**, 151–156. PMID: 17625558.
- Drachman, D. B. 1994. Myasthenia gravis. *New England Journal of Medicine* **330**, 1797–1810.
- Drosophila 12 Genomes Consortium, Clark, A.G., Eisen, M.B. et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167), 203–218. PMID: 17994087.
- Engel, S.R., Cherry, J.M. 2013. The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database. *Database (Oxford)* **2013**, bat012. PMID: 23487186.
- Engel, S.R., Balakrishnan, R., Binkley, G. et al. 2010. *Saccharomyces* Genome Database provides mutant phenotype data. *Nucleic Acids Research* **38**(Database issue), D433–436. PMID: 19906697.
- Fields, S. 2009. Interactive learning: lessons from two hybrids over two decades. *Proteomics* **9**(23), 5209–5213. PMID: 19834904.
- Fields, S., Song, O. 1989. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246.
- Flicek, P., Amode, M.R., Barrell, D. et al. 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Fraser, A. G., Kamath, R. S., Zipperlen, P. et al. 2000. Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330.
- Fuchs, H., Gailus-Durner, V., Adler, T. et al. 2011. Mouse phenotyping. *Methods* **53**(2), 120–135. PMID: 20708688.
- Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M. et al. 2008. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research* **36**, D120–124. PMID: 18158297.
- Gavin, A. C., Bösche, M., Krause, R. et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147. PMID: 11805826.
- Gavin, A. C., Aloy, P., Grandi, P. et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636. PMID: 16429126.
- Geer, L.Y., Marchler-Bauer, A., Geer, R.C. et al. 2010. The NCBI BioSystems database. *Nucleic Acids Research* **38**(Database issue), D492–496. PMID: 19854944.
- Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L. et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**(6012), 1775–1787. PMID: 21177976.
- Geurts, A.M., Moreno, C. 2010. Zinc-finger nucleases: new strategies to target the rat genome. *Clinical Science (London)* **119**(8), 303–311. PMID: 20615201.
- Giaever, G., Chu, A.M., Ni, L. et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**(6896), 387–391. PMID: 12140549.
- Gillis, J., Pavlidis, P. 2013. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics* **14** Suppl 3, S15. PMID: 23630983.
- Giot, L., Bader, J.S., Brouwer, C. et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736. PMID: 14605208.
- Guan, C., Ye, C., Yang, X., Gao, J. 2010. A review of current large-scale mouse knockout efforts. *Genesis* **48**(2), 73–85. PMID: 20095055.
- Guénet, J. L. 2005. The mouse genome. *Genome Research* **15**, 1729–1740.

- Gunn, T.M. 2012. Functional annotation and ENU. *BMC Research Notes* **5**, 580. PMID: 23095518.
- Gunsalus, K. C., Yueh, W. C., MacMenamin, P., Piano, F. 2004. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Research* **32**, D406–410.
- Hamp, T., Kassner, R., Seemayer, S. *et al.* 2013. Homology-based inference sets the bar high for protein function prediction. *BMC Bioinformatics* **14** Suppl 3, S7. PMID: 23514582.
- Hansen, J., Floss, T., Van Sloun, P. *et al.* 2003. A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proceedings of the National Academy of Science, USA* **100**, 9918–9922.
- Harris, T.W., Baran, J., Bieri, T. *et al.* 2014. WormBase 2014: new views of curated biology. *Nucleic Acids Research* **42**, D789–793. PMID: 24194605.
- Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N. *et al.* 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Computational Biology* **2**, e100.
- Hedges, S.B., Dudley, J., Kumar, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23), 2971–2972. PMID: 17021158.
- Henken, D. B., Rasooly, R. S., Javois, L., Hewitt, A. T. 2004. National Institutes of Health Trans-NIH Zebrafish Coordinating Committee. The National Institutes of Health and the Growth of the Zebrafish as an Experimental Model Organism. *Zebrafish* **1**, 105–110.
- Hentges, K. E., Justice, M. J. 2004. Checks and balancers: balancer chromosomes to facilitate genome annotation. *Trends in Genetics* **20**, 252–259.
- Hernández, H., Dziembowski, A., Taverner, T., Séraphin, B., Robinson, C. V. 2006. Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Reports* **7**, 605–610.
- Hirschman, J.E., Balakrishnan, R., Christie, K.R. *et al.* 2006. Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Research* **34**(Database issue), D442–445. PMID: 16381907.
- Ho, Y., Gruhler, A., Heilbut, A. *et al.* 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**(6868), 180–183. PMID: 11805837.
- Horn, T., Arziman, Z., Berger, J., Boutros, M. 2007. GenomeRNAi: a database for cell-based RNAi phenotypes. *Nucleic Acids Research* **35**, D492–497.
- Horner, V.L., Caspary, T. 2011. Creating a “hopeful monster”: mouse forward genetic screens. *Methods in Molecular Biology* **770**, 313–336. PMID: 21805270.
- Howe, D.G., Frazer, K., Fashena, D. *et al.* 2011. Data extraction, transformation, and dissemination through ZFIN. *Methods in Cell Biology* **104**, 311–325. PMID: 21924170.
- Howe, D.G., Bradford, Y.M., Conlin, T. *et al.* 2013a. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Research* **41**(Database issue), D854–860. PMID: 23074187.
- Howe, K., Clark, M.D., Torroja, C.F. *et al.* 2013b. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**(7446), 498–503. PMID: 23594743.
- Hu, J.C., Sherlock, G., Siegele, D.A. *et al.* 2014. PortEco: a resource for exploring bacterial biology through high-throughput data and analysis tools. *Nucleic Acids Research* **42**(Database issue), D677–684. PMID: 24285306.
- International Mouse Knockout Consortium, Collins, F. S., Rossant, J., Wurst, W. 2007. A mouse for all reasons. *Cell* **128**, 9–13.
- Ito, T., Chiba, T., Ozawa, R. *et al.* 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Science, USA* **98**(8), 4569–4574. PMID: 11283351.
- Joung, J.K., Sander, J.D. 2013. TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology* **14**(1), 49–55. PMID: 23169466.
- Kamath, R.S., Fraser, A.G., Dong, Y. *et al.* 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237. PMID: 12529635.
- Kanehisa, M., Araki, M., Goto, S. *et al.* 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* **36**, D480–484.

- Karp, P.D. 2001. Pathway databases: a case study in computational symbolic theories. *Science* **293**(5537), 2040–2044. PMID: 11557880.
- Kellis, M., Wold, B., Snyder, M.P. et al. 2014. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Science, USA* **111**(17), 6131–6138. PMID: 24753594.
- Kerrien, S., Orchard, S., Montecchi-Palazzi, L. et al. 2007. Broadening the horizon: level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology* **5**, 44. PMID: 17925023.
- Kerrien, S., Aranda, B., Breuza, L. et al. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Research* **40**, D841–846. PMID: 22121220.
- Keseler, I.M., Mackie, A., Peralta-Gil, M. et al. 2013. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research* **41**(Database issue), D605–612. PMID: 23143106.
- Kettleborough, R.N., Busch-Nentwich, E.M., Harvey, S.A. et al. 2013. A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496**(7446), 494–497. PMID: 23594742.
- Kile, B.T., Hentges, K.E., Clark, A.T. et al. 2003. Functional genetic analysis of mouse chromosome 11. *Nature* **425**, 81–86. PMID: 12955145.
- Kilpinen, H., Barrett, J.C. 2013. How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics* **29**(1), 23–30. PMID: 23103023.
- Kim, J.K., Gabel, H.W., Kamath, R.S. et al. 2005. Functional genomic analysis of RNA interference in *C. elegans*. *Science* **308**, 1164–1167. PMID: 15790806.
- Kim, S. K., Lund, J., Kiraly, M. et al. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092. PMID: 11557892.
- Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**, 624–626.
- Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
- Knowlton, M.N., Li, T., Ren, Y. C. et al. 2008. A PATO-compliant zebrafish screening database (MODB): management of morpholino knockdown screen information. *BMC Bioinformatics* **9**, 7.
- Koornneef, M., Meinke, D. 2010. The development of Arabidopsis as a model plant. *Plant Journal* **61**(6), 909–921. PMID: 20409266.
- Koscielny, G., Yaikhom, G., Iyer, V. et al. 2014. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research* **42**(Database issue), D802–809. PMID: 24194600.
- Koutsos, A. C., Blass, C., Meister, S. et al. 2007. Life cycle transcriptome of the malaria mosquito *Anopheles gambiae* and comparison with the fruitfly *Drosophila melanogaster*. *Proceedings of the National Academy of Science, USA* **104**, 11304–11309.
- Krogan, N. J. et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643.
- Lamesch, P., Berardini, T.Z., Li, D. et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**(Database issue), D1202–D1210. PMID: 22140109.
- Latendresse, M., Paley, S., Karp, P.D. 2012. Browsing metabolic and regulatory networks with BioCyc. *Methods in Molecular Biology* **804**, 197–216. PMID: 22144155.
- Le Cong, F., Ran, F.A., Cox, D. et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**(6121), 819–823. PMID: 23287718.
- Lee, T., Shah, C., Xu, E.Y. 2007. Gene trap mutagenesis: a functional genomics approach towards reproductive research. *Molecular Human Reproduction* **13**(11), 771–779. PMID: 17890780.
- Lehner, B., Fraser, A. G., Sanderson, C. M. 2004. Technique review: how to use RNA interference. *Briefings in Functional Genomics and Proteomics* **3**, 68–83.
- Leonelli, S., Ankeny, R.A. 2012. Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Biological and Biomedical Science* **43**(1), 29–36. PMID: 22326070.
- Li, D., Li, J., Ouyang, S. et al. 2006. Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: large-scale organization and robustness. *Proteomics* **6**(2), 456–461. PMID: 16317777.

- Li, X.J., Li, S.H., Sharp, A.H. *et al.* 1995. A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378**(6555), 398–402. PMID: 7477378.
- Logan, R.W., Robledo, R.F., Recla, J.M. *et al.* 2013. High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. *Genes, Brain and Behavior* **12**(4), 424–437. PMID: 23433259.
- Lourenço, A., Carneiro, S., Rocha, M., Ferreira, E.C., Rocha, I. 2011. Challenges in integrating Escherichia coli molecular biology data. *Briefings in Bioinformatics* **12**(2), 91–103. PMID: 21059604.
- Low, T.Y., van Heesch, S., van den Toorn, H. *et al.* 2013. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Reports* **5**(5), 1469–1478. PMID: 24290761.
- Lucretius. 1772. *De Rerum Natura Libri Sex*. John Baskerville, Birmingham.
- Ma, Y., Creanga, A., Lum, L., Beachy, P. A. 2006. Prevalence of off-target effects in *Drosophila* RNA interference screens. *Nature* **443**, 359–363.
- Mahajan, M. C., Weissman, S. M. 2006. Multi-protein complexes at the beta-globin locus. *Briefings in Functional Genomics and Proteomics* **5**, 62–65.
- Mali, P., Yang, L., Esvelt, K.M. *et al.* 2013. RNA-guided human genome engineering via Cas9. *Science* **339**(6121), 823–826. PMID: 23287722.
- Martin, F. 2012. Fifteen years of the yeast three-hybrid system: RNA-protein interactions under investigation. *Methods* **58**(4), 367–375. PMID: 22841566.
- Martin, S. E., Caplen, N. J. 2007. Applications of RNA interference in mammalian systems. *Annual Review of Genomics and Human Genetics* **8**, 81–108.
- Mathivanan, S., Periaswamy, B., Gandhi, T. K. *et al.* 2006. An evaluation of human protein–protein interaction data in the public domain. *BMC Bioinformatics* **7** Suppl 5, S19.
- McQuilton, P., St. Pierre, S.E., Thurmond, J. 2012. FlyBase Consortium. FlyBase 101: the basics of navigating FlyBase. *Nucleic Acids Research* **40**(Database issue), D706–714. PMID: 22127867.
- Mika, S., Rost, B. 2006. Protein–protein interactions more conserved within species than across species. *PLoS Computational Biology* **2**, e79.
- modENCODE Consortium, Roy, S., Ernst, J. *et al.* 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**(6012), 1787–1797. PMID: 21177974.
- Molloy, M. P., Witzmann, F. A. 2002. Proteomics: Technologies and applications. *Briefings in Functional Genomics and Proteomics* **1**, 23–39.
- Muller, H. J. 1918. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3**, 422–499.
- Nord, A. S., Chang, P. J., Conklin, B. R. *et al.* 2006. The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse. *Nucleic Acids Research* **34**, D642–648. PMID: 16381950.
- Novick, P., Schekman, R. 1979. Secretion and cell-surface growth are blocked in a temperature-sensitive mutant of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Science, USA* **76**, 1858–1862.
- Novick, P., Field, C., Schekman, R. 1980. Identification of 23 complementation groups required for post-translational events in the yeast secretory pathway. *Cell* **21**, 205–215.
- O'Brien, T. P., Frankel, W. N. 2004. Moving forward with chemical mutagenesis in the mouse. *Journal of Physiology* **554**, 13–21.
- Ooi, H.S., Schneider, G., Chan *et al.* 2010. Databases of protein–protein interactions and complexes. *Methods in Molecular Biology* **609**, 145–159. PMID: 20221918.
- Ooi, S. L., Pan, X., Peyser, B. D. *et al.* 2006. Global synthetic-lethality analysis and yeast functional profiling. *Trends in Genetics* **22**, 56–63.
- Orchard, S. 2012. Molecular interaction databases. *Proteomics* **12**(10), 1656–1662. PMID: 22611057.
- Palcy, S., Chevet, E. 2006. Integrating forward and reverse proteomics to unravel protein function. *Proteomics* **6**, 5467–5480.

- Pan, X., Ye, P., Yuan, D. S. *et al.* 2006. A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124**, 1069–1081. PMID: 16487579.
- Pan, X., Yuan, D. S., Ooi, S. L. *et al.* 2007. dSLAM analysis of genome-wide genetic interactions in *Saccharomyces cerevisiae*. *Methods* **41**, 206–221.
- Perez-Pinera, P., Kocak, D.D., Vockley, C.M. *et al.* 2013. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods* **10**(10), 973–976. PMID: 23892895.
- Pickart, M. A., Sivasubbu, S., Nielsen, A. L. *et al.* 2004. Functional genomics tools for the analysis of zebrafish pigment. *Pigment Cell Research* **17**, 461–470.
- Probst, F.J., Justice, M.J. 2010. Mouse mutagenesis with the chemical supermutagen ENU. *Methods in Enzymology* **477**, 297–312. PMID: 20699147.
- Radivojac, P., Clark, W.T., Oron, T.R. *et al.* 2013. A large-scale evaluation of computational protein function prediction. *Nature Methods* **10**(3), 221–227. PMID: 23353650.
- Reed, J. L., Famili, I., Thiele, I., Palsson, B. O. 2006. Towards multidimensional genome annotation. *Nature Reviews Genetics* **7**, 130–141.
- Reguly, T., Breitkreutz, A., Boucher, L. *et al.* 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *Journal of Biology* **5**, 11. PMID: 16762047.
- Rezwan, M., Auerbach, D. 2012. Yeast “N”-hybrid systems for protein–protein and drug–protein interaction discovery. *Methods* **57**(4), 423–429. PMID: 22728036.
- Riley, H.P. 1948. *Introduction to Genetics and Cytogenetics*. John Wiley & Sons, New York.
- Robinson, C. V., Sali, A., Baumeister, W. 2007. The molecular sociology of the cell. *Nature* **450**, 973–982.
- Ross-Macdonald, P. 2005. Forward in reverse: how reverse genetics complements chemical genetics. *Pharmacogenomics* **6**, 429–434.
- Ross-Macdonald, P. *et al.* 1999. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418.
- Russell, S. 2012. From sequence to function: the impact of the genome sequence on Drosophila biology. *Briefings in Functional Genomics* **11**(5), 333–335. PMID: 23023662.
- Russell, W. L., Kelly, E. M., Hunsicker, P. R. *et al.* 1979. Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. *Proceedings of the National Academy of Science USA* **76**, 5818–5819.
- Ryan, C.J., Cimerman i , P., Szpiech, Z.A. *et al.* 2013. High-resolution network biology: connecting sequence with function. *Nature Reviews Genetics* **14**(12), 865–879. PMID: 24197012.
- Sachidanandam, R. 2004. RNAi: design and analysis. *Current Protocols in Bioinformatics* **12**, 12.3.1–12.3.10.
- Saito, R., Smoot, M.E., Ono, K. *et al.* 2012. A travel guide to Cytoscape plugins. *Nature Methods* **9**(11), 1069–1076. PMID: 23132118.
- Sander, J.D., Maeder, M.L., Reyen, D. *et al.* 2010. ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Research* **38**(Web Server issue), W462–468. PMID: 20435679.
- Schächter, V. 2002. Bioinformatics of large-scale protein interaction networks. *Computational Proteomics supplement* **32**, 16–27 (2002).
- Scherens, B., Goffeau, A. 2004. The uses of genome-wide yeast mutant collections. *Genome Biology* **5**, 229.
- Schmidt, E.E., Pelz, O., Buhlmann, S. *et al.* 2013. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Research* **41**(Database issue), D1021–1026. PMID: 23193271.
- Schulze, T.G., McMahon, F.J. 2004. Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping. *Human Heredity* **58**, 131–138.
- Sharom, J. R., Bellows, D. S., Tyers, M. 2004. From large networks to small molecules. *Current Opinion in Chemical Biology* **8**, 81–90.

- Shehee, W. R., Oliver, P., Smithies, O. 1993. Lethal thalassemia after insertional disruption of the mouse major adult  $\beta$ -globin gene. *Proceedings of the National Academy of Science USA* **90**, 3177–3181.
- Sims, D., Bursteinas, B., Gao, Q., Zvelebil, M., Baum, B. 2006. FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. *Nucleic Acids Research* **34**, D479–483.
- Skarnes, W.C., von Melchner, H., Wurst, W. et al. 2004. A public gene trap resource for mouse functional genomics. *Nature Genetics* **36**, 543–544. PMID: 15167922.
- Smith, V., Botstein, D., Brown, P. O. 1995. Genetic footprinting: A genomic strategy for determining a gene's function given its sequence. *Proceedings of the National Academy of Science USA* **92**, 6479–6483.
- Smith, V., Chou, K. N., Lashkari, D., Botstein, D., Brown, P. O. 1996. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074.
- Sönnichsen, B., Koski, L.B., Walsh, A. et al. 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434**, 462–469. PMID: 15791247.
- St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P., The FlyBase Consortium. 2014. FlyBase 102: advanced approaches to interrogating FlyBase. *Nucleic Acids Research* **42**(Database issue), D780–788. PMID: 24234449.
- Stanford, W.L., Epp, T., Reid, T., Rossant, J. 2006. Gene trapping in embryonic stem cells. *Methods in Enzymology* **420**, 136–162.
- Stottmann, R.W., Beier, D.R. 2010. Using ENU mutagenesis for phenotype-driven analysis of the mouse. *Methods in Enzymology* **477**, 329–348. PMID: 20699149.
- Stynen, B., Tournu, H., Tavernier, J., Van Dijck, P. 2012. Diversity in genetic in vivo methods for protein–protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiology and Molecular Biology Review* **76**(2), 331–382. PMID: 22688816.
- Su, C., Peregrin-Alvarez, J. M., Butland, G. et al. 2008. Bacteriome.org: an integrated protein interaction database for *E. coli*. *Nucleic Acids Research* **36**, D632–636.
- Sun, M., Mondal, K., Patel, V. et al. 2012. Multiplex chromosomal exome sequencing accelerates identification of ENU-induced mutations in the mouse. *G3 (Bethesda)* **2**(1), 143–150. PMID: 22384391.
- Sutandy, F.X., Qian, J., Chen, C.S., Zhu, H. 2013. Overview of protein microarrays. *Current Protocol in Protein Science Chapter* **27**, Unit 27.1. PMID: 23546620.
- Suthram, S., Shlomi, T., Rupp, E., Sharan, R., Ideker, T. 2006. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics* **7**, 360.
- Thompson, O., Edgley, M., Strasbourger, P. et al. 2013. The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Research* **23**(10), 1749–1762. PMID: 23800452.
- Tong, A. H., Boone, C. 2006. Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods in Molecular Biology* **313**, 171–192.
- Tong, A. H., Evangelista, M., Parsons, A. B. et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368. PMID: 11743205.
- Tong, A. H., Lesage, G., Bader, G. D. et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813. PMID: 14764870.
- Uetz, P., Giot, L., Cagney, G. et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**(6770), 623–627. PMID: 10688190.
- Ullrich, M., Schuh, K. 2009. Gene trap: knockout on the fast lane. *Methods in Molecular Biology* **561**, 145–159. PMID: 19504070.
- van der Weyden, L., Adams, D. J., Bradley, A. 2002. Tools for targeted manipulation of the mouse genome. *Physiological Genomics* **11**, 133–164.
- van Zon, A., Mossink, M. H., Scheper, R. J., Sonneveld, P., Wiemer, E. A. 2003. The vault complex. *Cellular and Molecular Life Sciences* **60**, 1828–1837.
- Varshney, G.K., Huang, H., Zhang, S. et al. 2013. The Zebrafish Insertion Collection (ZInC): a web based, searchable collection of zebrafish mutations generated by DNA insertion. *Nucleic Acids Research* **41**(Database issue), D861–864. PMID: 23180778.

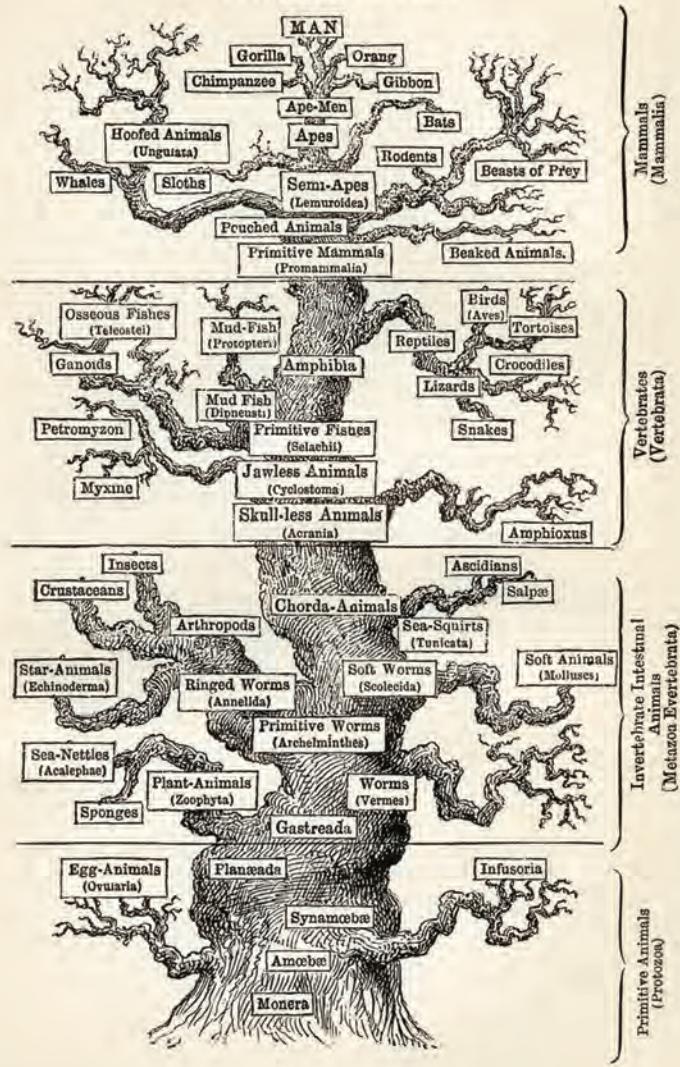
- Velasco-García, R., Vargas-Martínez, R. 2012. The study of protein–protein interactions in bacteria. *Canadian Journal of Microbiology* **58**(11), 1241–1257. PMID: 23145822.
- Verhage, M., Maia, A.S., Plomp, J.J. *et al.* 2000. Synaptic assembly of the brain in the absence of neurotransmitter secretion. *Science* **287**, 864–869. PMID: 10657302.
- Walhout, A.J. 2011. Gene-centered regulatory network mapping. *Methods in Cell Biology* **106**, 271–288. PMID: 22118281.
- Welsh, C.E., Miller, D.R., Manly, K.F. *et al.* 2012. Status and access to the Collaborative Cross population. *Mammalian Genome* **23**(9–10), 706–712. PMID: 22847377.
- White, J.K., Gerdin, A.K., Karp, N.A. *et al.* 2013. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**(2), 452–464. PMID: 23870131.
- Williamson, M.P., Sutcliffe, M.J. 2010. Protein–protein interactions. *Biochemistry Society Transactions* **38**(4), 875–878. PMID: 20658969.
- Wilming, L. G., Gilbert, J. G., Howe, K. *et al.* 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Research* **36**, D753–760.

# Genome Analysis

PART

III

## PEDIGREE OF MAN.



The tree of life from Ernst Haeckel (1879). The figure shows mammals (with humans at the top shown ascending from apes), vertebrates, invertebrates, and primitive animals at the bottom, including Monera (bacteria).

Source: [http://en.wikipedia.org/wiki/File:Tree\\_of\\_life\\_by\\_Haeckel.jpg](http://en.wikipedia.org/wiki/File:Tree_of_life_by_Haeckel.jpg).

In the final part of this book we explore life on Earth from a genomics perspective, using the tools of bioinformatics we learned in Parts I and II. We provide an overview (Chapter 15) then discuss viruses (Chapter 16), bacteria and archaea (Chapter 17), fungi as an introduction to eukaryotes (Chapter 18), and eukaryotes from parasites to primates (Chapter 19). We conclude with the human genome (Chapter 20) and human disease (Chapter 21). Just as contemporary genomics cannot be understood without bioinformatics, bioinformatics cannot fulfill its potential until it informs us about genomes and hence biology.

# Genomes Across the Tree of Life

# CHAPTER 15

*The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species. ... The limbs divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups. ... From the first growth of the tree, many a limb and branch has decayed and dropped off, and these lost branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only from having been found in a fossil state. ... As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all a feebler branch, so by generation I believe it has been with the Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.*

—Charles Darwin, *The Origin of Species* (1859)

Consider the scenario where a new *E. coli* sequence has been obtained from a futuristic handheld device (like a Star Trek tricorder) that generates the complete genome in seconds. While the genome sequence may only be slightly different from strains already in the public databases, the metadata associated with this bug is both unique and crucial. Where and when was the *E. coli* isolated? Was it transmitted as a food-borne pathogen? Did it hospitalize the patient from whom it was isolated? Was it part of a larger infectious outbreak? Knowledge that a pathogen was isolated from diseased patients or healthy controls will readily assist in intervention strategies derived from machine-readable data.

—Dawn Field et al. (2011), writing in support of the Genomic Standards Consortium.

## LEARNING OBJECTIVES

After reading this chapter you should be able to:

- compare and contrast approaches to generating a tree of life;
- briefly describe a chronology of genome sequencing projects;
- describe the process of genome sequencing; and
- describe genome annotation.

## INTRODUCTION

A genome is the collection of DNA that comprises an organism. Each individual organism's genome contains the genes and other DNA elements that ultimately define its identity. Genomes range in size from the smallest viruses, which encode fewer than 10 genes, to eukaryotes such as humans that have billions of base pairs of DNA encoding tens of thousands of genes.

The recent sequencing of genomes from all branches of life – including viruses, bacteria, archaea, fungi, nematodes, plants, and humans – presents us with an extraordinary moment in the history of biology. By analogy, this situation resembles the completion of the periodic table of the elements in the nineteenth century. As it became clear that the periodic table could be arranged in rows and columns, it became possible to predict the properties of individual elements. A logic emerged to explain the properties of the elements, but it still took another century to grasp the significance of the elements and to realize the potential of the organization inherent in the periodic table.

Today we have sequenced the DNA from thousands of genomes, and we are now searching for a logic to explain their organization and function. This process will take decades. A variety of tools must be applied, including bioinformatics approaches, biochemistry, genetics, and cell biology.

This chapter introduces the tree of life and the sequencing of genomes. There are seven sections: (1) in the remainder of this first section, we introduce perspectives on genomics, the tree of life, and taxonomy; (2) we then introduce major web resources; (3) we survey the chronology of genome sequencing projects; (4) we introduce genome analysis projects and (5) we explore sequence data and how they are stored; finally, we discuss (6) the assembly of genomes and (7) their annotation.

After this chapter we assess the progress in studying the genomes of viruses (Chapter 16); bacteria and archaea (Chapter 17); fungi, including the yeast *Saccharomyces cerevisiae* (Chapter 18); eukaryotes from parasites to primates (Chapter 19); and finally the human genome (Chapters 20 and 21).

For definitions of several key terms related to the tree of life, see **Table 15.1**.

**TABLE 15.1 Nomenclature for tree of life. Name refers to the name adopted in this book. Adapted from Woese et al. (1990).**

Name	Synonym(s)	Definition
Archaea (singular: archaeon)	Archaeabacteria	One of the three "urkingdoms" or "domains" of life
Bacteria	Eubacteria; Monera (obsolete name)	One of the three "urkingdoms" or "domains" of life; unicellular organisms characterized by lack of a nuclear membrane
Eukaryotes	Eucarya	One of the three "urkingdoms" or "domains" of life; cells characterized by a nuclear membrane
Microbe	—	Microorganisms that cause disease in humans; microbes include bacteria and eukaryotes such as protozoa and fungi
Microorganism	—	Unicellular life forms of microscopic size, including bacteria, archaea, and some eukaryotes
Progenote	Last universal common ancestor	The ancient, unicellular life form from which the three domains of life are descended
Prokaryotes	Prokaryotes; formerly synonymous with bacteria	Organism lacking a nuclear membrane; bacteria and archaea

## Five Perspectives on Genomics

For a course on genomics that I have taught, we discuss genomes across the tree of life from the following five perspectives. Each student selects any genome of interest and writes a report describing the genome according to these approaches. Students may identify an outstanding research problem and describe how genomics approaches are being applied to solve it. A related project is to select a single gene of interest and analyze it in depth, again following these five areas.

*Perspective 1: Catalog genomic information.* What are the basic features of each genome? These include its size; the number of chromosomes; the guanine plus cytosine (GC) content; the presence of isochores (described in Chapter 20); the number of genes, both coding and noncoding; repetitive DNA; and unique features of each genome. The techniques used to answer these questions include genomic DNA sequencing (Chapter 9); assembly; and genome annotation including gene prediction. Genome browsers represent a major resource to access catalogs of genomic information, organized into categories such as raw underlying DNA data as well as models of genes, regulatory elements, and other features of the genomic landscape.

*Perspective 2: Catalog comparative genomic information.* Our understanding of any genome is dramatically enhanced through comparisons to related genomes (Miller *et al.*, 2004). When did a given species diverge from its relatives? Which genes or other DNA elements are orthologous, or share conserved synteny (Chapter 8)? To what extent did lateral gene transfer (Chapter 17) occur in each genome? Techniques of comparative genomics used to address these issues include whole-genome alignment and analyses with databases such as Ensembl Genomes (Kersey *et al.*, 2013) and the UCSC Genome Browser (Karolchik *et al.*, 2014). This approach also includes phylogenetic reconstruction (Chapter 7).

*Perspective 3: Biological principles.* For each genome, what are the functions of the organism (e.g., with respect to development, metabolism, and behavior) and how are they served by the genome? What are the mechanisms of evolution of the genome? This includes consideration of how genome size is regulated, whether there is polyploidization (Chapters 18, 19), how the birth and death of genes occurs, and what forces operate on DNA whether they involve positive or negative selection or neutral evolution. What forces shape speciation? What is the role of epigenetics? Some of the many techniques used to address these issues include molecular phylogeny (Chapter 7) and BLAST or related tools (Chapters 4 and 5).

*Perspective 4: Human disease relevance.* What are the mechanisms by which organisms such as viruses or protozoan pathogens cause disease in humans or plants? What are the types of genomic responses and defenses that organisms have to prevent or adapt to avoid becoming subject to disease? A variety of techniques are applied to these questions, including the study of single-nucleotide polymorphisms (SNPs, Chapters 8 and 20) and linkage and association studies (Chapter 21).

*Perspective 5: Bioinformatics aspects.* What are some of the key databases and websites associated with each genome, and what command line or web-based software programs have been developed to facilitate the analysis and visualization of data? The functionality of genome browsers has been greatly enhanced in recent years, providing a system with which to store, analyze, and interpret hundreds of categories of genomic data.

Web Document 15.1 at <http://www.bioinfbook.org/chapter15> presents a table of these perspectives on genomics, and Web Document 15.2 outlines the details of a project to analyze a gene in depth from a genomics perspective.

## Brief History of Systematics

Throughout recorded history, philosophers and scientists have grappled with questions regarding the diversity of life on Earth (Mayr, 1982). Aristotle (384–322 BCE) was an active biologist, describing over 500 species in his zoological works. He did not create a

general classification scheme for life, but he did describe animals as “blooded” or “bloodless” in his *Historia animalium*. (Eventually, Lamarck (1744–1829) renamed these categories “vertebrates” and “invertebrates.”) Aristotle’s division of animals into genera and species provides the origin of the taxonomic system we use today.

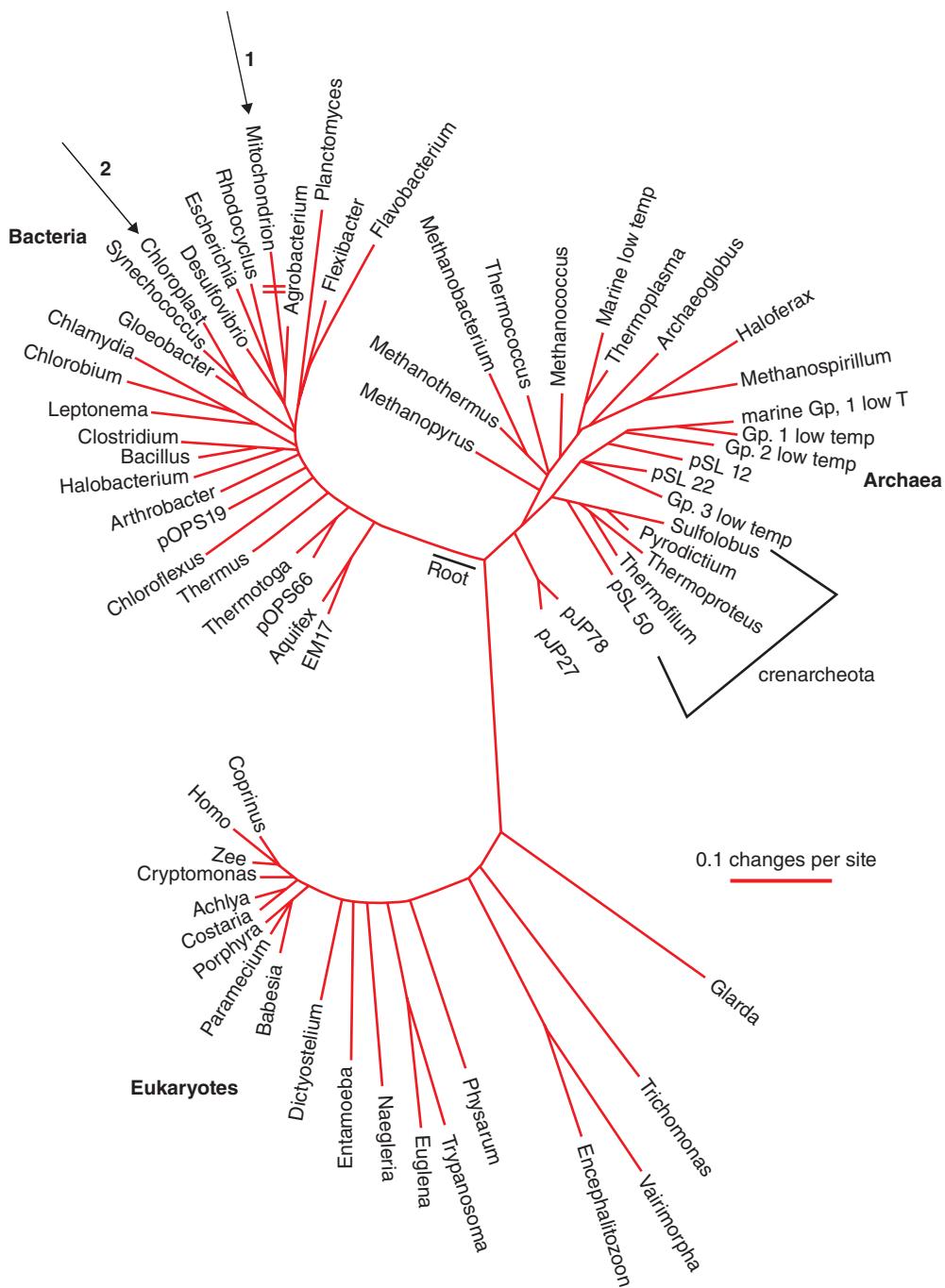
The greatest advocate of this binomial nomenclature system of genus and species for each organism was the Swedish naturalist Carl Linnaeus (1707–1778). Linnaeus also introduced the notion of the three kingdoms of Animaliae, Plantae, and Mineraliae; in his hierarchical system four levels were class, order, genus, and species. Ernst Haeckel (1834–1919), who described over 4000 new species, enlarged this system. He described life as a continuum from mere complex molecules to plants and animals, and he described the Moner as formless clumps of life. The monera were later named bacteria, and in 1937 Edouard Chatton made the distinction between prokaryotes (bacteria that lack nuclei) and eukaryotes (organisms with cells that have nuclei). By the end of the 1960s the work of Haeckel (1879), Copeland, Whittaker (1969), and many others led to the standard five-kingdom system of life: animals, plants, single-celled protists, fungi, and monera. Whittaker’s 1969 scheme shows monera at the base of the tree representing the prokaryotes, and then eukaryotes (either unicellular or multicellular) represented by the Protista, Plantae, Fungi, and Animalia. An example of the tree of life from an 1879 book by Haeckel is shown in the frontis to this chapter.

The tree of life was rewritten in the 1970s and 1980s by Carl Woese and colleagues (Fox *et al.*, 1980; Woese *et al.*, 1990; Woese, 1998). They studied a group of prokaryotes that were presumed to be bacteria because they were single-celled life forms that lack a nucleus. The researchers sequenced small-subunit ribosomal RNAs (SSU rRNA) and performed phylogenetic analyses. This revealed that archaea are as closely related to eukaryotes as they are to bacteria. A phylogenetic analysis of SSU rRNA sequences, which are present in all known life forms, provides one version of the tree of life (Fig. 15.1). There are three main branches. While the exact root of the tree is not known, the deepest branching bacteria and archaea are thermophiles, suggesting that life may have originated in a hot environment.

The term “prokaryotes” (or “procaryotes”) is used by many people to mean single-celled organisms that are not eukaryotes. Norman Pace (2009) has argued that the term prokaryote should be eliminated altogether. He notes the following. (1) “Prokaryote” is defined in terms of what it is not (i.e., not eukaryotic). (2) Earlier, now obsolete models of the history of life suggested that prokaryotes preceded more complex eukaryotes (Fig. 15.2a). The current three-domain tree of life (Fig. 15.2b) contains no phylogenetically coherent group of prokaryotes. None of the three primary domains is derived from another; each is equally old. (3) The root of the universal tree, where the origin of life is placed, separates bacteria and archaea. (We will see in Chapter 17 that archaea share some distinct properties with eukaryotes such as: reliance on histones to package DNA; they share some properties with bacteria; and in some ways they are unique, such as their reliance on ether-linked lipids rather than ester-linked lipids to make membranes.) (4) The term “prokaryote” therefore connotes an incorrect model of evolution. Pace’s point of view has not (yet) been adopted. This is almost entirely because others in the research community like the conciseness of the term (arguing that it is tedious to repeatedly refer to “archaea and bacteria”), and because “prokaryote” is generally understood to mean bacteria and archaea. In this book I have largely removed the term prokaryote while acknowledging its continued widespread usage.

A recent, alternative model is that there are just two domains of life: archaea and bacteria (Williams *et al.*, 2013). In Woese’s model the archaea are monophyletic, with eukaryotes as an outgroup (Fig. 15.2b). According to the Williams *et al.* model the archaea are paraphyletic (Fig. 15.2c). Eukaryotic genes, including ribosomal RNA genes and genes encoding proteins that function in protein translation, are placed within the archaea by phylogenetic analysis, originating from the Eocytes (Crenarchaeota), Thaumarchaeota,

A species is a group of similar organisms that only breed with one another under normal conditions. A genus may consist of between one and hundreds of species.

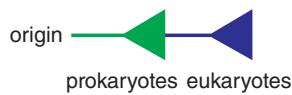


**FIGURE 15.1** A global tree of life, based upon phylogenetic analysis of small-subunit rRNA sequences. Life is thought to have originated about 3.8 BYA in an anaerobic environment. The primordial life form (progenote) displayed the defining features of life (self-replication and evolution). The eukaryotic mitochondrion (arrow 1) and chloroplast (arrow 2) are indicated, showing their bacterial origins. Data from Barns *et al.* (1996), Hugenholtz and Pace (1996), and Pace (1997).

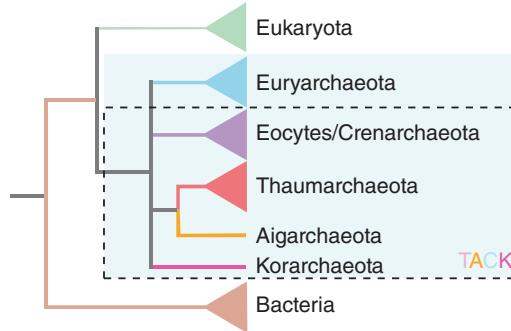
Aigarchaeota, and/or Korarchaeota. Eukaryotes emerged from an archaeal lineage. These eukaryotes then served as hosts for the bacterial endosymbiont that shed many of its genes to become the modern mitochondrion.

Many groups have reconstructed the tree of life using large number of taxa and/or concatenations of large number of protein (or DNA or RNA) sequences (e.g., Driskell

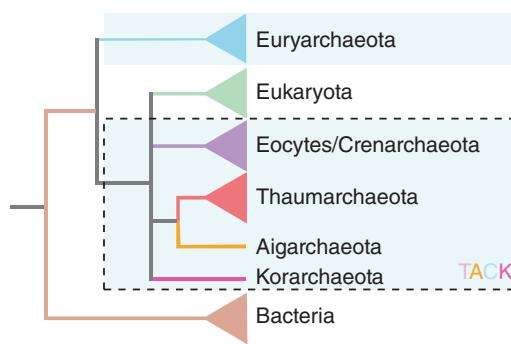
(a) Prokaryote-eukaryote model of evolution



(b) Three-domains hypothesis: monophyletic archaea



(c) Two-domains, eocyte hypothesis: paraphyletic archaea



**FIGURE 15.2** Models for the origin of eukaryotes invoking three or two domains of life for the origin of the eukaryotic host cell. Wedges represent radiations of species. (a) A model (not favored) of prokaryotes sequentially followed by eukaryotes. Redrawn from Pace (2009). Reproduced with permission from American Society for Microbiology. (b) The rooted three-domain model divides early cellular life into three major monophyletic groups: bacteria, archaea, and eukaryotes (the host lineage that acquired a bacterial endosymbiont that became the mitochondrion). According to this model, archaea and eukaryotes are most closely related and share a common ancestor that is not shared with bacteria. TACK refers to a group of archaea: Thaumarcheoata, Aigarchaeota, Eocytes/Crenarchaeota, and Korarchaeota. (c) A model in which there are two domains: bacteria and archaea. The closest lineage of the eukaryotes is one (or more) of the TACK group of archaea. Trees in both (b) and (c) are rooted on the bacterial stem. (b, c) Redrawn from Williams *et al.* (2013). Reproduced with permission from Macmillan Publishers.

A remarkable tree of life by Ciccarelli *et al.* (2006), from the group of Peer Bork, is available online at the Interactive Tree of Life webpage at <http://itol.embl.de/> (WebLink 15.1; see Letunic and Bork, 2007). Another extraordinary tree based on ribosomal RNA from about 3000 species is available from David Hillis and James Bull at <http://www.zo.utexas.edu/faculty/antisense/DownloadfilesToL.html> (WebLink 15.2).

*et al.*, 2004; Ciccarelli *et al.*, 2006). While the tree of life provides an appealing metaphor, there are other global descriptions of life forms such as a bush or reticulated tree (Doolittle, 1999) or a ring of life (Rivera and Lake, 2004). William Martin, Eugene Koonin, and others have emphasized that some fundamental evolutionary processes are not tree-like including the lateral transfer of genetic material among bacteria and archaea (discussed in Chapter 17); the endosymbiotic transfer of genes from organellar to nuclear genomes among eukaryotes; and the fusion of ancient genomes (Dagan and Martin, 2006; Martin, 2011; O’Malley and Koonin, 2011; Koonin, 2012).

Viruses do not meet the definition of living organisms, and are therefore excluded from most trees of life. Although they replicate and evolve, viruses only survive by commandeering the cell of a living organism (see Chapter 16).

## History of Life on Earth

Our recent view of the tree of life (Fig. 15.1) is accompanied by new interpretations of the history of life on Earth. All life forms share a common origin and are part of the tree of life. A species has an average half-life of 1–10 million years (Graur and Li, 2000), and more than 99% of all species that ever lived are now extinct (Wilson, 1992). In principle, there is one single tree of life that accurately describes the evolution of species. The object of phylogeny is to try to deduce the correct trees both for species and for homologous families of genes and proteins. Another object of phylogeny is to infer the time of divergence between organisms since the time they last shared a common ancestor.

The earliest evidence of life is from about 4 billion years ago (BYA), just 0.5 billion years after the formation of Earth. This earliest life was centered on RNA (rather than DNA or protein; reviewed in Joyce, 2002). Earth's atmosphere was anaerobic throughout much of early evolution, and this early life form was possibly a unicellular bacterium or bacterial-like organism. The first fossil evidence of life is dated about 3.5–3.8 BYA (e.g., Allwood *et al.*, 2006). The last common ancestor of life, predating the divergence of the lineage that leads to modern bacteria and modern archaea, was probably a hyperthermophile. This is suggested by the deepest branching organisms of trees (see Fig. 15.1), such as the bacterium *Aquifex* and the hyperthermophilic crenarcheota (Chapter 17). Eukaryotes appeared between 3 and 2 BYA and remained unicellular until almost 1 BYA. Plants and animals diverged approximately 1.5 BYA, as did fungi, from the lineage that gave rise to metazoans (animals; see Fig. 19.12). The most recent billion years of life has seen the evolution of an enormous variety of multicellular organisms. The so-called Cambrian explosion of 550 million years ago (MYA) witnessed a tremendous increase in the diversity of animal life forms. In the past 250 million years, the continents coalesced into the giant continent Pangaea (Fig. 15.3). When Pangaea separated into northern and southern supercontinents (Laurasia and Gondwana), this created natural barriers to reproduction and influenced subsequent evolution of life. The dinosaurs were extinct by 60 MYA, and the mammalian radiation was well underway.

The lines leading to modern *Homo sapiens*, chimpanzees, and bonobos diverged about 5 MYA (Chapter 19). The genomes of all three of these primates have now been sequenced, as described in chronology below. The earliest human ancestors include "Lucy," the early *Australopithecus*, and early hominids used stone tools over 2 MYA. Genomes from two extinct hominins, the Neandertals and the Denisovans, have now been sequenced (discussed in "Ancient DNA Projects" below). An overview of the history of life is shown in Figure 15.4.

## Molecular Sequences as the Basis of the Tree of Life

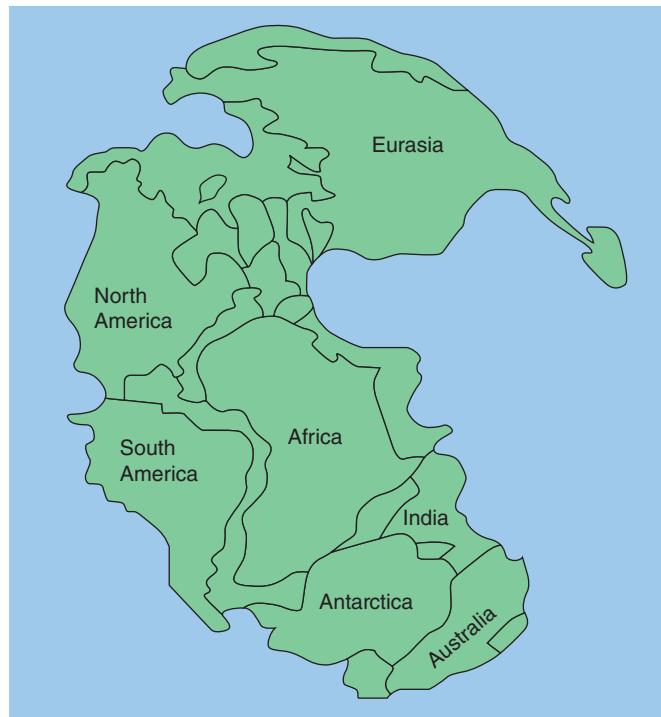
In past decades and centuries, the basis for proposing models of the tree of life was primarily morphology. Linnaeus divided animals into six classes (mammals, birds, fish, insects, reptiles, and worms), subdividing mammals according to features of their teeth, fish according to their fins, and insects by their wings. Early microscopic studies revealed that bacteria lack nuclei, allowing a fundamental separation of bacteria from the four other kingdoms of life. Bacteria could be classified based upon biochemical properties (e.g., by Albert Jan Kluyver (1888–1956)), and from a morphological perspective bacteria can be classified into several major groups. However, such criteria are insufficient to appreciate the dazzling diversity of millions of microbial species. Physical criteria by which to discover archaea as a distinct branch of life were therefore unavailable.

The advent of molecular sequence data has transformed our approach to the study of life. Such data were generated beginning in the 1950s and 1960s and, by 1978, Dayhoff's Atlas used several hundred protein sequences as the basis for PAM matrices (Chapter 3).

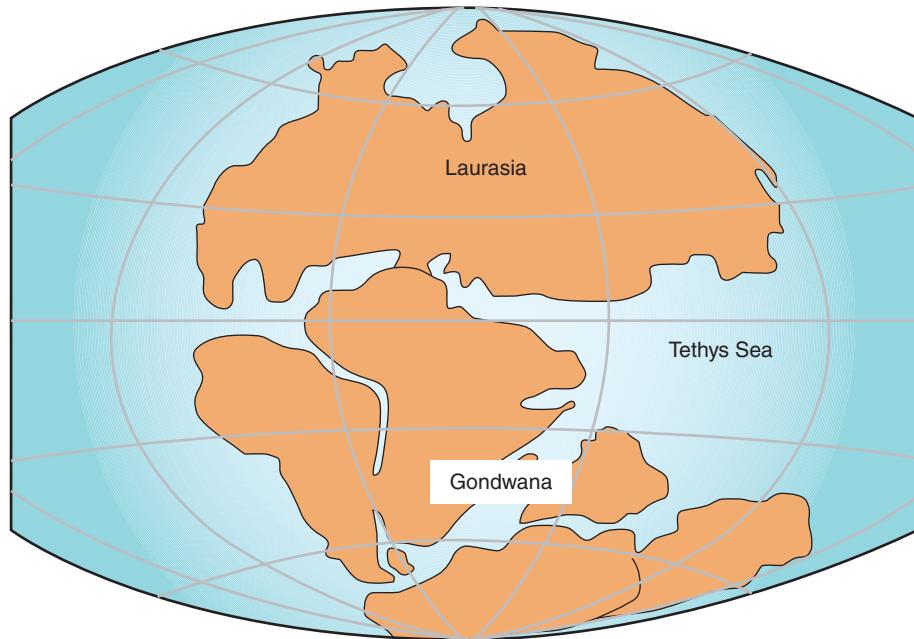
For another view of Pangaea, see Figure 16.12.

Multicellular organisms evolved independently many times. A variety of multicellular bacteria evolved several billion years ago, allowing selective benefits in feeding and in dispersion from predators (Kaiser, 2001; Chapter 17).

(a) Pangaea

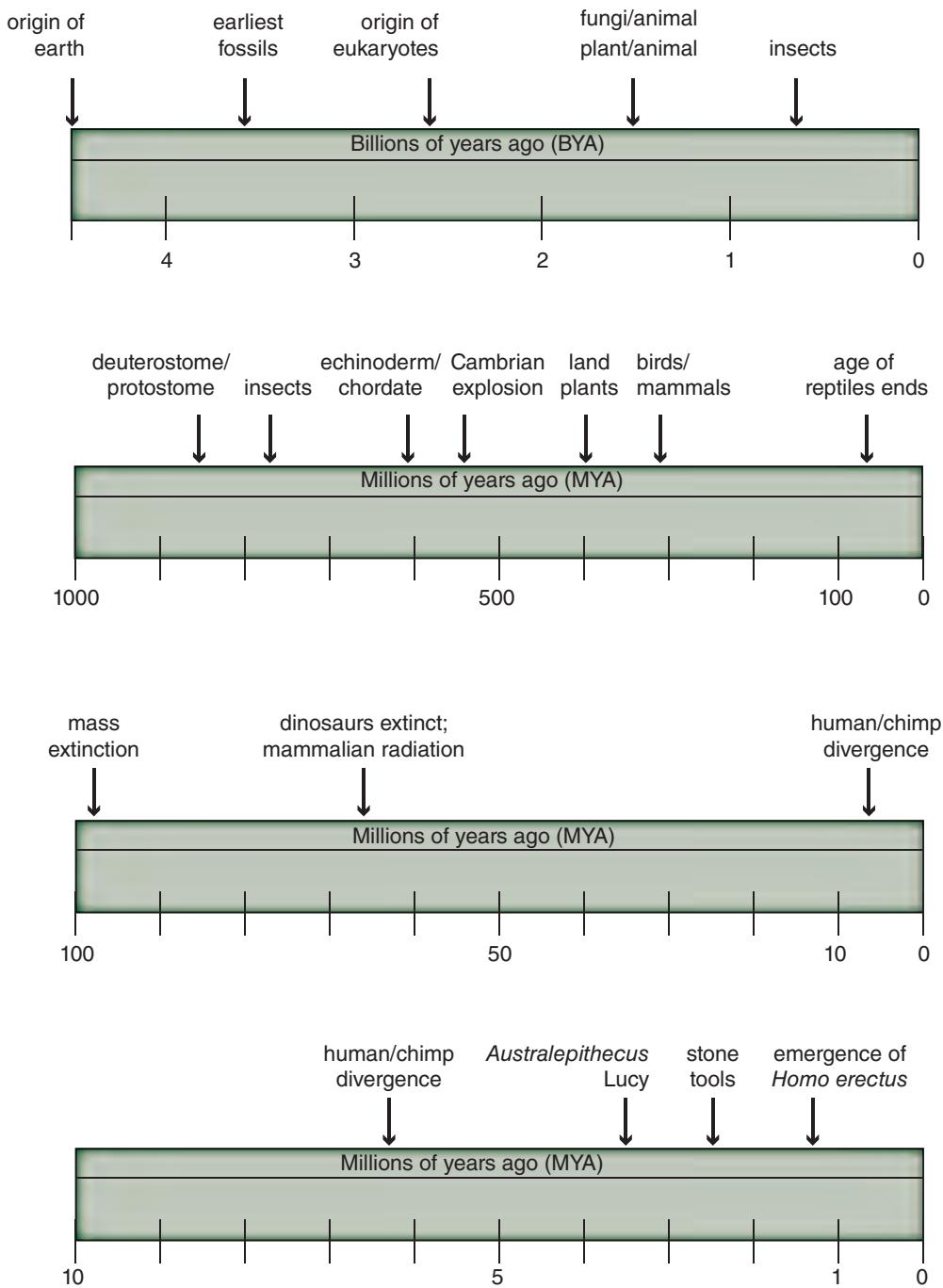


(b) Laurasia (modern Asia and North America) and Gondwana (modern Africa and South America)



**FIGURE 15.3** (a) Geological history of the earth from 225 MYA. At that time, there was one super-continent, Pangaea. By 165 MYA, Pangaea had separated into Laurasia (modern Asia and North America) and Gondwana (modern Africa and South America). (b) Near the end of the Triassic (~200 MYA), Laurasia and Gondwana had both begun separations that led to the present divisions among continents.

*Source:* (a) Kieff, licensed under the Creative Commons Attribution-Share Alike 3.0 Generic license and (b) Lenny222, licensed under the Creative Commons Attribution-Share Alike 3.0 Generic license.



**FIGURE 15.4** History of life on the planet. Data in part from Kumar and Hedges (1998), Hedges *et al.* (2001), and Benton and Ayala (2003).

There has been a rapid rise in available DNA sequences of the past several years, including metagenomic data (introduced in “Metagenomics Projects” below). Phylogenetic analyses are now possible based upon both phenotypic characters and gene sequences. The most widely used sequences are small subunit (SSU) rRNA molecules, which are present across virtually all extant life forms. The slow rate of evolution of SSU rRNAs and their convenient size makes them appropriate for phylogenetic analyses. Genome-sequencing efforts are now reshaping the field of evolutionary studies, providing thousands of DNA and protein sequences for phylogenetic trees. Major resources include the Ribosomal

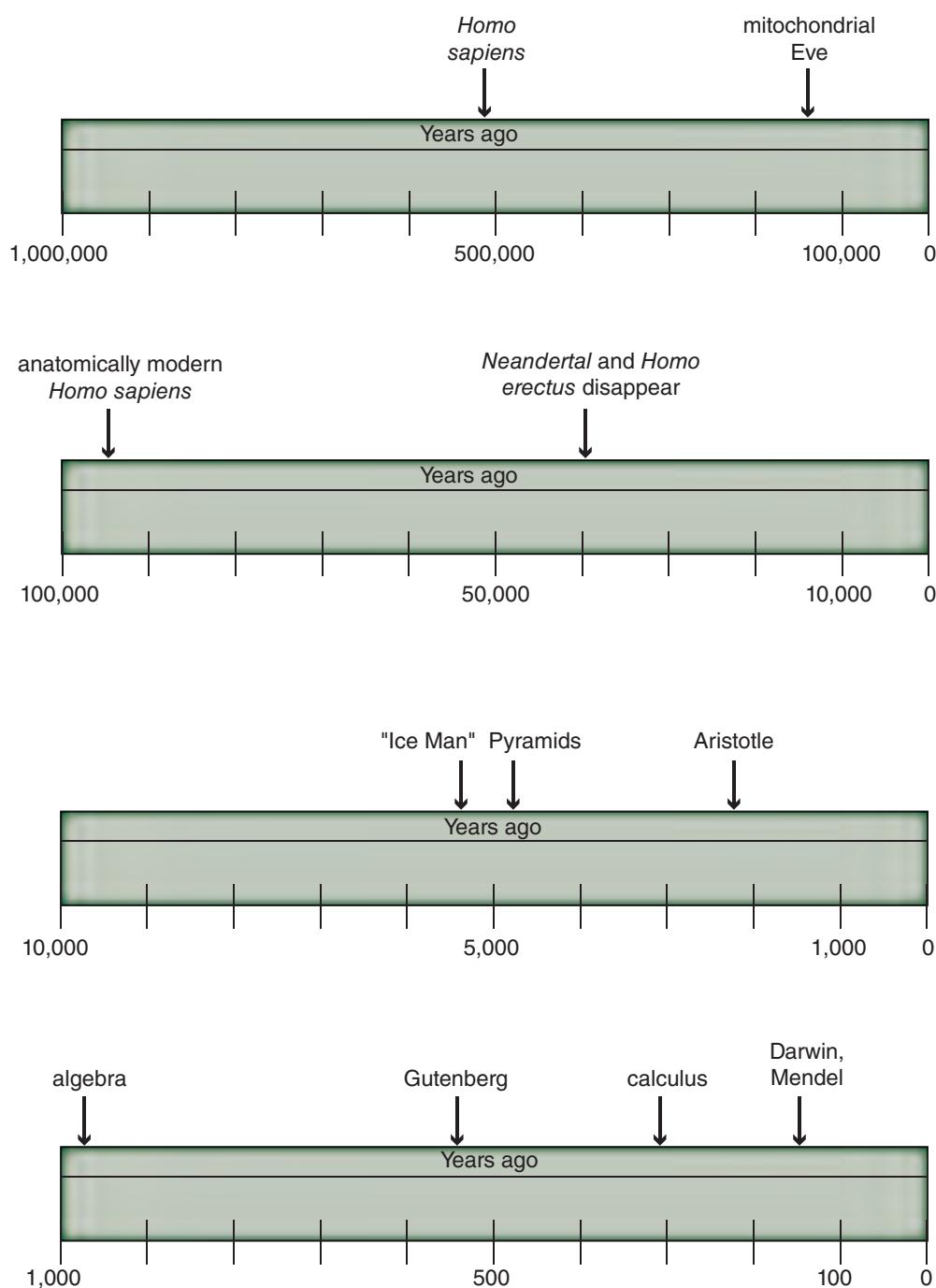


FIGURE 15.4 (Continued)

The Ribosomal Database Project can be viewed at <http://rdp.cme.msu.edu/> (WebLink 15.3). It currently includes 2.9 million 16S rRNA sequences. SILVA is at <http://www.arb-silva.de/> (WebLink 15.4). Both these resources offer large databases and extensive tools for sequence analysis. The "All-Species Living Tree" project at SILVA offers 16S and 23S rRNA datasets and phylogenetic trees spanning all sequenced type strains of archaea and bacteria.

Database Project (Cole *et al.*, 2014) and SILVA (Pruesse *et al.*, 2012; Quast *et al.*, 2013), and several large-scale genomics initiatives focus on rRNA sequencing (Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012; Yarza *et al.*, 2013).

Over 1100 complete bacterial and 100 archaeal genomes have now been sequenced (Table 15.2). We are now beginning to appreciate lateral gene transfer (Chapter 17), a phenomenon in which a species does not acquire a particular gene by descent from an ancestor. Instead, it acquires the gene horizontally (or laterally) from another unrelated

**TABLE 15.2 Summary of currently sequenced genomes (excluding viruses and organellar genomes).**

Organism	Complete	Draft assembly	In progress	Total
Prokaryotes	1117	966	595	2678
Archaea	100	5	48	153
Bacteria	1017	961	547	2525
Eukaryotes	36	319	294	649
Animals	6	137	106	249
Mammals	3	41	25	69
Birds		3	13	16
Fishes		16	16	32
Insects	2	38	17	57
Flatworms		3	3	6
Roundworms	1	16	11	28
Amphibians		1		1
Reptiles		2		2
Other animals		20	24	44
Plants	5	33	80	118
Land plants	3	29	73	105
Green Algae	2	4	6	12
Fungi	17	107	59	183
Ascomycetes	13	83	38	134
Basidiomycetes	2	16	11	29
Other fungi	2	8	10	20
Protists	8	39	46	93
Apicomplexans	3	11	16	30
Kinetoplasts	4	3	2	9
Other protists	1	24	28	53
Total	1153	1285	889	3327

Source: NCBI Genome, NCBI (<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>).

species; genes can therefore be exchanged between species (Eisen, 2000; Sousa and Hey, 2013). As a consequence, the use of different individual genes in molecular phylogeny often results in distinctly different tree topologies. Because of the phenomena of lateral gene transfer and gene loss, it might never be possible to construct a single tree of life that reflects the evolution of life on the planet (Wolf *et al.*, 2002).

The Genomics Standards Consortium has been established to provide community-driven standards for genomics research (Field *et al.*, 2008). When a molecule such as rRNA or any other gene is studied, researchers should provide minimal information describing experimental details of what was found along with relevant methods (Yilmaz *et al.*, 2011).

You can visit the GSC website at <http://gensc.org/> (WebLink 15.5).

## Role of Bioinformatics in Taxonomy

The field of bioinformatics is concerned with the use of computer algorithms and computer databases to elucidate the principles of biology. The domain of bioinformatics includes the study of genes, proteins, and cells in the context of organisms across the

The Catalogue of Life (<http://www.catalogueoflife.org/>, WebLink 15.6) lists >1.5 million species with >140 contributing databases. The Microbial Earth Project lists ~11,000 strains of archaea and bacteria, many thousands having genome projects, at <http://www.microbial-earth.org> (WebLink 15.7). The Earth Microbiome Project emphasizes metagenomics projects (<http://www.earthmicrobiome.org>, WebLink 15.8). NCBI taxonomy (<http://www.ncbi.nlm.nih.gov/taxonomy>, WebLink 15.9) includes 300,000 species. The Convention on Biological Diversity (<http://www.cbd.int>, WebLink 15.10) is an organization that address issues of global biodiversity. The Tree of Life is at <http://www.panspermia.org/tree.htm> (WebLink 15.11) and the Tree of Life Web Project (created by David R. Maddison) is at <http://tolweb.org/tree/phylogeny.html> (WebLink 15.12).

Metazoa are animals. Vertebrate metazoan include horses and fish. Invertebrate metazoan include worms and insects.

Ensembl Genome is accessible from <http://ensemblgenomes.org/> (WebLink 15.13). Ensembl Bacteria currently features over 20,000 genomes and has its own site <http://bacteria.ensembl.org/> (WebLink 15.14).

You can access the Genomes page at NCBI via <http://www.ncbi.nlm.nih.gov/genome> (WebLink 15.15), or from the home page of NCBI. Select All Databases then Genomes.

tree of life. Some have advocated a web-based taxonomy intended to catalog an inventory of life (Blackmore, 2002). Several projects attempt to create a tree of life (see sidebar). Others suggest that while web-based initiatives are useful, the current system is adequate: zoological, botanical, or other specimens are collected, named, and studied according to guidelines established by international conventions (Knapp *et al.*, 2002).

Databases such as the Catalogue of Life list the number of named species. Mora *et al.* (2011) have estimated that the total number of eukaryotic species is ~8.7 million ± 1.3 million (standard error). Other recent estimates have been as high as 100 million species. For bacteria and archaea the estimates are more uncertain.

## PROMINENT WEB RESOURCES

We now introduce several main web resources for the study of genomes in this and the following chapters.

### Ensembl Genomes

The European Bioinformatics Institute (EBI)/Ensembl offers a variety of genome resources. We have encountered the Ensembl website (e.g., Chapters 2, 8) with its particular focus on vertebrate species. Ensembl Genomes offers a complementary set of web interfaces for five nonvertebrate groups: bacteria, protists, fungi, plants, and invertebrate metazoan (Kersey *et al.*, 2014).

### NCBI Genome

The genomes section of the National Center for Biotechnology Information (NCBI) is organized with features to search eukaryotes, bacteria, archaea, and viruses with additional specialized genomics resources. The current NCBI holdings include over 3300 eukaryotic, bacterial, and archaeal genomes of which ~1200 have been completely sequenced (Table 15.2). There are an additional ~5000 completely sequenced organellar genomes (discussed in “1981: First Eukaryotic Organellar Genome” below).

### Genome Portal of DOE JGI and the Integrated Microbial Genomes

The US Department of Energy Joint Genome Institute (DOE JGI) has had a prominent role in supporting the Human Genome Project and in sequencing genomes from plants, fungi, microbes, and metagenomes. Its Genome Portal offers access to 4000 projects (Grigoriev *et al.*, 2012).

Within DOE JGI, the Integrated Microbial Genomes (IMG) system supports storage, analysis, and distribution of microbial genomes (Markowitz *et al.*, 2014b) and metagenomes (Markowitz *et al.*, 2014a). It supports projects such as the Human Microbiome Project (Markowitz *et al.*, 2012), and we explore this resource in Chapter 17.

### Genomes On Line Database (GOLD)

The GOLD database monitors genome and metagenome sequencing projects (Pagani *et al.*, 2012). Led by Nikos Kyrpides and colleagues at the DOE JGI, it is an authoritative repository. In addition to traditional search features it offers Genome Map and Genome Earth views of genome projects.

### UCSC

The genome browser at the University of California, Santa Cruz has a particular emphasis on vertebrate genomes (see Chapters 8 and 19). It also has associated microbial and archaeal Genome and Table Browsers.

## GENOME-SEQUENCING PROJECTS: CHRONOLOGY

The advent of DNA-sequencing technologies in the 1970s, including Frederick Sanger's dideoxynucleotide methodology, enabled large-scale sequencing projects to be performed. This chapter provides a brief history of genome-sequencing projects, including the completion of the genomic sequence of the first free-living organism in 1995, *Haemophilus influenzae*. By 2001, a draft sequence of the human genome was reported by two groups. The most remarkable feature of current efforts to determine the sequence of complete genomes is the dramatic increase in data that are collected each year (Fig. 2.3). The ability to sequence  $>10^{17}$  nucleotides of genomic DNA presents the scientific community with unprecedented opportunities and challenges.

Several themes have emerged in the past several years:

- The amount of sequence data that are generated continues to accelerate rapidly.
- For many genomes, even unfinished genomic sequence data – that is, versions of genomic sequence that include considerable gaps and sequencing errors – are immediately available and useful to the scientific community. A finished sequence (defined in “Four Approaches to Genome Assembly” below) provides substantially better descriptions of genome features than an unfinished sequence.
- Low-coverage genomes are also useful. In reporting the genome sequence of an orangutan, Locke *et al.* (2011) also provide sequence analysis of 10 unrelated orangutans.
- Annotations have a major impact on the usefulness of sequence data (Klimke *et al.*, 2011).
- Comparative genome analysis is needed for solving problems such as identifying protein-coding genes in human and mouse or differences in virulent and nonvirulent strains of pathogens (Miller *et al.*, 2004). Comparative analyses are also useful to define gene regulatory regions and the evolutionary history of species through the analysis of conserved DNA elements.
- Polyploid genomes have now been sequenced, including the hexaploid bread wheat genome (17 Gb) and the loblolly pine (a conifer having a genome spanning 23.2 Gb; Neale *et al.*, 2014).

### Brief Chronology

The progress in completing many hundreds of genome-sequencing projects has been rapid, and we can expect the pace to accelerate in the future. In the following sections we present a chronological overview to provide a framework for these events. When the sequencing of the first bacterial genomes was completed in 1995, there were relatively few other genome sequences available for comparison. Now with thousands of completed genomes available (including organellar genomes), we are better able to annotate and interpret the biological significance of genome sequences.

### 1976–1978: First Bacteriophage and Viral Genomes

Bacteriophage are viruses that infect bacteria. Fiers *et al.* (1976) reported the first complete bacteriophage genome, MS2. This genome of 3569 base pairs encodes just 4 genes. The next complete virus genome was Simian Virus 40 (SV40) by Fiers *et al.* (1978). That genome contains 5224 base pairs and contains 8 genes (7 of which encode proteins).

Frederick Sanger and colleagues also sequenced the genome of bacteriophage φX174 (Sanger *et al.*, 1977a). They developed several DNA-sequencing techniques, including the dideoxynucleotide chain termination procedure (Sanger *et al.*, 1977b). Bacteriophage φX174 is 5386 bp encoding 11 genes (see GenBank accession NC\_001422.1). A depiction of the NCBI Nucleotide and Genome entries for this viral genome is provided in Figure 15.5. At the time, a surprising result was the unexpected presence of overlapping genes that are transcribed on different reading frames.

Visit the Genome Portal at

<http://genome.jgi.doe.gov/>  
(WebLink 15.16).

IMG is online at <https://img.jgi.doe.gov> (WebLink 15.17).

GOLD is available at <http://genomesonline.org/> (WebLink 15.18). Currently (February 2015) it lists ~60,000 complete and ongoing genome projects.

The UCSC Genome Browser and Table Browser are at <http://genome.ucsc.edu> (WebLink 15.19). Its European mirror is <http://genome-euro.ucsc.edu> (WebLink 15.20). The UCSC Microbial Genome Browser is at <http://microbes.ucsc.edu> (WebLink 15.21) or <http://archaea.ucsc.edu> (WebLink 15.22) for archaea.

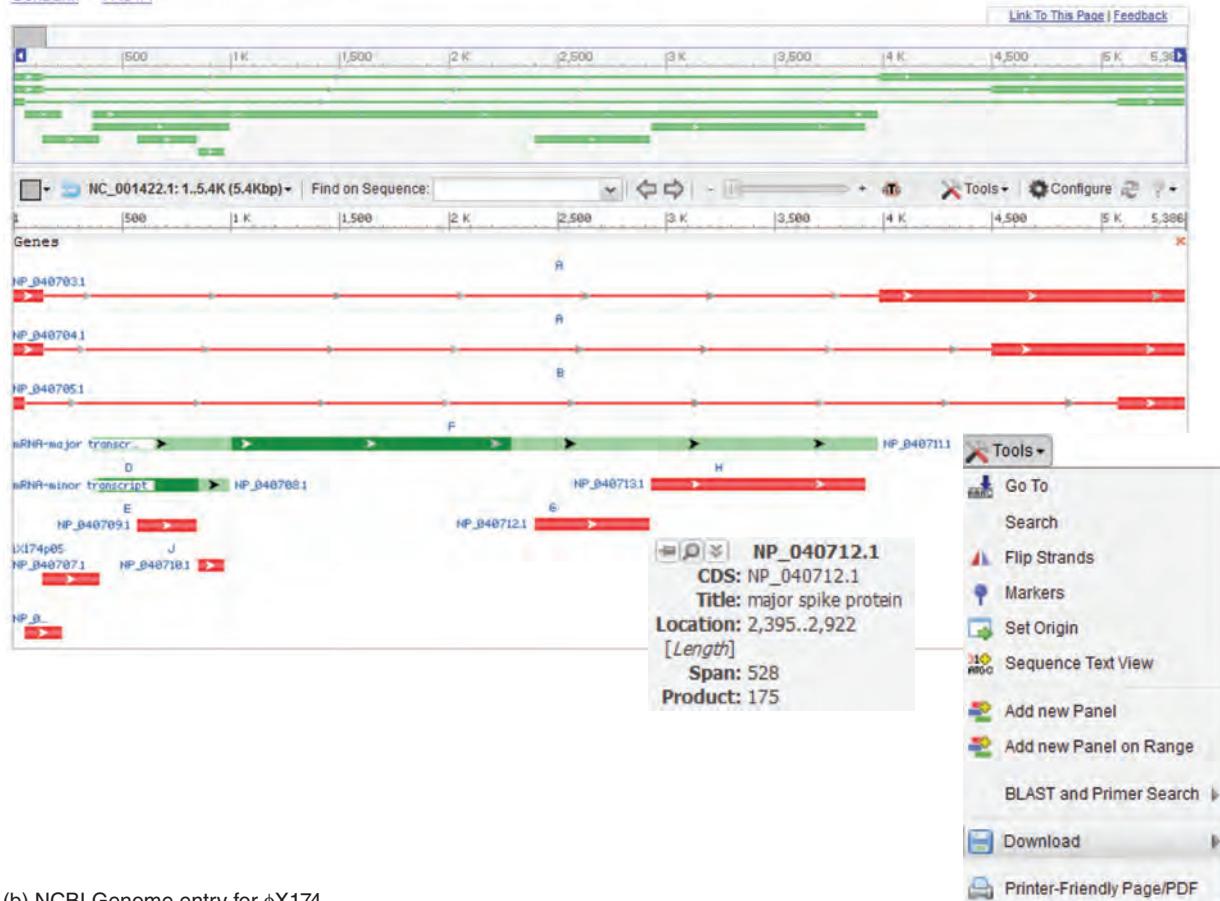
The bacteriophage MS2 genome has RefSeq accession NC\_001417.2. The SV40 RefSeq accession is NC\_001669.1.

(a) NCBI Nucleotide graphic view

### Enterobacteria phage phiX174 sensu lato, complete genome

NCBI Reference Sequence: NC\_001422.1

[GenBank](#) [FASTA](#)



(b) NCBI Genome entry for φX174

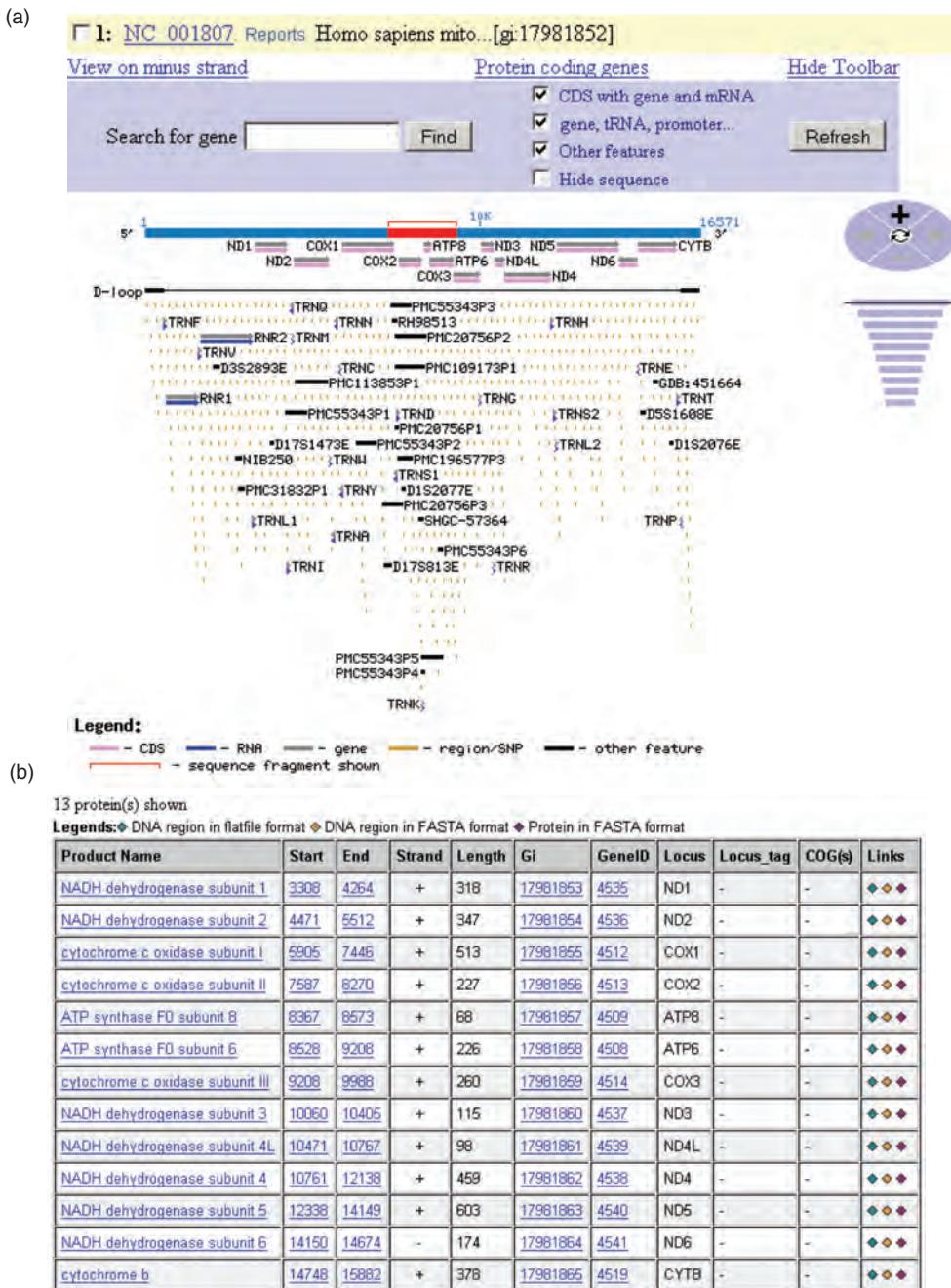


**FIGURE 15.5** NCBI data for bacteriophage φX174. (a) The Nucleotide record was obtained by viewing the entry for accession NC\_001422.1 in the graphics display format. This provides an overview of the predicted open reading frames (ORFs). Mousing over an entry displays information as shown for major spike protein. The various options of the tools menu are also shown. (b) The NCBI Genome record includes a summary of the accession number, length, number of proteins (11), sequence neighbors ( $n = 77$ ), and host species.

Source: NCBI.

### 1981: First Eukaryotic Organellar Genome

The first complete organellar genome to be sequenced was the human mitochondrial (Anderson *et al.*, 1981). The genome is characterized by extremely little noncoding DNA. The great majority of metazoan (i.e., multicellular animal) mitochondrial



**FIGURE 15.6** NCBI Genome includes entries for completed organellar genomes including the mitochondrial genome. There are generally 13 or 14 protein-coding genes encoded by the mitochondrial genome. You can access the protein sequences in the FASTA format, as a multiple alignment, or in protein clusters. This list includes the reference human mitochondrial genome (rCRS/Mitomap sequence NC\_012920.1; 16,569 base pair circular genome).

Source: NCBI Genome, NCBI.

genomes are about 15–20 kb (kilobase) circular genomes. The human mitochondrial genome is 16,569 base pairs and encodes 13 proteins, 2 ribosomal RNAs, and 22 transfer RNAs. It can be accessed through the NCBI Genome site (Fig. 15.6). DNA and corresponding protein sequences of all the mitochondrial genes are accessible in graphical or tabular forms.

We discuss the human mitochondrial genome in Chapters 20 and 21. Its accession is NC\_012920.1 (revised Cambridge reference sequence).

**TABLE 15.3 Selected mitochondrial genomes arranged by size. As of March 2015, ~5000 metazoan (multicellular animal) organellar genomes have been sequenced, ~180 fungi, and ~125 plants. Note that human mitochondrial genome NC\_001807.4 has been replaced by a new rCRS/Mitomap reference mitochondrial genome, NC\_012920.1. Reanalysis of the Cambridge reference sequence is described at <http://www.mitomap.org/MITOMAP/CambridgeReanalysis>.**

Kingdom	Species	Accession no.	Size (bp)
Eukaryote	<i>Plasmodium falciparum</i> (malaria parasite)	NC_002375.1	5,967
Metazoa (Bilateria)	<i>Caenorhabditis elegans</i> (worm)	NC_001328.1	13,794
Plant (Chlorophyta)	<i>Chlamydomonas reinhardtii</i> (green alga)	NC_001638.1	15,758
Metazoa (Bilateria)	<i>Mus musculus</i>	NC_005089.1	16,299
Metazoa (Bilateria)	<i>Pan troglodytes</i> (chimpanzee)	NC_001643.1	16,554
Metazoa (Bilateria)	<i>Homo sapiens</i>	NC_012920.1	16,569
Metazoa (Cnidaria)	<i>Metridium senile</i> (sea anemone)	NC_000933.1	17,443
Metazoa (Bilateria)	<i>Drosophila melanogaster</i>	NC_001709.1	19,517
Fungi (Ascomycota)	<i>Schizosaccharomyces pombe</i>	NC_001326.1	19,431
Fungi	<i>Candida albicans</i>	NC_002653.1	40,420
Eukaryote (stramenopiles)	<i>Pylaiella littoralis</i> (brown alga)	NC_003055.1	58,507
Fungi (Chytridiomycota)	<i>Rhizophyllum</i> sp. 136	NC_003053.1	68,834
Eukaryote	<i>Reclinomonas americana</i> (protist)	NC_001823.1	69,034
Fungi (Ascomycota)	<i>Saccharomyces cerevisiae</i>	NC_001224.1	85,779
Plant (Streptophyta)	<i>Arabidopsis thaliana</i>	NC_001284.2	366,924
Plant (Streptophyta)	<i>Zea mays</i> (corn)	NC_008332.1	680,603
Plant (Streptophyta)	<i>Tripsacum dactyloides</i>	NC_008362.1	704,100
Plant (Streptophyta)	<i>Cucurbita pepo</i>	NC_014050.1	982,833

Source: NCBI Genome, NCBI (<http://www.ncbi.nlm.nih.gov/Genomes/>).

Information on organellar genomes is available at <http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html> (WebLink 15.23).

Today, there are nearly 5000 completed mitochondrial genome sequences. Several of these are listed in **Table 15.3**, including several exceptionally large cases. While the thale cress *Arabidopsis thaliana* has a large mitochondrial genome (367 kb), those of some other plants approach or exceed a megabase. There is therefore a tremendous diversity of mitochondrial genomes (Lang *et al.*, 1999). Molecular phylogenetic approaches suggest that mitochondria are descendants of an endosymbiotic  $\alpha$ -proteobacterium.

## 1986: First Chloroplast Genomes

The first chloroplast genomes reported were *Nicotiana tabacum* (Shinozaki *et al.*, 1986), followed by the liverwort *Marchantia polymorpha* (Ohyama *et al.*, 1988). Most plant chloroplast genomes are 60,000–200,000 bp in size. Kua *et al.* (2012) compared 174 chloroplast genomes, describing duplications in an inverted repeat region even in highly reduced parasitic orchid genome or in a massive algal chloroplast.

There are other chloroplast-like organelles in eukaryotic organisms. Unicellular protozoan parasites of the phylum Apicomplexa, such as *Toxoplasma gondii* (**Table 15.4**), have smaller plastid genomes of which >400 are currently listed at NCBI.

We discuss chloroplasts and other plastids in the plant section of Chapter 19.

**TABLE 15.4 Selected chloroplast genomes.**

Species	Common name	Accession no.	Size (bp)
<i>Arabidopsis thaliana</i>	Thale cress	NC_000932.1	154,478
<i>Guillardia theta</i>	Red alga	NC_000926.1	121,524
<i>Marchantia polymorpha</i>	Liverwort; moss	NC_001319.1	121,024
<i>Nicotiana tabacum</i>	Tobacco	NC_001879.2	155,943
<i>Oryza sativa</i>	Rice	NC_001320.1	134,525
<i>Porphyra purpurea</i>	Red alga	NC_000925.1	191,028
<i>Toxoplasma gondii</i>	Apicomplexan parasite	NC_001799.1	34,996
<i>Zea mays</i>	corn	NC_001666.2	140,384

### 1992: First Eukaryotic Chromosome

The first eukaryotic chromosome was sequenced in 1992: chromosome III of the budding yeast *S. cerevisiae* (Oliver *et al.*, 1992). There were 182 predicted open reading frames (for proteins larger than 100 amino acids), and the size of the sequenced DNA was 315 kb. Of the 182 open reading frames that were identified, only 37 corresponded to previously known genes and 29 showed similarity to known genes. We explore this genome in Chapter 18.

### 1995: Complete Genome of Free-Living Organism

The first genome of a free-living organism to be completed was the bacterium *Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995; NC\_000907.1). Its size is 1,830,138 bp (i.e., 1.8 Mb or megabase pairs). This organism was sequenced at The Institute for Genomic Research using the whole-genome shotgun sequencing and assembly strategy (see “Four Approaches to Genome Assembly” below).

To study this genome in NCBI, go to the Genome page and “browse by organism.” Entering this bacterial name leads to links such as the lineage, a dendrogram of related bacteria, related BioProjects, and publications.

By the end of 1995, the complete DNA sequence of a second bacterial genome had been obtained, *Mycoplasma genitalium* (Fraser *et al.*, 1995). Notably, this was one of the smallest known genomes of any free-living organism (we introduce several smaller genomes in Chapter 17).

We describe this bacterial genome as derived from a “free-living organism” to distinguish it from a viral genome or an organellar genome. Viruses (Chapter 16) exist on the borderline of the definition of life, and organellar genomes are derived from bacteria that are no longer capable of independent life.

The *M. genitalium* accession is NC\_000908.2 and its size is 580,076 base pairs.

### 1996: First Eukaryotic Genome

The complete genome of the first eukaryote, *S. cerevisiae* (a yeast; see Chapter 18; Goffeau *et al.*, 1996), was sequenced by 1996. This was accomplished by a collaboration of over 600 researchers in 100 laboratories spread across Europe, North America, and Japan.

In 1996, The Institute of Genomic Research (TIGR) researchers reported the first complete genome sequence for an archaeon, *Methanococcus jannaschii* (Bult *et al.*, 1996). This offered the first opportunity to compare the three main divisions of life, including the overall metabolic capacity of bacteria, archaea, and eukaryotes. Other genomes sequenced in 1996 are listed in Web Document 15.3.

### 1997: *Escherichia coli*

In 1997, the complete genomic sequences of two archaea were reported (Klenk *et al.*, 1997; Smith *et al.*, 1997; Web Document 15.3). Of the five bacterial genomes that were reported, the most well-known is that of *Escherichia coli* (Blattner *et al.*, 1997;

Koonin, 1997), which has served as a model organism in bacteriology for decades. Its 4.6 Mb genome encodes over 4200 proteins, of which 38% had no identified function at the time. We explore this further in Chapter 17.

### 1998: First Genome of Multicellular Organism

The nematode *Caenorhabditis elegans* was the first multicellular organism to have its genome sequenced, although technically the sequencing is still not complete (because of the presence of repetitive DNA elements that have been difficult to resolve). The sequence spans 97 Mb and is predicted to encode over 20,000 genes (the *C. elegans* Sequencing Consortium, 1998).

Two more archaea brought the total to four sequenced genomes by 1998 (Web Document 15.3). Six more bacterial genomes were also sequenced. The genome of *Rickettsia prowazekii*, the  $\alpha$ -proteobacterium that causes typhus and was responsible for tens of millions of deaths in the twentieth century, is very closely related to the eukaryotic mitochondrial genome (Andersson *et al.*, 1998).

### 1999: Human Chromosome

In 1999, the sequence of the euchromatic portion of human chromosome 22 was published (Web Document 15.3; Dunham *et al.*, 1999). This was the first human chromosome to be essentially completely sequenced. We discuss each of the human chromosomes in Chapter 20.

### 2000: Fly, Plant, and Human Chromosome 21

We describe these and other eukaryotic genomes in Chapter 19.

In 2000, the completed genome sequences of the fruit fly *Drosophila melanogaster* and the plant *A. thaliana* were reported, bringing the number of eukaryotic genomes to four (with a yeast and a worm; Web Document 15.3). The *Drosophila* sequence was obtained by scientists at Celera Genomics and the Berkeley Drosophila Genome Project (BDGP; Adams *et al.*, 2000). There are approximately 14,000 protein-coding genes (according to current Ensembl annotation). *Arabidopsis* is a thale cress of the mustard family. Its compact genome serves as a model for plant genomics (*Arabidopsis* Genome Initiative, 2000).

Also in the year 2000, human chromosome 21 was the second human chromosome sequence to be reported (Hattori *et al.*, 2000). This is the smallest of the human autosomes. An extra copy of this chromosome causes Down syndrome, the most common inherited cause of intellectual disability.

Meanwhile, bacterial genomes continued to be sequenced, and many surprising properties emerged. The genome of *Neisseria meningitidis*, which causes bacterial meningitis, contains hundreds of repetitive elements (Parkhill *et al.*, 2000; Tettelin *et al.*, 2000). Such repeats are more typically associated with eukaryotes. The *Pseudomonas aeruginosa* genome is 6.3 Mb, making it the largest of the sequenced bacterial genomes at that time (Stover *et al.*, 2000).

We discuss lateral gene transfer in Chapter 17.

Among the archaea, the genome of *Thermoplasma acidophilum* was sequenced (Ruepp *et al.*, 2000). This organism thrives at 59°C and pH 2. Remarkably, it has undergone extensive lateral gene transfer with *Sulfolobus solfataricus*, an archaon that is distantly related from a phylogenetic perspective but occupies the same ecological niche in coal heaps.

### 2001: Draft Sequences of Human Genome

Two groups published the completion of a draft version of the human genome. This was accomplished by the International Human Genome Sequencing Consortium (2001) and by a consortium led by Celera Genomics (Web Document 15.3; Venter *et al.*, 2001). The

reports both arrive at the conclusion that there are about 30,000–40,000 protein-coding genes in the genome, an unexpectedly small number. Subsequently, the number of human genes was estimated to be 20,000 to 25,000 (International Human Genome Sequencing Consortium, 2004), while currently the Ensembl estimate is ~20,300. Analysis of the human genome sequence will have vast implications for all aspects of human biology.

The bacterial genomes that are sequenced continue to have interesting features. *Mycoplasma pulmonis* has one of the lowest guanine–cytosine (GC) contents that have been described: 26.6% (Chambaud *et al.*, 2001). The genome of *Mycobacterium leprae*, the bacterium that causes leprosy, has undergone massive gene decay with only half the genome coding for genes (Cole *et al.*, 2001). Analysis of the *Pasteurella multocida* genome suggests that the radiation of the  $\gamma$  subdivision of proteobacteria, which includes *H. influenzae* and *E. coli* and other pathogenic gram-negative bacteria, occurred at about 680 MYA (May *et al.*, 2001). The *Sinorhizobium meliloti* genome consists of a circular chromosome and two additional megaplasmids (Galibert *et al.*, 2001). Together these three elements total 6.7 Mb, expanding our view of the diversity of bacterial genome organization.

Cryptomonads are a type of algae that contain one distinct eukaryotic cell (a red alga, with a nucleus) nested inside another cell (see Fig. 19.9). This unique arrangement derives from an ancient evolutionary fusion of two organisms. That red algal nucleus, termed a nucleomorph, is the most gene-dense eukaryotic genome known. Its genome was sequenced (Douglas *et al.*, 2001) and found to be dense (1 gene per 977 base pairs) with ultrashort noncoding regions.

### 2002: Continuing Rise in Completed Genomes

In the year 2002, dozens more microbial genomes were sequenced. Of the eukaryotes (Web Document 15.3), the fission yeast *Schizosaccharomyces pombe* was found to have the smallest number of protein-coding genes (4824; Wood *et al.*, 2002). The genomes of both the malaria parasite *Plasmodium falciparum* and its host, the mosquito *Anopheles gambiae*, were reported (Holt *et al.*, 2002). Additionally, the genome of the rodent malaria parasite *Plasmodium yoelii yoelii* was determined and compared to that of *P. falciparum* (Carlton *et al.*, 2002).

### 2003: HapMap

In 2003 the Human Genome Project was concluding on the 50th anniversary of Watson and Crick's 1953 report on the double helix. That year the International HapMap Consortium (2003) launched a project to catalog common patterns of DNA sequence variation in the human genome. We describe the rich results of this project in Chapter 20. In part, it was significant because the focus began shifting from where humans are placed in the tree of life (relative to species such as mouse, nematode, and plants) to what genetic and genomic differences occur within the human species.

### 2004: Chicken, Rat, and Finished Human Sequences

The red jungle fowl *Gallus gallus*, better known as chicken, last shared a common ancestor with humans ~310 MYA and it is a descendant of the dinosaurs. The sequencing of its genome (International Chicken Genome Sequencing Consortium, 2004) revealed many surprising similarities (e.g., long blocks of conserved synteny with humans, and both coding and noncoding regions that are highly conserved between the two species) as well as notable differences (e.g., a relative paucity of retroposed pseudogenes). The rat genome also offered a wealth of information, in particular allowing three-way comparisons with the mouse and human genomes (Rat Genome Sequencing Project Consortium, 2004).

The International Human Genome Sequencing Consortium (2004) reported a draft sequence of the euchromatic portion of the human genome. This included 341 gaps and a low error rate of 1 per 100,000 bases. This assembly corresponded to Build 35 (GRCh35) and was to be followed by GRCh36 (in 2006), GRCh37 (2009), and GRCh38 (2013).

### 2005: Chimpanzee, Dog, Phase I HapMap

The genome of the chimpanzee *Pan troglodytes* was reported (Chimpanzee Sequencing and Analysis Consortium, 2005). These apes are among humans' closest relatives and the relatively few differences between our genomes provide a fascinating glimpse into what makes us human. That year Lindblad-Toh *et al.* reported the 4.4 Gb dog genome with its 38 pairs of autosomes and two sex chromosomes. Many dog breeds are susceptible to diseases that also afflict humans, and the genome sequence facilitates comparative studies.

The International HapMap Consortium (2005) released its first findings, characterizing more than 1 million SNPs in several geographic populations, including allele frequency distributions.

### 2006: Sea Urchin, Honeybee, dbGaP

Highlights included analyzing the honeybee genome (Honeybee Genome Sequencing Consortium, 2006) and the sea urchin (Sea Urchin Genome Sequencing Consortium *et al.*, 2006). The NIH introduced the Database of Genotypes and Phenotypes (dbGaP; Mailman *et al.*, 2007). dbGaP has emerged as a major repository for SNP and sequence data (Chapter 21).

### 2007: Rhesus Macaque, First Individual Human Genome, ENCODE Pilot

The public consortium that sequenced the human genome analyzed DNA from a pool of anonymous individuals, while the Celera effort relied mostly on DNA from J. Craig Venter. In 2007 Venter's became the first genome of an individual person to be sequenced (Levy *et al.*, 2007). While next-generation sequencing had become available, that study used the longer reads available from Sanger sequencing. That year the *Macaca mulatta* genome was sequenced (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007), and the ENCODE Project Consortium *et al.* (2007) released its findings on 1% of the human genome (Chapter 8).

### 2008: Platypus, First Cancer Genome, First Personal Genome Using NGS

In 2008 the era of individual human genome sequencing using next-generation sequencing began with the report of James Watson's genome (Wheeler *et al.*, 2008) as well as that of an Asian individual (Wang *et al.*, 2008) and a cancer genome (Ley *et al.*, 2008). The genome of the remarkable platypus was reported (Warren *et al.*, 2008), combining reptilian features (such as egg-laying) with mammalian features (such as female lactation). Such genome sequences provide essential resources for comparative mammalian analyses.

### 2009: Bovine, First Human Methlyome Map

The sequencing of the cattle genome may lead to genetic improvement for meat and milk production (Bovine Genome Sequencing and Analysis Consortium *et al.*, 2009). By 2009 maps of the human methylome emerged (Lister *et al.*, 2009). DNA methylation represents a heritable epigenetic modification often involving methylation of CpG dinucleotides. Aberrant methylation has been linked to disease.

## 2010: 1000 Genomes Pilot, Neandertal

Neandertals were the closest hominid relatives of humans until they became extinct about 30,000 years ago. Green *et al.* (2010) reported a draft sequence of the Neandertal genome. Surprisingly, humans of European and Asian ancestry share up to 3% of their genomic variants with Neandertals, while individuals of African ancestry do not. This may be explained by interbreeding between Neandertals and those early humans who migrated north from Africa.

The International HapMap 3 Consortium *et al.* (2010) continued its characterization of SNPs, expanding to >1100 individuals in 11 geographic sites across the world. Meanwhile the 1000 Genomes Project Consortium *et al.* (2010) reported ~15 million SNPs, 1 million short indels, and 20,000 structural variants, describing common variation based on next-generation sequencing. We discuss their findings in Chapters 20 and 21.

## 2011: A Vision for the Future of Genomics

Eric Green (director of the National Human Genome Research Institute or NHGRI) and colleagues articulated a vision for accomplishments across five domains of genomics research (Green and Guyer, 2011). These are: (1) understanding the structure of genomes, largely accomplished during the Human Genome Project from 1990 to 2003; (2) understanding the biology of genomes, largely spanning 2004–2010; (3) understanding the biology of disease, projected to extend to 2020; and, beyond 2020, (4) advancing the science of medicine; and (5) improving the effectiveness of healthcare.

## 2012: Denisovan Genome, Bonobo, and 1000 Genomes Project

The two primate species most closely related to humans are the chimpanzee (*Pan troglodytes*) and the bonobo (*Pan paniscus*). Prüfer *et al.* (2012) reported the genome of the bonobo, finding that some portions of the human genome are more closely related to either of these two apes. In addition to Neandertals, humans are also closely related to the extinct group of the Denisovans. Meyer *et al.* (2012) reported the first Denisovan genome. 1000 Genomes Project Consortium *et al.* (2012) reported genetic variation from 1092 human genomes across 14 populations.

## 2013: The Simplest Animal and a 700,000-Year-Old Horse

The ctenophores are the simplest animals, including comb jellies, sea walnuts, and sea gooseberries. Andy Baxevanis and colleagues described the genome of the comb jelly *Mnemiopsis leidyi*, showing that its lineage branched off from that leading to other animals at an early stage (Ryan *et al.*, 2013).

Orlando *et al.* (2013) sequenced the oldest genome to date. This came from a horse's foot bone (dated 780,000 to 560,000 years ago from the Middle Pleistocene).

## 2014: Mouse ENCODE, Primates, Plants, and Ancient Hominids

In 2014 the Mouse ENCODE Consortium reported its characterization of DNA elements in the mouse genome (Yue *et al.*, 2014), complementing the human ENCODE project. The genome of the gibbon was published by Carbone *et al.* (2014). The genome of the sugar beet (a eudicot) was described by Dohm *et al.* (2014), as was the genome of the hardwood tree *Eucalyptus* (Myburg *et al.*, 2014).

Ancient hominid genomes continue to be sequenced, including a Late Pleistocene human from Montana (Rasmussen *et al.*, 2014), a 45,000-year-old human from Siberia (Fu *et al.*, 2014), and an Upper Paleolithic Siberian human (Raghavan *et al.*, 2014), a Neandertal woman's genome (whose parents were likely half-siblings; Prüfer *et al.*, 2014).

## 2015: Diversity in Africa

The bulk of genetic diversity in humans is in African individuals. The African Genome Variation Project reported genotypes from 1481 Africans and whole-genome sequences from 320 sub-Saharan Africans (Gurdasani *et al.*, 2015).

A genome-wide association study (GWAS, introduced in Chapter 21) studied obesity in ~339,000 individuals (Locke *et al.*, 2015), while a companion paper characterized genetic loci associated with body fat distribution (Shungin *et al.*, 2015). Each of these papers includes >400 authors and thousands of collaborators.

Neafsey *et al.* (2015) sequenced and assembled the genomes and transcriptomes of 16 anopheline mosquitoes from three continents, spanning 100 million years of evolution and displaying genomic rates of change distinct from those of *Drosophila*.

These are examples of the thousands of genome projects that have been undertaken. We next examine different types of projects and the process by which genome sequences are assembled and annotated.

## GENOME ANALYSIS PROJECTS: INTRODUCTION

We have surveyed completed genome projects from a chronological point of view. We consider three main aspects of genome sequencing: generating the sequence; assembly (gathering all the reads into a coherent model of the DNA sequence across the chromosomes); and annotation (identifying features such as genes, regulatory regions, and repetitive elements).

First there are many questions associated with genome sequencing. Which genomes are sequenced? How big are genomes? What types of sequencing experiments are performed? Even the goals of sequence analysis are evolving as we learn what questions to ask and what tools are available to address those questions. Four main types of genome analysis projects are outlined in **Table 15.5**.

1. *De novo sequencing* involves determining the DNA sequence of an organism, as described chronologically in the sections above. While many more *de novo* genome sequencing projects are underway, two recently developed, specialized categories are the sequencing of ancient DNA (often from extinct organisms) and metagenomics (sampling the genomes of many organisms from a particular environmental site such as the human gut or an ocean region).

**TABLE 15.5 Applications of genome sequencing.**

Purpose	Template	Example
<i>De novo</i> sequencing	Genome sequencing	Sequencing >1000 influenza genomes
	Ancient DNA	Extinct Neandertal genome
	Metagenomics	Human gut
Resequencing	Whole genomes	Individual humans
	Genomic regions	Assessment of genomic rearrangements or disease-associated regions
Transcriptome	Somatic mutations	Sequencing mutations in cancer
	Full-length transcripts	Defining regulated messenger RNA transcripts
	Noncoding RNAs	Identifying and quantifying microRNAs in samples
Epigenetics	Methylation changes	Measuring methylation changes in cancer

2. *Resequencing a genome* permits the variation between individuals to be assessed. For example, the sequences of James Watson (a co-discoverer of the double helical nature of DNA) and J. Craig Venter (a pioneer in genome sequencing) were determined by 2008. Currently (in early 2015) it is thought that over 100,000 human exomes have been sequenced, and one major sequencing center (at the Broad Institute; see “Genome-Sequencing Centers” below) sequences an entire human genome every 30 minutes. Applications of resequencing include the assessment of genomic changes in disease-associated regions, the sequencing of all human exons in multiple individuals, or the sequencing of large sets of genes associated with cancer.
3. *RNA-seq* (Chapter 11) uses next-generation sequencing technology to measure mRNA transcript levels as well as other types of RNA.
4. *Epigenetics* refers to heritable changes other than those involving the four DNA sequences *per se*. Such epigenetic changes include the modification of DNA or chromatin through DNA methylation (the addition of methyl groups to cytosine residues in CpG dinucleotides) and/or through the post-translational modification of histones. High-throughput sequencing can be used to assess the methylation status of a genome.

## Large-Scale Genomics Projects

The completion of the human genome project coupled with the dramatic emergence of next-generation sequencing has led to several large-scale genomics projects.

- Human-centered projects include the 1000 Genomes Project, HapMap, and ENCODE (see Chapters 8 and 20) as well as disease-focused projects (**Table 15.6**).
- The Human Microbiome Project is characterizing the bacteria and archaea that inhabit the human body (Human Microbiome Project Consortium, 2012a, b).
- The Genomic Encyclopedia of Bacteria and Archaea (GEBA) is characterizing the diversity of those domains of life.
- For nonhuman species, projects are underway for sequencing large numbers of strains, inbred lines, or related organisms (**Table 15.7**). The Million Mutation Project involves sequencing 2000 mutagenized *C. elegans* strains, yielding nearly a million single-nucleotide variants and indels (Thompson *et al.*, 2013).

The HMP website is <http://www.hmpdacc.org> (WebLink 15.24), and is discussed further in Chapter 17.

Visit GEBA at <http://www.jgi.doe.gov/programs/GEBA> (WebLink 15.25).

**TABLE 15.6 Large-scale human sequencing projects (on-going and proposed).**

Project	URL	Goal
The 1000 Genomes Project	<a href="http://www.1000genomes.org/">http://www.1000genomes.org/</a>	Find human genetic variants having frequencies >1%
International Cancer Genome Consortium (ICGC)	<a href="http://www.icgc.org/">http://www.icgc.org/</a>	Catalog mutations in tumors from 50 cancer types
UK10K	<a href="http://www.sanger.ac.uk/about/press/2010/100624-uk10k.html">http://www.sanger.ac.uk/about/press/2010/100624-uk10k.html</a>	Sequence the genomes of 10,000 UK individuals
100,000 Genomes Project	<a href="http://www.genomicsengland.co.uk/">http://www.genomicsengland.co.uk/</a>	Sequence 100,000 individuals in the UK
Autism Genome 10K Project	<a href="http://autismgenome10k.org/">http://autismgenome10k.org/</a>	Sequence 10,000 autism-related genomes
Personal Genome Project	<a href="http://www.personalgenomes.org/">http://www.personalgenomes.org/</a>	Effort to sequence 100,000 human genomes

**TABLE 15.7 Large-scale model organism sequencing projects (on-going and proposed).**

Project	URL	Goal
1001 Genomes Project	✉ <a href="http://www.1001genomes.org/">http://www.1001genomes.org/</a>	Find whole-genome sequence variation in 1001 strains of <i>Arabidopsis thaliana</i>
Genome 10K project	✉ <a href="https://genome10k.soe.ucsc.edu/">https://genome10k.soe.ucsc.edu/</a>	Assemble sequences from 10,000 vertebrate species
Drosophila Genetic Reference Panel	✉ <a href="http://dgrp2.gnets.ncsu.edu/">http://dgrp2.gnets.ncsu.edu/</a>	Sequence the genomes of 192 inbred lines from <i>Drosophila</i>
1000 Fungal Genomes Project	✉ <a href="http://1000.fungalgenomes.org/home/">http://1000.fungalgenomes.org/home/</a>	Sequence 1000 fungal genomes
Mouse Genomes Project	✉ <a href="http://www.sanger.ac.uk/resources/mouse/genomes/">http://www.sanger.ac.uk/resources/mouse/genomes/</a>	Sequence 17 mouse strains
Million Mutation Project	✉ <a href="http://genome.sfu.ca/mmp/">http://genome.sfu.ca/mmp/</a>	<i>C. elegans</i>

The NHGRI large-scale genome sequencing program is described at ✉ <http://www.genome.gov/10001691> (WebLink 15.26). A list of white papers and sequencing targets is available online at ✉ <http://www.genome.gov/10002154> (WebLink 15.27).

## Criteria for Selection of Genomes for Sequencing

The choice of which genome to sequence depends on several main factors. The selection criteria change over time as technological advances reduce costs and as genome-sequencing centers gain experience in this new endeavor. One set of criteria is offered by the NHGRI at the National Institutes of Health. These include projects to study comparative genome evolution, to survey human structural variation, to annotate the human genome, and to perform medical sequencing. The process of selecting sequencing targets includes the submission of proposals (“white papers,” available on the NHGRI website) as well as working groups that help to set priorities.

### Genome Size

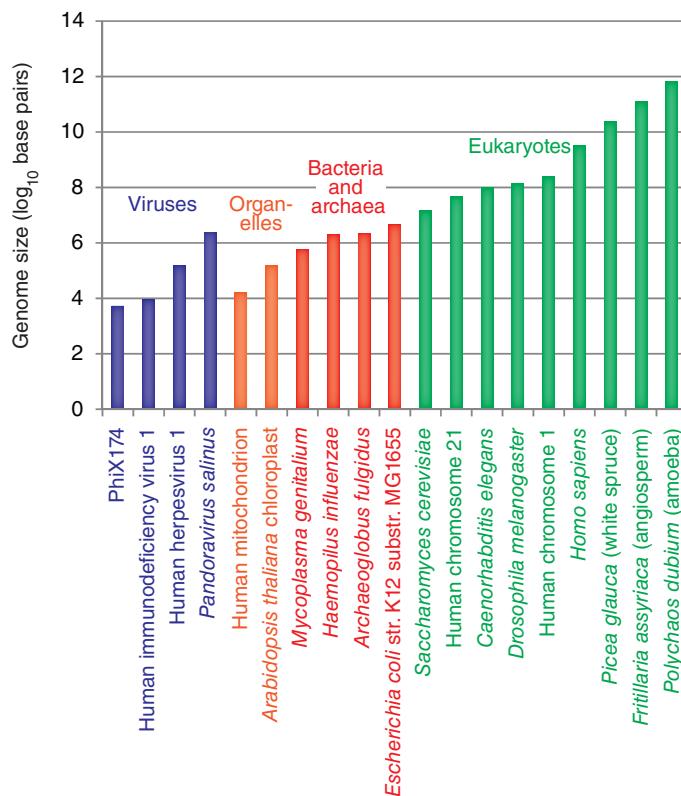
For a microbial genome, the size is typically several megabases (millions of base pairs), and a single lab can now have the resources to complete the entire project. For example, a small MiSeq (Illumina) benchtop sequencer can produce many billions of base pairs (gigabases) in a single run. For larger (typically eukaryotic) genomes, international collaborations are often established to share the effort.

A graphical overview of the sizes of various genomes is presented in **Figure 15.7**. Viral genomes range from 1 to an astounding 2.5 megabases (Chapter 16). In haploid genomes such as bacteria (Chapter 17), the genome size (or *C* value) is the total amount of DNA in the genome. Most bacterial genomes range from about 500,000 bp (*M. genitalium*; in Chapter 17 we describe a few even smaller bacterial genomes) to ~15 Mb (currently, the largest sequenced bacterial genome *Sorangium cellulosum* is 14.8 Mb).

In diploid or polyploid organisms, the genome size is the amount of DNA in the unreplicated haploid genome (such as the sperm cell nucleus). Among eukaryotes, there is about a 75,000-fold range in genome sizes from 8 Mb for some fungi to 686 Gb (gigabases) for some amoebae. The so-called *C* value paradox is that some organisms with extremely large *C* values are morphologically simple and appear to have a modest number of protein-coding genes. We explored this paradox in Chapter 8.

### Cost

The cost of sequencing has declined dramatically in recent years (**Fig. 9.3**). The total worldwide cost of producing a draft sequence of the human genome by a public



**FIGURE 15.7** Comparison of the sizes of various genomes. Virus genomes range from <10,000 base pairs to >2.5 megabases. Organellar genomes, derived from ancient bacterial endosymbionts, are reduced in size relative to present-day bacteria. Bacterial and archaeal genomes are commonly 2–5 megabases, with small genome sizes including 580,000 base pairs (*M. genitalium*, with 470 protein-coding genes) or less; larger bacterial genomes (e.g., cyanobacteria) exceed 13 Mb. For eukaryotic genomes, the range is from the 8 Mb for some fungi to 686 Gb for some amoebae. This has been called the *C* value paradox (see Chapter 8). The *C* value is the total amount of DNA in the genome, and the paradox is the relation between complexity of a eukaryote and its amount of genomic DNA.

consortium was about US\$ 300 million (or US\$ 3 billion including development costs). In contrast, completion of a draft sequence of another primate, the rhesus macaque, cost US\$ 22 million in 2006. By 2008 the cost of sequencing a human genome by Sanger technology was approximately US\$ 1–10 million, although Venter's sequence cost ~US\$ 70 million. A stated goal of the NHGRI is to promote the development of technology to reduce the cost of sequencing a human genome to US\$ 1000. This is close to the current cost of whole-exome sequencing. In 2014 Illumina introduced a sequencing machine (the HiSeq X Ten) that large sequencing centers can purchase and operate for >US\$ 10 million that, when operated at capacity, produces each human genome at a cost of under US\$ 1000.

Read about the NHGRI Genome Technology Program at <http://www.genome.gov/10000368> (WebLink 15.28).

#### Relevance to Human Disease

All genome projects have yielded information about how an organism causes disease and/or is susceptible to disease. For example, by sequencing the chimpanzee genome, we may learn why these animals are not susceptible to diseases that afflict humans, such as malaria and AIDS. We discuss genomics aspects of human disease in Chapter 21 and we consider the disease relevance of all parts of the tree of life in Chapters 16–19.

### ***Relevance to Basic Biological Questions***

Each genome is unique and its analysis enables basic questions about evolution and genome organization to be addressed. As an example, the chicken provides a nonmammalian vertebrate system that is widely used in the study of development. The analysis of protozoan genomes can illustrate the evolutionary history of the eukaryotes.

### ***Relevance to Agriculture***

Analyses of the chicken, cow, and honeybee genome sequences are expected to benefit agriculture in a variety of ways, such as leading to strategies to protect these organisms from disease. By 2050, 90% of the world's population will live in developing countries where agriculture is the most important activity. Raven *et al.* (2006) therefore suggest this should guide the choice of genome projects towards those that may benefit resource-poor farmers.

### ***Sequencing of One Versus Many Individuals from a Species***

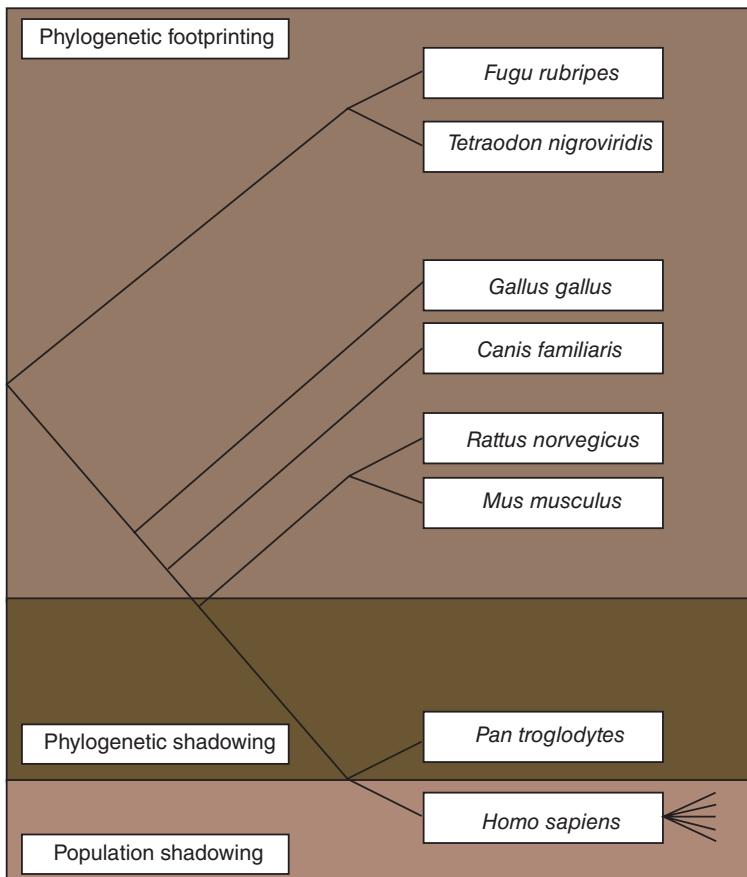
It is important to determine the entire genomic sequence from multiple individuals of a species in order to define variation and to correlate the genotype with the phenotype. In the case of humans (Chapter 20), the International HapMap Project and 1000 Genomes Project involved the genotyping and sequencing of genomic DNA from individuals from many geographic (ethnic) backgrounds, both male and female. Many genome projects of eukaryotes feature deep sequencing coverage of one individual and lighter coverage of genomes from additional individuals.

For viruses such as human immunodeficiency virus (HIV-1 and HIV-2), the virus rapidly undergoes enormous numbers of DNA changes, making it necessary to sequence many thousands of independent isolates (Chapter 16). This is practical to achieve because the genome is extremely small (<10 kb). In many cases, comparison of different bacterial strains reveals why one is harmless to humans while another is highly pathogenic. Such comparisons have been performed for harmless strains of *E. coli* that normally inhabit the human gut and other strains that cause severe, sometimes fatal, disease (Chapter 17).

## **Role of Comparative Genomics**

Comparative genomics involves the comparison of genome sequences from multiple species, or in some cases from individuals within a species. Miller *et al.* (2004) have reviewed this discipline and described how genome comparisons have aided the annotation of genomes (discussed in “Genome Analysis Projects: Annotation” below), particularly for the prediction of genes and conserved regulatory elements. They also discuss the impact on evolutionary analysis and function: through comparative analyses we can define DNA segments that are under positive or negative selection (Chapter 7).

The use of whole-genome comparisons at various evolutionary distances provides a powerful technique for applying many genomic analyses (Fig. 15.8, adapted from Miller *et al.*, 2004; see also Alfoldi and Lindblad-Toh, 2013). Phylogenetic footprinting refers to comparisons of genomic sequences from distantly related organisms, such as humans relative to fish, chicken, dog, and rodents. This is especially useful to identify conserved elements (under negative selection), emphasizing the relatively rare coding and noncoding segments of the genome that remain shared even after hundreds of millions of years since species such as human and fish diverged. Phylogenetic shadowing permits comparisons of more closely related species such as humans and chimpanzees that diverged about 6 MYA. These comparisons between closely related species allow the identification of regions that are different between the two, such as genes under positive selection. Population shadowing refers to sampling multiple genomes from one species (as discussed above for resequencing the human genome from many individuals). We adopt a comparative genomic approach throughout our exploration of the tree of life in Chapters 16–20.



**FIGURE 15.8** Comparative genomics allows the comparison of a genome (such as human) to other genomes of varying evolutionary distance. In phylogenetic footprinting, this includes genomes from organisms that diverged a relatively long time ago, such as fish (*Fugu rubripes*, *Tetraodon nigroviridis*) that diverged from the human lineage >400 MYA), chicken (*Gallus gallus*), dog (*Canis familiaris*), rat and mouse (*Rattus norvegicus* and *Mus musculus*) that diverged from the human lineage ~90–100 MYA). In phylogenetic shadowing, more closely related genomes are compared (e.g., the chimpanzee *Pan troglodytes*). In population shadowing, multiple genomes from one species are compared, permitting analyses of genotype–phenotype correlations. Redrawn from Miller *et al.* (2004). Reproduced with permission from Annual Reviews.

## Resequencing Projects

In studying genomic variation between individual humans, one approach is to resequence the entire human genome (reviewed in Bentley, 2006). This has been accomplished for perhaps nearly 100,000 individuals. One goal of such an endeavor is to use genomic information to guide medical decisions. As an alternative strategy it may be cost-effective to resequence portions of the genome that are of particular interest, such as globin loci in patients with thalassemia. Another approach is to sequence all human exons, since this focuses on protein-coding regions rather than the ~98% of the genome composed of noncoding regions (including introns, intergenic regions, and large expanses of repetitive DNA).

## Ancient DNA Projects

The study of ancient DNA presents a fascinating glimpse into the history of life on Earth. It is now possible to isolate genomic and/or mitochondrial DNA from museum specimens, fossils, and other sources of organisms that are now extinct. Svante Pääbo is a

pioneer in this field, in which researchers must address special challenges (Pääbo *et al.*, 2004; Willerslev and Cooper, 2005; Dabney *et al.*, 2013; Shapiro and Hofreiter, 2014):

- Ancient DNA is often degraded by nucleases. The size fragments of ancient DNA are therefore often small (100–500 base pairs) and the nucleotides are often damaged by strand breaks (induced by microorganisms or endogenous nucleases), oxidation (resulting in fragmentation of bases and/or deoxyribose groups), cross-linking of nucleotides, or deamination. There are many strategies available to address these issues, including performing multiple independent PCR or sequencing reactions from ancient DNA extracts. C to T and G to A substitutions are particularly prevalent, as shown for example in studies of 11 European cave bears (Hofreiter *et al.*, 2001).
- DNA isolated from ancient samples derives from unrelated organisms such as bacteria that invaded the specimen after death.
- DNA isolated from ancient specimens is easily contaminated by modern human DNA. Extraordinary measures must be taken to minimize laboratory or other sources of human contamination.
- A large number of criteria must be applied to demonstrate authenticity of ancient DNA samples. These include the use of appropriate control extracts and negative controls; analysis of multiple extracts independently isolated from each specimen; quantitation of the number of amplifiable molecules; and inverse correlation between amplification efficiency and the length of amplification which is expected to occur because of the fragmented nature of ancient DNA.

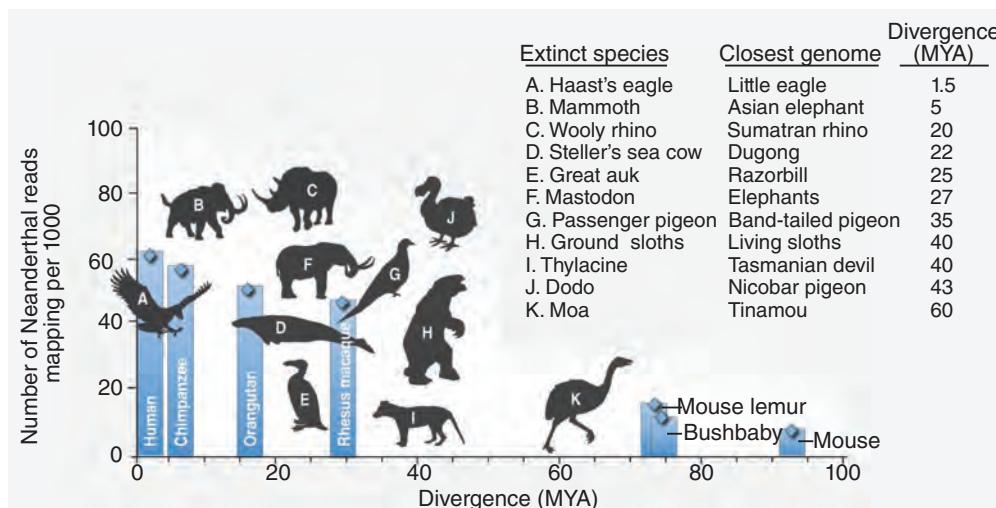
Despite the considerable technical challenges, dramatic progress has been made in the field of ancient DNA analysis. An example is the Neandertals, hominids that thrived from about 400,000 years ago until 30,000 years ago, and who represent the closest-known relative of modern humans. Mitochondrial DNA has been extracted and sequenced from over one dozen Neandertal fossils. Green *et al.* (2006, 2008) isolated genomic DNA from the 38,000-year-old fossil of a Neandertal bone found in modern Croatia. They recovered a complete Neandertal mitochondrial genome sequence and dated the divergence of Neandertal and modern human lineages at  $660,000 \pm 140,000$  years ago.

In 2010 Pääbo and colleagues reported a draft genome sequence of a Neandertal genome (Green *et al.*, 2010), followed by another (Prüfer *et al.*, 2014). Surprisingly, comparisons to present-day humans indicated that individuals of European and Asian (but not African) descent share ~3% of their genomes with Neandertals, possibly due to interbreeding between Neandertals and Eurasians who had migrated out of Africa. The Denisovans (named based on a cave in southern Siberia) are an extinct relative of the Neandertals who also admixed with the lineage, leading to present-day humans (Meyer *et al.*, 2012). Between 4% and 6% of the genomic DNA of present-day Melanesians has a Denisovan origin (Reich *et al.*, 2010).

Other ancient DNA projects include the sequencing of mitochondrial genomes from the moa (flightless birds from New Zealand; Cooper *et al.*, 2001) and woolly mammoth (Krause *et al.*, 2006), and from hair shafts of the Siberian mammoth *Mammuthus primigenius* (Gilbert *et al.*, 2007). For all these projects, the availability of a closely related extant genome greatly facilitates the assembly and annotation efforts for the extinct genome (Fig. 15.9). A list of DNA sequences available from extinct organisms is available at the NCBI Taxonomy website. While ancient DNA can be extracted, ancient RNA and proteins have not been extracted. As a notable exception, Schweitzer *et al.* (2007) found evidence of collagen in the extracellular matrix of bone from a *Tyrannosaurus rex* fossil based on immunohistochemistry (with antisera developed against avian collagen) and mass spectrometry.

Tracks at UCSC are available for the Neandertal genome (<http://genome.ucsc.edu/> Neandertal, WebLink 15.29) and for a Denisovan genome.

To see DNA entries from extinct organisms, visit the Taxonomy home at <http://www.ncbi.nlm.nih.gov/taxonomy> (WebLink 15.30) then follow the link to extinct organisms. Currently (October 2014) there are DNA (or protein) data available from 67 mammals, 47 birds, assorted plants, lizards, insects, amphibian, and two dinosaurs, *Brachylophosaurus canadensis* and *Tyrannosaurus rex*.



**FIGURE 15.9** Relationship between evolutionary distance and the usefulness of extant taxons as references genomes for genome assembly from extinct organisms. Eleven extinct genomes are depicted, each of which could be sequenced (based on sample availability). The closest extant genome for each extinct species is given, as well as the divergence time. Divergence times (MYA) are plotted on the x axis and read mapping is on the y axis. As the evolutionary distance decreases, the number of mappable reads to the reference genome increases (blue bars). Redrawn from Shapiro and Hofreiter (2014). Used with permission.

## Metagenomics Projects

The great majority of organisms on the planet are viruses and bacteria. Of these various organisms, most (probably >99%) are not cultivatable, making them extremely difficult to study. Metagenomics is the functional and sequence-based analysis of microorganisms that occur in an environmental sample (Riesenfeld *et al.*, 2004; Hunter *et al.*, 2012). Genomic sequencing efforts have been directed to a variety of environmental samples. We discuss these for viruses in Chapter 16 and for bacteria and archaea in Chapter 17, including the microbiome of organisms inhabiting the human body (Human Microbiome Project Consortium, 2012a, b).

Metagenomics projects may be grouped into two broad areas: environmental (also called ecological) and organismal. Environmental projects address the genomic community in an ecological site such as a hot spring, an ocean, sludge, or soil. As an example of an environmental project, Robert Edwards and colleagues (2006) obtained sequence data from two neighboring sites of an iron-rich mine in Minnesota. The samples were characterized by unexpectedly distinct sets of bacterial microorganisms, based principally on the analysis of 16S ribosomal DNA sequences.

Organismal metagenomics projects include such sites as human or mouse gut, feces, or lung. For example, it is estimated that the human intestinal tract contains on the order of  $10^{13}$  to  $10^{14}$  microorganisms (Gill *et al.*, 2006).

One primary source of information on metagenomics projects is NCBI, including BioProject (Barrett *et al.*, 2012). BioProject centralizes information about datasets, organizing and classifying project data submitted to NCBI, EBI, and DDBJ databases. The related BioSample database stores descriptions of biological materials including a range of types from cell lines to biopsies to environmental isolates.

Another primary source of information on metagenomics is the Genomes On Line Database (GOLD; Lioylios *et al.*, 2008), which we explore in Chapter 17.

To browse metagenomics projects at NCBI visit BioProject at <http://www.ncbi.nlm.nih.gov/bioproject/> (WebLink 15.31), browse by project attributes, and select metagenome for the project data type. There are currently ~900 projects, typically linking to GOLD.

The GOLD database is available at <http://www.genomesonline.org> (WebLink 15.32), listing ~500 studies and >4600 samples.

A list of genome-sequencing centers is offered at the NCBI (<http://www.ncbi.nlm.nih.gov/genomes/static/lcenters.html>, WebLink 15.33). URLs of the largest centers are <http://genome.jgi.doe.gov/> (JGI, WebLink 15.34), <http://www.jcvi.org/> (J. Craig Venter Institute, WebLink 15.35), and <http://www.broad.mit.edu/> (the Broad Institute, WebLink 15.36).

The Trace Archive is at <http://www.ncbi.nlm.nih.gov/Traces/> (WebLink 15.37). A specialized Trace Archive BLAST server is available at that site or from the NCBI BLAST home page.

An example of the accession for a Trace Archive record from an HBB search is gnl|ti|981051509.

The query Perl script is available from <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=faq> (WebLink 15.38), and is also provided at Web Document 15.4.

## GENOME ANALYSIS PROJECTS: SEQUENCING

### Genome-Sequencing Centers

Large-scale sequencing projects are conducted at centers around the world. Twenty sequencing centers contributed to the production of a draft version of the human genome in 2001 (see Chapter 20). These centers were also supported by the NIH and the EBI. All of these centers have also been involved in sequencing the genomes of other organisms. Currently, the five largest genome sequencing centers account for over half the sequencing that is being performed (Fig. 15.10).

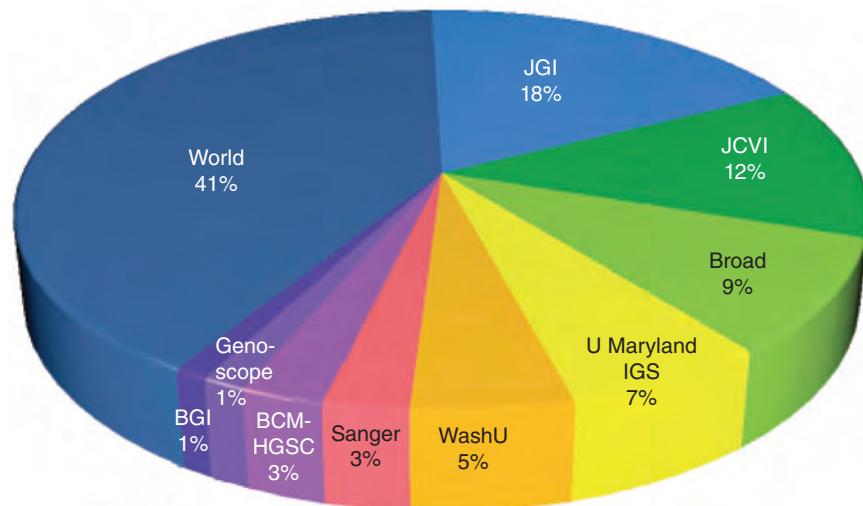
### Trace Archive: Repository for Genome Sequence Data

Raw sequence data for the genome-sequencing projects of several organisms have been deposited in the Trace Archive located at NCBI. All entries in this archive are given a Trace Identifier (Ti) number. The archive can be searched by several criteria (such as query by Ti or sequencing center or by BLAST).

Search the mouse Trace Archive (human WGS division) with our familiar human beta globin mRNA sequence (NM\_000518.4) and the output contains several Ti matches (as shown in Fig. 9.2). By clicking on the link to a Ti record, the sequence data can be obtained in the FASTA format or as a trace of the dye termination reaction used to sequence the DNA.

We can also search the Trace Archive on the command line using a Perl script. Navigate to your home directory and use `mkdir trace` to make a new directory. There, create a text document using an editor such as `vim`, `emacs`, or `nano`. Include the following script.

```
#!/usr/bin/perl -w
use strict;
use LWP::UserAgent;
use HTTP::Request::Common 'POST' ;
```



**FIGURE 15.10** Major genome sequencing centers. JGI: Joint Genome Institute; JCVI: J. Craig Venter Institute; Broad: Broad Institute; University of Maryland-IGS: Institute for Genome Sciences; WashU: Washington University in St Louis; Sanger: Wellcome Trust Sanger Institute; BCM-HGSC: Baylor College of Medicine, Human Genome Sequencing Center; BGI: Beijing Genomics Institute. Data are from >11,000 genomic and metagenomic projects in the GOLD database, 2011. Redrawn from Pagani *et al.* (2012). Reproduced with permission from Oxford University Press.

```
$ENV{ 'LANG' }='C';
$ENV{ 'LC_ALL' }='C';

my $query = join ' ', @ARGV;
$query = 'help' if $query =~ /^(\-h|\-\-help|\-)$/;
$query = join('', <STDIN>) if ! $query;

my $req = POST 'http://trace.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=raw',
[query=>$query];
my $res = LWP::UserAgent->new->request($req, sub { print $_[0] });
die "Couldn't connect to TRACE server\n" if ! $res->is_success;
```

Use `ls -lh` to list the file(s) in your directory, with `h` making them “human readable.” The permissions include one character that is `d` for directory or otherwise `-` as in in this case; `rwx` for read/write/executable in three groups describe permissions first for the user, then for group members, and then for others.

```
$ ls -lh
total 8
-rw-r--r- 1 pevsner 1357801299 642B Apr 5 08:27 query_tracedb
```

This pattern shows that you as the user can read or write to this file, but you cannot execute it. You can then make this script executable using the `chmod` utility:

```
$ chmod +x query_tracedb
$ ls -lh
total 8
-rwxr-xr-x 1 pevsner 1357801299 642B Apr 5 08:27 query_tracedb
```

Note how the permissions have changed and the script is executable. To see the help document, enter:

```
$ ./query_tracedb usage
```

You can also copy the executable Perl script to your `~/bin` directory so you can invoke the script from any directory (you will not need the `./` prefix):

```
$ cp query_tracedb ~/bin/
```

How many Trace Archive records are there for several species?

```
$ query_tracedb "query count species_code='homo sapiens'"
273924157
```

There are therefore ~274 million Trace Archive records for human; similar searches show ~208 million for the mouse *Mus musculus*, 52 million for *Rattus norvegicus*, 47 for the chimpanzee *Pan troglodytes*, and 16 million for the yellow fever mosquito *Aedes aegypti*.

You can also use this script to retrieve data. For example, we can use the Trace Archive identifier to retrieve a clone overlying beta globin in the FASTA format (I show the first few of 783 bases in this record). Other retrieval options allow you to retrieve quality scores, mate pair data, xml information, and more.

```
$ query_tracedb "retrieve fasta 981051509"
>gnl|ti|981051509 name:17000177953277
TTTCGAATAATTAAATACATCATTGCAATGAAAATAATGTTTTTATTAGGCAGAACCTTGCTCA
AGGCCCTCATATAATATCCCCAGTTAGTAGTTGGACTAGGAAACAAAGGAACCTTAATAGAAATTGG
```

In an innovative approach to using these raw data, Salzberg *et al.* (2005) studied the genomic DNA records from *Drosophila ananassae*, *D. simulans*, and *D. mojavensis* and

searched for matches to bacterial species that might colonize these fruit flies. They identified three new species of the bacterial endosymbiont *Wolbachia pipiensis* and were able to assemble sequences that covered substantial portions of the genomes.

### HTGS Archive: Repository for Unfinished Genome Sequence Data

Visit the HTG Sequences division of NCBI at <http://www.ncbi.nlm.nih.gov/genbank/htgs> (WebLink 15.39).

For examples of phases 1, 2, and 3 sequences in GenBank, see <http://www.ncbi.nlm.nih.gov/HTGS/examples.html> (WebLink 15.40).

We have seen that DNA sequence data are deposited in databases such as the Trace Archive at NCBI. At NCBI unfinished, raw genomic DNA data are made available through the high-throughput genomic (HTG) sequence division. Accession numbers are assigned to each entry. The HTG database contains sequence data in four phases.

Phase 0 data are typically sequences derived from a single cosmid or bacterial artificial chromosome (BAC). They are likely to have sequencing errors and gaps of indeterminate size. However, the data may still have tremendous usefulness to the scientific community even in this form. For example, if you are performing BLAST searches and are looking for novel homologs to your query, the HTG division may contain useful information.

Phase 1 data may consist of sequencing reads from contigs derived from a larger clone (e.g., a BAC clone) in which the order of the contigs is unknown and their orientation (top strand or bottom strand) is also unknown. The sequence is defined as unfinished, and still contains gaps.

In the finished state (phase 2), the contigs are ordered and oriented properly and the error rate must be  $10^{-4}$  or less. Finally, phase 3 data are transferred from HTG to a primary division. These sequences are finished and have no gaps.

## GENOME ANALYSIS PROJECTS: ASSEMBLY

It is remarkable to reflect on the size of the human genome – each chromosome is from ~50 Mb to ~250 Mb in length – and the fact that next-generation sequencing strategies used to sequence human and other genomes often produce reads that are only a few hundred nucleotides in length. Assembly is the process by which the reads are stitched together to build a comprehensive model of the sequence of a chromosome.

We introduced genome assembly strategies in Chapter 9 with respect to DNA sequence analysis, including the overlap/layout/consensus approach and de Bruijn graphs (Flicek and Birney, 2009). We next describe general strategies for sequencing and assembly.

NCBI offers information on genome assembly at <http://www.ncbi.nlm.nih.gov/assembly/basics/> (WebLink 15.41). We follow that document in this section.

### Four Approaches to Genome Assembly

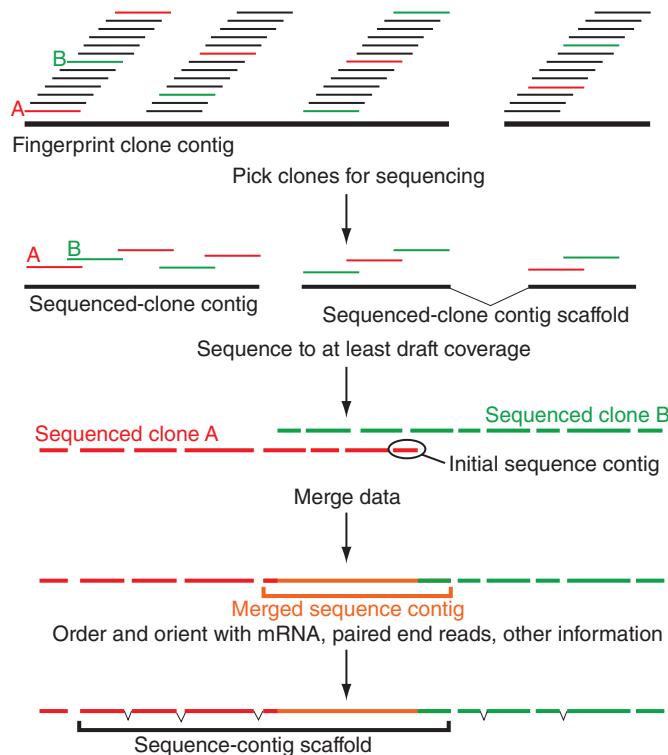
There are four main approaches to genome assembly outlined at NCBI. **Table 5.8** introduces some of the terminology associated with genome sequencing and assembly.

1. Hierarchical assembly (or clone-based assembly) relies on mapping large-insert clones such as bacterial artificial chromosomes (BACs) or fosmids. These clones are created by digesting genomic DNA then subcloning fragments into vectors, creating libraries with large inserts (e.g., 100–500 kb). Alternatively, smaller cosmid libraries (with insert sizes of about 50 kb) or plasmid libraries (2–10 kb inserts) are generated. This hierarchical strategy employs clones that are mapped to known chromosomal locations. Sequence assembly is therefore focused on a small region of the genome of known chromosomal location. Each large clone is fragmented, sequenced, and assembled into overlapping consensus sequences contigs. As these contigs become ordered and oriented they are further built into scaffolds. This approach has been taken for many large, eukaryotic genomes, including the public consortium's version of the Human Genome Project (International Human Genome Sequence Consortium,

**TABLE 15.8 Terminology used in genome-sequencing projects. Adapted from** <http://www.ncbi.nlm.nih.gov/genome/guide/build.html>, <http://www.ncbi.nlm.nih.gov/projects/genome/glossary.shtml>, and <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/info/definitions.shtml>.

Term	Definition
Alternate locus	A sequence that provides an alternate representation of a locus found in a largely haploid assembly. These sequences don't represent a complete chromosome sequence, although there is no hard limit on the size of the alternate locus; currently these are less than 1 Mb.
Assembly	A set of chromosomes, unlocalized and unplaced (random) sequences and alternate loci used to represent an organism's genome. Most current assemblies are a haploid representation of an organism's genome, although some loci may be represented more than once (see Alternate locus). This representation may be obtained from a single individual (e.g., chimp or mouse) or multiple individuals (e.g., human reference assembly). Except in the case of organisms which have been bred to homozygosity, the haploid assembly does not typically represent a single haplotype, but rather a mixture of haplotypes. As sequencing technology evolves, it is anticipated that diploid sequences representing an individual's genome will become available.
BAC end sequence	The ends of a bacterial artificial chromosome (BAC) have been sequenced and submitted to GenBank; the internal BAC sequence may not be available. When both end sequences from the same BAC are available, this information can be used to order contigs into scaffolds.
Contig	A set of overlapping clones or sequences from which a sequence can be obtained. NCBI contig records represent contiguous sequences constructed from many clone sequences. These records may include draft and finished sequences and may contain sequence gaps (within a clone) or gaps between clones when the gap is spanned by another clone which is not sequenced.
Draft sequence	At least three- to four-fold of the estimated clone insert is covered in Phred Q20 bases in the shotgun sequencing stage, as defined for the human genome sequencing project. Note that the exact definition of "draft" may be different for other genome projects. Clone sequence may contain several pieces of the sequence separated by gaps. The true order and orientation of these pieces may not be known.
Finished sequence	The clone insert is contiguously sequenced with a high-quality standard of error rate of 0.01%. There are usually no gaps in the sequence.
Fragment	A contiguous stretch of a sequence within a clone sequence that does not contain a gap, vector, or other contaminating sequence.
Meld	When two or more fragments overlap in the entire alignable region, these sequences are merged together to make a single longer sequence.
Order and orientation	Sequence overlap information is used to order and orient (ONO) fragments within a large clone sequence.
Scaffold	Ordered and oriented set of contigs placed on the chromosome. A scaffold will contain gaps, but there is typically some evidence to support the contig order, orientation, and gap size estimates.

- 2001). **Figure 15.11** shows a figure from that paper describing the process. Clone sequences may be unfinished, and deposited in HTGS until finished.
2. Whole-genome assembly (WGA) is the most commonly used strategy today. Notably clones are not mapped; instead, genomic DNA is fragmented, packaged into libraries, sequenced, and assembled. Frederick Sanger first applied this approach in the sequencing of bacteriophage  $\phi$ X174: Randomly selected fragments of genomic DNA were isolated, sequenced, and then assembled to derive a complete sequence. The



**FIGURE 15.11** Schematic of the hierarchical shotgun sequencing strategy. Genomic DNA is isolated from an organism of interest, fragmented, and inserted into a BAC library. Each BAC clone is 100–500 kb. BACs are ordered (mapped). Individual BAC clones are fragmented into smaller cDNA clones and sequenced. Individual sequencing reactions are typically 300–700 nucleotides. These “shotgun sequences” are assembled. Adapted from IHGSC (2001, p. 863). Reproduced with permission from Macmillan Publishers.

application of this approach to an entire organismal genome was pioneered by Hamilton O. Smith of Johns Hopkins and J. Craig Venter of the J. Craig Venter Institute who used this strategy to sequence *H. influenzae* (Fleischmann *et al.*, 1995).

The whole-genome assembly method was initially used for most small genomes (i.e., viruses, bacteria and archaea, and eukaryotic genomes that lack large portions of repetitive DNA). Genomic DNA is isolated from an organism and mechanically sheared (or digested with restriction enzymes). The fragments are subcloned into small-insert libraries (e.g., 2 kb fragments) and large-insert libraries (e.g., 10–20 kb). Clones are sequenced from both ends (i.e., both “top” strand and “bottom” strand), and the sequences are assembled. A typical sequencing reaction generates about 500–800 bp of sequence data. These small amounts of sequence are assembled into contiguous transcripts (“contigs”) and then into a map of the complete genome.

The WGA approach requires the computationally difficult task of fitting contigs together, regardless of which chromosomal region they are derived from. It was thought by some that this approach could not be practically applied to large eukaryotic genomes. However, it was successfully applied to the 120 Mb *D. melanogaster* genome (Adams *et al.*, 2000) in combination with a hierarchical approach and to the human genome (Weber and Myers, 1997; Venter *et al.*, 2001). The WGS data are processed at GenBank but are not distributed with GenBank releases. Instead, beginning with GenBank release 129 in 2002, WGS entries have been available from GenBank on a per-project basis (and are searchable by BLAST). Release 206 (February 2015) contains  $\sim$ 870 billion base pairs ( $8.7 \times 10^{11}$  bp), surpassing the  $\sim$ 187 billion base pairs in the corresponding traditional GenBank release (Fig. 2.3).

Whole-genome shotgun (WGS) contigs are deposited in the WGS division of GenBank (<http://www.ncbi.nlm.nih.gov/genbank/wgs/>, WebLink 15.42). Regions of heterochromatin contain large segments of highly repetitive DNA (Chapter 8) and, in some cases, cannot be effectively sequenced using WGS or hierarchical approaches. Skaletsky *et al.* (2003) applied an alternative technique of iterative mapping and sequencing to determine the extremely repetitive sequence of the human Y chromosome.

3. Hybrid methods combine whole-genome and hierarchical assembly. For example, sequencing of the cattle genome used a combination of BAC and whole-genome shotgun sequences (Bovine Genome Sequencing and Analysis Consortium *et al.*, 2009).
4. Comparative assembly uses a finished reference genome from a relatively closely related species to guide assembly. Here assemblers use an alignment-consensus algorithm rather than overlap-layout-consensus (see Chapter 9).

## Genome Assembly: From FASTQ to Contigs with Velvet

The assembly process involves the collection of individual sequences, the closing of gaps, and the lowering of the error rate. This process can be performed using a variety of software packages, such as Phrap (and its graphical viewer, Consed), Assembler, and Sequencher. For either the whole-genome sequencing or the hierarchical approach, after the shotgun phase is complete the next step is to assemble contigs. This is accomplished in a process called finishing. The goal of finishing is to identify gaps in the tile path and to close them. Ideally, this process results in a single contiguous DNA sequence that spans all the contigs.

To illustrate genome assembly we use Velvet software and analyze sequence data from a pathogenic *E. coli* strain. We follow an excellent tutorial by Edwards and Holt (2013). First let's obtain *E. coli* sequences for assembly. We select *E. coli* O14:H4 strain TY-2482 sequence reads, obtained from the European Nucleotide Archive (ENA). Enter a query for SRR292770 and save the FASTQ files to your local computer. The ENA entry provides additional information such as the sequencing machine (Illumina HiSeq 2000) and experimental details.

While you can proceed in Unix, Mac, or PC operating systems, we will proceed on a Mac using the terminal. We navigate to our home directory, and create a new folder called `assemblytutorial`.

```
$ cd ~ # This navigates to the home directory
$ mkdir assemblytutorial # This creates a new directory
# Next, the mv utility moves our downloaded FASTQ files into the newly
# created directory
$ mv ~/Downloads/SRR292770_1.fastq ~/assemblytutorial/
$ mv ~/Downloads/SRR292770_2.fastq ~/assemblytutorial/
$ head -4 SRR292770_1.fastq # We display the first four rows
@SRR292770_1 FCB0671ABXX:4:1101:1155:2103/1
GGAGTCATCATACGGCGCTGATCAGACCGCAACGACTTAAAGGTCGCA
+
FFFFCFGDCGGFCGBGFFFAEGFG;B7A@GEFBFGFFGFEFCFFF
```

Phred and Phrap are available at <http://www.phrap.org/> (WebLink 15.43) and operate on UNIX-based systems. Many other assembly software programs are available, including Arachne from the Broad Institute (<http://www.broadinstitute.org/science/programs/genome-biology/computational-rd/computational-research-and-development>) (WebLink 15.44).

The ENA website is <http://www.ebi.ac.uk/ena/> (WebLink 15.45) The FASTQ files are also available as Web Documents 15.5 and 15.6 (250 MB each, corresponding to forward and reverse reads).

How many entries are there? We can use the word count utility `wc`, and that shows us that each file has ~20 million rows; a FASTQ file has four rows, however. If we use `grep` to extract the rows that include the symbol `@SRR` (found in every FASTQ record for this dataset), we see that there are ~5.1 million entries. The `-c` modifier produces a count of the entries in each file. Note that using `grep -c` to search for and count the pattern `@` (rather than the pattern `@SRR`) gives us >6.8 million entries. That is the number of FASTQ entries plus all the instances of `@` as a symbol for a base quality score; it is not the expression we want for counting entries.

```
$ wc -l SRR*
20408164 SRR292770_1.fastq # Forward reads
20408164 SRR292770_2.fastq # Reverse reads
40816328 total
$ grep -c '@SRR' SRR292770_1.fastq
5102041
$ grep -c '@' SRR292770_1.fastq
6886214
```

Visit the FastQC home page at the Babraham Institute, <http://www.bioinformatics.babraham.ac.uk/projects/download.html> (WebLink 15.46). FastQC is a java application that requires Java Runtime Environment (JRE). To determine whether your computer has JRE installed, at the command prompt type `$ java -version`.

The Velvet homepage is <https://www.ebi.ac.uk/~zerbino/velvet/> (WebLink 15.47), linking to a GitHub repository for software downloads.

Next we can examine the quality of the reads with FastQC. We introduced FastQC in Chapter 9 as a tool that can be used in Galaxy or on the command line to assess the quality of FASTQ files. You can also download FastQC for the Linux, Windows, or Mac platforms. We can open FastQC, select a FASTQ file, and obtain quality reports.

The next step is to assemble reads into contigs. We do this using the Velvet program (Zerbino and Birney, 2008; Zerbino, 2010), and we continue to follow the Edwards and Holt (2013) tutorial in an abbreviated form. We download Velvet, move its directory to within our `assemblytutorial` directory, compile it with the `make` command, and add it to the `~/bin` directory so that it can be deployed from any directory location.

We first use `velveth` to specify that we want to create a hash table of the reads from the paired FASTQ files, using a  $k$ -mer length of 29. The results are stored in a folder called `out_data_29`.

```
$ velveth out_data_29 29 -fastq -shortPaired -separate SRR292770_1.fastq  
SRR292770_2.fastq
```

The output of `velveth` includes a Roadmap file and a Sequences file which are required for `velvetg`, a program that creates and manipulates a de Bruijn graph (Chapter 9).

```
$ velvetg out_data_29 -clean yes -exp_cov 21 -cov_cutoff 2.81 -min_contig_lgth 200
```

Here `-exp_cov` is the expected coverage of unique regions; this can be useful to exclude highly covered reads (e.g., abundant mitochondrial sequences) from an assembly. `-clean` refers to removing intermediary files that are not needed. `-min_contig_lgth` is the minimum contig length exported to `contigs.fa`; the default is the hash length  $\times$  2, but here it is set to 200 nucleotides.

The main output of Velvet is a set of contigs that have been assembled. Let's look at the first five lines of the `contigs.fa` FASTA file; it contains sequences of the contigs that are longer than  $2k$  (where  $k$  is the word length used in `velveth`). There may be N residues for gaps between scaffolded contigs, although in this example there are no Ns.

```
$ head -5 contigs.fa  
>NODE_1_length_17146_cov_33.514290  
ATAAGACGCGCAAGCGTCGCATCAGGCAACACCACGTATGGATAGAGATCGTGAGTACAT  
TAGAACAAACAATAGGCAATACGCCCTGGTGAAGTTGCAGCGAATGGGGCGGATAACG  
GCACTGAAGTGTGGTAAACTGGAAGGCAATAACCCGGCAGGTTCGGTGAAAGATCGTG  
CGGCACCTTCGATGATCGTCGAGCGGAAAGCGCGGGATTAAACCGGGTGTATGTCT
```

`velvetg` also creates a `stats.txt` file describing the nodes of the assembly.

```
$ head -5 stats.txt  
ID lgth out in long_cov short1_cov short1_Ocov short2_cov short2_Ocov  
long_nb short1_nb short2_nb  
1 17146 1 1 0.000000 33.514289 33.507640 0.000000 0.000000 0 15303 0  
2 31995 1 1 0.000000 33.554680 33.535396 0.000000 0.000000 0 28629 0  
3 7935 1 1 0.000000 32.280403 32.253560 0.000000 0.000000 0 7050 0  
4 72906 1 1 0.000000 32.900516 32.889899 0.000000 0.000000 0 64526 0
```

## Comparative Genome Assembly: Mapping Contigs to Known Genomes

Genomes can be assembled *de novo* (“anew,” without referring to other completed genomes) or by mapping reads onto a reference genome. We continue to use examples from the tutorial by Edwards and Holt (2013), and now use Mauve software (Darling *et al.*, 2010, 2011) to map contigs to known genomes.

Mauve computes and visualizes multiple genome alignments. It searches for conserved segments between two genomes that are called “locally collinear blocks” (LCBs). Its strategy is anchored alignment (Chapter 6) using inexact, ungapped matches with the seed-and-extend method of PatternHunter introduced by Ma *et al.* (Fig. 5.13). Progressive Mauve further uses three different seed patterns. It then performs progressive alignment.

We first demonstrate the alignment of two complete genomes by Mauve. We select *E. coli* strain K12 substr. MG1655 (a standard reference strain; Fig. 15.12a, upper genome) and a related *Shigella flexneri* genome (Fig. 15.12a, lower genome). Locally collinear blocks are indicated by colored blocks, and inverted regions are indicated below the genome’s center line (Fig. 15.12a, magenta stars). There are 42 LCBs in this example, and by adjusting the LCB weight the sensitivity can be adjusted (e.g., by increasing the weight the number of LCB declines, reducing the number of true positive as well as spurious rearrangements).

Next we use FTP to download several thousand *E. coli* contigs from NCBI. These sequences were obtained by sequencing a stool sample from a patient with hemolytic uremic syndrome during an *E. coli* outbreak in Germany in 2011. We align the unplaced whole-genome sequencing contigs to the genome of another sequenced *E. coli* strain. When we open Mauve, we use the option under “Tools” to “Move Contigs.” Select an output folder, select a reference genome, then select your contigs in the FASTA format. Mauve produces a graphical output (Fig. 15.12b) and a set of output files including a FASTA formatted file with the ordered and oriented contigs.

### Finishing: When Has a Genome Been Fully Sequenced?

The redundancy of genomic coverage is a function of the number of reads, the average read length, and the length of the region (e.g., genome) being sequenced. We described this in Chapter 9 (see Equations (9.4) and (9.5); Table 9.5).

### Genome Assembly: Measures of Success

There are several main approaches to quantifying the success of an assembly.

- The coverage estimate is relatively high. The need for coverage varies based on the sequencing technology (e.g., Sanger technology generating 750 base pair reads requires less coverage than a next-generation sequencing technology producing several hundred base pair reads).
- The N50 value is the length of contigs which contain half the bases in a given assembly. An assembly is more complete with longer N50 values. If an average human gene is ~50 kb then a contig N50 of that size will have half its contigs spanning the length of a gene.
- The scaffold N50 is also a measure of assembly completeness.
- As the assembly becomes more complete, the absolute number of contigs and scaffolds becomes smaller.
- The assembly is subjected to some form of annotation (discussed in “Genome Analysis Projects: Annotation” below), typically including a catalog of protein-coding gene models. The extent to which an assembly spans ESTs and cDNAs is a measure of completeness.
- When appropriate, the extent to which gene models overlap a core set of vertebrate genes is measured (see “CEGMA” below).

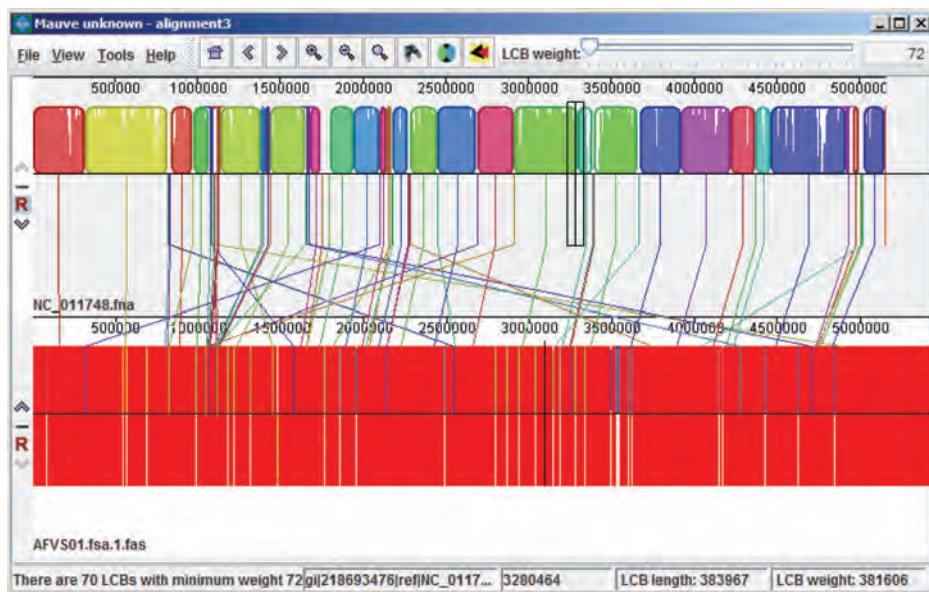
### Genome Assembly: Challenges

Errors in assembly are important because we rely on each assembly for all aspects of the genomic landscape, including the locations of genes. We can illustrate some of the challenges in genome assembly using the example of the cattle genome.

The Mauve homepage is <http://gel.ahabs.wisc.edu/mauve/> (WebLink 15.48). It is available for PC, Mac, or Unix platforms. You may need to also install Java Runtime Environment (JRE) for Mauve to function.

You can download the *E. coli* sequence in the FASTA format from <ftp://ftp.ncbi.nlm.nih.gov/genomes/> (WebLink 15.49). Browse to “Bacteria” then select the genomes of interest. We’ll choose *E. coli* strain K12 substr. MG1655 (NC\_000913.fna to obtain it in the FASTA format) and *Shigella flexneri* 2a 301 (NC\_004337.fna; selected as an example from the Mauve user’s guide). See [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia\\_coli\\_55989\\_uid59383/NC\\_011748.fna](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_55989_uid59383/NC_011748.fna) (WebLink 15.50).

We use files suggested by Edwards and Holt (2013). The reference genome is *E. coli* Ec55989 obtained from the NCBI Genomes FTP site. The unplaced contigs are from *E. coli* O104:H4, downloaded from <http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AFVS01> (WebLink 15.51).

(a) Comparison of *E. coli* strain K12 substr. MG1655 (upper) and *Shigella flexneri* (lower)(b) Alignment of *E. coli* Ec55989 (upper) and a set of *E. coli* O104:H4 contigs (lower)

**FIGURE 15.12** Genome comparisons with Mauve software. (a) Two (or more) genomes can be aligned by Mauve, enabling visualization of conserved syntenic loci (colored blocks) and inversions (blocks near magenta stars). *E. coli* strain K12 substr. MG1655 (upper portion) is aligned to *Shigella flexneri* (lower portion). (b) When genome sequences are assembled into contigs, these contigs can then be mapped to a completed reference genome. Here unordered contigs from *E. coli* O104:H4 are mapped to a completed *E. coli* Ec55989 genome.

Source: Based on software described by Darling *et al.* (2010).

The Bovine Genome Sequencing and Annotation Consortium *et al.* (2009) reported the genome sequence of taurine cattle. The cattle lineage diverged from the human lineage ~97 MYA and emerged ~60 MYA as the suborder Ruminantia. Humans began domesticating cattle 8000 to 10,000 years ago. The genome sequencing involved bacterial artificial chromosomes (BACs) and whole-genome shotgun (WGS) sequencing. The

contig N50 was ~49 kb and the scaffold N50 was 1.9 Mb, and this consortium provided detailed genome annotation.

That same year Zimin *et al.* (2009) reported significant errors in the consortium assembly (called BCM4). Ten of the 30 cattle chromosomes had large (>500 kb) inversions, deletions, or translocations. Zimin *et al.* obtained raw sequence reads from the Trace Archive and produced a different assembly (UMD2) that was substantially more accurate and complete. For example, UMD2 placed 136 Mb of sequence on the X chromosome, while the BCM4 assembly placed only 83 Mb. The differences included the strategies employed by the assembly software and the helpful reliance by UMD2 on conserved synteny with the human genome assembly. Florea *et al.* (2011) subsequently produced an even more accurate assembly (UMD3) having fewer contigs, larger contig N50 values, fewer scaffolds, larger scaffold N50s, and fewer gaps.

The effects of improved assembly can be seen in better annotation of protein-coding genes and SNPs. A conclusion from these studies is that assembly remains challenging and should not be viewed as producing a static, final determination of the genome sequence. Instead it is a process that requires re-evaluation and improvement, whether by incrementally improving an existing assembly or by creating a *de novo* assembly.

As another example, Zhang *et al.* (2012) assessed a published rhesus macaque draft genome and reported that half the gene models are missing, incomplete, or incorrect. They suggest that this magnitude of error is common to any draft vertebrate genome that is subject to an automated gene annotation pipeline.

Compounding these deep concerns, most papers reporting draft genome sequences do not include detailed methods for assembly and annotation. This makes the challenge of improving these areas even greater.

## GENOME ANALYSIS PROJECTS: ANNOTATION

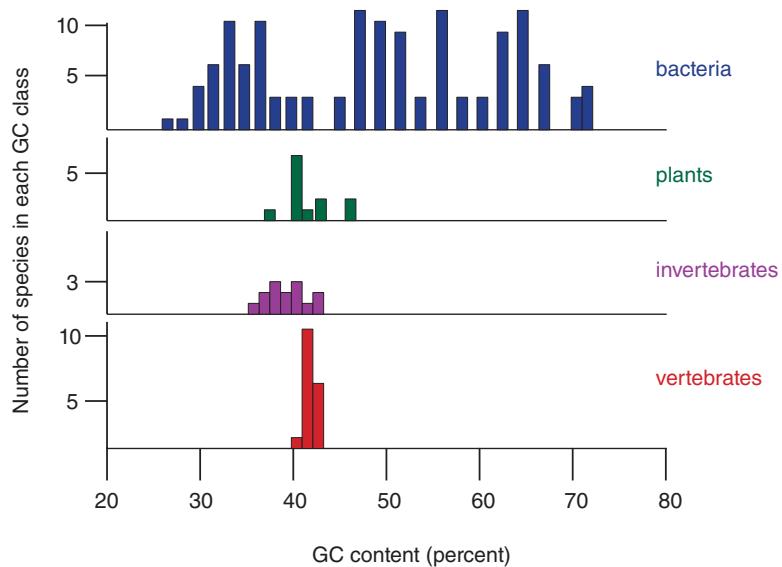
When a genome is sequenced, we learn its size and we obtain the complete (or nearly complete) nucleotide sequence as associated with particular chromosomes. Genome annotation is the process by which the landscape of genomic DNA is surveyed, and key features of the DNA are described (Yandell and Ence, 2012). These basic features of genomic DNA include the following:

- The number of chromosomes to which the genomic DNA is mapped is known for many species. In some cases the chromosome number is not yet known, and within some species the chromosome number and/or length varies greatly between isolates.
- The overall GC content or other nucleotide composition has been assessed since the pioneering work of Noboru Sueoka in the 1960s. Many eukaryotic genomes are characterized by a GC content of about 35–45%, while bacteria display a far wider range (**Fig. 15.13**).
- The repetitive elements of a genome can constitute well over 50% of the DNA. These can be identified and classified with software such as RepeatMasker, incorporated into many analysis pipelines and software tools.
- The identification of genes is a major concern of annotation efforts.

We showed examples of repetitive DNA, and the software used to identify and mask it, in Chapter 8.

A first approach to identifying protein-coding genes is by the alignment of expressed sequence tags (ESTs) to the genome. Transcripts that are expressed (i.e., RNA molecules) are converted to cDNA, incorporated into libraries, and sequenced. Such cDNAs are ESTs. While they do not inherently reveal information about the corresponding genomic DNA, such as the sequence of introns or the chromosomal locus, they are invaluable in identifying expressed genes (see “Annotation of Genes in Eukaryotes: Ensembl Pipeline” below).

A second “intrinsic” approach to predicting gene structures (exons and introns) is through analysis of genomic DNA, searching for features such as open reading frames,



**FIGURE 15.13** Guanine plus cytosine (GC) content of bacteria, plants, invertebrates, and vertebrates. Note that most eukaryotic genomes have 40–45% GC content, while bacteria and archaea have a far wider range. This figure is adapted from Bernardi and Bernardi (1990) based on studies in the 1960s to 1980s. Recent eukaryotic genome sequencing projects (described in Chapter 19) reveal that GC content for various organisms includes 19.4% (*P. falciparum*), 22.2% (the slime mold *Dictyostelium discoideum*), 34.9% (*A. thaliana*), 36% (*C. elegans*), 38.3% (*S. cerevisiae*), 41.1% (human), 42% (*M. musculus*), and 43.3% (*O. sativa*). For sequenced bacteria, GC content values range from 26% (*Ureaplasma urealyticum parvum*) to 72% (*Streptomyces coelicolor*). Adapted from Bernardi and Bernardi (1990), with permission from Springer Science and Business Media.

exon/intron boundaries, start and stop codons, and codon usage typical of coding regions. A third approach is comparative, mapping genes from one organism to conserved syntenic regions of a closely related organism whose genome has previously been sequenced.

The features of genomic DNA are substantially different between bacteria (and archaea) and eukaryotes. We consider them in more detail in Chapters 17 (on bacteria and archaea) and 18–19 (on eukaryotes).

The e!62 and e!63 human GRCh37 assembly pipeline is described at Ensembl (<http://www.ensembl.org/info/genome/genebuild/assembly.html>, WebLink 15.52) and is also available as Web Document 15.7. The estimates of how long each step requires are included in the documentation.

### Annotation of Genes in Eukaryotes: Ensembl Pipeline

The Ensembl website provides descriptions of its gene annotation pipeline for a variety of organisms (Curwen *et al.*, 2004; Potter *et al.*, 2004). For human gene annotation there are 12 steps.

1. In the raw computes stage (requiring 3 weeks), genomic sequence data are screened for sequence patterns with RepeatMasker (Chapter 8), Tandem Repeats Finder, and other software.
2. (7 weeks) Coding models are generated using evidence such as UniProt and RefSeq for proteins and ENA/GenBank/DDBJ and RefSeq for complementary DNAs.
3. (2 weeks) Additional coding models are generated based on database searches of mammalian (or other vertebrate) UniProt entries from other species. These analyses also include EST and cDNA evidence.
4. (2–3 weeks) cDNA and EST sequences are downloaded, poly(A)+ tails are clipped from the 3' ends, and they are aligned to the genome using Exonerate software. Alignments of cDNA required 98% nucleotide identity (and 97% identity with 90% coverage for ESTs which are often shorter and more fragmented than cDNAs).

5. (2 weeks) Coding models are manually filtered to remove dubious matches to proteins or cDNAs.
6. (2 weeks) Coding models are extended by adding untranslated regions.
7. (4–5 weeks) Redundant transcript models are collapsed, and unique sets of transcript models are clustered into multi-transcript genes. (Each transcript in a gene has at least one coding exon that overlaps a coding exon from another transcript within the same gene.)
8. (3 weeks) The gene set is screened for pseudogenes and retrotransposed genes. Immunoglobulin genes are annotated using a specialized workflow.
9. (10 weeks) The completed Ensembl gene set is finalized by merging (at the transcript level) manual annotations from the Vega database. Long intergenic noncoding RNA genes (lncRNAs) are annotated. As a quality control step, Ensembl protein-coding transcripts are translated and these proteins are aligned against NCBI RefSeq and UniProt/Swiss-Prot protein sequences.
10. (4 weeks) Annotations are added including cross-references to external databases. Stable accessions are assigned to each gene, transcript, exon, and translation.
11. (1–2 weeks) Haplotype regions are annotated, particularly on chromosomes 6, 14, and 17.
12. (3–4 weeks) Post genebuild filtering is used to remove poorly supported models. This stage includes comparative genomics analyses.

This Ensembl annotation pipeline is applied to the human genome, and you can also find documentation for the Ensembl annotation of other genomes. Every gene model is supported by biological sequence evidence, and you can view this information in the “Supporting evidence” link on the sidebar of any gene or transcript page. Note the long number of weeks required for each of the 12 steps; creating a full assembly is complex and time-consuming. Builds (assemblies) for each organism are therefore released infrequently.

### Annotation of Genes in Eukaryotes: NCBI Pipeline

The NCBI eukaryotic genome annotation pipeline is conceptually similar to that of Ensembl. A chart shows the integration of data from assemblies including nucleotide and protein databases and the Sequence Read Archive (**Fig. 15.14**).

The NCBI workflow includes the alignment of transcripts, proteins, short reads, and RefSeq genomic sequences to the assembled genome.

- Masking employs RepeatMasker or WindowMasker software.
- Transcript alignment may include transcripts from other organisms; RefSeq transcripts; ESTs; and other sources. These are mapped to the genome sequence with Splign. Kapustin *et al.* (2008) benchmarked this against five related software tools, finding that Splign is accurate yet tolerant to sequencing errors and polymorphic sites.
- Short reads from RNA-seq are also aligned using Splign.
- Proteins are aligned using ProSplign.
- NCBI uses Gnomon for gene prediction. This combines homology searching with *ab initio* modeling.

Different annotation pipelines produce results that can differ greatly (e.g., see Rice Annotation Project *et al.*, 2008, which compares Gnomon to a separate workflow). As a result, many groups strive to validate annotation data.

### Core Eukaryotic Genes Mapping Approach (CEGMA)

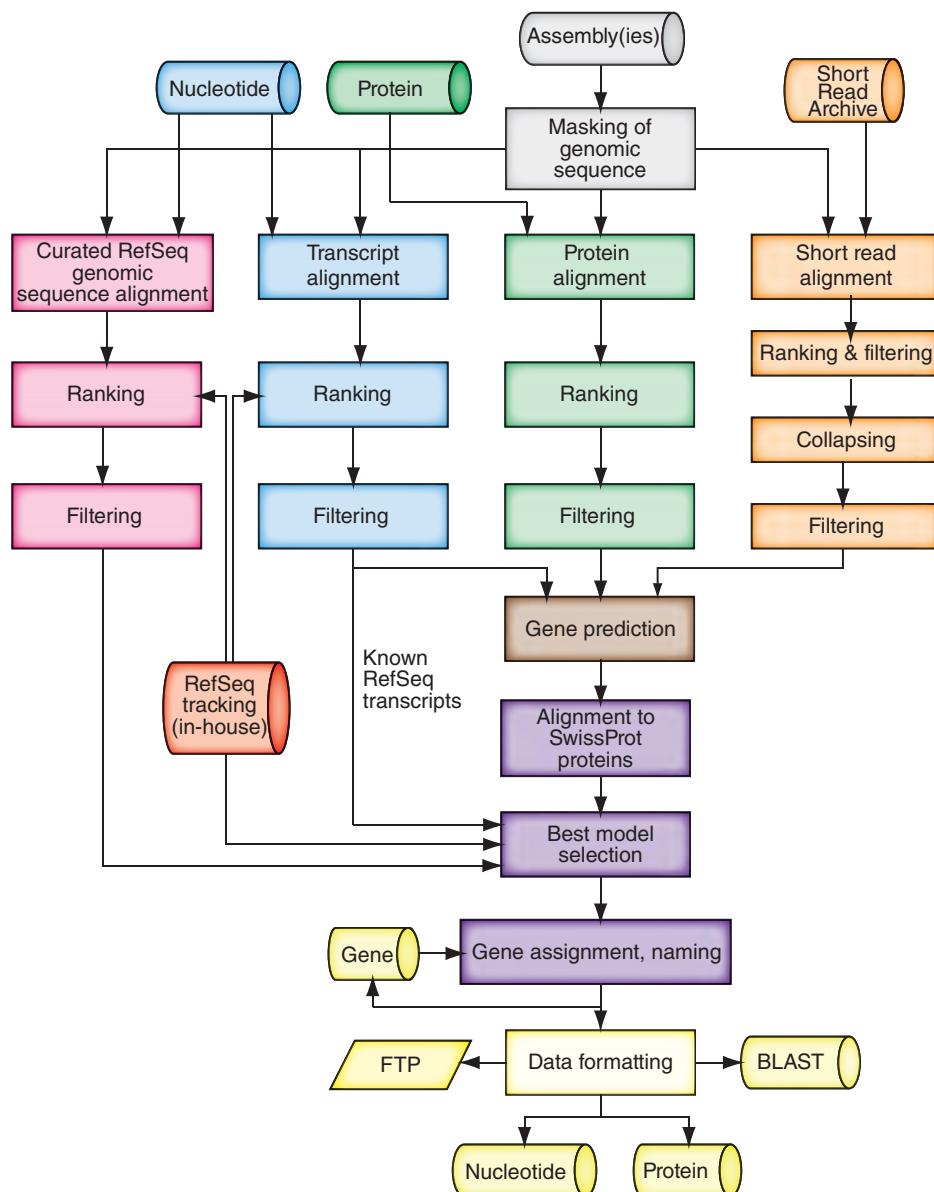
Eukaryotic genomes include sets of genes that are highly conserved across species, as first catalogued in detail by Margaret Dayhoff and colleagues (Chapter 3). Parra *et al.*

**Figure 15.14** is adapted from  
[http://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/process/](http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/) (WebLink 15.53). The NCBI annotation pipeline is also described in an online NCBI book at <http://www.ncbi.nlm.nih.gov/books/NBK169439/> (WebLink 15.54).

Splign is available as an online tool at <http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi> (WebLink 15.55). Try it using alpha 2 globin (*HBA2*) mRNA as a query (NM\_000517.4) against the human genome, and see how it aligns to the closely related *HBA1* locus.

Visit the ProSplign site at <http://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html> (WebLink 15.56).

You can learn more about Gnomon at <http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml> (WebLink 15.57).



**FIGURE 15.14** The eukaryotic genome annotation pipeline from NCBI. Genomic sequences are masked (gray). Transcripts (blue), proteins (green), and short reads (orange) are aligned to the genome, as well as RefSeq genomic sequences (when available; pink). Next, gene model prediction is performed (brown). Models are selected, named, and assigned accessions (purple). Annotated entries are formatted and made publicly available (yellow).

Source: Redrawn from NCBI.

(2007) (from the group of Ian Korf) introduced CEGMA to build a highly reliable set of gene annotations from sequenced eukaryotic genomes. They selected protein families from the eukaryotic orthologous groups (KOGs) project at NCBI, aligned them with T-COFFEE (Chapter 6), added quality control steps, and selected 458 protein groups (“core eukaryotic genes” that are conserved between *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*). The CEGMA method allows the exon-intron structures of the genes encoding these core proteins to be mapped to a novel genome sequence.

CEGMA has been useful in assessing the completeness of draft and finished genomes including *Anopheles gambiae*, *Ciona intestinalis*, and *Toxoplasma gondii*. When core genes are missing, this likely reflects false negative findings from an annotation pipeline.

CEGMA can be downloaded from  
 ↗ <http://korflab.ucdavis.edu/datasets/cegma/> (WebLink 15.58).

## Assemblies from the Genome Reference Consortium

The Genome Reference Consortium (GRC) provides assemblies for human, mouse, and zebrafish genomes. The initial sequencing and assembly of these genomes focused on identifying a single tiling path (sometimes called the “golden path” for the human genome) to represent the genome. The GRC now focuses on representing the regions of complex allelic diversity, such as the major histocompatibility locus (MHC) on human chromosome 6 or improving the annotation of the complex pericentric region of human chromosome 9.

## Assembly Hubs and Transfers at UCSC, Ensembl, and NCBI

The focus of UCSC Genome Browser is on vertebrate genomes, although many additional genomes are supported. An option is to create and view assembly hubs. These are available by clicking “track hubs” from a standard genome browser page.

Within the standard UCSC Genome Browser you can apply tracks (from the Mapping and Sequencing group) to display differences between assemblies (e.g., “Hg38 Diff”). You can also switch between assemblies with the View > “In other genomes (convert)” tool. The LiftOver tool also converts genome coordinates between assemblies.

Similarly, NCBI and Ensembl offer re-mapping services to navigate between assemblies for a variety of organisms. For example, we can take a BED file that was generated with GRCh37/hg19 coordinates of the human genome and convert it to GRCh38/hg38.

The GRC (↗ <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/>, WebLink 15.59), hosted at NCBI, consists of four groups: The Wellcome Trust Sanger Institute (↗ <http://www.sanger.ac.uk/research/areas/bioinformatics/grc>, WebLink 15.60), The Genome Institute at Washington University (↗ <http://genome.wustl.edu>, WebLink 15.61), The European Bioinformatics Institute (↗ <http://www.ebi.ac.uk>, WebLink 15.62), and the National Center for Biotechnology Information (↗ <http://www.ncbi.nlm.nih.gov>, WebLink 15.63).

## Annotation of Genes in Bacteria and Archaea

Bacterial and archaeal genomes have both genes and additional, relatively small intergenic regions. Typically, these genomes are circular, and there is about one gene in each kilobase of genomic DNA. For bacteria and archaea, genes are most simply identified by the presence of long open reading frames (ORFs) that are greater in length than some cutoff value such as 90 nucleotides (30 amino acids; a protein of about 3 kilodaltons). Programs such as GLIMMER and GenMark efficiently locate genes in bacterial genomic sequence. We describe this in Chapter 17, as well as RAST annotation software. As with eukaryotic genomes, there are challenges associated with annotating draft and finished microbial genomes (Mavromatis *et al.*, 2012).

UCSC assembly hubs are described at ↗ [http://genomewiki.ucsc.edu/index.php/Assembly\\_Hubs](http://genomewiki.ucsc.edu/index.php/Assembly_Hubs) (WebLink 15.64). After selecting an assembly hub you can access it as a group from the Browser Gateway.

## Genome Annotation Standards

An archival genome record reflects a particular state of a genome sequence and its annotation. To ensure high-quality annotation of genomes, investigators at several institutions proposed a series of standards (Klimke *et al.*, 2011) described as “Minimum Information about an Environmental Sequence” (MIENS; Yilmaz *et al.*, 2011).

Liftover at UCSC is available at  
 ↗ <http://genome.ucsc.edu/cgi-bin/hgLiftOver> (WebLink 15.65). The input is a BED file. LiftOver can also be downloaded as an executable for Linux systems.

1. A complete bacterial or archaeal genome should include ribosomal RNAs (at least one each of 5S, 16S, and 32S), tRNAs (at least one per amino acid), protein-coding genes at the expected density (based on precedence of similar genomes), and annotation of core genes.
2. Annotations should follow guidelines of the International Nucleotide Sequence Database Collaboration (INSDC; Chapter 2).
3. Methodologies and standard operating procedures should be documented.
4. Exceptions (unusual annotations such as atypical GC content) should be documented and given strong supporting evidence.

Visit the NCBI Genome Remapping Service at ↗ <http://www.ncbi.nlm.nih.gov/genome/tools/remap> (WebLink 15.66). An Assembly Converter is available from Ensembl (↗ [http://www.ensembl.org/Homo\\_sapiens/Tools/AssemblyConverter?db=core](http://www.ensembl.org/Homo_sapiens/Tools/AssemblyConverter?db=core), WebLink 15.67).

5. Pseudogenes should be annotated following accepted formats.
6. Enriched annotations should follow INSDC guidelines.
7. A set of databases, tools, and resources for annotation (given by Klimke *et al.*) should be used.
8. Validation checks should be performed using recently developed software.

Other large-scale sequencing centers (such as JCVI) also provide standard operating procedures for genome annotation (Tanenbaum *et al.*, 2010). Another example is the Metadata Coverage Index introduced by Liolios *et al.* (2012) as a metric to assess metadata availability and utility.

## PERSPECTIVE

In 1995 we entered an era in which the completed genome sequences could be determined. Thousands of complete genome sequences are now available. Since the completion of the human genome sequencing in the year 2003, some call the present state of biology the “postgenomic era.”

In recent years, the number of completed eukaryotic, archaeal, and bacterial genomes has continued to increase, with a particularly large number of genome projects that are currently in the assembly phase (near completion) or otherwise in progress. There are tens of thousands of such projects, excluding thousands of ongoing and completed viral and organellar genome projects. Several trends contribute to the rapid development of this field. (1) In sequencing a genome of interest, the availability of completed genomes of closely related organisms greatly aids the assembly and annotation process. For example, the assembly of the chimpanzee genome relied heavily on using the very closely related human reference genome as a template (Chapter 19). (2) Sequencing technologies have continuously improved; an entire bacterial genome can be sequenced in just several hours using technologies described in Chapter 9. (3) There has been progress in selecting, obtaining, and preparing genomic DNA from a spectacular range of biological sources. This has led to the creation of the new disciplines of the genomics of ancient, extinct organisms (such as the Neandertal and Denisovan genomes) to metagenomics projects that define the community of organisms living in sites such as the oceans or the human gut.

A major consequence of genome-sequencing projects is that molecular phylogeny has been revolutionized. The present version of the tree of life includes three main branches (bacteria, archaea, and eukaryotes). In the coming years, molecular data will help to clarify some of the key questions about life on Earth:

- How many species exist on the planet?
- How did life evolve from 4 BYA up to the present time?
- Why are some organisms pathogenic while close relatives are harmless?
- What mutations cause disease in humans and other organisms?

## PITFALLS

While the research community is generating massive amounts of DNA sequence data, there are many pitfalls associated with interpretation of those data. There is an error rate associated with genome sequences (typically less than one nucleotide per 10,000 in finished DNA). In evaluating possible polymorphisms or mutations in genomic DNA sequences, it is therefore important to assess the quality of the sequence data. Even if the sequence is correct, algorithms do not yet have complete success in problems such as finding protein-coding genes in eukaryotic DNA; there are many examples of genome-sequencing projects (such as cattle, *Drosophila*, rice and human) in which the predicted exons and gene models improve dramatically with each subsequent revision of the genome

assembly. (For bacterial genomes, which generally lack introns, the success rate is much higher.) Once protein-coding genes or other types of genes are identified, there are very large numbers of errors in genome annotation. It will be important to carefully assess the basis of functional annotation of genes; ultimately, the problem of gene function must be assessed by biological as well as computational criteria.

## ADVICE FOR STUDENTS

While Parts I and II of this book focused on bioinformatics, this third and final part covers the tree of life from a genomics perspective. Think about how the tools of bioinformatics inform the discipline of genomics. Try to get a sense of the broad scope of the tree of life, including the history of life on Earth. What species thrived 100 MYA, 1 billion years ago, or even earlier?

In this chapter we discussed genome assembly and annotation. Several studies have suggested that half the gene models in some assemblies are incorrect. Try to go further by becoming familiar with software for some stage of assembly and/or annotation. Read the published literature and the manual, download and use the software, and try to understand why assembly and annotation are so challenging.



## Discussion Questions

**[15.1]** What would a tree of life look like if it included all species (both extant and extinct) since the first life emerged to the present?

**[15.2]** If you could sequence the genomes of 100 individuals from any species, which species would you choose? What hypotheses would you test, how would you perform data analyses, and what resources would you require in terms of hardware, software, and collaborators? What ethical issues might arise in sequencing human genomes?

### PROBLEMS/COMPUTER LAB

**[15.1]** Figure 15.1 shows a tree of life based on rRNA sequences. Construct a tree of life based on glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein sequences. One approach is to identify this family in Pfam, which includes two different GAPDH domains. For the NAD binding domain (PF00044) there are currently 112 seed proteins and >14,000 full proteins. Export either some or all of the sequences in an aligned format. (Alternatively, perform a multiple sequence alignment using MUSCLE or Clustal Omega; see Chapter 6). Create and evaluate a neighbor-joining tree using MEGA (Chapter 7). How similar is your tree to that depicted in Figure 15.1? What might account for their differences?

**[15.2]** Obtain approximately 1000 bases of DNA sequence in the FASTA format from the bacterium *Escherichia coli* K12 (the accession number of the complete genome is

NC\_000913). Use this as a query in a BLASTN search of the Trace Archive at NCBI. Can you identify a eukaryotic sequencing project that includes bacterial DNA? For example, search against human whole-genome shotgun (WGS) sequences. How would you determine the total amount of bacterial DNA in any given eukaryotic entry in the Trace Archive?

**[15.3]** We have seen that some mitochondrial and chloroplast genomes are exceptionally large. Lilly and Harvey (2001) have described repetitive DNA in some plant organellar genomes. The *Zea mays* chloroplast genome (NC\_001666.2; Table 15.4) is 140 kb. What major repeats does it contain? Use MegaBLAST to search it against itself, then use RepeatMasker to characterize its repeats. Separately, use MegaBLAST to examine repeats within the *Saccharomyces cerevisiae* mitochondrial genome (NC\_001224.1, Table 15.3). How do you interpret the dot matrix view of the yeast organellar genome? One of its repeat units follows:

```

ATTATTATTATAGTAATAATAAAAATATTCTAAATATATTATATATTAT
TATTTTTTTATTATTAAT
AAAATATTATAATAAAATTAAATAAGTTATAATTGGATAAGTATTGTT
ATATTTTTATTCCAAT
ATATAACTCCCGTTCTTACGAAACCGGACCTCGGAGACGTAATAGGGG
GAGGGGGTGGGTGATAAGA
ACCAAACATTCAATAAAATATAGAGCACACATTAGTTAATATTAAATA
TAACTAATATATAATAATT
ATAAAATAATTAAATTATAATAATATAAAAGTCCCCGCCGGCGGGGA
CCCCAAAGGAGTATTAAACA
ATATAATATTGTATAAAATAATTATAAAATTAAATATTAAATAAAAAACCAAATA
AATAATATAATAATGATA
AACAAAGAAGATATCCGGTCCAATAATAATTATTATTGAAAATAATAATT
GGGACCCCCATCTAAAATA
TATATATAACTAATAATATATTATATAATATAATATAATAATATTATTA

```

```

AAATATAATATTATTTAAAAA
AAAAAGTATATATAAAAATAAGATATATATATAAATATATATTCTTAA
TAAATATTATATAATAAA
TAATAAATTATTCATAATAAAATTATTCTTTTATTAATAA

```

**[15.4]** Use the Splign tool to align an mRNA sequence (from human alpha 2 globin *HBA2*) against the human genome. *HBA2* is adjacent to and very closely related to *HBA1*; these genes encode proteins with 100% amino acid identity. How does the *HBA2* mRNA map to the human genome?

**[15.5]** Which RefSeq genomes include a gene named HBB? Use E-Direct (Chapter 2). This problem is adapted from <http://www.ncbi.nlm.nih.gov/books/NBK179288/>. Use the following code:

```

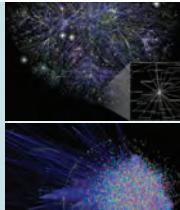
$ esearch -db nuccore -query "HBB [GENE] AND
NC_0:NC_99999999 [PACC]" | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element TaxId
| \
sort -n | uniq | \
epost -db taxonomy | \

```

```

efetch -format docsum | \
xtract -pattern DocumentSummary -element
ScientificName | \
sort
Borrelia afzelii HLJ01
Borrelia afzelii PKo
Borrelia burgdorferi B31
Borrelia burgdorferi ZS7
Bos taurus
Callithrix jacchus
Equus caballus
Felis catus
Gallus gallus
Gorilla gorilla gorilla
Homo sapiens
Macaca fascicularis
Macaca mulatta
Nomascus leucogenys
Oryctolagus cuniculus
Pan troglodytes
Papio anubis
Pongo abelii
Rattus norvegicus
Sus scrofa

```



## Self-Test Quiz

**[15.1]** The first complete genome to be sequenced was:

- (a) *Saccharomyces cerevisiae* chromosome III;
- (b) *Haemophilus influenzae*;
- (c) a bacteriophage; or
- (d) the human mitochondrial genome.

**[15.2]** A typical eukaryotic mitochondrial genome encodes about how many proteins (excluding RNAs)?

- (a) from 5 to 20;
- (b) from 50 to 100;
- (c) from 500 to 1000; or
- (d) 10,000.

**[15.3]** Thousands of genomes have now been completely sequenced. The majority of these are:

- (a) viral;
- (b) bacterial;
- (c) archaeal;
- (d) organellar (mitochondrial and plastid); or
- (e) eukaryotic

**[15.4]** Ancient DNA projects allow the sequencing of historical samples. A special challenge is:

- (a) the DNA is often fragmented;
- (b) the DNA is often contaminated by modern human DNA;
- (c) the DNA is often contaminated by bacterial DNA; or
- (d) all of the above.

**[15.5]** Velvet:

- (a) maps reads onto contigs;
- (b) assembles reads into contigs;
- (c) merges contigs into reads; or
- (d) reads contigs into assemblies.

**[15.6]** The term “whole-genome shotgun sequencing” refers to a strategy to sequence an entire genome by:

- (a) breaking up DNA and sequencing using oligonucleotide primers that span the genomic DNA;
- (b) breaking up DNA, cloning it into libraries, and sequencing using oligonucleotide primers that correspond to known chromosomal locations (contigs);

- (c) breaking up DNA, cloning it into libraries, hybridizing small fragments, then reassembling the fragments into a complete map; or
- (d) breaking up DNA, cloning it into libraries, sequencing small fragments, then reassembling the fragments into a complete map.

**[15.7]** The biggest problem in predicting protein-coding genes from genomic sequences using algorithms is that:

- (a) the software is difficult to use;
- (b) the false negative rate is high: many exons are missed;

- (c) the false positive rate is high: many exons are falsely assigned; or
- (d) the false positive rate is high: many exons have unknown function.

**[15.8]** For finished DNA sequence, the error rate must be:

- (a) 0.01 or less;
- (b) 0.001 or less;
- (c) 0.0001 or less; or
- (d) 0.00001 or less.

## SUGGESTED READING

The Green and Guyer (2011) paper on the future of genomics is highly recommended. For a review of ancient DNA, see Shapiro and Hofreiter (2014).

In this chapter we followed a guide to comparative bacterial genome analysis by David Edwards and Kathryn Holt (2013), including an accompanying tutorial. Yandell and Ence (2012) provide an overview of eukaryotic genome annotation.

## REFERENCES

- 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D. *et al.* 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073. PMID: 20981092.
- 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A. *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65. PMID: 23128226.
- Adams, M. D., Celtniker, S.E., Holt, R.A. *et al.* 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195. PMID: 10731132.
- Alföldi, J., Lindblad-Toh, K. 2013. Comparative genomics as a tool to understand evolution and disease. *Genome Research* **23**(7), 1063–1068. PMID: 23817047.
- Allwood, A.C., Walter, M.R., Kamber, B.S., Marshall, C.P., Burch, I.W. 2006. Stromatolite reef from the Early Archaean era of Australia. *Nature* **441**, 714–718.
- Anderson, S., Bankier, A.T., Barrell, B.G. *et al.* 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465. PMID: 7219534.
- Andersson, S. G., Zomorodipour, A., Andersson, J.O. *et al.* 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140. PMID: 9823893.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.
- Barns, S.M., Delwiche, C.F., Palmer, J.D., Pace, N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Science, USA* **93**(17), 9188–9193. PMID: 8799176.
- Barrett, T., Clark, K., Gevorgyan, R. *et al.* 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research* **40**(Database issue), D57–63. PMID: 22139929.
- Bentley, D.R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics and Development* **16**, 545–552.
- Benton, M. J., Ayala, F. J. 2003. Dating the tree of life. *Science* **300**, 1698–1700.

- Bernardi, G., Bernardi, G. 1990. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *Journal of Molecular Evolution* **31**, 282–293.
- Blackmore, S. 2002. Environment biodiversity update: progress in taxonomy. *Science* **298**, 365.
- Blattner, F. R., Plunkett, G. 3rd, Bloch, C.A. *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474. PMID: 9278503.
- Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L. *et al.* 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**(5926), 522–528. PMID: 19390049.
- Bult, C. J., White, O., Olsen, G.J. *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073. PMID: 8688087.
- C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018.
- Carbone, L., Harris, R.A., Gnerre, S. *et al.* 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**(7517), 195–201. PMID: 25209798.
- Carlton, J. M., Angiuoli, S.V., Suh, B.B. *et al.* 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512–519. PMID: 12368865.
- Chambaud, I., Heilig, R., Ferris, S. *et al.* 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Research* **29**, 2145–2153. PMID: 11353084.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055), 69–87. PMID: 16136131.
- Ciccarelli, F.D., Doerks, T., von Mering, C. *et al.* 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**(5765), 1283–1287. PMID: 16513982.
- Cole, J.R., Wang, Q., Fish, J.A. *et al.* 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**(Database issue), D633–642. PMID: 24288368.
- Cole, S. T., Eiglmeier, K., Parkhill, J. *et al.* 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011. PMID: 11234002.
- Cooper, A., Lalueza-Fox, C., Anderson, S. *et al.* 2001. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**, 704–707.
- Curwen, V., Eyras, E., Andrews, T.D. *et al.* 2004. The Ensembl automatic gene annotation system. *Genome Research* **14**, 942–950.
- Dabney, J., Meyer, M., Pääbo, S. 2013. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology* **5**(7), pii:a012567. PMID: 23729639.
- Dagan, T., Martin, W. 2006. The tree of one percent. *Genome Biology* **7**(10), 118. PMID: 17081279.
- Darling, A.E., Mau, B., Perna, N.T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**(6), e11147. PMID: 20593022.
- Darling, A.E., Tritt, A., Eisen, J.A., Facciotti, M.T. 2011. Mauve assembly metrics. *Bioinformatics* **27**(19), 2756–2757. PMID: 21810901.
- Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life*. J. Murray, London.
- Dohm, J.C., Minoche, A.E., Holtgräwe, D. *et al.* 2014. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* **505**(7484), 546–549. PMID: 24352233.
- Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129.
- Douglas, S., Zauner, S., Fraunholz, M. *et al.* 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091–1096. PMID: 11323671.
- Driskell, A.C., Ané, C., Burleigh, J.G. *et al.* 2004. Prospects for building the tree of life from large sequence databases. *Science* **306**, 1172–1174.
- Dunham, I., Shimizu, N., Roe, B.A. *et al.* 1999. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495. PMID: 10591208.
- Edwards, D.J., Holt, K.E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation* **3**(1), 2. PMID: 23575213.

- Edwards, R.A., Rodriguez-Brito, B., Wegley, L. *et al.* 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57.
- Eisen, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics and Development* **10**, 606–611.
- ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A. *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146), 799–816. PMID: 17571346.
- Field, D., Garrity, G., Gray, T. *et al.* 2008. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology* **26**(5), 541–547. PMID: 18464787.
- Field, D., Amaral-Zettler, L., Cochrane, G. *et al.* 2011. *The Genomic Standards Consortium. PLoS Biology* **9**(6), e1001088. PMID: 21713030.
- Fiers, W., Contreras, R., Duerinck, F. *et al.* 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507. PMID: 1264203.
- Fiers, W., Contreras, R., Haegemann, G. *et al.* 1978. Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113–120. PMID: 205802.
- Fleischmann, R. D., Adams, M. D., White, O. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512. PMID: 7542800.
- Flicek, P., Birney, E. 2009. Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**(11 Suppl), S6–S12. PMID: 19844229.
- Florea, L., Souvorov, A., Kalbfleisch, T.S., Salzberg, S.L. 2011. Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies. *PLoS One* **6**(6), e21400. PMID: 21731731.
- Fox, G. E., Stackebrandt, E., Hespell, R.B. *et al.* 1980. The phylogeny of prokaryotes. *Science* **209**, 457–463. PMID: 6771870.
- Fraser, C. M., Gocayne, J.D., White, O. *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403. PMID: 7569993
- Fu, Q., Li, H., Moorjani, P. *et al.* 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**(7523), 445–449. PMID: 25341783.
- Galibert, F., Finan, T.M., Long, S.R. *et al.* 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668–672. PMID: 11474104.
- Gilbert, M.T., Tomsho, L.P., Rendulic, S. *et al.* 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **317**, 1927–1930. PMID: 17901335.
- Gill, S.R., Pop, M., Deboy, R.T. *et al.* 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359.
- Goffeau, A., Barrell, B.G., Bussey, H. *et al.* 1996. Life with 6000 genes. *Science* **274**, 546, 563–577. PMID: 8849441.
- Graur, D., Li, W.-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Green, E.D., Guyer, M.S. 2011. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204–213. PMID: 21307933.
- Green, R.E., Krause, J., Ptak, S.E. *et al.* 2006. Analysis of one million base pairs of Neandertal DNA. *Nature* **444**, 330–336.
- Green, R.E., Malaspinas, A.S., Krause, J. *et al.* 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**(3), 416–426. PMID: 18692465.
- Green, R.E., Krause, J., Briggs, A.W. *et al.* 2010. A draft sequence of the Neandertal genome. *Science* **328**(5979), 710–722. PMID: 20448178.
- Grigoriev, I.V., Nordberg, H., Shabalov, I. *et al.* 2012. The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* **40**(Database issue), D26–32. PMID: 22110030.
- Gurdasani, D., Carstensen, T., Tekola-Ayalele, F. *et al.* 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**(7534), 327–332. PMID: 25470054.

- Haeckel, E. 1879. *The Evolution of Man: A Popular Exposition of the Principal Points of Human Ontogeny and Phylogeny*. D. Appleton and Company, New York.
- Hattori, M., Fujiyama, A., Taylor, T.D. *et al.* 2000. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405**, 311–319. PMID: 10830953.
- Hedges, S. B., Chen, H., Kumar, S. *et al.* 2001. Genomic timescale for the origin of eukaryotes. *BMC Evolutionary Biology* **1**, 1–10. PMID: 11580860.
- Hofreiter, M., Jaenicke, V., Serre, D., Haeseler, Av A., Pääbo, S. 2001. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research* **29**, 4793–4799.
- Holt, R. A., Subramanian, G.M., Halpern, A. *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149. PMID: 12364791.
- Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**(7114), 931–949. PMID: 17073008.
- Hugenholtz, P., Pace, N.R. 1996. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends in Biotechnology* **14**, 190–197.
- Human Microbiome Project Consortium. 2012a. A framework for human microbiome research. *Nature* **486**(7402), 215–221. PMID: 22699610.
- Human Microbiome Project Consortium. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402), 207–214. PMID: 22699609.
- Hunter, C.I., Mitchell, A., Jones, P. *et al.* 2012. Metagenomic analysis: the challenge of the data bonanza. *Briefings in Bioinformatics* **13**(6), 743–746. PMID: 22962339.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**(7018), 695–716. PMID: 15592404.
- International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**(6968), 789–796. PMID: 14685227.
- International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**(7063), 1299–1320. PMID: 16255080.
- International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A. *et al.* 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311), 52–58. PMID: 20811451.
- International Human Genome Sequence Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945. PMID: 15496913.
- Joyce, G. F. 2002. The antiquity of RNA-based evolution. *Nature* **418**, 214–221.
- Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 2012. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* **7**(6):e39315. PMID: 22720093.
- Kaiser, D. 2001. Building a multicellular organism. *Annual Review of Genetics* **35**, 103–123.
- Kapustin, Y., Souvorov, A., Tatusova, T., Lipman, D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct* **3**, 20. PMID: 18495041.
- Karolchik, D., Barber, G.P., Casper, J. *et al.* 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research* **42**, D764–770. PMID: 24270787.
- Kersey, P.J., Allen, J.E., Christensen, M. *et al.* 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research* **42**, D546–552. PMID: 24163254.
- Klenk, H. P., Clayton, R.A., Tomb, J.F. *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370. PMID: 9389475.
- Klimke, W., O'Donovan, C., White, O. *et al.* 2011. Solving the problem: genome annotation standards before the data deluge. *Standards in Genomic Science* **5**(1), 168–193. PMID: 22180819.

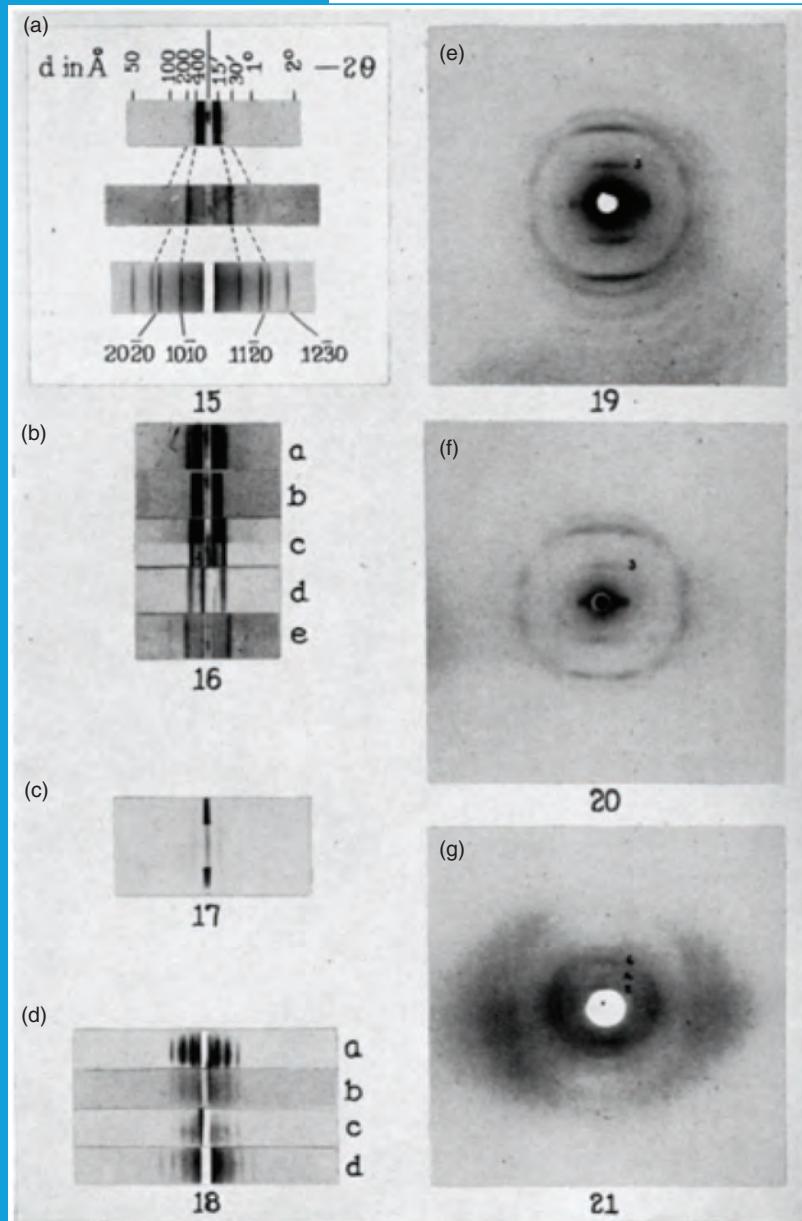
- Knapp, S., Bateman, R.M., Chalmers, N.R. *et al.* 2002. Taxonomy needs evolution, not revolution. *Nature* **419**, 559. PMID: 12374947.
- Koonin, E. V. 1997. Genome sequences: Genome sequence of a model prokaryote. *Current Biology* **7**, R656–659.
- Koonin, E.V. 2012. *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT Press, Upper Saddle River, New Jersey.
- Krause, J., Dear, P.H., Pollack, J.L. *et al.* 2006. Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* **439**, 724–727.
- Kua, C.S., Ruan, J., Harting, J. *et al.* 2012. Reference-free comparative genomics of 174 chloroplasts. *PLoS One* **7**(11), e48995. PMID: 23185288.
- Kumar, S., Hedges, S. B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920.
- Lang, B. F., Gray, M. W., Burger, G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annual Review of Genetics* **33**, 351–397.
- Letunic, I., Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128.
- Levy, S., Sutton, G., Ng, P.C. *et al.* 2007. The diploid genome sequence of an individual human. *PLoS Biology* **5**(10), e254 (2007). PMID: 17803354
- Ley, T.J., Mardis, E.R., Ding, L. *et al.* 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**(7218), 66–72. PMID: 18987736.
- Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S. *et al.* 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(7069), 803–819. PMID: 16341006.
- Liolios, K., Mavromatis, K., Tavernarakis, N., Kyripides, N.C. 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **36**(Database issue), D475–479.
- Liolios, K., Schriml, L., Hirschman, L. *et al.* 2012. The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness. *Standards in Genomic Science* **6**(3), 438–447. PMID: 23409217.
- Lister, R., Pelizzola, M., Dowen, R.H. *et al.* 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271), 315–322. PMID: 19829295.
- Locke, A.E., Kahali, B., Berndt, S.I. *et al.* 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**(7538), 197–206. PMID: 25673413.
- Locke, D.P., Hillier, L.W., Warren, W.C. *et al.* 2011. Comparative and demographic analysis of orangutan genomes. *Nature* **469**(7331), 529–533. PMID: 21270892.
- Mailman, M.D., Feolo, M., Jin, Y. *et al.* 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* **39**(10), 1181–1186. PMID: 17898773.
- Markowitz, V.M., Chen, I.M., Chu, K. *et al.* 2012. IMG/M-HMP: a metagenome comparative analysis system for the Human Microbiome Project. *PLoS One* **7**(7), e40151. PMID: 22792232.
- Markowitz, V.M., Chen, I.M., Chu, K. *et al.* 2014a. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research* **42**(Database issue), D568–573. PMID: 24136997.
- Markowitz, V.M., Chen, I.M., Palaniappan, K. *et al.* 2014b. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42**(Database issue), D560–567. PMID: 24165883.
- Martin, W.F. 2011. Early evolution without a tree of life. *Biology Direct* **6**, 36. PMID: 21714942.
- Mavromatis, K., Land, M.L., Brettin, T.S. *et al.* 2012. The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLoS One* **7**(12), e48837. PMID: 23251337.
- May, B. J. *et al.* 2001. Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proceedings of the National Academy of Science USA* **98**, 3460–3465.
- Mayr, E. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Belknap Harvard, Cambridge, MA.

- Meyer, M., Kircher, M., Gansauge, M.T. *et al.* 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**(6104), 222–226. PMID: 22936568.
- Miller, W., Makova, K.D., Nekrutenko, A., Hardison, R.C. 2004. Comparative genomics. *Annual Review of Genomics and Human Genetics* **5**, 15–56.
- Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G., Worm, B. 2011. How many species are there on Earth and in the ocean? *PLoS Biology* **9**(8), e1001127. PMID: 21886479.
- Myburg, A.A., Grattapaglia, D., Tuskan, G.A. *et al.* 2014. The genome of *Eucalyptus grandis*. *Nature* **510**(7505), 356–362. PMID: 24919147.
- Neafsey, D.E., Waterhouse, R.M., Abai, M.R. *et al.* 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347**(6217), 1258522. PMID: 25554792.
- Neale, D.B., Wegrzyn, J.L., Stevens, K.A. *et al.* 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**(3), R59. PMID: 24647006.
- Ohyama, K., Fukuzawa, H., Kohchi, T. *et al.* 1988. Structure and organization of *Marchantia polymorpha* chloroplast genome. I. Cloning and gene identification. *Journal of Molecular Biology* **203**(2), 281–298. PMID: 2462054.
- Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone M.L. *et al.* 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**, 38–46. PMID: 1574125.
- O’Malley, M.A., Koonin, E.V. 2011. How stands the Tree of Life a century and a half after The Origin? *Biology Direct* **6**, 32. PMID: 21714936.
- Orlando, L., Ginolhac, A., Zhang, G. *et al.* 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**(7456), 74–78. PMID: 23803765.
- Pääbo, S., Poinar, H., Serre, D. *et al.* 2004. Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645–679.
- Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740.
- Pace, N.R. 2009. Problems with “prokaryote”. *Journal of Bacteriology* **191**(7), 2008–2010. PMID: 19168605.
- Pagani, I., Liolios, K., Jansson, J. *et al.* 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40**(Database issue), D571–579. PMID: 22135293.
- Parkhill, J., Achtman, M., James, K.D. *et al.* 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506. PMID: 10761919.
- Parra, G., Bradnam, K., Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9), 1061–1067. PMID: 17332020.
- Potter, S.C., Clarke, L., Curwen, V. *et al.* 2004. The Ensembl analysis pipeline. *Genome Research* **14**, 934–941.
- Pruesse, E., Peplies, J., Glöckner, F.O. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829.
- Prüfer, K., Munch, K., Hellmann, I. *et al.* 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**(7404), 527–531. PMID: 22722832.
- Prüfer, K., Racimo, F., Patterson, N. *et al.* 2014. The complete genome sequence of a Neandertal from the Altai Mountains. *Nature* **505**(7481), 43–49. PMID: 24352235.
- Quast, C., Pruesse, E., Yilmaz, P. *et al.* 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41** (D1), D590–D596. PMID: 23193283.
- Raghavan, M., Skoglund, P., Graf, K.E. *et al.* 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**(7481), 87–91. PMID: 24256729.
- Rasmussen, M., Anzick, S.L., Waters, M.R. *et al.* 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**(7487), 225–229. PMID: 24522598.
- Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521. PMID: 15057822.

- Raven, P., Fauquet, C., Swaminathan, M.S., Borlaug, N., Samper, C. 2006. Where next for genome sequencing? *Science* **311**, 468.
- Reich, D., Green, R.E., Kircher, M. *et al.* 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**(7327), 1053–1060. PMID: 21179161.
- Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R.A., Rogers, J. *et al.* 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**(5822), 222–234. PMID: 17431167.
- Rice Annotation Project, Tanaka, T., Antonio, B.A. *et al.* 2008. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Research* **36**(Database issue), D1028–1033. PMID: 18089549
- Riesenfeld, C.S., Schloss, P.D., Handelsman, J. 2004. Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* **38**, 525–552.
- Rivera, M.C., Lake, J.A. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155.
- Ruepp, A., Graml, W., Santos-Martinez, M.L. *et al.* 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**, 508–513. PMID: 11029001.
- Ryan, J.F., Pang, K., Schnitzler, C.E. *et al.* 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342**(6164), 1242592. PMID: 24337300.
- Salzberg, S.L., Hotopp, J.C., Delcher, A.L. *et al.* 2005. Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biology* **6**, R23.
- Sanger, F., Air, G.M., Barrell, B.G. *et al.* 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–895. PMID: 870828.
- Sanger, F., Nicklen, S., Coulson, A.R. 1977b. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science, USA* **74**, 5463–5467.
- Schweitzer, M.H., Suo, Z., Avci, R. *et al.* 2007. Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. *Science* **316**, 277–280.
- Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G.M. *et al.* 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**(5801), 941–952. PMID: 17095691.
- Shapiro, B., Hofreiter, M. 2014. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science* **343**(6169), 1236573. PMID: 24458647.
- Shinozaki, K. M., Ohme, M., Tanaka, M. *et al.* 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO Journal* **5**, 2043–2049. PMID: 16453699.
- Shungin, D., Winkler, T.W., Croteau-Chonka, D.C. *et al.* 2015. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**(7538), 187–196. PMID: 25673412.
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J. *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837. PMID: 12815422.
- Smith, D. R., Doucette-Stamm, L.A., Deloughery, C. *et al.* 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *Journal of Bacteriology* **179**, 7135–7155. PMID: 9371463.
- Sousa, V., Hey, J. 2013. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics* **14**(6), 404–414. PMID: 23657479.
- Stover, C. K., Pham, X.Q., Erwin, A.L. *et al.* 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964. PMID: 10984043.
- Tanenbaum, D.M., Goll, J., Murphy, S. *et al.* 2010. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Standards in Genomic Science* **2**(2), 229–237. PMID: 21304707.
- Tettelin, H., Saunders, N.J., Heidelberg, J. *et al.* 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815. PMID: 10710307.
- Thompson, O., Edgley, M., Strasbourger, P. *et al.* 2013. The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. *Genome Research* **23**(10), 1749–1762. PMID: 23800452.

- Venter, J. C., Adams, M.D., Myers, E.W. *et al.* 2001. The sequence of the human genome. *Science* **291**, 1304–1351. PMID: 11181995.
- Wang, J., Wang, W., Li, R. *et al.* 2008. The diploid genome sequence of an Asian individual. *Nature* **456**(7218), 60–65. PMID: 18987735.
- Warren, W.C., Hillier, L.W., Marshall Graves, J.A. *et al.* 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**(7192), 175–183. PMID: 18464734.
- Watson, J.D., Crick, F.H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356), 737–738. PMID: 13054692.
- Weber, J. L., Myers, E. W. 1997. Human whole-genome shotgun sequencing. *Genome Research* **7**, 401–409.
- Wheeler, D.A., Srinivasan, M., Egholm, M. *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189), 872–876. PMID: 18421352.
- Whittaker, R.H. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science* **163**, 150–160.
- Willerslev, E., Cooper, A. 2005. Ancient DNA. *Proceedings of the Royal Society B: Biological Science* **272**, 3–16.
- Williams, T.A., Foster, P.G., Cox, C.J., Embley, T.M. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**(7479), 231–236. PMID: 24336283.
- Wilson, E. O. 1992. *The Diversity Of Life*. W. W. Norton, New York.
- Woese, C. R. 1998. Default taxonomy: Ernst Mayr's view of the microbial world. *Proceedings of the National Academy of Science, USA* **95**, 11043–11046.
- Woese, C. R., Kandler, O., Wheelis, M. L. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Science USA* **87**, 4576–4579.
- Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Koonin, E. V. 2002. Genome trees and the tree of life. *Trends in Genetics* **18**, 472–479.
- Wood, V., Gwilliam, R., Rajandream, M.A. *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880. PMID: 11859360.
- Yandell, M., Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13**(5), 329–342.
- Yarza, P., Spröer, C., Swiderski, J. *et al.* 2013. Sequencing orphan species initiative (SOS): Filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. *Systematic and Applied Microbiology* **36**(1), 69–73. PMID: 23410935.
- Yilmaz, P., Kottmann, R., Field, D. *et al.* 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MiXs) specifications. *Nature Biotechnology* **29**(5), 415–420. PMID: 21552244.
- Yue, F., Cheng, Y., Breschi, A. *et al.* 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**(7527), 355–364. PMID: 25409824.
- Zerbino, D.R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocol in Bioinformatics Chapter 11*, Unit 11.5. PMID: 20836074.
- Zerbino, D.R., Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**(5), 821–829. PMID: 18349386.
- Zhang, X., Goodsell, J., Norgren, R.B. Jr. 2012. Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics* **13**, 206. PMID: 22646658.
- Zimin, A.V., Delcher, A.L., Florea, L. *et al.* 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology* **10**(4), R42. PMID: 19393038.





As early as 1885 Adolf Mayer showed that mosaic disease of the tobacco plant is contagious; we now know that it is caused by tobacco mosaic virus. Martinus Beijerinck (1851–1931) further isolated a “contagium vivum fluidum” (virus) from tobacco leaves, distinguishing the causative agent from bacteria. Due to their small size, almost all viruses cannot be visualized by conventional microscopy. Beginning in the 1930s Helmut Ruska pioneered the use of the electron microscope to visualize viruses (Kruger *et al.*, 2000). Early studies of the structure of viruses based on X-ray crystallography were performed by John D. Bernal (1901–1971). He also trained Maurice Wilkins and Rosalind Franklin (who confirmed the structure of the double helix of DNA) and Nobel laureate Dorothy Crowfoot Hodgkin (who solved the structure of vitamin B<sub>12</sub>). Together with Rosalind Franklin, Bernal studied tobacco mosaic virus in the 1950s. This figure shows a variety of purified viruses and X-ray analyses from Bernal and Fankuchen (1941, table 4). This set of images shows: (a) shifts of intermolecular reflections; (b) varying concentrations of viruses; (c) enation mosaic virus; (d) dry gels of various virus proteins; (e) tobacco mosaic virus; (f) cucumber mosaic virus; and (g) potato virus X.

Source: Bernal and Fankuchen (1941).

# Completed Genomes: Viruses

# CHAPTER 16

*Probably, the outstanding feature of the evolutionary process in parasitic microorganisms is the unimportance of the individual. A few influenza-virus particles initiate infection in one individual of a susceptible human community, and an epidemic of some thousands of cases results. From the point of view of the virus, we have a series of precipitate population increases, followed by catastrophic destruction. In each individual infected, the peak population of virus particles probably exceeds  $10^{10}$ , but it is certainly rare for even 10 of these to find opportunity for continued multiplication. When an active epidemic is in progress over a populated area, we might conceivably have  $10^{17}$  virus particles in a viable state. A few weeks later, there may be no viable particles whatever in this particular environment.*

—Sir MacFarlane Burnet, 1953 (p. 385).

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

- define viruses;
- explain the basis of the classification of viruses;
- describe the genomes of HIV, influenza, measles, Ebola, and herpesviruses;
- describe bioinformatics approaches to determining the function of viral genes and proteins;
- describe key bioinformatics resources for studying viruses; and
- compare and contrast DNA and RNA viruses.

## INTRODUCTION

In this chapter we will consider bioinformatic approaches to viruses. Viruses are small, infectious, obligate intracellular parasites. They depend on host cells for their ability to replicate. The virion (virus particle) consists of a nucleic acid genome surrounded by coat proteins (capsid) that may be enveloped in a lipid bilayer (derived from the host cell) studded with viral glycoproteins. Unlike other genomes, viral genomes can consist of either DNA or RNA. Furthermore, they can be single, double, or partially double stranded, and can be circular, linear, or segmented (having different genes on distinct nucleic acid segments).

Viruses lack the biochemical machinery that is necessary for independent existence. This is the fundamental distinction between viruses and free-living organisms. While they

replicate and evolve, viruses therefore exist on the borderline of the definition of life. The largest virus has a genome size of almost 2.5 megabases (*Pandoravirus salinus*; see “Giant Viruses” below), and other large viruses (such as pox viruses and Mimivirus) have genome sizes of several hundred kilobases to over a megabase. Some of these exceed the genome sizes of the smallest archaeal and bacterial genomes (e.g., *Nanoarchaeum equitans* and *Mycoplasma genitalium*; Chapter 17). It is not a coincidence that those smallest bacterial genomes are from organisms that (like viruses) are small, infectious, obligate intracellular agents. The largest *Pandoravirus* genomes even exceed eukaryotic genome sizes such as those of the microsporidian *Encephalitozoon* (Chapter 18). Notably, many of these very small bacterial genomes are in the process of transferring their genes to some host genome as they forego their capacity for independent existence.

While there may be tens or hundreds of millions of species of bacteria and archaea, only a few thousand species of virus are known. This disparity probably reflects their specialized requirement for invading a host. Also, recent metagenomics projects (described in “Metagenomics and Virus Diversity” below) suggest that we have an extremely limited understanding of both the number of virus species and the diversity of viral genes and genomes. Viruses infect all forms of life, including bacteria, archaea (Prangishvili *et al.*, 2006), and eukaryotes from plants to humans to fungi. A virus has even been found to infect a second virus (see “Giant Viruses” section below). Although we have catalogued relatively few viral species, viruses are nonetheless the most abundant biological entities on the earth (Edwards and Rohwer, 2005).

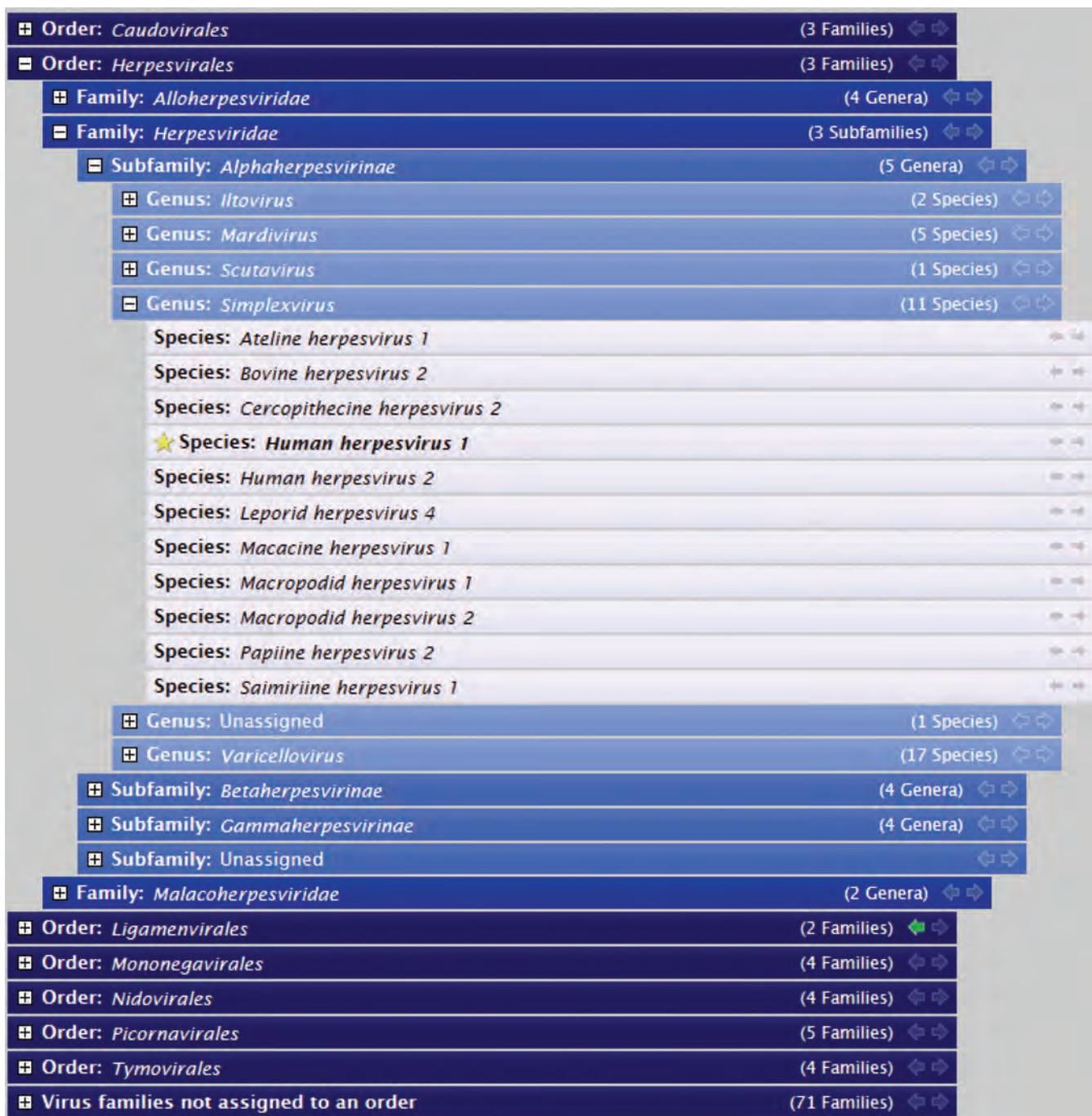
In this chapter we first discuss virus taxonomy, then classification of viruses (based on morphology, nucleic acid composition, genome size, and disease relevance). We describe the diversity and evolution of viruses, including viral metagenomics. After introducing bioinformatics approaches to problems in virology, we address specific viruses from small to large: influenza virus, human immunodeficiency virus (HIV), Ebola virus, measles virus, herpesvirus, and giant viruses. Each of these viruses allows us to gain insight into genomic principles of viruses. We also explore a series of bioinformatics tools to study viruses.

### International Committee on Taxonomy of Viruses (ICTV) and Virus Species

Established in 1971, the ICTV is a committee of the International Union of Microbiological Societies whose purpose is to classify viruses into taxa (King *et al.*, 2011). These have followed the Linnaean system of order, family, subfamily, genus and species. The ICTV database (2012 report) subdivides viruses into 7 orders, 96 families, 420 genera, and >2600 species of viruses. An example of the current taxonomy scheme is shown in **Figure 16.1** (note that there are 2–5 families per order as well as 71 families not assigned to an order). In the case of the genus *Simplexvirus*, the species *Human herpesvirus 1* is indicated with a yellow star as the type species of that genus.

The ICTV website is at  <http://ictvonline.org/> (WebLink 16.1).

According to its 1991 definition, “A virus species is a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche” (cited in Van Regenmortel *et al.*, 2013 who expand upon the meaning of “polythetic” as members having some properties in common but not necessarily a single common shared property). The ICTV recently introduced changes to the way viruses are defined. A species is “the lowest taxonomic level in the hierarchy approved by the ICTV. A species is a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria” (Adams *et al.*, 2013). These criteria may include natural and experimental host range, pathogenicity, antigenicity, vector specificity, cell and tissue tropism, and degree of relatedness of the genomes or genes. A species is monophyletic in that it is derived from a common ancestor; species are therefore discrete, nonoverlapping groups and phylogenetic analysis is explicitly required in identifying a new species.



**FIGURE 16.1** Virus taxonomy from the ICTV website (2012 release). The menus are opened to show human herpesvirus species.  
Source: ICTV. Reproduced with permission from The International Committee on Taxonomy of Viruses (ICTV).

Gibbs (2013) notes that ICTV-defined species and genera comprise virus groups that share most of their genes, but the broader categories of families and orders are not so discrete with some genes present in multiple families or orders. Van Regenmortgel *et al.* (2013) strongly critique the idea of monophyletic virus species. Additional detailed proposals for virus nomenclature are being actively developed (Kuhn *et al.*, 2013).

The ICTV has also recently changed the way viruses are named (Adams *et al.*, 2013). Virus species names are italicized with the first letter capitalized (e.g., *Rabbit hemorrhagic disease virus*). In contrast to virus species names, virus names are not italicized and are

According to George Gaylord Simpson (1963, p. 7), "Species are groups of actually or potentially inbreeding populations, which are reproductively isolated from other such groups. An evolutionary species is a lineage (an ancestral-descendant sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies."

Electron micrographs of viruses are available at sites such as All the Virology on the WWW (<http://www.virology.net/>, WebLink 16.2).

given in lower case (e.g., rabbit hemorrhagic disease virus which may be abbreviated RHDV). Kuhn and Jahrling (2010) and Van Regenmortel *et al.* (2010) discuss this distinction of virus and virus species.

The National Center for Biotechnology Information (NCBI) held an Annotation Workshop to guide the development of viral genome annotation standards (including nomenclature issues raised by ICTV). Tatiana Tatusova and colleagues emphasized the importance of consistent, comprehensive annotation, particularly as next-generation sequencing enables the determination of thousands of viral genome sequences (Brister *et al.*, 2010).

## CLASSIFICATION OF VIRUSES

We present four approaches to classifying viruses based on morphology, nucleic acid composition, genome size, and disease relevance.

### Classification of Viruses Based on Morphology

Before the sequencing era, morphology was an important criterion for the classification of viruses. Since 1959, electron microscopy has been employed to describe the structure of over 5500 bacteriophages (viruses that invade bacteria; Ackermann, 2007) as well as additional viruses that invade plants and animals. Ninety-six percent of bacteriophages are tailed viruses, with the remainder having filamentous, icosahedral, or pleiomorphic shapes. Many electron microscopic images of viruses are available online, and several are presented in **Figure 16.2**.

### Classification of Viruses Based on Nucleic Acid Composition

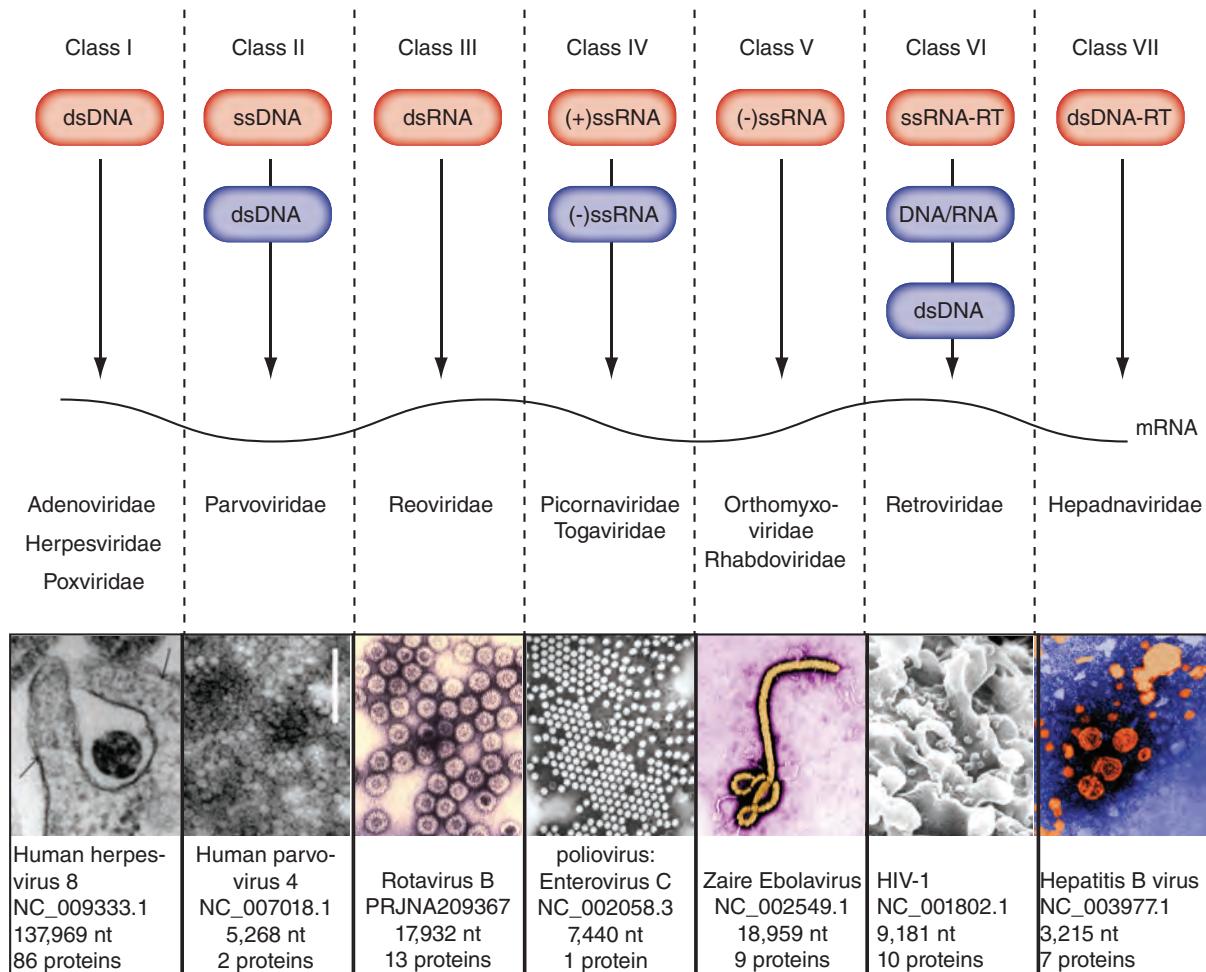
Another fundamental basis for classifying viruses is to define the type of nucleic acid genome that is packaged into the virion. Virions contain DNA or RNA; the nucleic acid may be single or double stranded, and translation may occur from the sense strand, the antisense strand, or both (Fig. 16.2). Double-stranded viral genomes replicate by using the individual strands of the DNA or RNA duplex as a template to synthesize daughter strands. Single-stranded DNA or RNA viruses use their strand of nucleic acid as a template for a polymerase to copy a complementary strand. Replication may involve the stable or transient formation of double-stranded intermediates. Some viruses with single-stranded RNA genomes convert the RNA strand to DNA using reverse transcriptase (RNA-dependent DNA polymerase). In the case of HIV-1, the *pol* gene encodes a reverse transcriptase.

### Classification of Viruses Based on Genome Size

Some of the major groups of viruses are shown in **Figure 16.2** and **Table 16.1**. Some have a very small genome size, such as rubella and hepatitis B (~2–3 kb). The first complete virus genome (Simian Virus 40 or SV40, 5243 bp) and first complete bacteriophage genome (bacteriophage MS2, 3569 bp), sequenced in the 1970s, are relatively small. Others are over 350 kb in size. A decade ago a giant virus (called Mimivirus for *Mimicking microbe*) was described, having a double-stranded circular genome of 1,181,404 base pairs (1.2 megabases) (La Scola *et al.*, 2003; Raoult *et al.*, 2004). Since then even larger related members of this group, now proposed to be called the order *Megavirales*, have been discovered. The largest has a genome size of 2.4 Mb (a pandoravirus, discussed in "Giant Viruses" below).

Although viruses are relatively simple agents, they are more complex than two other pathogenic agents: viroids and prions. Viroids are small, circular RNA molecules

SV40 was sequenced by Fiers *et al.* (1978) while MS2 was sequenced by Fiers *et al.* (1976).



**FIGURE 16.2** Classification of viruses. According to the classification system of David Baltimore, there are seven groups that vary in nucleic acid (DNA or RNA), strandedness (single- or double-stranded), sense (+ or – strand), and replication method. **Class I** viruses (e.g., the families Adenoviridae, Herpesviridae, and Poxviridae) have genomes composed of double-stranded DNA. The transmission electron microscopy (TEM) image of human herpesvirus 8 (HHV-8) shows a virion being internalized into a human ocular cell, with protrusions indicated by arrowheads. **Class II** viral genomes have single-stranded DNA that becomes double-stranded, as in the family Parvoviridae. The TEM image of purified human parvovirus-4-like particles has a scale bar of 200 nm. **Class III** viruses have double-stranded RNA and include the family Reoviridae. Rotaviruses are members of this family. This TEM image at 455,882 $\times$  magnification shows rotavirus icosahedral protein capsid particles. **Class IV** viruses have single-stranded RNA on the + (sense) strand. These include the families Picornaviridae and Togaviridae. A prominent example is Enterovirus, the cause of polio. **Class V** viruses have negative-, single-stranded RNA. Examples are the families Orthomyxoviridae (including influenza viruses) and Rhabdoviridae. We show an image of Ebola virus. **Class VI** includes retroviruses that use single-stranded RNA genomes with reverse transcription (RT) to form DNA or RNA intermediates. Human immunodeficiency virus-1 (HIV-1) is an example, shown in a scanning electron micrograph (SEM) with virions on the surface of cultured lymphocytes. **Class VII** viruses have double-stranded DNA genomes that use reverse transcription. They include the family Hepadnaviridae, including hepatitis viruses. TEM of hepatitis B is shown; this infects 300 million people worldwide and is responsible for 1 million deaths annually.

*Source:* the upper portion of the figure is adapted from a Wikipedia article on viruses (<http://en.wikipedia.org/wiki/Virus>). Image sources: HHV8, NIH ([http://openi.nlm.nih.gov/detailedresult.php?img=3312246\\_CDI2012-651691.002&req=4](http://openi.nlm.nih.gov/detailedresult.php?img=3312246_CDI2012-651691.002&req=4)); human parvovirus 4, NIH ([http://openi.nlm.nih.gov/detailedresult.php?img=3204632\\_10-0750-F1&query=parvovirus&it=xg&req=4&npos=15](http://openi.nlm.nih.gov/detailedresult.php?img=3204632_10-0750-F1&query=parvovirus&it=xg&req=4&npos=15)); Rotavirus, Dr Erskine L. Palmer of Centers for Disease Control (CDC) in 1978 (<http://phil.cdc.gov/phil/details.asp>); Enterovirus, CDC (<http://www2c.cdc.gov/podcasts/rssiframe.asp?c=303>); Ebola virus, CDC via NIH (<http://www.niaid.nih.gov/news/newsreleases/2010/Pages/EbolaImage.aspx>); HIV, CDC (<http://www2c.cdc.gov/podcasts/rssiframe.asp?c=303>); Hepatitis B, CDC (<http://www.cdc.gov/nchhstp/newsroom/DiseaseAgents.htm>). Adapted from Thomas Splettstoesser ([www.scistyle.com](http://www.scistyle.com)) under the terms of the Creative Commons CC-BY-SA-3.0 licence. Images from NLM, NIH and CDC.

**TABLE 16.1 Classification of viruses based on nucleic acid composition. Note that NCBI BioProject accessions begin PRJNA and typically encompass several segments. Adapted from Schaechter *et al.* (1999) with permission from Wolters Kluwer and with data from NCBI.**

Nucleic acid	Strands	Family	Example	Accession	Base pairs
RNA	Single	Picornaviridae	Human poliovirus 1	NC_002058.3	7,440
		Togaviridae	Rubella virus	NC_001545.2	9,762
		Flaviviridae	Yellow fever virus	NC_002031.1	10,862
		Coronaviridae	Coronavirus	NC_002645.1	27,317
		Rhabdoviridae	Rabies virus	NC_001542.1	11,932
		Paramyxoviridae	Measles virus	NC_001498.1	15,894
		Orthomyxoviridae	Influenza A virus	PRJNA14892	13,498
		Bunyaviridae	Tula virus (a hantavirus)	PRJNA14936	12,066
		Arenaviridae	Lassa fever virus	PRJNA14864	10,681
		Retroviridae	HIV	NC_001802.1	9,181
DNA	Double	Reoviridae	Rotavirus C	PRJNA16140	17,910
	Single	Parvoviridae	Parvovirus H1	NC_001358.1	5,176
	Mixed	Hepadnaviridae	Hepatitis B	NC_003977.1	3,215
	Double	Papovaviridae	JC virus	NC_001699.1	5,130
		Adenoviridae	Human adenovirus, type 17	AC_000006.1	35,100
		Herpesviruses	Human herpesvirus 1	NC_001806.1	152,261
		Poxviridae	Vaccinia	NC_006998.1	194,711

NCBI currently lists RefSeq accessions for 44 RNA viroid genomes, almost all <400 nucleotides. To see them visit NCBI Genomes (<http://www.ncbi.nlm.nih.gov/genome>, WebLink 16.3) and follow the links to viruses then viroids.

Stanley Prusiner won the Nobel Prize in Physiology or Medicine 1997 "for his discovery of Prions - a new biological principle of infection." See ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1997/](http://nobelprize.org/nobel_prizes/medicine/laureates/1997/) (WebLink 16.4).

The National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health offers information on viral and other diseases at (<http://www.niaid.nih.gov/topics/Pages/default.aspx> (WebLink 16.5).

of 200–400 nucleotides that cause diseases in plants (Flores, 2001; Daròs *et al.*, 2006; Ding, 2010). This minuscule genome does not encode any proteins, and the noncoding RNA itself has enzymatic activity. Gago *et al.* (2009) measured the extraordinarily high mutation rate of a hammerhead viroid (Fig. 16.3). We consider the mutation rates of other viruses in "Diversity and Evolution of Viruses" below.

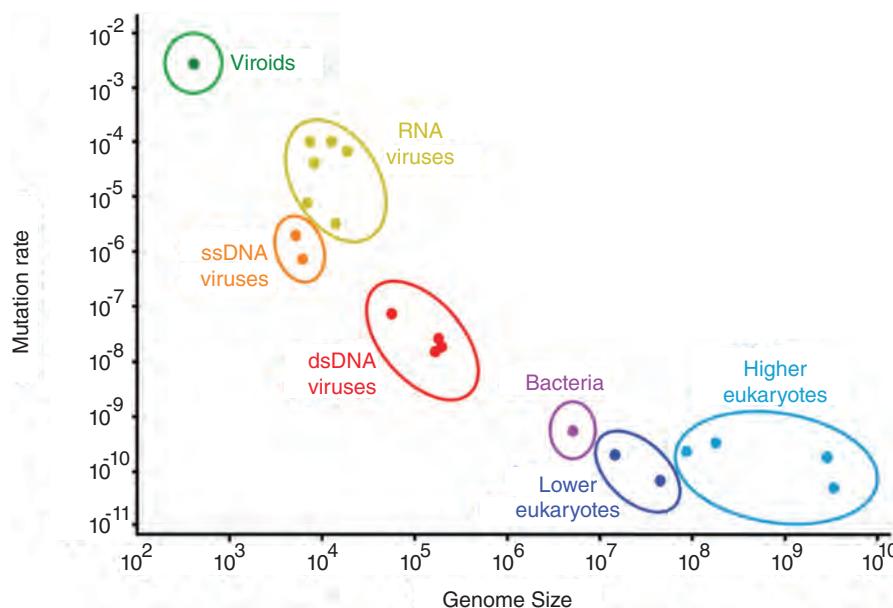
Prions are infectious protein molecules (Prusiner, 1998; DeArmond and Prusiner, 2003). Cruetzfeld–Jakob disease is the most common human prion disease (Johnson and Gibbs, 1998). It has a worldwide incidence of one in one million individuals and usually presents as dementia. Scrapie in sheep and bovine spongiform encephalopathy (BSE; "mad cow" disease) are the most common prion diseases in animals.

### Classification of Viruses Based on Disease Relevance

A different approach to classifying viruses is to identify those that cause human disease. Many viral diseases can be prevented by vaccination (Table 16.2). Others, such as smallpox, are of concern because of their potential use by bioterrorists (Cieslak *et al.*, 2002). Smallpox, caused by the variola virus, was eradicated in 1977; routine vaccination was discontinued in 1972 in the United States.

In general, RNA viruses (such as influenza virus, measles virus, Ebola virus, and HIV) present a greater disease burden in humans than DNA viruses (such as herpesviruses; Holmes, 2008).

Seven viruses are now known to cause cancer, collectively accounting for 10–15% of all cancers worldwide (Moore and Chang, 2010; Table 16.3). These include human herpesvirus 4 (HHV4, also called Epstein-Barr virus) found in cell lines from patients with Burkitt's lymphoma. Moore and Chang note that, surprisingly, human cancer viruses derive from a multitude of viral classes. They include exogenous retroviruses, positive-stranded RNA viruses, and double-stranded DNA viruses. All have close relatives that do not cause cancer.



**FIGURE 16.3** Per-site mutation rate as a function of genome size. The small viroid with the extremely high mutation rate is hammerhead viroid CChMVD. RNA viruses include tobacco mosaic virus, human rhinovirus, poliovirus, vesicular stomatitis virus, bacteriophage Φ6, and measles virus. Single-stranded DNA viruses are bacteriophage ΦX174 and bacteriophage m13. Double-stranded DNA viruses are bacteriophage λ, herpes simplex virus, bacteriophage T2, and bacteriophage T4. Bacteria is *Escherichia coli*. Lower eukaryotes are *Saccharomyces cerevisiae* and *Neurospora crassa*. Higher eukaryotes are *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*. Redrawn from Gago *et al.* (2009). Reproduced with permission from AAAS.

Viruses infect plants and inflict disease, causing great economic losses. Some plant viruses have limited economic impact but are considered important scientifically as model systems to understand the biology of viruses, plants, and/or their interactions. Scholthof *et al.* (2011) polled the virology community and proposed a list of the most important plant viruses, including the following. (1) *Tobacco mosaic virus*, identified as an infectious entity in 1898, is an important model. It was the first plant virus to be

**TABLE 16.2** Vaccine-preventable viral diseases. Data from <http://www.cdc.gov/vaccines/vpd-vac/default.htm> (WebLink 16.26).

Disease	Virus	Comment
Hepatitis A	Hepatitis A virus	Causes liver disease
Hepatitis B	Hepatitis B virus	Causes liver disease
Influenza	Influenza type A or B	Causes 20,000 deaths per year (US)
Measles	Measles virus	See below
Mumps	Rubulavirus	A disease of the lymph nodes
Poliomyelitis	Poliovirus (three serotypes)	Inflammation of the gray matter of the spinal cord; kills neurons
Rotavirus	Rotavirus	Most common cause of diarrhea in children; kills 600,000 children annually worldwide
Rubella	Genus rubivirus	Also called German measles
Smallpox	Variola virus	Eradicated in 1977
Varicella	Varicella-zoster virus	About 75% of all children contract varicella by age 15

**TABLE 16.3 Seven viruses that cause cancer in humans. Note that EBV is also called human herpesvirus 4 (HHV-4). Adapted from Moore and Chang (2010) with permission from Macmillan Publishers.**

Virus	Genome	Notable cancers	Year first described
Epstein–Barr virus (EBV)	Double-stranded DNA herpesvirus	Most Burkitt's lymphoma and nasopharyngeal carcinoma; most lymphoproliferative disorders; some Hodgkin's disease; some nonHodgkin's lymphoma; some gastrointestinal lymphoma	1964
Hepatitis B virus (HBV)	Single-stranded and double-stranded DNA hepadenovirus	Some hepatocellular carcinoma	1965
Human T-lymphotropic virus-I (HTLV-I)	Positive-strand, single-stranded RNA retrovirus	Adult T cell leukaemia	1980
High-risk human papillomaviruses (HPV) 16 and HPV 18 (some other α-HPV types are also carcinogens)	Double-stranded DNA papillomavirus	Most cervical cancer and penile cancers and some other anogenital and head and neck cancers	1983–1984
Hepatitis C virus (HCV)	Positive-strand, single-stranded RNA flavivirus	Some hepatocellular carcinoma and some lymphomas	1989
Kaposi's sarcoma herpesvirus (KSHV; also known as human herpesvirus 8 (HHV-8))	Double-stranded DNA herpesvirus	Kaposi's sarcoma, primary effusion lymphoma and some multicentric Castleman's disease	1994
Merkel cell polyomavirus (MCV)	Double-stranded DNA polyomavirus	Most Merkel cell carcinoma	2008

sequenced. (2) *Tomato spotted wilt virus* causes >US\$ 1 billion in crop losses annually. (3) *Tomato yellow leaf curl virus*, transmitted by the whitefly *Bemisia tabaci*, is a rapidly emerging disease of tomatoes. (4) *Cucumber mosaic virus* infects >1200 plant species in >100 families including tomato, tobacco and pepper. (5) *Potato virus Y*, transmitted by over 40 aphid species, infects Solanaceae including potatoes. Listing these viruses helps to establish research priorities; see Scholthof *et al.* for more information about these particular agents.

## Diversity and Evolution of Viruses

A practical way to access the diversity of known viruses is through the NCBI website. We introduced the NCBI Genome resources in Chapter 15. This site includes dedicated resources for viruses (Fig. 16.4) as well as specialized sites for influenza virus, retroviruses, SARS, Ebola virus, and links to the ICTV database.

A premise of taxonomy is that it should represent phylogeny. In the case of viruses, their unique, elusive, and sometimes fragile nature makes it difficult to trace their evolution in as comprehensive a fashion as can be accomplished with archaea, bacteria, and eukaryotes. Like living organisms, viruses are subject to mutation (genetic variability) and selection. Viral genomes present special difficulties for evolutionary studies, however:

- Viruses tend not to survive in archeological or historical samples. There is considerable evidence for the existence of viruses over 10,000 years ago, based upon human skeletal remains, historical accounts, and other historical artifacts. However, ancient viral DNA or RNA has not been recovered. As discussed in “Influenza Virus”, influenza virus from the deadly 1918 pandemic has been isolated, sequenced, and

Currently (October 2014) there are ~4200 viral genomes and additional phage genomes listed at the NCBI Genome site. The NCBI Genome homepage is  
 <http://www.ncbi.nlm.nih.gov/genome/> (WebLink 16.3) with a link to virus resources.

**Viral Genomes**

**FAQs**

- ▶ How to retrieve nucleotide and protein sequences of viral reference genomes?
- ▶ How to retrieve non-RefSeq nucleotide sequences of complete viral genomes?
- ▶ Why is a particular full-length genomic sequence missing in Entrez Genomes? Is there a RefSeq genome for this virus?

**hypothesis**  
read »

**Did viruses invent DNA?**

**Influenza A virus replication scheme.** Click on it for explanations.

A **viral genome** consists of either single-stranded or double-stranded DNA or RNA in either linear or circular form, and can comprise one or more segments.

**FIGURE 16.4** The viral genomes page at NCBI provides information and resources for the study of viruses. There are links to tools (such as PASC for comparisons of viral genomes) and to specialized NCBI sites on retroviruses, SARS, and influenza viruses.

Source: Viral Genomes, NCBI.

functionally analyzed. We also describe below a giant virus recovered from 30,000-year-old permafrost.

- The great diversity of viral genomes precludes us from making comprehensive phylogenetic trees based upon molecular sequence data that span the entire set of viruses. This reflects the complex molecular evolutionary events that form viral genomes (McClure, 2000).
- Many viral genomes are segmented. This allows segments to be shuffled among progeny, producing a great diversity of viral subtypes (see influenza virus and HIV sections below). Pond *et al.* (2012) discuss recombination and HyPhy software for the detection of recombination and selection.

For a variety of viral families, phylogenetic trees have been generated. These are indispensable in establishing the evolution, host specificity, virulence, and other biological properties of viral species. We examine phylogenetic reconstructions of the herpesviruses and HIV. Phylogenetic trees have been generated for other viruses from measles to hepatitis.

HyPhy (Hypothesis testing using Phylogenies) software is available from <http://hyphy.org/> (WebLink 16.6). Some HyPhy analyses are incorporated in MEGA (Chapter 7).

We showed the exceptionally large mutation rate for a viroid (Fig. 16.3). There are progressively lesser mutation rates in RNA viruses, single-stranded DNA viruses, double-stranded DNA viruses, bacteria, and eukaryotes. As shown in the figure, higher mutation rates tend to be associated with smaller genome sizes. Duffy *et al.* (2008) reviewed the rates of evolutionary change in viruses. The mutation rate also correlates with the fidelity of the polymerase used in replication. RNA viruses use RNA-dependent RNA polymerases, and these typically lack proofreading activity. This leads to a mutation rate that may be 1 million to 10 million times greater than that of DNA genomes (McClure, 2000). Retroviruses such as HIV use RNA-dependent DNA polymerases, that is, reverse transcriptases, also with low fidelity. DNA viruses use DNA polymerases, whether encoded by virus or host.

In addition to a high mutation rate, many viruses also have an extremely high rate of replication. A single host cell can produce 10,000 poliovirus particles, and an HIV-infected individual can produce  $10^9$  virus particles per day. For hepatitis C,  $10^{12}$  virions per day can be produced (Neumann *et al.*, 1998). This can lead to the formation of quasi-species (a population of related but nonidentical viruses).

Viruses are often subjected to intense selective pressures such as host immune responses or antiviral drug therapies. The rapid mutation rate of HIV-1 ensures that some versions of the virus are likely to contain mutations conferring resistance to retroviral drugs, and these HIV-1 molecules will be selected for.

## Metagenomics and Virus Diversity

Historically, we have classified viruses based on observation of their effects (e.g., by studying plant or human diseases caused by viruses), based on morphology, or based on the nature of the nucleic acid in purified virus particles. Metagenomics projects survey large amounts of genomic sequence from environmental samples or from host organisms (Chapter 15). Several metagenomics studies have resulted in the identification of large numbers of virus genomes (reviewed in Edwards and Rohwer, 2005; Mokili *et al.*, 2012; Willner and Hugenholtz, 2013). Rosario and Breitbart (2011) summarize 24 published viral metagenomes, noting that half the sequences are previously unknown. Other novel viral genome sequences unexpectedly include metabolic genes found in cellular organisms.

A major metagenomics approach is to characterize DNA sequences in environmental samples. J. Craig Venter and colleagues surveyed marine planktonic microbiota in a Global Ocean Sampling expedition (Rusch *et al.*, 2007). Forty-one samples were collected over a range of 8000 km, and 7.7 million sequencing reads were obtained. Combining their results with the previous Sargasso Sea survey (Venter *et al.*, 2004), they reported the identification of 6.1 million proteins. There was a disproportionately large number of novel protein sequences assigned to viral genomes, consistent with the view that we have not yet achieved a broad sampling of viral diversity. Venter's group extended their study to the Indian Ocean, sampling viral fractions of different size classes and identifying putative host genera such as *Prochlorococcus* and *Acanthochlois* (Williamson *et al.*, 2012). Culley *et al.* (2006) also reported a diverse set of previously unknown RNA viruses in seawater.

Another metagenomic approach is to sample genomic DNA from individual organisms. In particular the human gut is colonized by hundreds or thousands of microbial species, including bacteria and archaea. Many of these species are infected by viruses (Reyes *et al.*, 2012). One goal is to identify viral pathogens in patients (Bibby, 2013). This can be done for clinical diagnostics, to detect and respond to viral pathogen outbreaks, or to discover new viruses. For example, acute diarrhea causes ~1.8 million deaths in children each year (see Fig. 21.3), and causes are known in ~60% of cases. For the remaining 40% of cases the etiology is undetermined and could involve unknown viruses. Various groups have therefore performed sequencing of viral particles from feces of those who are healthy or have diarrhea (Breitbart *et al.*, 2003; Zhang *et al.*, 2006; Finkbeiner *et al.*, 2008).

Correlation does not imply causation: the presence of a novel virus in the feces of a patient does not imply that it necessarily causes diarrhea. According to Koch's postulates, there are several criteria needed to establish a causal relationship between a microbe and a disease. Jakob Henle and his student Robert Koch developed these rules in the late nineteenth century in studies of anthrax and tuberculosis, and they have also been applied to viruses. The postulates, quoted from Evans (1976), are:

1. The parasite occurs in every case of the disease in question and under circumstances which can account for the pathological changes and clinical course of the disease.
2. It occurs in no other disease as a fortuitous and nonpathogenic parasite.
3. After being fully isolated from the body and repeatedly grown in pure culture, it can induce the disease anew.

This third postulate was often difficult to achieve because many bacteria could not be grown in culture, and often the disease could not be reproduced in an animal model. For viruses, which require a host in which to propagate, the Henle–Koch postulates were even harder to fulfill. In the late 1950s Robert Huebner suggested guidelines for establishing a virus as a cause of a human disease, including the following (adapted from Evans, 1976):

1. The virus must be a “real” entity established by passage in animal or tissue cultures.
2. The virus must originate from human specimens (as opposed to representing a viral contaminant of experimental animals, cells, or media it is grown in).
3. Active infection should produce an antibody response.
4. A new virus should be fully characterized and compared with other agents (e.g., host and host-cell ranges, pathologic lesions).
5. The virus must be constantly associated with a specific illness.
6. Human volunteers inoculated with the newly recognized agent in double-blind studies should reproduce the clinical syndrome. (Such studies may be prohibited today on ethical grounds.)
7. Epidemiological studies should identify patterns of infection and disease.
8. A specific vaccine should prevent the disease, therefore establishing an agent as the cause.

Today metagenomics studies can identify viruses in patients with a disease (Tang and Chiu, 2010). The Huebner guidelines may be helpful in evaluating the relevance of the virus to the clinical phenotype.

## BIOINFORMATICS APPROACHES TO PROBLEMS IN VIROLOGY

The tools of bioinformatics are well suited to address some of the outstanding problems in virology:

- Why does a virus such as HIV-1 infect one species selectively (human) while a closely related virus (simian immunodeficiency virus) infects monkeys but not humans? Analysis of the sequence of the viruses as well as the host cell receptors can address this question.
- Why do some viruses change their natural host? In 1997 a chicken influenza virus infected 18 humans, killing 6. Are there changes in the genome of the virus, of the host, or both that facilitate cross-species changes in specificity?
- Why are some viral strains deadlier than others? We explore the properties of the 1918 influenza virus that killed as many as 50 million people.
- What are the mechanisms of viral evasion of host immune systems? We see below (“Herpesvirus”) how some herpesviruses acquire viral homologs of human immune system molecules and therefore interfere with human antiviral mechanisms.

- Where did viruses originate? There are several theories (Holmes, 2011). Their origin could be ancient, even predating the last universal cellular ancestor (often abbreviated LUCA). The discovery of giant viruses such as mimivirus (described in “Giant Viruses” below) could support this model. Alternatively, viruses could have emerged relatively recently, having been derived from more complex intracellular parasites that eliminated many nonessential features. They could also be derived from normal cellular components that now replicate autonomously. Phylogenetic analyses could help resolve these theories. Edward Holmes (2008) has reviewed the evolutionary history of DNA and RNA viruses.
- Which vaccines are most likely to be effective? There are two main approaches to developing vaccines for viruses that display a great amount of molecular sequence diversity. One approach is to select isolates of a particular subtype based on regional prevalence. A second approach is to deduce an ancestral sequence or a consensus sequence for use as an antigen in vaccine development (Gaschen *et al.*, 2002). These approaches depend on molecular phylogeny.

In the remainder of this chapter we examine six specific viruses of interest, presented in order from smallest to largest genome size. (1) Human Immunodeficiency Virus (HIV) is a retrovirus associated with Acquired Immune Deficiency Syndrome (AIDS). (2) Influenza viruses cause human sickness and death each year, with a constant threat of potentially causing a pandemic such as that which caused tens of millions of deaths one century ago. (3) Measles virus, responsible for killing as many as half a million children each year, is an example of a virus that changes its antigenicity very slowly. (4) We introduce Ebola virus, a recently emerging threat. (5) We explore herpesviruses, and introduce PASC for pairwise comparison of viruses. (6) We explore pandoraviruses which have relatively enormous sizes and mysterious roles in biology, introducing MUMmer software to compare two large genome sequences.

## HUMAN IMMUNODEFICIENCY VIRUS (HIV)

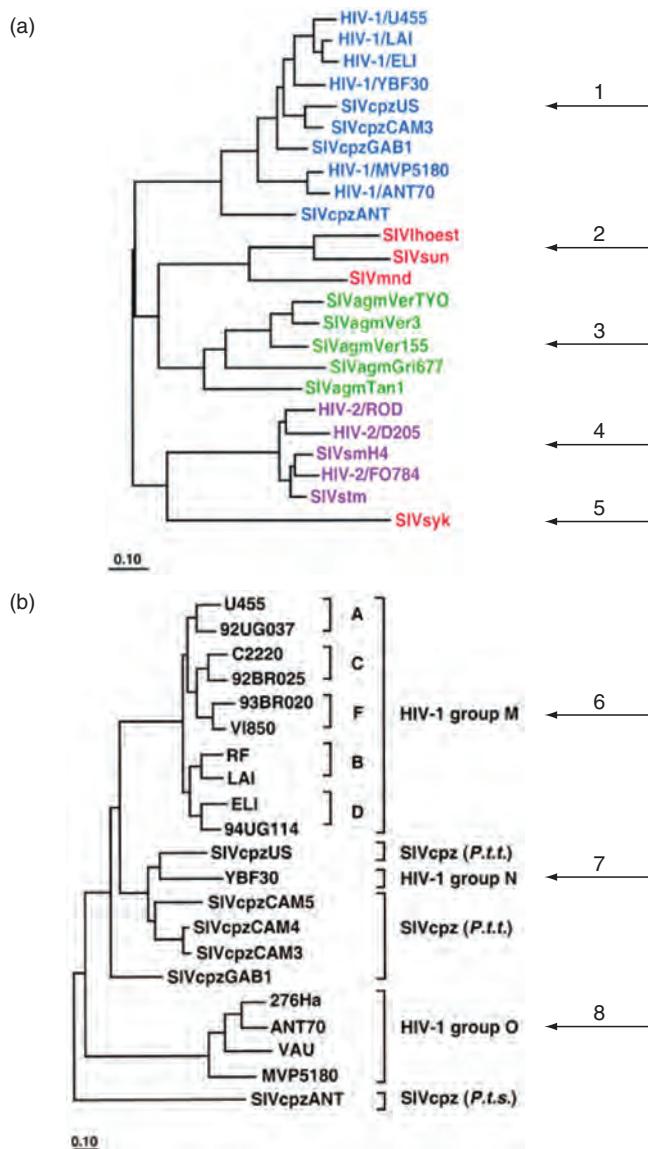
Information about AIDS is available at <http://www.niaid.nih.gov/topics/HIV/AIDS/Pages/Default.aspx> (WebLink 16.7), an NIH website. Information on prevalence is from the Centers for Disease Control and Prevention at <http://www.cdc.gov/hiv/library/factsheets/> (WebLink 16.8) and UNAIDS and the World Health Organization at <http://www.unaids.org/> (WebLink 16.9). We discuss DALYS in Chapter 21.

Human immunodeficiency virus is the cause of AIDS (reviewed in Meissner and Coffin, 1999). Early after its prominent emergence in the 1980s, HIV was uniformly fatal. Most of the symptoms of AIDS are not caused directly by the virus, but instead are a consequence of the ability of the virus to compromise the host immune system. HIV infection therefore leads to disease caused by opportunistic organisms.

Currently, over 34 million people are infected with HIV worldwide, with 2.5 million new cases in 2011. Nearly 30 million people have died from AIDS since the 1980s, with the greatest burden in sub-Saharan Africa. The prevalence of AIDS is increasing by about 3% per year. Although mortality rates are declining, HIV/AIDS still ranks as the fifth leading cause of global disability-adjusted life years (DALYs) in 2010 (Ortblad *et al.*, 2013). There have been many multinational efforts to combat HIV/AIDS across disciplines from treatment to prevention (Piot *et al.*, 2004). For broad surveys of the state of HIV policy and research, see the compendia of articles in *Science* and *Nature* (Mandavilli, 2010; Roberts, 2012). Barré-Sinoussi *et al.* (2013) and Ciuffi and Telenti (2013) review aspects of HIV research.

HIV-1 and HIV-2 are retroviruses of the group lentivirus. The viruses probably originated in sub-Saharan Africa, where the diversity of viral strains is greatest and the infection rates are highest (Sharp *et al.*, 2001). The primate lentiviruses occur in five major lineages, as shown by a phylogenetic tree based on full-length pol protein sequences (Fig. 16.5a; see arrows 1–5; Hahn *et al.*, 2000; see also Rambaut *et al.*, 2004; Heeney *et al.*, 2006 and a review by Castro-Nallar *et al.*, 2012). These five lineages are:

1. Simian immunodeficiency virus (SIV) from the chimpanzee *Pan troglodytes* (SIVcpz), together with HIV-1;



**FIGURE 16.5** Evolutionary relationships of primate lentiviruses. (a) Full-length Pol protein sequences were aligned and a tree was created using the maximum-likelihood method. There are five major lineages (arrows 1–5). The scale bar indicates 0.1 amino acid replacements per site after correction for multiple hits. (b) The HIV-1/SIVcpz lineage is displayed based on a maximum-likelihood tree using Env protein sequences. Note that the three major HIV-1 groups (M, N, O; arrows 6–8) are distinguished. The scale bar is the same as in (a). From Hahn *et al.* (2000).

2. SIV from the sooty mangabeys *Cercocebus atys* (SIVsm), together with HIV-2 and SIV from the macaques (genus *Macaca*; SIVmac);
3. SIV from African green monkeys (genus *Chlorocebus*; SIVagm);
4. SIV from Sykes' monkeys, *Cercopithecus albogularis* (SIVsyk); and
5. SIV from l'Hoest monkeys, *Cercopithecus lhoesti* (SIVlhoest); SIV from suntailed monkeys (*Cercopithecus solatus*; SIVsun); and SIV from a mandrill (*Mandrillus sphinx*; SIVmnd).

A prominent feature of phylogenetic analyses such as those in Fig. 16.5a is that viruses appear to have evolved in a host-dependent manner, as we discuss below in “Herpesvirus.” HIV-related viruses infecting any particular nonhuman primate species are more closely

Prevalence of a disease (or infection) is the proportion of individuals in a population who have a disease at a particular time. Prevalence does not describe when individuals contracted a disease. Incidence is the frequency of new cases of a disease that occur over a particular time. For example, the incidence of a disease might be described as 10 new cases per 1000 people in the general population in a given year.

related to one another than to viruses from other species. For HIV-2, transmission from the sooty mangabees was indicated by five lines of evidence (Hahn *et al.*, 2000):

1. similarities in the genome structures of HIV-2 and SIVsm;
2. phylogenetic relatedness of HIV-2 and SIVsm (see Fig. 16.5, arrow 4);
3. prevalence of SIVsm in the natural host;
4. geographic coincidence of those affected and the natural host; and
5. plausible routes of transmission, such as exposure of humans to chimpanzee blood in markets.

Similar arguments have been applied to HIV-1, which probably appeared in Africa in 1930–1940 as a cross-species contamination by SIVcpz. HIV-1 occurs in three major subtypes, called M, N, and O. This is consistent with the occurrence of three separate SIVcpz transmissions to humans: M is the main group of HIV-1 viruses; O is an outlier group; and N is also distinct from M and O. The three main HIV-1 subtypes are apparent in a phylogenetic tree generated from full-length Env protein sequences (Fig. 16.5b, arrows 6–8; Hahn *et al.*, 2000).

SIV has an unexpectedly deep history. Worobey *et al.* (2010) identified SIV in primates living on an island of Equatorial Guinea that was isolated by rising sea levels 12,000 years ago. Their phylogenetic analyses suggested that SIV has been present in primates for some 32,000 years. Human may therefore have encountered SIV many times in the past, at least sporadically.

We discussed the NCBI entry for HIV-1 in Chapter 2 (“HIV-1 pol”); the genome is 9181 bases and contains 9 open reading frames that encode proteins. The structure and function of these proteins have been characterized (Briggs and Kräusslich, 2011; Engelmann and Cherepanov, 2012). While the HIV-1 genome is small and there are few gene products, GenBank currently has ~600,000 nucleotide sequence records and ~700,000 protein records. The reason for this enormous quantity of data is that HIV-1 mutates extremely rapidly, producing many subtypes of the M, N, and O variants. Researchers therefore sequence HIV variants very often. A major challenge for virologists is to learn how to manipulate such large amounts of data and how to use those data to find meaningful approaches to treating or curing AIDS. As shown in Figure 16.3, other RNA viruses have even higher mutation rates. This highlights the additional important roles of recombination and natural selection from immune evasion (Holmes, 2009).

While HIV-1 exhibits great diversity both globally and within each infected individual, it is possible to characterize the genome using next-generation sequencing. Gall *et al.* (2012) developed a method to amplify, sequence, and assemble any HIV-1 genome regardless of sequence or subtype. This approach also reveals mutations associated with drug resistance in clinical samples.

We next describe two bioinformatics resources for the study of HIV molecular sequence data: NCBI and LANL.

### NCBI and LANL resources for HIV-1

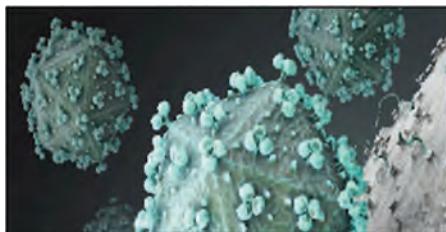
The NCBI website offers several ways to study retroviruses, including HIV. You can access information on HIV-1 via the Genome site at NCBI, as we also describe below in the relevant sections for influenza viruses, HHV-8, and megaviruses.

NCBI also offers a dedicated resource for the study of retroviruses (Fig. 16.6). This site includes the following:

- a genotyping tool based upon BLAST searching;
- a multiple sequence alignment tool specific to retroviral sequences;
- a reference set of retroviral genomes;
- specific pages with tools to study HIV-1, HIV-2, SIV, human T-cell lymphotropic virus type 1 (HTLV), and STLV;
- a listing of the previous week’s publications on retroviruses;

To see NCBI Nucleotide records for HIV-1, visit the Taxonomy browser page and enter HIV-1. If you limit the output in a search of the Entrez nucleotide database to RefSeq entries, there is only one entry: the complete HIV-1 genome (NC\_001802.1).

The NCBI Genome section (<http://www.ncbi.nlm.nih.gov/genome>, WebLink 16.3) has a virus link as well as a browse feature to link to HIV-1 (NC\_001802). Clicking on the name of the virus provides a link to the NCBI taxonomy browser, which includes information on the lineage of HIV-1 (Viruses; Retrovirus; Retroviridae; Lentivirus; Primate lentivirus group) as well as links to dozens of HIV-1 variants. From the Genome page, by clicking on the accession number NC\_001802 a link to the Nucleotide (GenBank) entry for HIV-1 is provided.



## Retroviruses

Information about retroviruses and specialized tools for the analysis of retroviral proteins and genomes

<b>Using the Resource</b>	<b>Retrovirus Tools</b>	<b>Retrovirus Genomes and Taxonomy</b>
<a href="#">About</a>	<a href="#">HIV-1 Human Interaction Database</a>	<a href="#">Reference retrovirus genomes</a>
<a href="#">Help</a>	<a href="#">Retrovirus genotyping tool</a>	<a href="#">Browse retrovirus genomes by species</a>
<a href="#">Questions</a>	<a href="#">Retrovirus nucleotide Blast</a>	<a href="#">Retrovirus taxonomy browser</a>
<b>External Retrovirus Resources</b>	<b>Healthcare and Education</b>	<b>Other NCBI Virus Resources</b>
<a href="#">Los Alamos National Laboratory HIV Databases</a>	<a href="#">Centers for Disease Control</a>	<a href="#">Viral Genomes Resource</a>
<a href="#">Stanford HIV Drug Resistance Mutation Database</a>	<a href="#">National Institute of Allergy and Infectious Disease</a>	<a href="#">Virus Variation Resource</a>
<a href="#">HIV Drug Resistance Program</a>	<a href="#">Retroviruses Textbook</a>	<a href="#">Influenza Virus Resource</a>
<a href="#">NIH AIDS Reagent Program</a>	<a href="#">PubMed Retrovirus Articles</a>	
<a href="#">HIV BioAfrica</a>		

**FIGURE 16.6** Retroviruses resource.

Source: Retroviruses, NCBI (<http://www.ncbi.nlm.nih.gov/retroviruses/>).

- a listing of the previous week's GenBank releases (many hundreds of new HIV-1 sequences are deposited weekly); and
- links to external retroviral website resources.

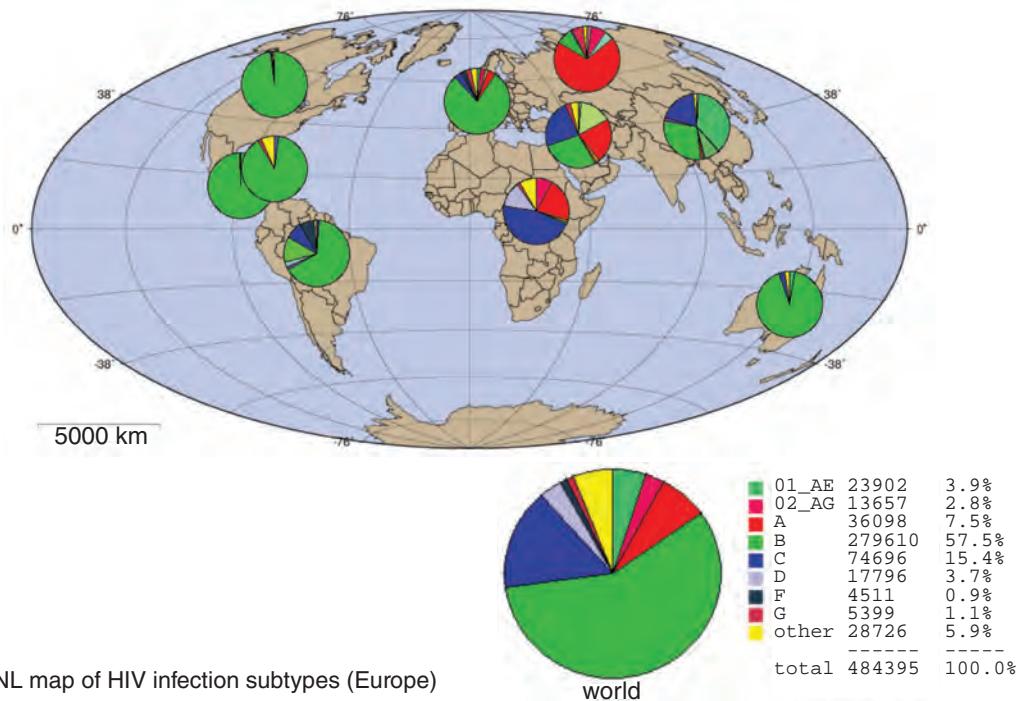
A fundamental resource for the study of several virus types including HIV is the Los Alamos National Laboratory (LANL) which operates a group of HIV databases. The HIV Sequence Database is an important, comprehensive repository of HIV sequence data. It allows searches for sequences by common names, accession number, PubMed identifier, country in which each case was sampled, and likely country in which infection occurred. Sequences may be retrieved as part of a multiple sequence alignment or unaligned, and groups of sequences derived from an individual patient may be retrieved. The site includes a variety of specialized tools, including:

- an HIV BLAST server;
- SNAP (Synonymous/Nonsynonymous Analysis Program), a program that calculates synonymous and nonsynonymous substitution rates;
- Recombinant Identification Program (RIP), a program that identifies mosaic viral sequences that may have arisen through recombination;
- a multiple alignment program called MPAlign (Gaschen *et al.*, 2001) that uses HMMER software (Chapter 6);
- PCoord (Principal Coordinate Analysis), a program that performs a procedure similar to principal components analysis (Chapter 11) on sequence data based on distance scores; and
- a geography tool that shows both total HIV infection levels (either worldwide or by continent) as well as the subtype distribution of HIV (Fig. 16.7a, b).

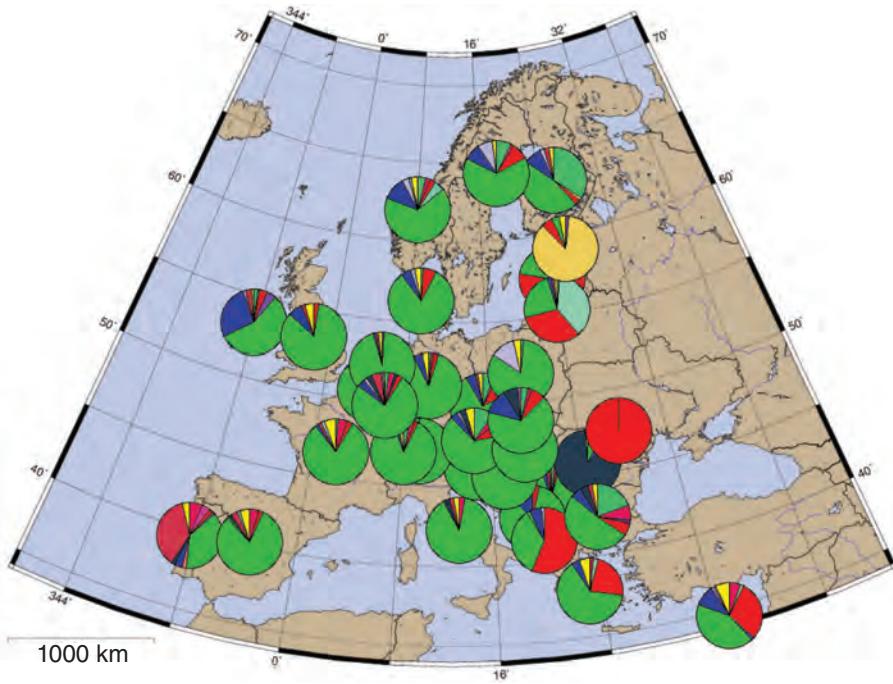
"Retrovirus Resources" are available at <http://www.ncbi.nlm.nih.gov/retroviruses/> (WebLink 16.10). NCBI also offers a database of interactions between HIV and human proteins (<http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/>, WebLink 16.11).

The LANL HIV databases are available at <http://hiv-web.lanl.gov/> (WebLink 16.12). This site offers three databases: sequence, immunology, and vaccine trials. In the HIV Sequence Database you can find the geography tool by selecting "Tools" and then "Geography."

(a) LANL map of HIV infection subtypes (worldwide)



(b) LANL map of HIV infection subtypes (Europe)



**FIGURE 16.7** The geography tool at LANL allows you to view HIV infection subtypes (a) globally or (b) by continent (Europe is shown). In (a), the total and dominant subtypes are indicated. The subtype distribution is displayed using pie charts.

Source: Los Alamos National Security, LLC, for the US Department of Energy (<http://www.hiv.lanl.gov/>).

## INFLUENZA VIRUS

The “Spanish” influenza pandemic of 1918–1919 infected hundreds of millions of people, and is estimated to have killed 50 million people. The death rate among otherwise healthy young adults was especially high. Why was it so deadly? Influenza virus pandemics returned in the 1957 “Asian” flu and the 1968 “Hong Kong” flu. More recently, the avian influenza subtype H5N1 infected over 300 humans and killed over 200, and also led to the slaughter of millions of birds. In 2013 an H7N9 influenza virus of avian origin struck in China, killing 37 of 132 infected people. Many wild birds such as ducks, geese, swans, and gulls are infected with influenza A (Olsen *et al.*, 2006), often not causing symptoms in birds. Will an avian influenza virus such as H5N1 or H7N9 infect humans globally? What are the properties of the influenza genome, and how can genome analyses help us to predict the next epidemic and devise strategies to prevent and/or treat its effects? In addition to the deadly avian flu strains, other subtypes of influenza virus are estimated to cause 250,000–500,000 deaths annually (36,000 deaths annually in the United States).

The influenza virus (of the family *Orthomyxoviridae*) presents in three types (A, B, and C) based on genetic and antigenic differences (Pleschka, 2013; **Table 16.4**). Influenza virus A is most responsible for human disease. Each influenza virus strain consists of about 12,500–14,500 bases of single-stranded negative-sense RNA and encoding 9–12 genes (**Table 16.5**). The genome of influenza A consists of eight segments (ranging in length from 890 to 2341 nucleotides) named PB1, PB2, PA, HA, NP, NA, M, and NS (**Fig. 16.8**). The hemagglutinin (HA) and neuraminidase (NA) segments encode two key surface glycoproteins that together define viral subtypes. The HA and NA segments occur

For information on influenza virus see the Centers for Disease Control website <http://www.cdc.gov/flu/about/viruses/> (WebLink 16.13). The World Health Organization (WHO) maintains a listing of confirmed human cases of avian influenza A (H5N1) (<http://www.who.int/topics/influenza/en/>, WebLink 16.14) with links to updates including maps of global influenza cases. From 2003 to 2009 there were 489 cases and 289 deaths attributed to H5N1.

**TABLE 16.4** Influenza viruses: examples of family *Orthomyxoviridae* complete genomes.

Virus	Source information	Segments	Length (nt)	Proteins
<b>Influenzavirus A</b>				
Influenza A virus (A/Goose/Guangdong/1/1996(H5N1))	Strain: A/Goose/Guangdong/1/96(H5N1)	8	13,590	12
Influenza A virus (A/Hong Kong/1073/99(H9N2))	Serotype: H9N2; Strain: A/Hong Kong/1073/99	8	13,460	12
Influenza A virus (A/Korea/426/1968(H2N2))	Serotype: H2N2; Strain: A/Korea/426/68	8	13,460	12
Influenza A virus (A/New York/392/2004(H3N2))	Serotype: H3N2; Strain: A/New York/392/2004	8	13,627	12
Influenza A virus (A/Puerto Rico/8/1934(H1N1))	Serotype: H1N1; Strain: A/Puerto Rico/8/34	8	13,588	12
<b>Influenzavirus B</b>				
Influenza B virus	Strain: B/Lee/40	8	14,452	11
<b>Influenzavirus C</b>				
Influenza C virus	Strain: C/Ann Arbor/1/50	7	12,555	9
<b>Isavirus</b>				
Infectious salmon anemia virus	Isolate: CCBB	8	12,716	10
<b>Thogotivirus</b>				
Thogoto virus	Strain: SiAr 126	6	10,461	7

Source: NCBI Genomes, NCBI (<http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=11308>, WebLink 16.27).

**TABLE 16.5 Genes in a representative Influenza A virus complete genome (A/Puerto Rico/8/34(H1N1)), taxonomy identifier 211044.**

Gene	Segment	Protein Accession	Length (amino acids)	Name
PB2	1	NP_040987	759	RNA-dependent RNA polymerase subunit PB2
PB1	2	NP_040985	757	RNA-dependent RNA polymerase subunit PB1
PB1-F2	2	YP_418248	87	PB1-F2 protein
PA	3	NP_040986	716	RNA-dependent RNA polymerase subunit PA
PA-X	3	YP_006495785	252	RNA-dependent RNA polymerase subunit PA-X
HA	4	NP_040980	566	Haemagglutinin
NP	5	NP_040982	498	Nucleocapsid protein
NA	6	NP_040981	454	Neuraminidase
M2	7	NP_040979	97	Matrix protein 2
M1	7	NP_040978	252	Matrix protein 1
NS1	8	NP_040984	230	Nonstructural protein NS1
NS2	8	NP_040983	121	Nonstructural protein NS2

Source: NCBI Genomes, NCBI ([http://www.ncbi.nlm.nih.gov/genome/proteins/10290?project\\_id=15521](http://www.ncbi.nlm.nih.gov/genome/proteins/10290?project_id=15521)).

Type	Name	RefSeq	INSDC	Size (Kb)	GC%	Protein	Gene
Segment	6	NC_004909.1	AJ404629.1	1.42	42.6	NA 1	1
Segment	1	NC_004910.1	AJ404630.1	2.34	43.2	PB2 1	1
Segment	3	NC_004912.1	AJ404637.1	2.23	44.0	PA 2 PA-X 2	2
Segment	8	NC_004906.1	AJ278649.1	0.89	43.3	NS2 2 NS1 2	2
Segment	7	NC_004907.1	AJ278646.1	1.03	47.8	M2 2 M1 2	2
Segment	5	NC_004905.2	AJ289871.1	1.56	47.3	NP 1	1
Segment	2	NC_004911.1	AJ404634.1	2.33	43.5	PB1 2 PB1-F2 2	2
Segment	4	NC_004908.1	AJ404626.1	1.71	42.5	HA 1	1

**FIGURE 16.8** Schematic of the eight segments from a typical Influenza A virus (from NCBI). Note the link to “protein details” which provides tabular and graphical overviews of the protein content of each genome. The gene names and their corresponding products are NA (neuraminidase), PB2 (polymerase Pb2), PA (polymerase PA), PA-X (PA-X protein), NS2 (nonstructural protein 2), M2 (matrix protein 2), M1 (matrix protein 1), NP (nucleoprotein), PB1 (polymerase Pb1), and HA (hemagglutinin).

Source: NCBI.

in particular combinations that account for the antigenic variation of the virus. These combinations include H1N1, H1N2, and H3N2. The 1918 pandemic was of the H1N1 subtype, while subsequent 1957 and 1968 pandemics were dominated by the H2N2 and H3N2 subtypes, respectively (Fig. 16.9). In the 1957 and 1968 pandemics the viruses resembled human strains into which avian HA, NA, and PB1 molecules became incorporated, while the recent Asian outbreaks are caused by avian strains that infected humans.

Further insight into the structure of influenza virus helps explain its mechanisms of viral transcription and replication (Ruigrok *et al.*, 2010). Moeller *et al.* (2012) performed cryogenic electron microscopy to determine the structure of ribonucleoprotein complexes (including the viral genome, polymerase, and nucleoprotein NP).

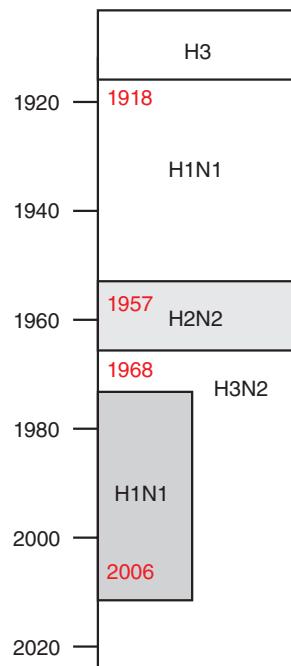
An NIH Influenza Genome Sequencing Project (IGSP) sequenced ~14,000 full influenza genomes (by early 2015), an extraordinary achievement in genomics. All the sequence data are available through GenBank (Bao *et al.*, 2008). This project provides an opportunity to address a range of fundamental questions about influenza viruses (summarized from Holmes, 2009; see also Janies *et al.*, 2010), as follows.

1. Reassortment of viral segments is common, often altering antigenic properties so that vaccines fail. In earlier studies, Ghedin *et al.* (2005) sequenced 209 human influenza A genomes taken from one geographic location (New York State) over a period of several years (1998–2004). They plotted the amino acid positions from 207 viruses as a function of year and presented evidence for segment exchange between viruses. Reassortment among recent H3N2 strains was also reported by Holmes *et al.* (2005). Large-scale surveillance through genome sequencing permits the frequency of mutations and segment exchanges to be estimated, both within human influenza strains and between avian and human subtypes.
2. Multiple, diverse lineages of the same subtype often circulate in human populations, with similar viral diversity in relatively isolated communities as in major cities.
3. East and Southeast Asia likely serve as a global source for human influenza A virus. Determining such sources may facilitate vaccine design.
4. Drug resistance can follow complicated patterns. For example, resistance to a class of antiviral drugs (adamantanes) was caused by a Ser31Asn amino acid substitution in the viral M2 protein encoding an ion channel. This mutation may have been linked to a second mutation elsewhere in the viral genome however, such that drug resistance emerged even in locations where these drugs are only rarely used.
5. It is of interest to relate the viral genomic sequence data (i.e., the genotype) to the clinical presentation (i.e., the phenotype) which may range from subclinical to severe. The continued sequencing of influenza genomes may facilitate genotype–phenotype studies.

Analysis of genomes of avian influenza isolates has yielded important information about the evolution of influenza A genes. Obenauer *et al.* (2006) analyzed 169 complete avian influenza genomes and reported strong positive selection for an alternatively spliced transcript of the PB1 gene (the nonsynonymous to synonymous substitution rate ratio dN/dS was over 9; see Chapter 7). In addition to performing phylogenetic analyses to distinguish emerging viral clades, Obenauer *et al.* described “proteotyping” in which unique amino acid signatures of viral proteins are determined.

Reverse genetics approaches to influenza virus involve introducing targeted mutations into the genome (reviewed in Engelhardt, 2013). In a dramatic effort to understand the nature of the 1918 influenza virus, Jeffery Taubenberger, Terrence Tumpey and colleagues isolated and determined its full genome sequence. Viral nucleic acid was purified from historic samples including an Alaskan woman and several soldiers who died of the 1918 flu. Taubenberger *et al.* (2005) proposed that the 1918 virus was entirely of avian origin (in contrast to the 1957 and 1968 strains that were reassortant viruses). Tumpey *et al.*

The National Institute of Allergy and Infectious Disease (NIAID) at NIH hosts the Influenza Genome Sequencing Project (<http://www.niaid.nih.gov/labsandresources/resources/dmid/gsc/influenza/Pages/default.aspx>, WebLink 16.15). Its genome sequences have been deposited in GenBank and can be accessed through the NCBI influenza virus resource (<http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>, WebLink 16.16) which currently includes 300,000 flu records. This NCBI resource includes tools to select and align influenza virus genome sequences and to produce phylogenetic trees.



**FIGURE 16.9** Summary of influenza A strains. Analysis of archived tissue samples indicates that the H3 strain predominated prior to 1918, while the great pandemic of 1918 was of the H1N1 subtype. Subsequent pandemics were associated with the H2N2 and H3N2 subtypes, while the H1N1 subtype has gained in recent decades. Adapted from Eserink (2006). Used with permission.

(2005) created a viral strain having the complete coding sequences of the eight viral segments of the 1918 virus. They introduced the 1918 virus into mice, where it caused a titer from 125 to 39,000 higher than in mice exposed to a contemporary, less virulent strain. Lethality was 100-fold greater, with all mice dying within six days of infection (but none dying from the less-virulent strain). This work carries considerable risk, but allows analysis of mutations that confer virulence. For example, a mutation found in the polymerase gene PB2 was also found in the virus isolated from a recent fatal case of bird flu involving the H7N7 subtype (von Bubnoff, 2005). Such analyses may aid surveillance efforts as we prepare for the next influenza pandemic (Taubenberger *et al.*, 2007).

## MEASLES VIRUS

Measles virus is one of the deadliest viruses in human history. Today, it is a leading cause of death in children in many countries. In 2008 all World Health Organization (WHO) members agreed to try to reduce measles mortality by 90% over the following decade. Simons *et al.* (2012) report progress from 535,300 deaths (95% confidence interval 347,200–976,400) in 2000 to 139,300 deaths (71,200–447,800) in 2010. Liu *et al.* (2012) described a similar estimate of global burden of death caused by measles of 114,000 (92,000–176,000) in 2010.

Vaccines have helped to reduce the mortality and morbidity rates, but the presence of an immature immune system and maternal antibodies prevent successful immunization in newborns before nine months of age. The virus spreads by respiratory droplets, infecting epithelial cells in the respiratory tract. This disease has been considered a leading vaccine-preventable cause of child mortality (Moss and Griffin, 2006).

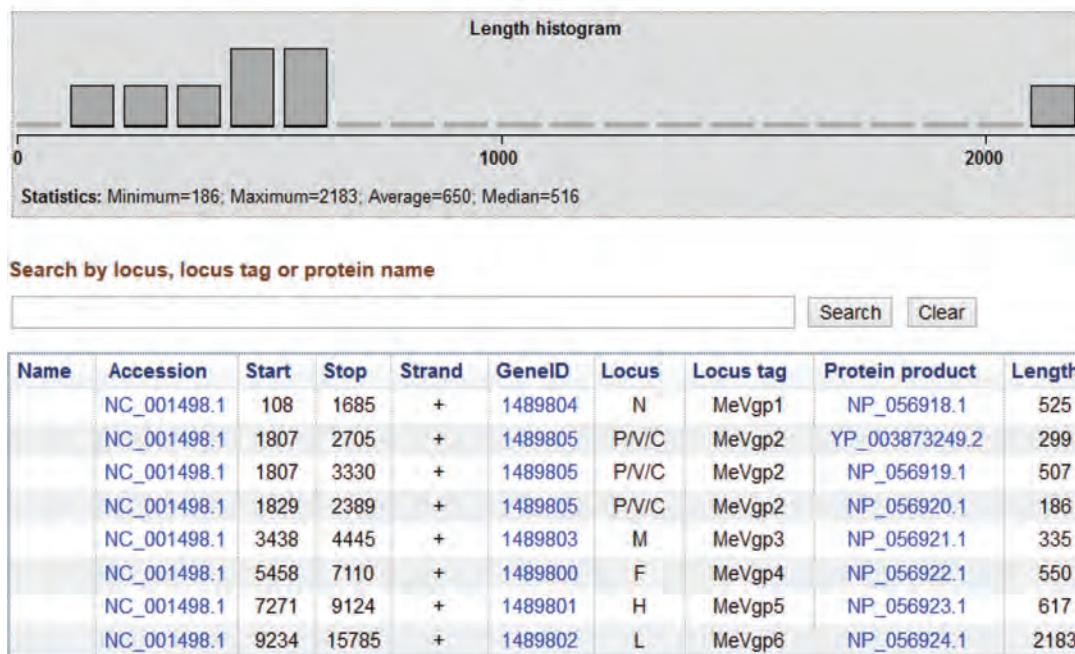
The measles virus is a Morbillivirus of the Paramyxoviridae family, which includes mumps and respiratory syncytial virus. Rota and Bellini (2003) reviewed the worldwide distribution of 14 different measles virus genotypes. You can access a reference genome through the NCBI genomes resource (accession NC\_001498.1). Measles virus consists of a nonsegmented, negative-sense RNA genome protected by nucleocapsids and an envelope. The genome has 15,894 bases and six genes that encode eight proteins. These sequences can be accessed from the NCBI record (Fig. 16.10). Six genes are designated N (nucleocapsid), P (phosphoprotein), M (matrix), F (fusion), H (hemagglutinin), and L (large polymerase). The P gene is predicted to encode a: (1) ~70 kDa phosphoprotein involved in transcription; (2) ~20 kDa protein (nonstructural C protein) using an alternative start site on a different reading frame; and (3) ~46 kDa protein consisting of the amino-terminal region of P and a different, cysteine-rich carboxy terminus. This third protein is generated by editing the measles genome to add one G to three G residues specified by the genome (Cattaneo *et al.*, 1989).

The functions of the measles virus proteins have been assigned: N binds to genomic RNA and surrounds it; P and L form a complex involved in RNA synthesis; L is responsible for replication as well as transcription; M links the ribonucleoprotein to the envelope glycoproteins H and F which are inserted in the virus membrane on the surface of the virion; H binds the cell surface receptor through which the virus enters its host; and F is a fusion protein that promotes insertion of the virus into the host cell membrane. Rima and Duprex (2009) describe the roles of these proteins in the measles virus replication cycle and in transcription. In our discussion of protein structure (Chapter 13) we described intrinsically disordered proteins as lacking a fixed three-dimensional shape. Ferron *et al.* (2006) and Bourhis *et al.* (2006) noted that N and P are both characterized by intrinsic disorder spanning regions of 50–230 residues, contributing to their multiple functions.

The functions of each of these proteins can also be inferred by performing BLAST searches. For the nonstructural C protein, a DELTA-BLAST nonredundant (nr) search reveals homology to proteins encoded by the genomes of rinderpest virus, canine and

Before the measles vaccination was introduced in the United States, there were 450,000 cases annually (and about 450 deaths). See <http://www.cdc.gov/nchs/fastats/measles.htm> (WebLink 16.17).

## Protein Details for Measles virus



**FIGURE 16.10** Eight proteins encoded by six genes of the measles virus genome.

Source: Genome Annotation Report, NCBI Genome, NCBI.

phocine distemper virus, and dolphin morbillivirus. A DELTA-BLAST nr search with the viral hemagglutinin reveals membership in a Pfam family (pfam00423, Hemagglutinin-neuraminidase), and there are several hundred matches to measles virus hemagglutinin. Repeat the search with the Entrez limit “hemagglutinin NOT measles virus[Organism]” and the results are reduced to hemagglutinins from the homologous morbilliviruses other than measles. A DELTA-BLAST search identifies hundreds of additional hemagglutinins from viruses such as human parainfluenza, mumps, and a turkey rhinotracheitis virus.

Another member of the Paramyxoviridae family is rinderpest virus. This causes rinderpest, an ancient plague of cattle and dozens of other domestic and wild artiodactyl species (Barrett and Rossiter, 1999). This virus has had a devastating impact, killing vast numbers of ruminants and leading to human famine. In May 2011 it was announced that rinderpest has been eradicated, making it the second disease (after smallpox) to ever be eradicated (reviewed in Morens *et al.*, 2011; Mariner *et al.*, 2012). You can study the rinderpest genome through nucleotide accession NC\_006296.2 or BioProject accession PRJNA15050.

## EBOLA VIRUS

Ebola is a filovirus that is transmitted between people by contact with body fluids. The first reported outbreak occurred in 1976. The largest outbreak began in 2014, centered initially in West Africa, and has generated worldwide concern about the spread of deadly epidemics. The virus causes hemorrhagic fever that is often fatal. Ebola virus is an enveloped, single-stranded RNA negative-strand virus of the family Filoviridae. The Zaire Ebola virus reference genome is 18,959 bases in length (accession NC\_002549.1), with seven genes encoding nine proteins. The longest of these proteins, named L (accession NP\_066251.1), is an RNA-dependent RNA polymerase sharing 44–73% identity with

The NCBI Ebola virus resource is at <http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola/> (WebLink 16.18). Visit the UCSC Ebola Genome Portal at <http://genome.ucsc.edu/ebolaPortal/> (WebLink 16.19). ExPASy (Chapter 12) offers a description of Ebola virus molecular biology at [http://viralzone.expasy.org/all\\_by\\_species/207.html](http://viralzone.expasy.org/all_by_species/207.html) (WebLink 16.20).

L proteins from other Ebola strains. Virus particles include a nucleocapsid consisting of the RNA genome and viral proteins L, NP and VP30 (two nucleoproteins), and VP35 (a polymerase complex protein). An outer viral envelope includes viral glycoproteins, with VP40 and VP24 (a matrix protein and a membrane protein) localized between the nucleocapsid and the envelope.

Several bioinformatics resources are available to study the Ebola virus genome. NCBI offers an Ebola virus resource including databases of nucleotide and protein sequences and a genome browser. The University of California, Santa Cruz (UCSC) Genome Bioinformatics site includes an Ebola genome browser. That includes Multiz multiple sequence alignments of 160 Ebola virus strains as well as the closely related Marburg virus. Other tracks include data on variants, immune epitope data, as well as links to three-dimensional protein structures. Knowledge of the structure and function of the few genes and proteins that comprise this virus may lead to accelerated vaccine development and antiviral drug development.

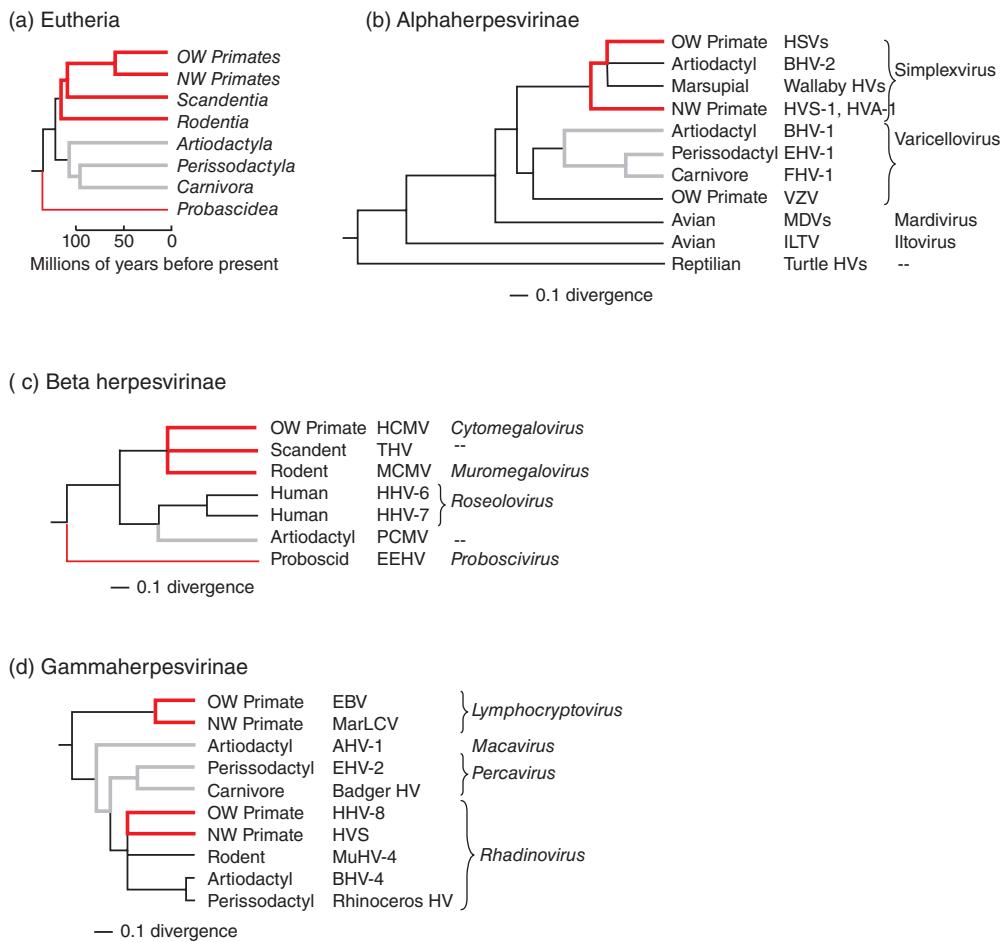
## HERPESVIRUS: FROM PHYLOGENY TO GENE EXPRESSION

RNA viruses such as influenza virus or HIV-1 tend to have small genome sizes with a high evolutionary rate (Fig. 16.3). Infections tend to be acute, virulence can be very high, and cross-species transmission is common (reviewed in Holmes, 2008). DNA viruses such as our next topic, herpesviruses, can have larger genome sizes. They tend to have lower evolutionary rates (e.g.,  $10^{-7}$  to  $10^{-9}$  substitutions per site per year), persistent rather than acute infections, long-term codivergence of viral subtypes with different species, and low virulence.

Herpesviruses are a diverse group of linear, double-stranded DNA viruses that include herpes simplex, cytomegalovirus, and Epstein-Barr virus (McGeoch *et al.*, 2006). The herpesviruses are morphologically distinct from other viruses, having a genome (125–290 kb) packaged in an icosahedral capsid that is further surrounded by a tegument (a proteinaceous matrix) and a lipid envelope.

The herpesviruses have recently been reclassified by the ICTV (Davison *et al.*, 2009; Davison, 2010). The order *Herpesvirales* includes three families: *Herpesviridae* (including mammal, bird, and reptile viruses); *Alloherpesviridae* (piscine and amphibian viruses); and *Malacoherpesviridae* (comprising a lone virus that infects invertebrate bivalves). There are a further 3 subfamilies, 17 genera, and 90 species.

McGeoch *et al.* (2006) analyzed well-conserved genes to deduce a phylogeny of the herpesviruses. Their phylogenetic reconstruction includes three subfamilies:  $\alpha$ -herpesviruses (formally called *Alphaherpesvirinae*);  $\beta$ -herpesviruses (*Betaherpesvirinae*); and  $\gamma$ -herpesviruses (*Gammaherpesvirinae*). This and similar analyses (Davison, 2002; McGeoch *et al.*, 1995) provide great insight into the origin, diversity, and function of herpesviruses. Each herpesvirus is associated with a single host species (although some hosts, including humans, are infected by a variety of herpesviruses). This specificity suggests that herpesviruses have coevolved with their hosts over millions of years. Within each of the three subfamilies, the branching order showing the emergence of various herpesvirus subtypes corresponds to the emergence of the corresponding host organisms (Fig. 16.11). This suggests coevolution of the virus and host lineages. Figure 16.11a shows the timescale for the emergence of major Eutherian (placental mammal) lineages. Figure 16.11b-d shows the three herpesvirus subfamilies with molecular clocks. Note for example that in Figure 16.11b there is a clade (thick red lines) of herpesviruses of the genus *Varicellovirus* (containing artiodactyls, perissodactyl, and carnivore viruses). There is a correspondence of this clade structure to the evolution of those host organisms in Figure 16.11a. McGeoch *et al.* (2006) estimate that the herpesviruses shown in Figure 16.11 arose about 400 million years ago (MYA). Grose (2012) described this co-evolution with



**FIGURE 16.11** Phylogeny of the herpesviruses and comparison to the evolution of host genomes. (a) Phylogenetic tree for eight orders of the *Eutheria* (placental mammals), all of which are hosts to herpesviruses. Three deep clades are indicated in thick red, thin red, and gray. (b) *Alpha-*, (c) *Beta-*, and (d) *Gammaherpesvirinae* are indicated with the hosts and examples of viruses. The divergence scales (in units of substitutions per site) are indicated. NW: New World; OW: Old World. For virus abbreviations see the source of this figure.

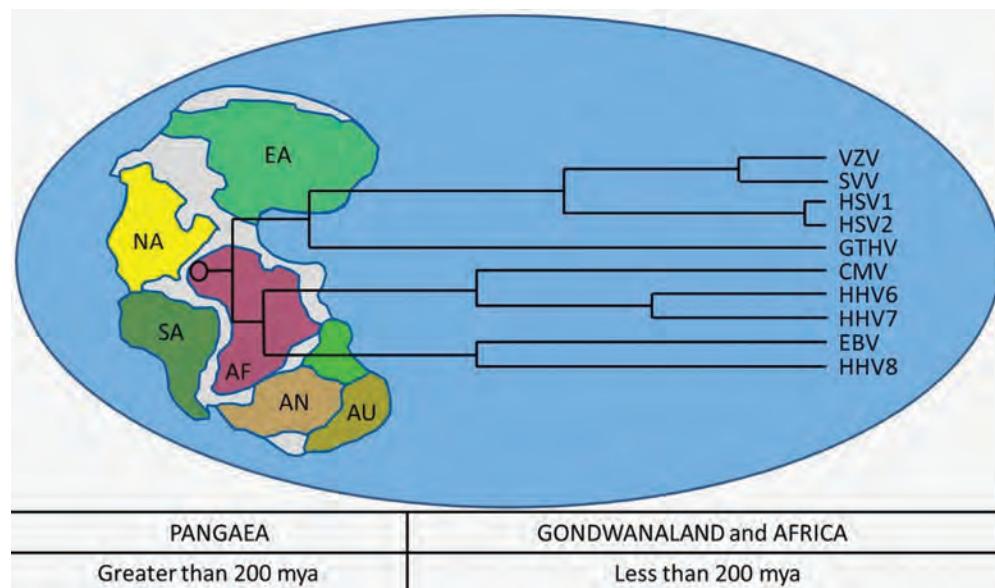
Source: Redrawn from McGeoch *et al.* (2006). Reproduced with permission from Elsevier.

a particular emphasis on varicella-zoster virus and its emergence from Africa (Fig. 16.12). The supercontinent Pangaea separated into Laurasia (to the north) and Gondwana (to the south) ~175 MYA. When Gondwanaland further separated to form Africa and other modern continents, ancestral alphaherpesviruses are hypothesized to have coevolved in primates and subsequently evolved into distinct forms with specificity for assorted primate hosts.

Consider human herpesvirus 8 (HHV-8), a  $\gamma$ -herpesvirus (Fig. 16.11d). HHV-8 is also called Kaposi's sarcoma-associated herpesvirus, and it was initially identified by representational difference analysis in Kaposi's sarcoma lesions of AIDS patients (Chang *et al.*, 1994). HHV-8 causes AIDS-associated Kaposi's sarcoma and other disorders such as primary effusion lymphoma and multicentric Castleman's disease. HHV-8 is closely related to rhesus rhadinovirus (RRV). The divergence of the HHV-8 and RRV may have coincided with speciation of humans and rhesus monkeys (Davison, 2002). The presence of both HHV-8 and an HHV-8-related virus in chimpanzees suggests that an additional RRV-like virus may be identified that infects humans.

What is the molecular basis for the cycle of latent and lytic infection by HHV-8? The genome is about 140,000 bp (NC\_009333.1) and encodes over 80 proteins (Russo

Kaposi's sarcoma is the most common tumor related to AIDS. It is a vascular malignancy that is typically first apparent in the skin.



**FIGURE 16.12** Map of Pangaea, a supercontinent that formed ~400 million years ago (MYA) and separated ~175 MYA into a northern supercontinent (Laurasia) and a southern supercontinent (Gondwanaland, from which Africa derived). According to an out-of-Africa hypothesis, herpesviruses infected marine invertebrates such as oyster and abalone as early as 500 MYA, and various members of the *Herpesviridae* (listed to the right) establishing host species specificity. EA: Europe/Asia; NA: North America; SA: South America; AF: Africa; AN: Antarctica; AU: Australia. Virus abbreviations include CMV (cytomegalovirus), EBV (Epstein-Barr virus), HHV-8 (human herpesvirus 8), and VZV (varicella-zoster virus).

Source: Grose (2012). Reproduced with permission from American Society for Microbiology.

*et al.*, 1996). We can explore the genome at the NCBI Genome (Chapter 15). From a genomes home page you can browse to HHV-8 (taxonomy identifier or txid: 37296) and view an organism summary (Fig. 16.13a). You can further view the open reading frames encoded by its genome in a graphic form or as a table (Fig. 16.13b).

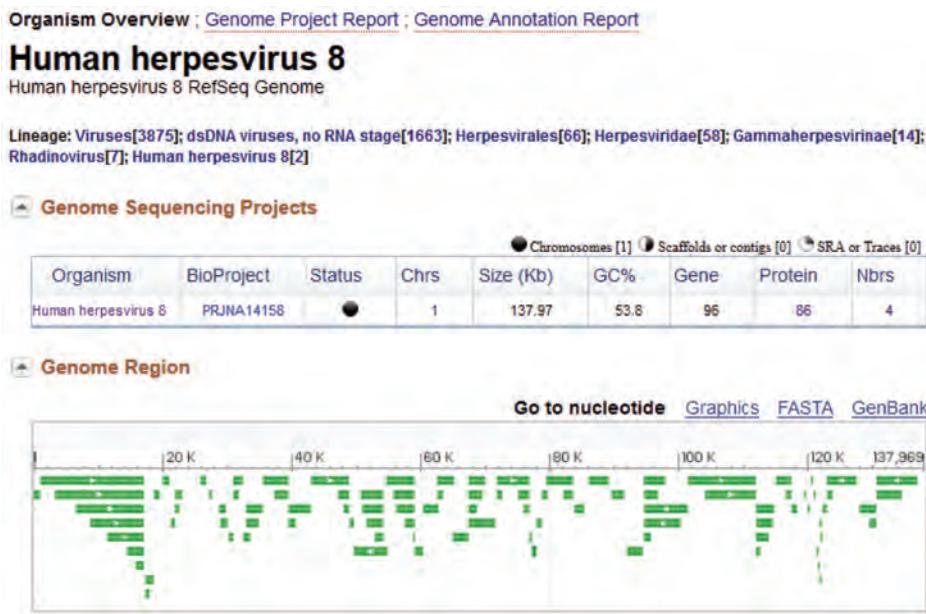
The HHV-8 proteins include virion structural and metabolic proteins. Interestingly, it also contains a variety of viral homologs of human host proteins such as complement-binding proteins, the apoptosis inhibitor Bcl-2, dihydrofolate reductase, interferon regulatory factors, an interleukin 8 (IL-8) receptor, a neural cell adhesion molecule-like adhesin, and a D-type cyclin.

How can viral genomes acquire a motif or an entire gene from a host organism? This can occur by a variety of mechanisms, including recombination, transposition, splicing, translocation, and inversion (McClure, 2000). Consider the IL-8 receptor, encoded by a eukaryotic gene that functions in cell growth and survival. This receptor is a member of the large family of G-protein-coupled receptors, including rhodopsin (that responds to light), the beta-adrenergic receptor (that binds adrenalin), and a variety of neurotransmitter receptors. A DELTA-BLAST search using HHV-8 ORF74 as a query shows that homologs of this protein exist across many vertebrates (Fig. 16.14). Separately, a DELTA-BLAST search restricted to viruses reveals viral homologs of this receptor, including a murine  $\gamma$ -herpesvirus. Presumably, when the virus infects a mammalian cell this viral IL-8 receptor is expressed and confers growth and survival that is advantageous to the virus (Wakeling *et al.*, 2001; Montaner *et al.*, 2013).

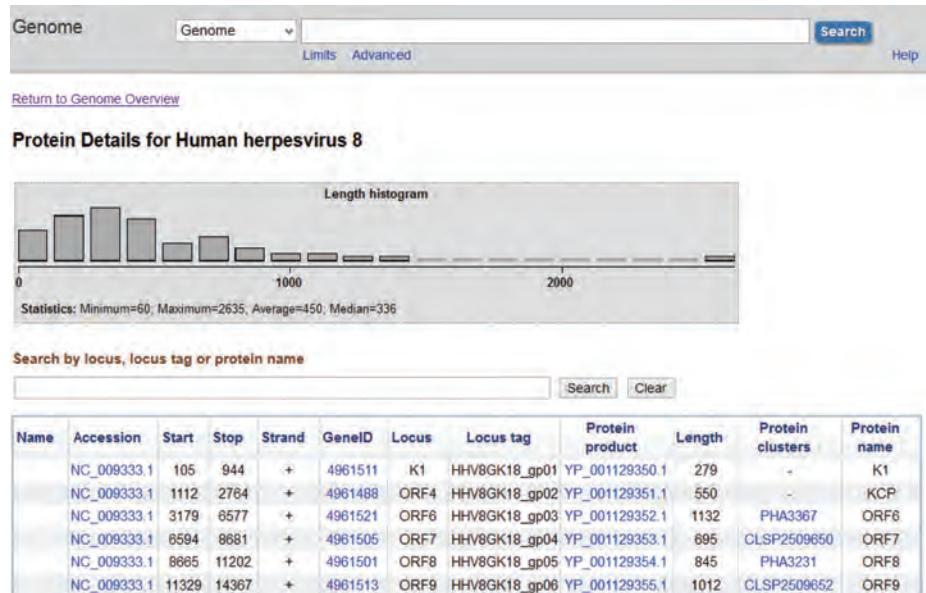
Two complementary approaches have been taken to further study the function of viral genes (such as v-IL-8 receptor) as well as mechanisms of HHV-8 infection. Paulose-Murphy *et al.* (2001) synthesized a microarray that represents 88 HHV-8 open reading frames and measured the transcriptional response of viral genes that are

See Chapter 5 for a description of DELTA-BLAST.

## (a) HHV8 genome overview (NCBI)



## (b) HHV8 proteins (graphic and tabular summaries)



**FIGURE 16.13** HHV-8 data at NCBI. (a) The organism overview for HHV-8 includes links to its BioProject and a graphical representation of its 96 genes. (b) The protein details (linked from the Genome Annotation Report) include a clickable histogram of the proteins as well as a table (showing 6 of the 86 proteins).

Source: NCBI.

activated during the lytic replication cycle of HHV-8 in human cells. They measured gene expression across a time series after inducing lytic infection and described clusters of genes that are coexpressed. Such genes may be functionally related. Clusters of genes coexpressed at early time points include several implicated in activation of the lytic viral cycle; another group of genes encode proteins that function in virion assembly. The viral homologs of human proteins were expressed throughout the induced lytic cycle. Gatherer *et al.* (2011), Marcinowski *et al.* (2012) and Stern-Ginossar *et al.* (2012) performed similar studies on human cytomegalovirus transcriptional profiles

	Description	Query cover	E value	Ident	Accession	
<input type="checkbox"/>	<a href="#">ORF74 [Human herpesvirus 8] &gt;sp Q98146.1 VGPCR_HHV8P RecName: Full=viral G</a>	100%	4e-95	100%	<a href="#">YP_001129433.1</a>	← 1
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 1 [Otolemur garnettii]</a> primate	82%	3e-74	31%	<a href="#">XP_003785131.1</a>	← 2
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Otolemur garnettii]</a>	93%	4e-74	30%	<a href="#">XP_003785015.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Bos mutus]</a> wild yak	91%	6e-74	28%	<a href="#">XP_005889704.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Ailuropoda melanoleuca]</a>	79%	3e-73	31%	<a href="#">XP_002913751.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 1 isoform 1 [Papio anubis] &gt;ref XP_0039</a>	78%	6e-73	30%	<a href="#">XP_003907987.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2 isoform X2 [Equus caballus]</a> horse	79%	1e-72	29%	<a href="#">XP_005610662.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2 isoform X1 [Lctidomys tridecemlineatus]</a>	77%	2e-72	30%	<a href="#">XP_005330510.1</a>	
<input type="checkbox"/>	<a href="#">interleukin 8 receptor alpha [Bos taurus]</a> domestic cow	91%	2e-72	28%	<a href="#">NP_001098508.1</a>	
<input type="checkbox"/>	<a href="#">C-X-C chemokine receptor type 1 [Macaca mulatta]</a> rhesus monkey	78%	2e-72	30%	<a href="#">NP_001035510.1</a>	
<input type="checkbox"/>	<a href="#">C-X-C chemokine receptor type 1 [Oryctolagus cuniculus]</a> rabbit	79%	2e-72	30%	<a href="#">NP_001164553.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 1 isoform X1 [Macaca fascicularis] &gt;ref X</a>	78%	2e-72	30%	<a href="#">XP_005574304.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 1 [Dasypus novemcinctus]</a>	84%	3e-72	29%	<a href="#">XP_004469648.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Ovis aries]</a> sheep	86%	3e-72	30%	<a href="#">XP_004004968.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2 [Jaculus jaculus]</a>	81%	4e-72	31%	<a href="#">XP_004662901.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Sorex araneus]</a>	80%	5e-72	31%	<a href="#">XP_004617206.1</a>	
<input type="checkbox"/>	<a href="#">PREDICTED: c-X-C chemokine receptor type 2-like [Capra hircus]</a> goat	86%	5e-72	30%	<a href="#">XP_005676587.1</a>	

**FIGURE 16.14** A viral protein (HHV-8 open reading frame 74 or ORF74; RefSeq accession YP\_001129433.1) is a G-protein coupled receptor that is homologous to a superfamily of mammalian G-protein coupled receptors, including a high-affinity interleukin 8 (IL-8) receptor. Database matches from a DELTA-BLAST search against the RefSeq database include HHV-8 ORF74 itself (arrow 1) and c-X-C chemokine receptor or interleukin 8 receptor from a variety of vertebrates including the primate *Otolemur garnettii* (arrow 2). The gene encoding this receptor was presumably of mammalian origin and integrated into the genomes of several viruses. Upon viral infection, this receptor may promote growth and survival of infected cells.

Source: NCBI.

Apoptosis is a type of programmed cell death in which the cell actively commits suicide. It serves as a mechanism by which a host cell can destroy infected cells, preventing a pathogen from spreading throughout the body. However, viruses have adapted to manipulate the cellular death pathway. Angiogenesis is the development of blood vessels. Infectious viruses (and cancerous tumors) require the presence of an adequate blood supply and sometimes promote angiogenesis.

using microarrays and RNA-seq. Stern-Ginossar *et al.* extended this approach by combining RNA-seq, which allowed them to identify hundreds of novel cytomegalovirus transcripts, with proteomics techniques such as mass spectrometry and transient expression assays to localize newly identified proteins.

In a second approach to characterizing RNA transcripts, Poole *et al.* (2002) characterized the host cell responses to infection. They infected human dermal microvascular endothelial cells with HHV-8 and measured the transcriptional response of human cells to both latent and lytic virus infection. HHV-8 transforms the endothelial cells from a cobblestone shape to a characteristic spindle shape. Kaposi's sarcoma is associated with many additional pathological features, including angiogenesis and immune dysregulation. The endothelial genes regulated by HHV-8 infection included those such as interferon-responsive genes involved in immune function and genes encoding proteins with roles in cytoskeletal function, apoptosis, and angiogenesis. Such studies may be useful in defining the cellular response to viral infection.

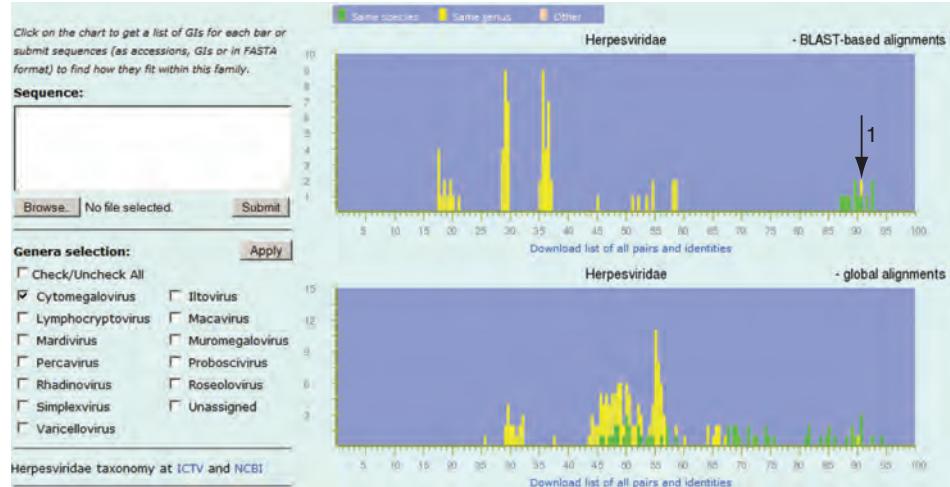
### The Pairwise Sequence Comparison (PASC) Tool

NCBI offers the Pairwise Sequence Comparison (PASC) tool to help classify viruses in a broad range of families or genera (Bao *et al.*, 2012). For a range of complete virus

genomes, it uses two methods to compute relatedness between viruses: local alignment using BLAST; and global alignment using the Needleman–Wunsch algorithm (Chapter 3).

We can demonstrate the use of PASC by selecting Herpesviridae and choosing the genus Cytomegalovirus. The genomes include members of the same genus and species (shaded green in Fig. 16.15a) or the same genus but a different species (shaded yellow;

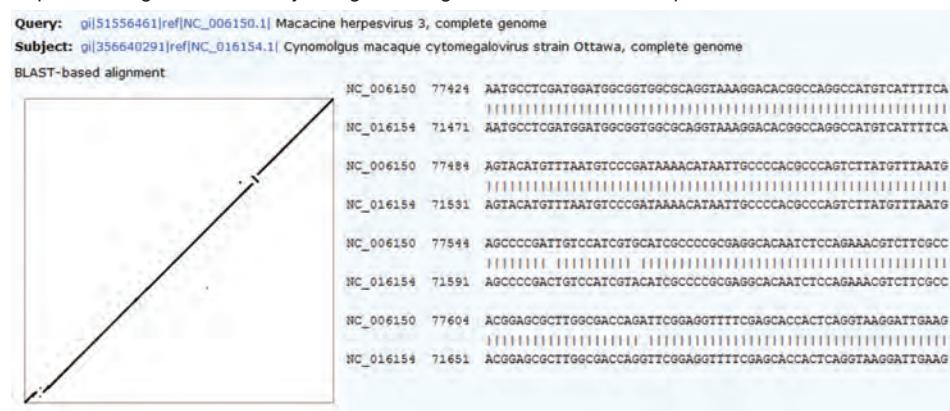
(a) PASC tool at NCBI for comparison of viral genomes



(b) Report on a genome pair (from arrow 1)



(c) Dot plot and alignment of two cytomegalovirus genomes from different species



**FIGURE 16.15** The NCBI PASC tool allows comparisons of complete viral genomes from a variety of families. (a) Data for the genus cytomegalovirus of the Herpesviridae. Relationships based on BLAST (upper panel) or Needleman–Wunsch global alignment (lower panel) are shown. The x axis shows percent identity of the alignments. Green bars represent pairs of genomes that are assigned to the same species (87–93% identity in the BLAST-based alignments). Yellow bars represent genome pairs of the same genus that are classified as different species (17–59% identity, with an outlier having 91% identity indicated by arrow 1). (b) That outlier can be clicked on, providing annotation data. (c) Pairwise alignment of the two genomes indicated by arrow 1, including a dotplot to visualize their relatedness.

Source: PASC, NCBI.

PASC is available from the NCBI Genome home page at <http://www.ncbi.nlm.nih.gov/genome> (WebLink 16.21).

arrow 1 highlights a genome comparison in that category involving two genomes). Clicking that entry shows that these two genomes share 91% identity (Fig. 16.15b), and the pairwise alignment can also be displayed (Fig. 16.15c). In this instance, it might be appropriate for these two genomes sharing such high nucleotide identity to be classified in the same species. The usefulness of the PASC tool is in exploring such relationships. Notably, the results for global pairwise alignment tend to inflate percent identities (at least partly because two random sequences of the same length are expected to share 25% nucleotide identity by chance). The scores therefore tend to be shifted to the right for global relative to local alignments in Figure 16.15a.

## GIANT VIRUSES

Historically, from the first characterization of viruses in the late 1800s to the modern era of molecular biology, we have thought of viruses as extremely small entities. That view has been firmly shifted by the recent description of a group of giant viruses. They are exceptional in terms of the size of the viral particle (sometimes 750 nanometers or nm in diameter), such that they are the only viruses visible by conventional light microscopy. They are also exceptional in terms of the size of the genome, which tends to be greater than 1 megabase (Table 16.6). Colson *et al.* (2013a) proposed the introduction of a new viral order, *Megavirales*; previously, these were known as nucleocytoplasmic large DNA viruses (NCLDVs). The first to be discovered was named Mimivirus for “microbe mimicking.” This giant virus infects phagocytic protists such as the amoebae *Acanthamoeba polyphaga*.

*Acanthamoeba polyphaga* Mimivirus has mature particles that are an enormous 400 nm in diameter, with an outer layer of dense fibrils that brings the total diameter to 750 nm. Seibert *et al.* (2011) determined its structure by X-ray diffraction, even though the virus represented a noncrystalline sample. Its genome size of 1.2 Mb is larger than that of many bacteria (the *Mycoplasma genitalium* genome is 580 kb) and archaea (the *Nanoarchaeum equitans* genome is 490 kb), and it is almost half the size of the smallest eukaryotic genome (that of *Encephalitozoon cuniculi*, 2.5 Mb). Raoult *et al.* (2004) characterized its genome. Of its 1262 open reading frames of length ≥100 amino acids, just 194 had similarity to proteins of known function.

The Mimivirus GenBank accession number is NC\_014649.1.

In the decade since its discovery, more giant viruses have been described and had their genomes sequenced (Table 16.6). These include *Acanthamoeba polyphaga moumouvirus* (Yoosuf *et al.*, 2012), Marseillesvirus (Boyer *et al.*, 2009), *Megavirus chilensis* (Arslan *et al.*, 2011), Lausannevirus (Thomas *et al.*, 2011), *Acanthamoeba castellanii mamavirus* (Colson *et al.*, 2011), and Courdo11 virus (Yoosuf *et al.*, 2014). The largest of

**TABLE 16.6 Largest virus genomes. All are double-stranded DNA (no RNA stage). The order *Megavirales* has been proposed to reflect the genome size of at least 1 megabase.**

Genus, species	Accession	Base pairs
<i>Acanthamoeba polyphaga moumouvirus</i>	NC_020104.1	1,021,348
<i>Acanthamoeba polyphaga mimivirus</i>	NC_014649.1	1,181,549
<i>Acanthamoeba castellanii mamavirus</i>	JF801956.1	1,191,693
<i>Megavirus chilensis</i>	NC_016072.1	1,259,197
<i>Pandoravirus dulcis</i>	NC_021858.1	1,908,524
<i>Pandoravirus salinus</i>	NC_022098.1	2,473,870

the recently described viruses are *Pandoravirus salinus* and *Pandoravirus dulcis* (Philippe *et al.*, 2013). *P. salinus* was discovered off the coast of central Chile where it infects *Acanthamoeba castellanii*. Using Illumina, 454 and Pacific Biosciences next-generation sequencing, the 2.5 Mb genome was sequenced and assembled, although the presence of repeats suggests a minimum genome size of 2.8 Mb. The *P. dulcis* virus, discovered in a freshwater pond in Australia, has a genome size of 1.9 Mb. These two viral genomes are predicted to encode ~2500 and ~1500 proteins, respectively.

The discovery of *Megavirales* is significant in several ways.

- It redefines the nature of viruses, both in terms of genome size and gene content. While these large viruses do not produce ribosomes, they do encode some aminoacyl-tRNA synthetases and related proteins, hinting at functions beyond those normally attributed to viruses. It is notable that smaller relatives of the family *Mimiviridae* have been identified as sharing gene families with their larger relatives (Yutin *et al.*, 2013).
- The question arises as to the origin of these viruses. Some of the *Megavirales* genes (perhaps 15%) have been acquired by lateral gene transfer (Chapter 17). Other genes might have been acquired from host (e.g., amoeba) genomes, but this does not seem to have occurred appreciably (<1%). Jean-Michel Claverie and colleagues therefore suggested that Mimivirus and related viruses could be descended from a cellular organism in a separate branch of life. That organism, it is hypothesized, lost most of its genes (Legendre *et al.*, 2012, reviewed in Pennisi, 2013). Those that remain include many with no detectable homology to eukaryotes, bacteria, archaea, or even to other viruses.
- In addition to giant viruses that infect amoebae, a pathogenic delta-proteobacterium called BABL1 (accession GQ495224.1) has been identified that invades the same amoebae (Slimani *et al.*, 2013). There is also a small virus that infects *Acanthamoeba polyphaga Mimivirus* within the amoeba. This virus is referred to as a virophage (a virus that infects another virus; Slimani *et al.*, 2013).
- As we discover more of these viruses globally, it is possible they will be found to infect other eukaryotes. Colson *et al.* (2013b) provide evidence for *Megavirales* in human stool and metagenome samples.

Jean-Michel Claverie and colleagues searched for giant viruses in a region of the Siberian permafrost that has been dated as 30,000 years old. Using *Acanthamoeba castellanii* as a bait they identified a virus named *Pithovirus sibericum* that has a 1.5 µm particle length (1500 nm) and a 610,033 base pair, AT-rich (64%) double-stranded DNA genome (Legendre *et al.*, 2014). This extends the possibilities of the locations of giant viruses, and demonstrates that viruses from the deep past can be revived. The consequence for the health of humans and other organisms susceptible to infection today is unknown.

The accession number of *P. sibericum* is NC\_023423.1.

## Comparing genomes with MUMmer

A major challenge in aligning genomes (whether viral, bacterial, archaeal, or eukaryotic) is the excessive amount of time required to perform an alignment of millions of base pairs using dynamic programming (Chapter 3). We introduced several fast algorithms such as BLAT in Chapter 5. However, additional tools to accomplish large-scale genome alignment are needed. MUMmer is a software package that offers rapid, accurate alignments of microbial genomes (Delcher *et al.*, 1999). It has been adapted to aligning eukaryotic sequences (Delcher *et al.*, 2002; Kurtz *et al.*, 2004).

MUMmer was written by Steven Salzberg and colleagues. You can download the software from <http://mummer.sourceforge.net/> (WebLink 16.22).

MUMmer accepts two sequences as input. The algorithm finds all subsequences that are longer than a specified minimum length  $k$  and that are perfectly matched. By definition, these matches are maximal because extending them further in either direction causes

a mismatch. The algorithm uses a suffix tree, which is a search structure that identifies all the maximal unique matches (“MUM’s) in the pairwise alignment. The MUMs are ordered, and the algorithm closes gaps by identifying large inserts, repeats, small mutated regions, and single-nucleotide polymorphisms (SNPs).

MUMmer output consists of a dot matrix plot showing the alignment of the two genomic sequences with some minimum alignment length (e.g., 15 or 100 bp). The kinds of results that can be obtained include:

1. SNPs;
2. regions where sequences diverge by more than a SNP;
3. large insertions (e.g., by transposition, sequence reversal, or lateral gene transfer);
4. repeats (e.g., a duplication in one genome); or
5. tandem repeats (in different copy number).

Let’s compare *Acanthamoebapolyphagamimivirus* (about 1.2 Mb) with *Acanthamoeba castellanii mamavirus* and then with *Acanthamoeba polyphaga moumouvirus* using MUMmer. First, install MUMmer on your Linux or Mac OS/X computer.

Next, we obtain FASTA formatted files containing the mimivirus, mamavirus, and moumouvirus DNA sequences. You can use the accession numbers given in **Table 16.6**. Sequences can also be downloaded from NCBI by FTP, or the NCBI Nucleotide page offers a “send to” option to send the FASTA format of the sequence to a file which can then be transferred to a directory on a Linux machine. We place them in a directory called data with the names `mimivirus.fasta`, `mamavirus.fasta`, and `moumouvirus.fasta`.

We then use MUMmer to compare two virus sequences. To see the basic commands, view the help documentation:

```
$ mummer -h
```

To compare the first two sequences, type:

```
$ mummer -mum -b -c ~/data/mimivirus.fasta ~/data/mamavirus.fasta > ~/data/mimimama.mums
```

Here the `-mum` command computes maximal matches that are unique in both sequences, the `-b` command computes both forward and reverse complement matches, the `-c` command reports query positions of a reverse complement relative to the original query sequence, and the `>` symbol indicates that we will specify the name of the output file. The output includes a dotplot showing that these two viral genomes are largely collinear, that is, there are MUMs shaded red with a positive slope indicating well-aligned segments (**Fig. 16.16a**). This is consistent with the analyses of Colson *et al.* (2011) who reported ~99% nucleotide identity between these aligned genomes.

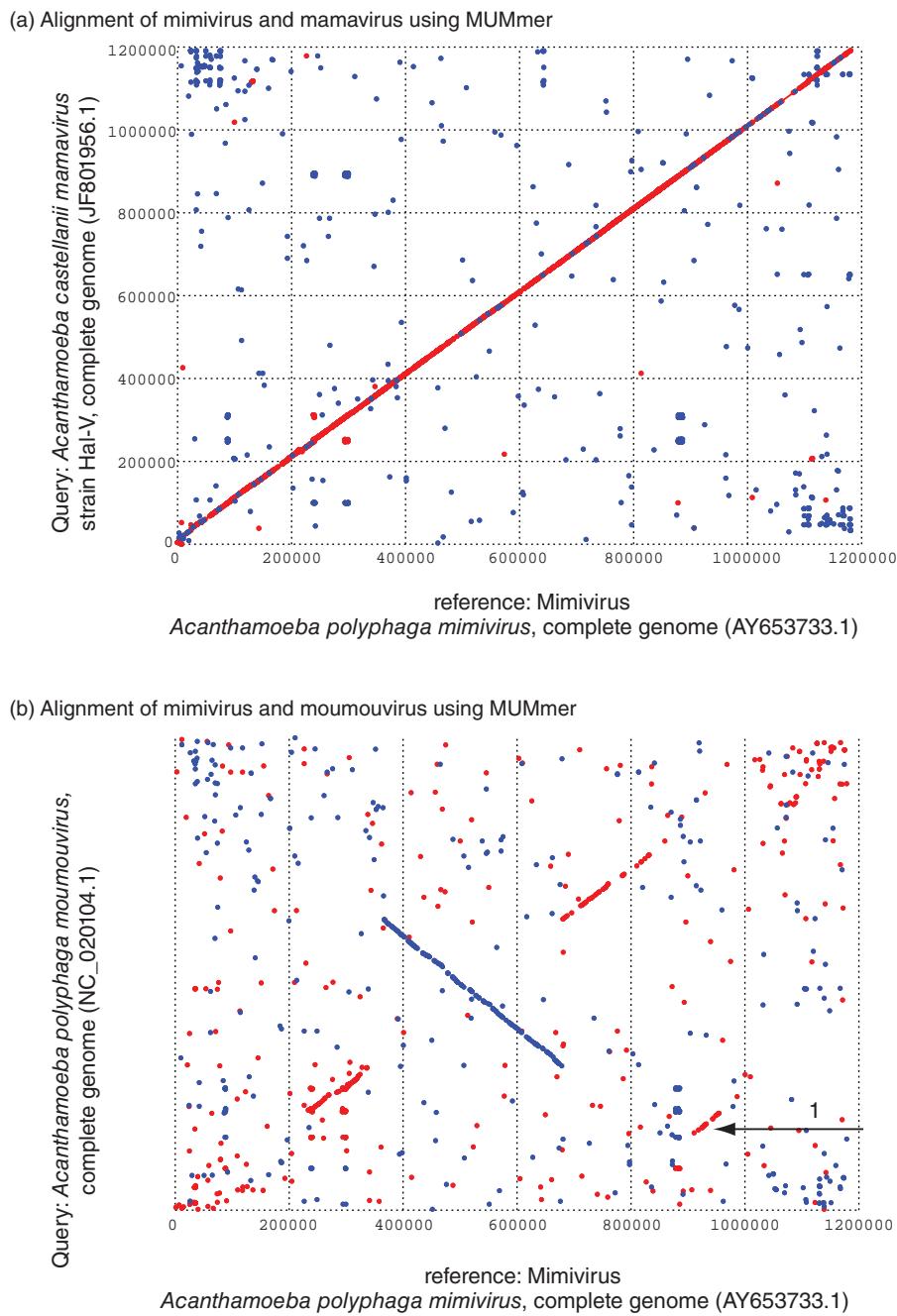
We can repeat the comparison of the genomes with mimivirus versus moumouvirus:

```
$ mummer -mum -b -c ~/data/mimivirus.fasta ~/data/moumouvirus.fasta > ~/data/mimimoumou.mums
```

Here, the output shows limited collinearity (red segments) with a prominent inversion (shaded blue; **Fig. 16.16b**). Additionally, a forward (red) segment that is displaced from the main diagonal indicates a translocation (arrow 1). Yoosuf *et al.* (2012) reported a similar plot that was instead based on orthologous proteins from BLASTP searches, revealing the same inversion and overall extent of collinearity.

MUMmer software is useful for many kinds of genomic comparisons, and we will return to it in Chapters 17 and 19.

If you are working on a Mac OS/X, after downloading you will need the `make` program and other dependencies (listed in the MUMmer documentation; WebLink 16.7). These are currently not included on Mac OS/X, and to obtain `make` and other programs you will first need to install Xcode software.



**FIGURE 16.16** Comparison of megabase-size viral genome sequences using MUMmer software.  
 (a) Comparison of *Acanthamoeba polyphaga* Mimivirus (reference, x axis) versus the query *Acanthamoeba castellanii* mamavirus (y axis). Note that the two genomes are largely collinear.  
 (b) Comparison of *Acanthamoeba polyphaga* Mimivirus (reference) versus *Acanthamoeba polyphaga* moumouvirus (query). Forward MUMs are indicated in red, while reverse MUMs are colored blue. A prominent inversion is evident near the middle of the two genomes as well as a translocation (arrow 1). Created using MUMmer.

## PERSPECTIVES

Several thousand species of viruses are known. In contrast, there may be tens or hundreds of millions of species of bacteria and archaea (Chapter 15) and perhaps tens of millions of eukaryotic species (Chapters 18–20). There are probably relatively few species of viruses because of their specialized requirements for replication in host cells.

Essentially, all the bioinformatic tools that are applied to eukaryotic or bacterial protein and nucleic acid sequences are also applicable to the study of viruses (Kellam, 2001).

- BLAST, DELTA-BLAST, and other database searches may be applied to define the homology of viral sequences to other molecules.
- Microarrays have been used to represent viral genes, and now RNA-seq also allows an assessment of viral gene transcription during different phases of the viral life cycle.
- In independent approaches, the transcriptional response of host cells to viral infection has begun to be characterized.
- Structural genomics approaches to viruses result in the identification of three-dimensional structures of viral proteins. Some structures are solved in the presence of pharmacological inhibitors. The Entrez protein division of NCBI currently includes over 6700 virus structural records.

For some viruses such as HIV, molecular studies have permitted detailed studies of phylogeny to complement knowledge of viral life cycles and pathogenesis. At the same time genomics has not yet contributed to the challenge of successfully creating a vaccine.

## PITFALLS

Viruses evolve extremely rapidly, in large part because some RNA virus polymerases tend to operate with low fidelity. It is for this reason that a person infected with HIV may harbor millions of distinct forms of the virus, each with its own unique RNA sequence. It may therefore be difficult to define a single canonical sequence for some viruses. This complicates attempts to study the evolution of viruses and gene function, or to develop vaccines.

While the tree of life has been described using rRNA or other sequences (Chapter 15), viruses are almost entirely absent from this tree. This is because there are no genes or proteins that all viruses share in common with other life forms or with each other.

## ADVICE FOR STUDENTS

Viruses impact all of our lives, and in a real sense they threaten us. Just a century ago, a single influenza epidemic killed many tens of millions of people. As discussed in this chapter, viral diseases such as measles continue to kill and cause immense suffering. To actively approach this topic, choose a viral genome that interests you the most, and try the following. (1) Read the primary literature on its genome. I recommend a review on the measles virus replication cycle by Rima and Duprex (2009); while the focus is more on biochemistry and virology than on bioinformatics, the article explains the extraordinary properties of the tiny measles virus and the challenges it faces. (2) Explore its genome in depth. For measles use NCBI; for HIV the LANL site offers a wide range of resources. Use the tools we learnt in Part I, such as database searching, to predict the function of poorly characterized genes.

A separate exercise is to integrate your studies of viruses with next-generation sequencing. Download the collection of all virus DNA sequences in the FASTA format. Choose one or more human whole-genome sequences, and align the short reads to your virus reference database. What do you find?

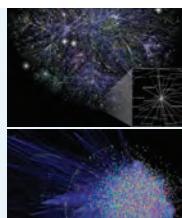
## WEB RESOURCES

We have focused on ICTVdb, NCBI, and LANL tools. Many specialized databases have been established for the study of viruses including those listed in **Table 16.7**. For example, the Viral Bioinformatics Resource Center offers software tools, including the Viral

**TABLE 16.7** Virus resources available on web.

Resource	Description	URL
ICTVdb	Universal virus database	<a href="http://ictvonline.org/">http://ictvonline.org/</a>
All the Virology on the WWW	Provides many virology links and resources	<a href="http://www.virology.net/">http://www.virology.net/</a>
The Big Picture Book of Viruses	General virus resource	<a href="http://www.virology.net/Big_Virology/BVHomePage.html">http://www.virology.net/Big_Virology/BVHomePage.html</a>
Virus Particle ExploreER (VIPER)	High-resolution virus structures in the Protein Data Bank (PDB)	<a href="http://viperdb.scripps.edu/">http://viperdb.scripps.edu/</a>
Viral Bioinformatics Resource Center	Databases and software to analyze viruses	<a href="http://athena.bioc.uvic.ca/">http://athena.bioc.uvic.ca/</a>
Virusworld	A research institute at the University of Wisconsin-Madison	<a href="http://www.virology.wisc.edu/virusworld/viruslist.php">http://www.virology.wisc.edu/virusworld/viruslist.php</a>
Stanford HIV Drug Resistance Database	A curated database with information on drug targets	<a href="http://hivdb.stanford.edu/">http://hivdb.stanford.edu/</a>

Genome Organizer, for the graphical display of viral sequences (Upton *et al.*, 2000). This site also contains a Viral Genome DataBase (VGDB) with analyses of the properties of viral genomes such as GC content. The Stanford HIV RT and Protease Sequence Database offers an algorithm that can be queried with an input viral DNA sequence (Rhee *et al.*, 2003). The output describes possible mutations in the viral gene and an interpretation of likely susceptibility of that protein to drug resistance.



## Discussion Questions

**[16-1]** There is no comprehensive molecular phylogenetic tree of all viruses. Why not?

**[16-2]** If you wanted to generate phylogenetic trees that are as comprehensive as possible, using DNA or RNA or protein sequences available in GenBank, what molecule(s) would you select? What database(s) would you search?

**[16-3]** In a metagenomic study, Cox-Foster *et al.* (2007) determined DNA sequences associated with colony collapse disorder, a recent phenomenon in which honey bee colonies collapse. This now affects about a quarter of bee-keeping operations in the United States. RNA samples were collected from hives that are either affected or not, and pyrosequencing was performed. In addition to bacterial and fungal sequences, a group of RNA viruses were identified including one (Israeli acute paralysis virus) associated with risk for colony collapse disorder. What criteria would you use to decide if this virus has a causal role in the disorder?

### PROBLEMS/COMPUTER LAB

**[16-1]** We mention colony collapse disorder in the discussion question above. The accession for Israeli acute paralysis virus is NC\_009025, and it is a picorna-like virus. How many proteins does this genome encode? What are their functions?

**[16-2]** NCBI offers the PopSet resource at <http://www.ncbi.nlm.nih.gov/popset> (WebLink 16.23). PopSet collects DNA sequences for evolutionary analyses of populations. Enter a query for megavirus in the home page of NCBI and link to a set of 22 megavirus-related polymerase sequences (“DsDNA viruses, no RNA stage B-family DNA polymerase gene, complete cds.”). Choose the “Send to” option to download the sequences in the FASTA format, and input them to MEGA (Chapter 7). Perform a multiple sequence alignment with MUSCLE within MEGA and perform phylogenetic analyses of this gene family.

**[16-3]** This problem introduces you to finding how many proteins are associated with viruses. (1) How many HIV-1 proteins are in the NCBI Protein resource? (2) Given the

tremendous heterogeneity of HIV-1, you might expect there to be thousands of variant forms of each protein. How many are actually assigned RefSeq accession numbers? (3) How many measles virus RefSeq proteins are there? (4) Query NCBI Genome with the search term “measles” and view the Genome Annotation Report. There is a link to “see protein details.” What are the sizes of the smallest and largest measles proteins?

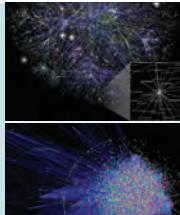
**[16-4]** Find an HIV-1 protein with a RefSeq identifier in NCBI Protein (such as the Vif protein, NP\_057851; you should select your own example). Perform a BLASTP search with it, and inspect the results using the taxonomy report. Next repeat the search, excluding HIV from the output. How broadly is the gene or protein you selected represented among viruses? Do you expect some genes to be HIV-specific while other genes are shared broadly by viruses?

**[16-5]** Analyze a set of influenza viruses using the NCBI Influenza Virus Resource (<http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>, WebLink 16.24). (1) Click tree to begin choosing sequences. Select the virus species (Influenza A), host (human), country/region (e.g., Europe), and segment (HA). Include the options of full-length sequences only, and remove identical sequences.

Click “Get sequences.” (2) Construct a multiple sequence alignment and phylogenetic tree. Use neighbor-joining. In the case of HA, does the tree form clades corresponding to H1N1, H3N2, and H7N7 subtypes? Optionally, export the sequences in the FASTA format, perform your own multiple sequence alignments using MAFFT or MUSCLE (Chapter 6), then import the alignment into MEGA (or other software) to perform phylogenetic analyses yourself.

**[16-6]** Analyze HIV sequences at the HIV Sequence Database (<http://www.hiv.lanl.gov/>, WebLink 16.25). Select the search interface, then choose genomic regions with the Vif coding sequence (Vif CDS). Restrict the output to ten sequences. Select these, and click “Make tree.” Include the reference sequences HXB2. Choose a distance model (the default is Felsenstein 1984) and either equal site rates or a gamma distribution. How many clades do you observe? What do these clades represent? Note that you can download the multiple sequence alignment used to generate the tree to perform further phylogenetic analyses.

**[16-7]** Select a reference sequence of a virus and a strain that has undergone reassortment. Align the genomic sequences using MUMmer as outlined in this chapter.



## Self-Test Quiz

**[16-1]** There are several thousand known viruses, while there are many millions of bacterial, archaeal, and eukaryotic species. The most likely explanation for the small number of viruses is that:

- (a) we have not yet learned how to detect most viruses;
- (b) we have not yet learned how to sequence most viruses;
- (c) there are few viruses because their needs for survival are highly specialized; or
- (d) viruses use an alternative genetic code.

**[16-2]** RNA viruses, when compared to DNA viruses, tend to:

- (a) be less virulent;
- (b) be less persistent;
- (c) be less mutable; or
- (d) have larger genome sizes.

**[16-3]** The HIV genome contains nine protein-coding genes. The number of GenBank nucleotide accession numbers for these nine genes is approximately:

- (a) 9;
- (b) 900;
- (c) 9000; or
- (d) >600,000.

**[16-4]** For functional genomics analyses of viruses, it is possible to measure gene expression:

- (a) of viral genes upon viral infection of human tissues;
- (b) of human genes upon viral infection of human tissues; or
- (c) of viral genes and human genes, measured upon viral infection of human tissue.

**[16-5]** Herpesviruses probably first appeared about:

- (a) 500 million years ago;
- (b) 5 million years ago;
- (c) 50,000 years ago; or
- (d) 500 years ago.

**[16-6]** HIV-1 in its present form probably first appeared about:

- (a) 70 million years ago;
- (b) 7 million years ago;
- (c) 700,000 years ago; or
- (d) 70 years ago.

**[16-7]** Phylogeny of HIV virus subtypes:

- (a) establishes that HIV emerged from a cattle virus;

- (b) can be used to develop vaccines directed against ancestral protein sequences; or
- (c) establishes which human tissues are most susceptible to infection.

**[16-8]** Specialized virus databases such as that at Oak Ridge National Laboratory offer resources for the study of HIV that are not available at NCBI or EBI. An example is:

- (a) a listing of thousands of variant forms for each HIV gene;
- (b) a listing of literature and citations from the previous week;
- (c) graphical displays of the genome; or
- (d) a description of where HIV variants have been identified across the world.

## SUGGESTED READING

Gibbs (2013) discusses new ICTV changes to virus species nomenclature. Duffy *et al.* (2008) provide an exceptional review on the evolution of viruses, including rates of mutation and substitution. Jeffrey Gordon and colleagues (Reyes *et al.* 2012) provide an outstanding overview of viral metagenomics with respect to the human microbiome. For viral metagenomics methods, see Willner and Hugenholtz (2013). Edward Holmes (2008) reviews the evolutionary history and phylogeography of human viruses, highlighting the differences between RNA viruses and DNA viruses. For an overview of HIV, particularly from a molecular phylogenetics perspective, see Castro-Nallar *et al.* (2012).

## REFERENCES

- Ackermann, H.W. 2007. 5500 Phages examined in the electron microscope. *Archive of Virology* **152**, 227–243. PMID: 17051420.
- Adams, M.J., Lefkowitz, E.J., King, A.M., Carstens, E.B. 2013. Recently agreed changes to the International Code of Virus Classification and Nomenclature. *Archive of Virology* **158**(12), 2633–2639. PMID: 23836393.
- Arslan, D., Legendre, M., Seltzer, V., Abergel, C., Claverie, J.M. 2011. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Science, USA* **108**(42), 17486–17491. PMID: 21987820.
- Bao, Y., Bolotov, P., Dernovoy, D. *et al.* 2008. The influenza virus resource at the National Center for Biotechnology Information. *Journal of Virology* **82**(2), 596–601. PMID: 17942553.
- Bao, Y., Chetverin, V., Tatusova, T. 2012. Pairwise Sequence Comparison (PASC) and its application in the classification of filoviruses. *Viruses* **4**(8), 1318–1327. PMID: 23012628.
- Barré-Sinoussi, F., Ross, A.L., Delfraissy, J.F. 2013. Past, present and future: 30 years of HIV research. *Nature Reviews Microbiology* **11**(12), 877–883. PMID: 24162027.
- Barrett, T., Rossiter, P. B. 1999. Rinderpest: The disease and its impact on humans and animals. *Advances in Virus Research* **53**, 89–110. PMID: 10582096.
- Bernal, J. D., Fankuchen, I. 1941. X-ray and crystallographic studies of plant virus preparations. *Journal of General Physiology* **25**, 111–146. PMID: 19873255.

- Bibby, K. 2013. Metagenomic identification of viral pathogens. *Trends in Biotechnology* **31**(5), 275–279. PMID: 23415279.
- Bourhis, J.M., Canard, B., Longhi, S. 2006. Structural disorder within the replicative complex of measles virus: functional implications. *Virology* **344**(1), 94–110. PMID: 16364741.
- Boyer, M., Yutin, N., Pagnier, I. *et al.* 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Science, USA* **106**(51), 21848–21853. PMID: 20007369.
- Breitbart, M., Hewson, I., Felts, B. *et al.* 2003. Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology* **185**, 6220–6223. PMID: 14526037.
- Briggs, J.A., Kräusslich, H.G. 2011. The molecular architecture of HIV. *Journal of Molecular Biology* **410**(4), 491–500. PMID: 21762795.
- Brister, J.R., Bao, Y., Kuiken, C. *et al.* 2010. Towards Viral Genome Annotation Standards, Report from the 2010 NCBI Annotation Workshop. *Viruses* **2**(10), 2258–2268. PMID: 21994619.
- Burnet, M. 1953. Virus classification and nomenclature. *Annals of the New York Academy of Sciences* **56**(3), 383–390. PMID: 13139240.
- Castro-Nallar, E., Pérez-Losada, M., Burton, G.F., Crandall, K.A. 2012. The evolution of HIV: inferences using phylogenetics. *Molecular Phylogenetics and Evolution* **62**(2), 777–792. PMID: 22138161.
- Cattaneo, R., Kaelin, K., Baczko, K., Billeter, M.A. 1989. Measles virus editing provides an additional cysteine-rich protein. *Cell* **56**(5), 759–764. PMID: 2924348.
- Chang, Y., Cesarman, E., Pessin, M.S. *et al.* 1994. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869. PMID: 7997879.
- Cieslak, T. J., Christopher, G. W., Ottolini, M. G. 2002. Biological warfare and the skin II: Viruses. *Clinical Dermatology* **20**, 355–364. PMID: 12208623.
- Ciuffi, A., Telenti, A. 2013. State of genomics and epigenomics research in the perspective of HIV cure. *Current Opinion in HIV and AIDS* **8**(3), 176–181. PMID: 23426238.
- Colson, P., Yutin, N., Shabalina, S.A. *et al.* 2011. Viruses with more than 1,000 genes: Mamavirus, a new *Acanthamoeba polyphaga mimivirus* strain, and reannotation of Mimivirus genes. *Genome Biology and Evolution* **3**, 737–742. PMID: 21705471.
- Colson, P., De Lamballerie, X., Yutin, N. *et al.* 2013a. “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Archive of Virology* **158**(12), 2517–2521. PMID: 23812617.
- Colson, P., Fancello, L., Gimenez, G. *et al.* 2013b. Evidence of the megavirome in humans. *Journal of Clinical Virology* **57**(3), 191–200. PMID: 23664726.
- Cox-Foster, D.L., Conlan, S., Holmes, E.C. *et al.* 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* **318**, 283–287. PMID: 17823314.
- Culley, A.I., Lang, A.S., Suttle, C.A. 2006. Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795–1798.
- Daròs, J.A., Elena, S.F., Flores, R. 2006. Viroids: an Ariadne's thread into the RNA labyrinth. *EMBO Reports* **7**, 593–598.
- Davison, A. J. 2002. Evolution of the herpesviruses. *Veterinary Microbiology* **86**, 69–88.
- Davison, A.J. 2010. Herpesvirus systematics. *Veterinary Microbiology* **143**(1), 52–69. PMID: 20346601.
- Davison, A.J., Eberle, R., Ehlers, B. *et al.* 2009. The order *Herpesvirales*. *Archive of Virology* **154**(1), 171–177. PMID: 19066710.
- DeArmond, S.J., Prusiner, S.B. 2003. Perspectives on prion biology, prion disease pathogenesis, and pharmacologic approaches to treatment. *Clinics in Laboratory Medicine* **23**, 1–41.
- Delcher, A.L., Kasif, S., Fleischmann, R.D. *et al.* 1999. Alignment of whole genomes. *Nucleic Acids Research* **27**, 2369–2376. PMID: 10325427.
- Delcher, A. L., Phillippy, A., Carlton, J., Salzberg, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30**, 2478–2483.
- Ding, B. 2010. Viroids: self-replicating, mobile, and fast-evolving noncoding regulatory RNAs. *Wiley Interdisciplinary Reviews: RNA* **1**(3), 362–375. PMID: 21956936.

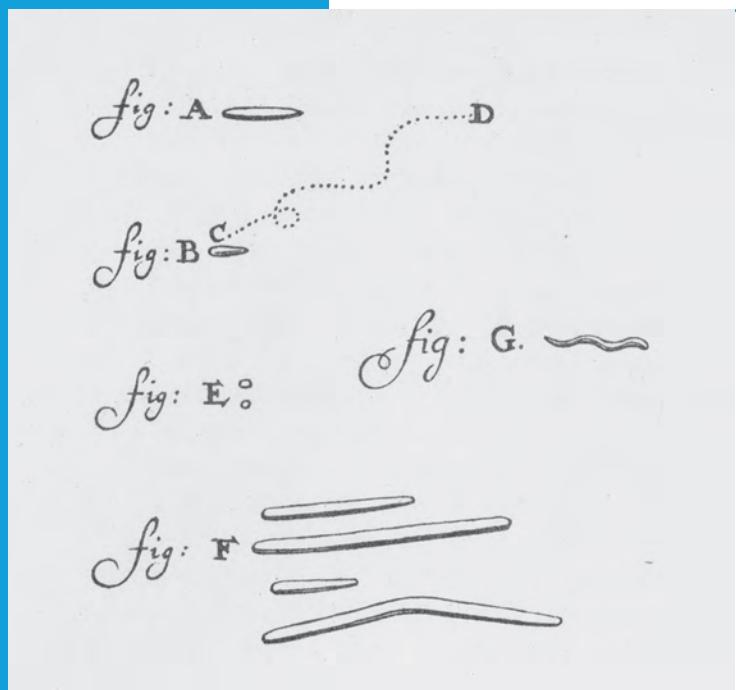
- Duffy, S., Shackelton, L.A., Holmes, E.C. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* **9**(4), 267–276. PMID: 18319742.
- Edwards, R.A., Rohwer, F. 2005. Viral metagenomics. *Nature Reviews Microbiology* **3**, 504–510.
- Engelhardt, O.G. 2013. Many ways to make an influenza virus: review of influenza virus reverse genetics methods. *Influenza Other Respir. Viruses* **7**(3), 249–256. PMID: 22712782.
- Engelman, A., Cherepanov, P. 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology* **10**(4), 279–290. PMID: 22421880.
- Enserink, M. 2006. Influenza. What came before 1918? Archaeovirologist offers a first glimpse. *Science* **312**, 1725.
- Evans, A.S. 1976. Causation and disease: the Henle-Koch postulates revisited. *Yale Journal of Biology and Medicine* **49**(2), 175–195. PMID: 782050.
- Ferron, F., Longhi, S., Canard, B., Karlin, D. 2006. A practical overview of protein disorder prediction methods. *Proteins* **65**(1), 1–14. PMID: 16856179.
- Fiers, W., Contreras, R., Duerinck, F. et al. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507. PMID: 1264203.
- Fiers, W., Contreras, R., Haegemann, G. et al. 1978. Complete nucleotide sequence of SV40 DNA. *Nature* **273**, 113–120. PMID: 205802.
- Finkbeiner, S.R., Allred, A.F., Tarr, P.I. et al. 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathogens* **4**(2), e1000011. PMID: 18398449.
- Flores, R. 2001. A naked plant-specific RNA ten-fold smaller than the smallest known viral RNA: The viroid. *Comptes Rendus de l'Academie des Sciences III* **324**, 943–952.
- Gago, S., Elena, S.F., Flores, R., Sanjuán, R. 2009. Extremely high mutation rate of a hammerhead viroid. *Science* **323**(5919), 1308. PMID: 19265013.
- Gall, A., Ferns, B., Morris, C. et al. 2012. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *Journal of Clinical Microbiology* **50**(12), 3838–3844. PMID: 22993180.
- Gaschen, B., Kuiken, C., Korber, B., Foley, B. 2001. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* **17**, 415–418.
- Gaschen, B., Taylor, J., Yusim, K. et al. 2002. Diversity considerations in HIV-1 vaccine selection. *Science* **296**, 2354–2360. PMID: 12089434.
- Gatherer, D., Seirafian, S., Cunningham, C. et al. 2011. High-resolution human cytomegalovirus transcriptome. *Proceedings of the National Academy of Science, USA* **108**(49), 19755–19760. PMID: 22109557.
- Ghedin, E., Sengamalay, N.A., Shumway, M. et al. 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437**, 1162–1166. PMID: 16208317.
- Gibbs, A.J. 2013. Viral taxonomy needs a spring clean; its exploration era is over. *Virology Journal* **10**, 254. PMID: 23938184.
- Grose, C. 2012. Pangaea and the Out-of-Africa Model of Varicella-Zoster virus evolution and phylogeny. *Journal of Virology* **86**(18), 9558–9565. PMID: 22761371.
- Hahn, B.H., Shaw, G.M., De Cock, K.M., Sharp, P. M. 2000. AIDS as a zoonosis: Scientific and public health implications. *Science* **287**, 607–614.
- Heeney, J.L., Dagleish, A.G., Weiss, R.A. 2006. Origins of HIV and the evolution of resistance to AIDS. *Science* **313**, 462–466.
- Holmes, E.C. 2008. Evolutionary history and phylogeography of human viruses. *Annual Review of Microbiology* **62**, 307–328. PMID: 18785840.
- Holmes, E.C. 2009. RNA virus genomics: a world of possibilities. *Journal of Clinical Investigation* **119**(9), 2488–2495. PMID: 19729846.
- Holmes, E.C. 2011. What does virus evolution tell us about virus origins? *Journal of Virology* **85**(11), 5247–5251. PMID: 21450811.
- Holmes, E.C., Ghedin, E., Miller, N. et al. 2005. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biology* **3**(9), e300. PMID: 16026181.

- Janies, D.A., Voronkin, I.O., Das, M. *et al.* 2010. Genome informatics of influenza A: from data sharing to shared analytical capabilities. *Animal Health Research Reviews* **11**(1), 73–79. PMID: 20591214.
- Johnson, R. T., Gibbs, C. J., Jr. 1998. Creutzfeldt-Jakob disease and related transmissible spongiform encephalopathies. *New England Journal of Medicine* **339**, 1994–2004.
- Kellam, P. 2001. Post-genomic virology: The impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Reviews in Medical Virology* **11**, 313–329.
- King, A.M.Q., Lefkowitz, E., Adams, M.J., Carstens, E.B. (eds) 2011. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier, San Diego, CA.
- Kruger, D. H., Schneck, P., Gelderblom, H. R. 2000. Helmut Ruska and the visualisation of viruses. *Lancet* **355**, 1713–1717.
- Kuhn, J.H., Jahrling, P.B. 2010. Clarification and guidance on the proper usage of virus and virus species names. *Archive of Virology* **155**(4), 445–453. PMID: 20204430.
- Kuhn, J.H., Radoshitzky, S.R., Bavari, S., Jahrling, P.B. 2013. The International Code of Virus Classification and Nomenclature (ICVCN): proposal for text changes for improved differentiation of viral taxa and viruses. *Archive of Virology* **158**(7), 1621–1629. PMID: 23417351.
- Kurtz, S., Phillippy, A., Delcher, A.L. *et al.* 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5**, R12.
- La Scola, B., Audic, S., Robert, C. *et al.* 2003. A giant virus in Amoebae. *Science* **299**, 2033. PMID: 12663918.
- Legendre, M., Arslan, D., Abergel, C., Claverie, J.M. 2012. Genomics of Megavirus and the elusive fourth domain of Life. *Communicative and Integrative Biology* **5**(1), 102–106. PMID: 22482024.
- Legendre, M., Bartoli, J., Shmakova, L. *et al.* 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Science, USA* **111**(11), 4274–4279. PMID: 24591590.
- Liu, L., Johnson, H.L., Cousens, S. *et al.* 2012. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* **379**(9832), 2151–2161. PMID: 22579125.
- Mandavilli, A. 2010. HIV/AIDS. *Nature* **466**(7304), S1. PMID: 20631695.
- Marcinowski, L., Lidschreiber, M., Windhager, L. *et al.* 2012. Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection. *PLoS Pathogens* **8**(9), e1002908. PMID: 22969428.
- Mariner, J.C., House, J.A., Mebus, C.A. *et al.* 2012. Rinderpest eradication: appropriate technology and social innovations. *Science* **337**(6100), 1309–1312. PMID: 22984063.
- McClure, M. A. 2000. The complexities of genome analysis, the Retrovirus agent perspective. *Bioinformatics* **16**, 79–95.
- McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E., Telford, E. A. 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *Journal of Molecular Biology* **247**, 443–458.
- McGeoch, D.J., Rixon, F.J., Davison, A.J. 2006. Topics in herpesvirus genomics and evolution. *Virus Research* **117**, 90–104.
- Meissner, C., Coffin, J. M. 1999. The human retroviruses: AIDS and other diseases. In *Mechanisms of Microbial Disease* (eds M.Schaechter, N. C.Engleberg, B. I.Eisenstein, G.Medoff), Lippincott Williams & Wilkins, Baltimore, MD, Chapter 38.
- Moeller, A., Kirchdoerfer, R.N., Potter, C.S., Carragher, B., Wilson, I.A. 2012. Organization of the influenza virus replication machinery. *Science* **338**(6114), 1631–1634. PMID: 23180774.
- Mokili, J.L., Rohwer, F., Dutilh, B.E. 2012. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2**(1), 63–77. PMID: 22440968.
- Montaner, S., Kufareva, I., Abagyan, R., Gutkind, J.S. 2013. Molecular mechanisms deployed by virally encoded G protein-coupled receptors in human diseases. *Annual Review of Pharmacology and Toxicology* **53**, 331–354 (2013). PMID: 23092247.
- Moore, P.S., Chang, Y. 2010. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nature Reviews Cancer* **10**(12), 878–889. PMID: 21102637.

- Morens, D.M., Holmes, E.C., Davis, A.S., Taubenberger, J.K. 2011. Global rinderpest eradication: lessons learned and why humans should celebrate too. *Journal of Infectious Diseases* **204**(4), 502–505. PMID: 21653230.
- Moss, W.J., Griffin, D.E. 2006. Global measles elimination. *Nature Reviews Microbiology* **4**, 900–908.
- Neumann, A.U., Lam, N.P., Dahari, H. *et al.* 1998. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science* **282**, 103–107.
- Obenauer, J.C., Denson, J., Mehta, P.K. *et al.* 2006. Large-scale sequence analysis of avian influenza isolates. *Science* **311**, 1576–1580.
- Olsen, B., Munster, V.J., Wallensten, A., Waldenström, J., Osterhaus, A.D., Fouchier, R.A. 2006. Global patterns of influenza a virus in wild birds. *Science* **312**, 384–388.
- Ortblad, K.F., Lozano, R., Murray, C.J. 2013. The burden of HIV: insights from the GBD 2010. *AIDS* **27**(13), 2003–2017. PMID: 23660576.
- Paulose-Murphy, M., Ha, N.K., Xiang, C. *et al.* 2001. Transcription program of human herpesvirus 8 (kaposi's sarcoma-associated herpesvirus). *Journal of Virology* **75**, 4843–4853. PMID: 11312356.
- Pennisi, E. 2013. Microbiology. Ever-bigger viruses shake tree of life. *Science* **341**(6143), 226–227. PMID: 23868995.
- Philippe, N., Legendre, M., Doutre, G. *et al.* 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**(6143), 281–286. PMID: 23869018.
- Piot, P., Feachem, R.G., Lee, J.W., Wolfensohn, J.D. 2004. Public health. A global response to AIDS: lessons learned, next steps. *Science* **304**, 1909–1910.
- Pleschka, S. 2013. Overview of influenza viruses. *Current Topics in Microbiology and Immunology* **370**, 1–20. PMID: 23124938.
- Pond, S.L., Murrell, B., Poon, A.F. 2012. Evolution of viral genomes: interplay between selection, recombination, and other forces. *Methods in Molecular Biology* **856**, 239–272. PMID: 22399462.
- Poole, L. J., Yu, Y., Kim, P.S. *et al.* 2002. Altered patterns of cellular gene expression in dermal microvascular endothelial cells infected with Kaposi's sarcoma-associated herpesvirus. *Journal of Virology* **76**, 3395–3420. PMID: 11884566.
- Prangishvili, D., Garrett, R.A., Koonin, E.V. 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Research* **117**, 52–67.
- Prusiner, S. B. 1998. Prions. *Proceedings of the National Academy of Science, USA* **95**, 13363–13383.
- Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C. 2004. The causes and consequences of HIV evolution. *Nature Reviews Genetics* **5**, 52–61.
- Raoult, D., Audic, S., Robert, C. *et al.* 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350.
- Reyes, A., Semenkovich, N.P., Whiteson, K., Rohwer, F., Gordon, J.I. 2012. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology* **10**(9), 607–617. PMID: 22864264.
- Rhee, S. Y., Gonzales, M. J., Kantor, R. *et al.* 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* **31**, 298–303.
- Rima, B.K., Duprex, W.P. 2009. The measles virus replication cycle. *Current Topics in Microbiology and Immunology* **329**, 77–102. PMID: 19198563.
- Roberts, L. 2012. HIV/AIDS in America. Introduction. *Science* **337**(6091), 167. PMID: 22798592.
- Rosario, K., Breitbart, M. 2011. Exploring the viral world through metagenomics. *Current Opinion in Virology* **1**(4), 289–297. PMID: 22440785.
- Rota, P.A., Bellini, W.J. 2003. Update on the global distribution of genotypes of wild type measles viruses. *Journal of Infectious Diseases* **187**, S270–276.
- Ruigrok, R.W., Crépin, T., Hart, D.J., Cusack, S. 2010. Towards an atomic resolution understanding of the influenza virus replication machinery. *Current Opinion in Structural Biology* **20**(1), 104–113. PMID: 20061134.
- Rusch, D.B., Halpern, A.L., Sutton, G. *et al.* 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology* **5**, e77. PMID: 17355176.

- Russo, J.J., Bohenzky, R.A., Chien, M.C. *et al.* 1996. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proceedings of the National Academy of Science USA* **93**, 14862–14867. PMID: 8962146.
- Schaechter, M., Engleberg, N. C., Eisenstein, B. I., Medoff, G. 1999. *Mechanisms of Microbial Disease*. Lippincott Williams & Wilkins, Baltimore, MD.
- Scholthof, K.B., Adkins, S., Czosnek, H. *et al.* 2011. Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology* **12**(9), 938–954. PMID: 22017770.
- Seibert, M.M., Ekeberg, T., Maia, F.R. *et al.* 2011. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature* **470**(7332), 78–81. PMID: 21293374.
- Sharp, P.M., Bailes, E., Chaudhuri, R.R. *et al.* 2001. The origins of acquired immune deficiency syndrome viruses: Where and when? *Philosophical Transactions of the Royal Society of London: B Biological Sciences* **356**, 867–876. PMID: 11405934.
- Simons, E., Ferrari, M., Fricks, J. *et al.* 2012. Assessment of the 2010 global measles mortality reduction goal: results from a model of surveillance data. *Lancet* **379**(9832), 2173–2178. PMID: 22534001.
- Simpson, G. G. 1963. The meaning of taxonomic statements. In *Classification and Human Evolution* (ed. S. L. Washburn). Aldine Publishing Co., Chicago, pp. 1–31.
- Slimani, M., Pagnier, I., Raoult, D., La Scola, B. 2013. Amoebae as battlefields for bacteria, giant viruses, and virophages. *Journal of Virology* **87**(8), 4783–4785. PMID: 23388714.
- Stern-Ginossar, N., Weisburd, B., Michalski, A. *et al.* 2012. Decoding human cytomegalovirus. *Science* **338**(6110), 1088–1093. PMID: 23180859.
- Tang, P., Chiu, C. 2010. Metagenomics for the discovery of novel human viruses. *Future Microbiology* **5**(2), 177–189. PMID: 20143943.
- Taubenberger, J.K., Reid, A.H., Lourens, R.M. *et al.* 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**, 889–893.
- Taubenberger, J.K., Morens, D.M., Fauci, A.S. 2007. The next influenza pandemic: can it be predicted? *JAMA* **297**, 2025–2027.
- Thomas, V., Bertelli, C., Collyn, F. *et al.* 2011. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology* **13**(6), 1454–1466. PMID: 21392201.
- Tumpey, T.M., Basler, C.F., Aguilar, P.V. *et al.* 2005. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* **310**, 77–80.
- Upton, C., Hogg, D., Perrin, D., Boone, M., Harris, N. L. 2000. Viral genome organizer: A system for analyzing complete viral genomes. *Virus Research* **70**, 55–64.
- Van Regenmortel, M.H., Burke, D.S., Calisher, C.H. *et al.* 2010. A proposal to change existing virus species names to non-Latinized binomials. *Archive of Virology* **155**(11), 1909–1919. PMID: 20953644.
- Van Regenmortel, M.H., Ackermann, H.W., Calisher, C.H. *et al.* 2013. Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. *Archive of Virology* **158**(5), 1115–1119. PMID: 23269443.
- Venter, J.C., Remington, K., Heidelberg, J.F. *et al.* 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74. PMID: 15001713.
- von Bubnoff, A. 2005. The 1918 flu virus is resurrected. *Nature* **437**, 794–795.
- Wakeling, M.N., Roy, D.J., Nash, A.A., Stewart, J.P. 2001. Characterization of the murine gammaherpesvirus 68 ORF74 product: A novel oncogenic G protein-coupled receptor. *Journal of General Virology* **82**, 1187–1197.
- Williamson, S.J., Allen, L.Z., Lorenzi, H.A. *et al.* 2012. Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS One* **7**(10), e42047. PMID: 23082107.
- Willner, D., Hugenholtz, P. 2013. From deep sequencing to viral tagging: recent advances in viral metagenomics. *Bioessays* **35**(5), 436–442. PMID: 23450659.
- Worobey, M., Telfer, P., Souquière, S. *et al.* 2010. Island biogeography reveals the deep history of SIV. *Science* **329**(5998), 1487. PMID: 20847261.

- Yoosuf, N., Yutin, N., Colson, P. *et al.* 2012. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. *Genome Biology and Evolution* **4**(12), 1324–1330. PMID: 23221609.
- Yoosuf, N., Pagnier, I., Fournous, G. *et al.* 2014. Complete genome sequence of Courdo11 virus, a member of the family Mimiviridae. *Virus Genes* **48**(2), 218–223. PMID: 24293219.
- Yutin, N., Colson, P., Raoult, D., Koonin, E.V. 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology Journal* **10**, 106. PMID: 23557328.
- Zhang, T., Breitbart, M., Lee, W.H. *et al.* 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biology* **4**, e3.



Antony van Leeuwenhoek (1622–1723) has been called the Father of protozoology and bacteriology. This figure shows bacteria he observed taken from his own mouth. Figure A indicates a motile *Bacillus*. Figure B shows *Selenomonas sputigena*, while C and D show the path of its motion. Figure E shows two micrococci, F shows *Leptotrichia buccalis*, and G shows a spirochete. He describes these "animalcules," found in his and others' mouths, in a letter written 17 September 1683:

While I was talking to an old man (who leads a sober life, and never drinks brandy or [smokes] tobacco, and very seldom any wine), my eye fell upon his teeth, which were all coated over; so I asked him when he had last cleaned his mouth? And I got for answer that he'd never washed his mouth in

all his life. So I took some spittle out of his mouth and examined it; but I could find in it nought but what I had found in my own and other people's. I also took some of the matter that was lodged between and against his teeth, and mixing it with his own spit, and also with fair water (in which there were no animalcules), I found an unbelievably great company of living animalcules, a-swimming more nimbly than any I had ever seen up to this time. The biggest sort (where of there were a great plenty) bent their body into curves in going forwards, as in Fig. G. Moreover, the other animalcules were in such enormous numbers, that all the water (notwithstanding only a very little of the matter taken from between the teeth was mingled with it) seemed to be alive.

Source: Leeuwenhoek, trans. Dobell (1932). Reproduced with permission from Dover Publications.

# Completed Genomes: Bacteria and Archaea

# CHAPTER 17

*And now you may be disposed to ask: To what end is this discourse on the anatomy of beings too minute for ordinary vision, and of whose very existence we should be ignorant unless it were revealed to us by a powerful microscope? What part in nature can such apparently insignificant animalcules play, that can in any way interest us in their organization, or repay us for the pains of acquiring a knowledge of it? I shall endeavour briefly to answer these questions. The Polygastric Infusoria, notwithstanding their extreme minuteness, take a great share in important offices of the economy of nature, on which our own well-being more or less immediately depends.*

*Consider their incredible numbers, their universal distribution, their insatiable voracity; and that it is the particles of decaying vegetable and animal bodies which they are appointed to devour and assimilate.*

*Surely we must in some degree be indebted to those ever active invisible scavengers for the salubrity of our atmosphere. Nor is this all: they perform a still more important office, in preventing the gradual diminution of the present amount of organized matter upon the earth. For when this matter is dissolved or suspended in water, in that state of comminution and decay which immediately precedes its final decomposition into the elementary gases, and its consequent return from the organic to the inorganic world, these wakeful members of nature's invisible police are every where ready to arrest the fugitive organized particles, and turn them back into the ascending stream of animal life.*

—Richard Owen (1843, p. 27)

## LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- define bacteria and archaea;
- explain the bases of their classification;
- describe the genomes of *Escherichia coli* and other bacteria;
- describe bioinformatics approaches to identifying and characterizing bacterial and archaeal genes and proteins; and
- compare bacterial genomes.

## INTRODUCTION

In this chapter we consider bioinformatic approaches to two of the three main branches of life: bacteria and archaea. Bacteria and archaea are grouped together because they are single-celled organisms (in most contexts) that lack nuclei. Bacteria and archaea are

William Martin and Eugene Koonin (2006) briefly discuss the definition of the term prokaryote. We contrasted eukaryotes to bacteria and archaea in Chapter 8.

It has been estimated that there are  $10^{30}$  bacteria, comprising the majority of the biomass on the planet (Sherratt, 2001).

sometimes also termed microorganisms. The term microbe refers to those microorganisms that cause disease in humans; microbes include many eukaryotes such as fungi and protozoa (Chapters 18 and 19) as well as some bacteria and archaea. In Chapter 15 we discussed a proposal by Norman Pace (2009) that the term “prokaryote” (or “procaryote”) should be abandoned because it implies an incorrect model of evolution. While the community continues to maintain the term (with its sense of referring to bacteria and archaea), in this book we limit its use in support of Pace’s argument.

It has been estimated that bacteria account for 60% of Earth’s biomass. Bacteria occupy every conceivable ecological niche in the planet, and there may be from  $10^7$  to  $10^9$  distinct bacterial species (Fraser *et al.*, 2000), although some suggest there may be fewer species (Schloss and Handelsman, 2004). The great majority of bacteria and archaea (>99%) have never been cultured or characterized (DeLong and Pace, 2001). A compelling reason to study bacteria is that many cause disease in humans and other animals.

This chapter provides an overview of bioinformatic approaches to the study of bacteria and archaea. We review aspects of bacterial and archaeal biology such as genome size and complexity, and tools for the analysis and comparison of these genomes. The analysis of whole genomes, bolstered by next-generation sequencing, has had profound effects on our understanding of bacteria and archaea (reviewed in Bentley and Parkhill, 2004; Fraser-Liggett, 2005; Ward and Fraser, 2005; Binnewies *et al.*, 2006; Medini *et al.*, 2008; Fournier and Raoult, 2011; Loman *et al.*, 2012; Mavromatis *et al.*, 2012; Stepanauskas, 2012). Some of the main issues are: (1) gaining an improved sampling of species diversity through genomic sequence analyses, along with improved phylogeny and classification; and (2) achieving a better understanding of the forces that shape microbial genomes. These forces include the following:

- loss of genes and reductions in genome size, especially in species that are dependent on their hosts for survival such as obligate intracellular parasites;
- gains in genome size, especially in free-living organisms that may require many genomic resources to cope with variable environmental conditions;
- lateral gene transfer, in which genetic material is transferred horizontally between organisms that share an environmental niche and not vertically through descent from ancestors; and
- chromosomal rearrangements such as inversions that often occur in related species or strains.

In this chapter we discuss these topics as well as bioinformatics tools that are available to investigate them.

## CLASSIFICATION OF BACTERIA AND ARCHAEA

In Chapter 15 we described many of the genome-sequencing projects for bacteria and archaea in chronological order, beginning with the sequencing of *Haemophilus influenzae* in 1995. We now consider the classification of bacteria and archaea by six different criteria: (1) morphology; (2) genome size; (3) lifestyle; (4) relevance to human disease; (5) molecular phylogeny using rRNA; and (6) molecular phylogeny using other molecules. There are many other ways to classify bacteria and archaea (Box 17.1).

We use bioinformatics tools to analyze individual microbial genomes and to compare two or more genomes. It is through comparative genomics that we are beginning to appreciate some of the important principles of microbial biology, such as the adaptation of bacteria and archaea to highly specific ecological niches, the lateral transfer of genes between organisms, genome expansion and reduction, and the molecular basis of pathogenicity (Bentley and Parkhill, 2004; Binnewies *et al.*, 2006).

Pathogenicity is the ability of an organism to cause disease. Virulence is the degree of pathogenicity.

### BOX 17.1 CLASSIFICATION OF BACTERIA AND ARCHAEA

While we choose six basic ways to classify bacteria and archaea, there are many other approaches. These include the energy source (respiration, fermentation, photosynthesis), their formation of characteristic products (e.g., acids), the presence of immunological markers such as proteins or lipopolysaccharides, their ecological niche (also related to lifestyle), and their nutritional growth requirements. The growth requirements include obligate and/or facultative aerobes (requiring oxygen) or anaerobes (growing in environments without oxygen), chemotrophs (deriving energy from the breakdown of organic molecules such as proteins, lipids, and carbohydrates), or autotrophs (synthesizing organic molecules through the use of an external energy source and inorganic compounds such as carbon dioxide and nitrates). Autotrophs (from the Greek for “self feeder”) are either photoautotrophs (obtaining energy through photosynthesis, requiring carbon dioxide and expiring oxygen) or chemautotrophs (obtaining energy from inorganic compounds and carbon from carbon dioxide). Heterotrophs, unlike autotrophs, must feed on other organisms to obtain energy.

Several resources provide broad information about the current state of bacterial and archaeal genomics.

- The National Center for Biotechnology Information (NCBI) Genome resource currently lists >2600 complete bacterial genomes and 168 archaeal genomes (NCBI Resource Coordinators, 2014). NCBI describes major divisions of bacteria (**Table 17.1**) as well as archaea (**Table 17.2**). These tables provide an overview as we begin to classify these organisms by various criteria.
- The Genomes Online Database (GOLD) includes >2600 complete and published bacterial genome projects, as well as >9000 permanent draft projects and >19,000 incomplete projects (Pagani *et al.*, 2012).
- EnsemblBacteria includes >9000 genome sequences from bacteria and archaea (Kersey *et al.*, 2014).

**TABLE 17.1 Classification of bacteria. Bacteria are described as a kingdom, followed by “intermediate ranks.”**

Intermediate rank 1	Intermediate rank 2	Genus, species, and strain (examples)	Genome size (Mb)	GenBank accession
Actinobacteria	Actinomycetidae	<i>Mycobacterium tuberculosis</i> CDC1551	4.4	NC_002755
Aquificae	Aquificales	<i>Aquifex aeolicus</i> VF5	1.5	NC_000918
Bacteroidetes	Bacteroides	<i>Porphyromonas gingivalis</i> W83	2.3	NC_002950.2
Chlamydiae	Chlamydiales	<i>Chlamydia trachomatis</i> serovar D	1.0	NC_000117
Chlorobi	Chlorobia	<i>Chlorobium tepidum</i> TLS	2.1	NC_002932
Cyanobacteria	Chroococcales	<i>Synechocystis</i> sp. PCC6803	3.5	NC_000911
	Nostocales	<i>Nostoc</i> sp. PCC 7120	6.4	NC_003272
Deinococcus-Thermus	Deinococci	<i>Deinococcus radiodurans</i> R1	2.6	NC_001263
Firmicutes	Bacillales	<i>Bacillus subtilis</i> 168	4.2	NC_000964
	Clostridia	<i>Clostridium perfringens</i> 13	3.0	NC_003366
	Lactobacillales	<i>Streptococcus pneumoniae</i> R6	2.0	NC_003098
	Mollicutes	<i>Mycoplasma genitalium</i> G-37	0.58	NC_000908
Fusobacteria	Fusobacteria	<i>Fusobacterium nucleatum</i> ATCC 25586	2.1	NC_003454
Proteobacteria	Alphaproteobacteria	<i>Rickettsia prowazekii</i> Madrid E	1.1	NC_000963
	Betaproteobacteria	<i>Neisseria meningitidis</i> MC58	2.2	NC_003112
	Epsilon subdivision	<i>Helicobacter pylori</i> J99	1.6	NC_000921
	Gamma subdivision	<i>Escherichia coli</i> K-12-MG1655	4.6	NC_000913
	Magnetotactic cocci	<i>Magnetococcus</i> sp. MC-1	NA	NC_008576
Spirochaetales	Spirochaetaceae	<i>Borrelia burgdorferi</i> B31	0.91	NC_001318
Thermotogales	Thermotoga	<i>Thermotoga maritima</i> MSB8	1.8	NC_000853

Source: NCBI (<http://www.ncbi.nlm.nih.gov>).

**TABLE 17.2 Classification of archaea. Archaea are described as a kingdom, followed by “intermediate ranks.”**

Intermediate rank 1	Intermediate rank 2	Genus, Species, and strain (example)	Genome size (Mb)	GenBank accession
Crenarchaeota	Thermoprotei	<i>Aeropyrum pernix</i> K1	1.6	NC_000854
Euryarchaeota	Archaeoglobi	<i>Archaeoglobus fulgidus</i> DSM4304	2.2	NC_000917
	Halobacteria	<i>Halobacterium</i> sp. NRC-1	2.0	NC_002607
	Methanobacteria	<i>Methanobacterium thermoautotrophicum</i> delta H	1.7	NC_000916
	Methanococci	<i>Methanococcus jannaschii</i> DSM2661	1.6	NC_000909
	Methanopyri	<i>Methanopyrus kandleri</i> AV19	1.6	NC_003551
	Thermococci	<i>Pyrococcus abyssi</i> GE5	1.7	NC_000868
	Thermoplasmata	<i>Thermoplasma volcanium</i> GSS1	1.5	NC_002689

Source: NCBI (<http://www.ncbi.nlm.nih.gov>).

An NCBI microbial resource is available at [http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial\\_taxtree.html](http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html) (WebLink 17.1). Genomes Online Database is available at <http://genomesonline.org/> (WebLink 17.2). EnsemblBacteria is online at <http://bacteria.ensembl.org/> (WebLink 17.3). PATRIC is available at <http://patrcbrc.vbi.vt.edu/> (WebLink 17.4). The IMG home page is <http://img.jgi.doe.gov/> (WebLink 17.5). Another major resource for bacterial, archaeal, and eukaryotic genomes is PEDANT at the Munich Information Center for Protein Sequences (MIPS; <http://pedant.gsf.de/>, WebLink 17.6).

- The Pathosystems Resource Integration Center (PATRIC) currently lists >4200 bacterial genomes (Gillespie *et al.*, 2011) and includes a set of analysis tools.
- The Integrated Microbial Genomes (IMG) system includes a broad set of genome and metagenome analysis tools (Markowitz *et al.*, 2014).

### Classification of Bacteria by Morphological Criteria

Most bacteria are classified into four main types: Gram-positive and Gram-negative cocci or rods (reviewed in Schaechter, 1999). Examples of these different bacteria are presented in **Table 17.3**. The Gram stain is absorbed by about half of all bacteria and reflects the protein and peptidoglycan composition of the cell wall. Many other bacteria do not fit the categories of Gram-positive or Gram-negative cocci or rods because they have atypical shapes or staining patterns. As an example, spirochetes such as the Lyme disease agent *Borrelia burgdorferi* have a characteristic outer membrane sheath, protoplasmic cell cylinder, and periplasmic flagella (Charon and Goldstein, 2002).

The classification of microbes based on molecular phylogeny is far more comprehensive, as described in “Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences” below. Molecular differences can reveal the extent of microbial diversity both between species (showing the breadth of the bacterial branch of the tree of life) and within species (e.g., showing molecular differences in pathogenic isolates and in closely related, nonvirulent strains). However, beyond molecular criteria there are many additional ways to differentiate bacteria based on microscopy and studies of physiology, for example distinguishing those microbes that are capable of oxygenic photosynthesis (Cyanobacteria) or those that produce methane.

**TABLE 17.3 Major categories of bacteria based on morphological criteria (the disease is indicated in parentheses).**

Type	Examples
Gram-positive cocci	<i>Streptococcus pyogenes</i> , <i>Staphylococcus aureus</i>
Gram-positive rods	<i>Corynebacterium diphtheriae</i> , <i>Bacillus anthracis</i> (anthrax), <i>Clostridium botulinum</i>
Gram-negative cocci	<i>Neisseria</i> , <i>Gonococcus</i>
Gram-negative rods	<i>Escherichia coli</i> , <i>Vibrio cholerae</i> , <i>Helicobacter pylori</i>
Other	<i>Mycobacterium leprae</i> (leprosy), <i>Borrelia burgdorferi</i> (Lyme disease), <i>Chlamydia trachomatis</i> (sexually transmitted disease), <i>Mycoplasma pneumoniae</i>

The diversity of morphologies in bacterial life forms is spectacular. We can provide examples of two predatory bacteria that prey on other bacteria. Each of these examples is intended to highlight both the diversity of morphologies that may occur, and the role that genome sequence analysis may have in elucidating mechanisms of structural change.

1. The Myxobacteria are single-celled  $\delta$ -proteobacteria that are highly successful, with millions of cells per gram of cultivated soil. Upon encountering low nutrient conditions, up to 100,000 individuals of *Myxococcus xanthus* join to form a fruiting body which is essentially a multicellular organism having a spherical shape and that is resistant to different kinds of stress. In favorable nutrient conditions, individual spores within the fruiting body germinate and thousands of *M. xanthus* spores swarm. This swarm can surround, lyse, and consume prey bacteria. Goldman *et al.* (2006) reported the complete genome sequence of *M. xanthus* and provided insight into genes that encode motor proteins and allow the organism to glide, use retractable pili, and secrete mucus. Also, the large genome size (9.1 megabases or Mb) contrasts with the much smaller size of other related  $\delta$  subgroup proteobacteria (3.7–5.0 Mb). Goldman *et al.* characterized the nature of the *M. xanthus* genome expansion and its possible relation to this organism's extraordinary behavior and morphology (reviewed by Kaiser, 2013).
2. *Bdellovibrio bacteriovorus* provides a second example of a bacterium with an extraordinary morphology. This is also a predatory delta-proteobacterium that eats Gram-negative bacteria. Its genome of about 3.8 Mb is predicted to encode over 3500 proteins (Rendulic *et al.*, 2004). The bacterium attacks its prey (by swimming to them at high speed), adheres irreversibly, opens a pore in the prey's outer membrane and peptidoglycan layer, then enters the periplasm and replicates. *B. bacteriovorus* then forms a structure called a bdelloplast in which the rod-shaped prey becomes rounded and the predator grows to several multiples of its normal size as it consumes the prey nutrients. Later, the predator exits the bdelloplast. The analysis of this genome allowed Rendulic *et al.* to identify genes encoding catabolic enzymes (e.g., proteases, nucleases, glycanases, and lipases) implicated in its lifestyle, as well as a host interaction locus containing genes implicated in pilus and adherence genes.

The *M. xanthus* DK 1622 complete, circular genome (length 9,139,763 nucleotides) has accession NC\_008095.1. Note that by entering that accession number into the Entrez search engine from the home page of NCBI, you can link to the Genome Project page that provides an overview of the organism. The slime mold *Dictyostelium discoideum*, a eukaryote, also includes a lifestyle that can alternate between single-celled and multicellular (Chapter 19).

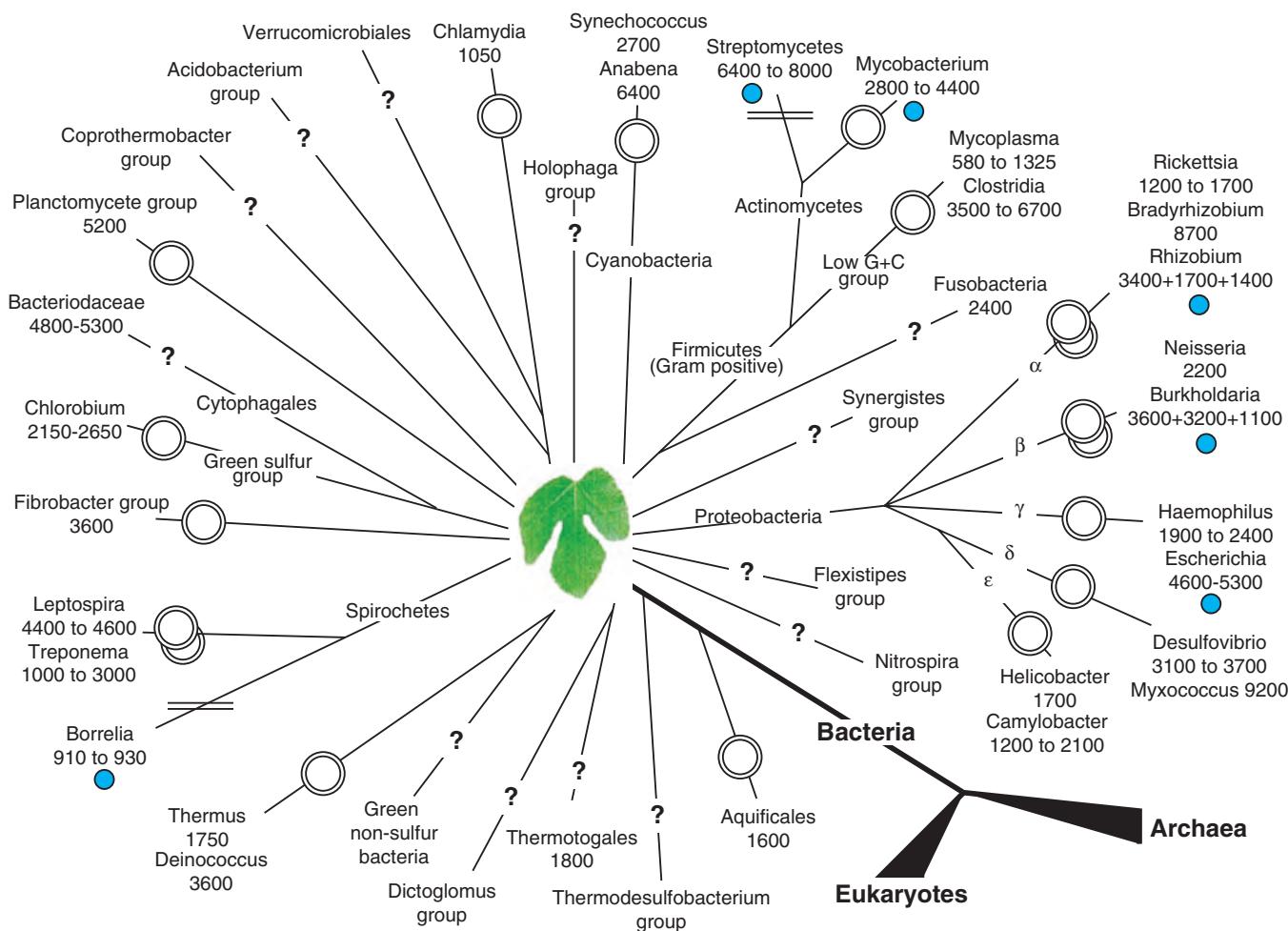
The *B. bacteriovorus* accession is NC\_005363.1. Its lifecycle is described at the NCBI Genome project page for this organism.

## Classification of Bacteria and Archaea Based on Genome Size and Geometry

In haploid organisms such as bacteria and archaea, the genome size (or C value) is the total amount of DNA in the genome. Bacterial and archaeal genomes vary in size from under 500,000 bp (0.5 Mb) to almost 15 Mb (Table 17.4) (Casjens, 1998). The genome sizes of 23 named major bacterial phyla and some of their subgroups are shown in Figure 17.1. As indicated in the figure, most bacterial genomes are circular although some are linear;

**TABLE 17.4 Range of genome sizes in bacteria and archaea. Adapted from Graur and Li (2000) with permission from Sinauer Associates.**

Taxon	Genome size range (Mb)	Ratio (highest/lowest)
Bacteria	0.16–13.2	83
Mollicutes	0.58–2.2	4
Gram negative	0.16–9.5	59
Gram positive	1.6–11.6	7
Cyanobacteria	3.1–13.2	4
Archaea	0.49–5.75	12



**FIGURE 17.1** Bacterial chromosome size and geometry. The 23 named major bacterial phyla are represented as well as some of their subgroups. The tree is based on rRNA sequences and is unrooted. The branch lengths do not depict phylogenetic distances, and the fig leaf at the center indicates uncertain branching patterns. The chromosome geometry (circular or linear, in some cases with multiple chromosomes) is indicated at the end of each branch. The chromosome sizes of representative genera are given (in kilobases). Linear extrachromosomal elements, common in borrelia and actinomycetes, are indicated. Adapted from Casjens (1998) with permission from Annual Reviews.

In diploid or polyploid organisms, the genome size is the amount of DNA in the unreplicated haploid genome (such as the sperm cell nucleus). We discuss eukaryotic genome sizes in Chapters 18–19.

some bacterial genomes consist of multiple circular chromosomes. Plasmids (small circular extrachromosomal elements) have been found in most bacterial phyla, although linear extrachromosomal elements are more rare.

Some bacterial genomes are comparable in size or even larger than eukaryotic genomes. The genome of the fungus *Encephalitozoon cuniculi* is just 2.5 Mb and encodes about 2000 proteins (see Chapter 18), and at least a dozen eukaryotic genomes that are currently being sequenced are under 10 Mb. The genomes of two strains of the myxobacterium *Sorangium cellulosum* have been sequenced and are among the largest bacterial genomes that have been sequenced to date. One is >13 Mb and includes over 9700 genes (Schneiker *et al.*, 2007), while another is ~14.8 Mb and includes >10,500 genes (Han *et al.*, 2013; Table 17.5). The cyanobacterium *Mastigocoleus testarum* BC008 has a genome size of 15.9 Mb. In general, those bacteria having notably large genome sizes exhibit great behavioral or phenotypic complexity, participating in complex social behavior (such as multicellular interactions) or processes such as differentiation.

Overall, the number of genes encoded in a bacterial genome ranges from the extraordinarily small number of 182 to >10,000 in exceptional cases. This range is comparable to the range in C values. For a large number of bacteria with completely sequenced

**TABLE 17.5** Genome size of selected bacteria and archaea having relatively large or small genomes. (A): archaeal; (B): eubacterial. Adapted from <http://www.sanger.ac.uk/Projects/Microbes/> with permission from Dr A. Bateman and adapted from the NCBI website (PubMed, NCBI Genome).

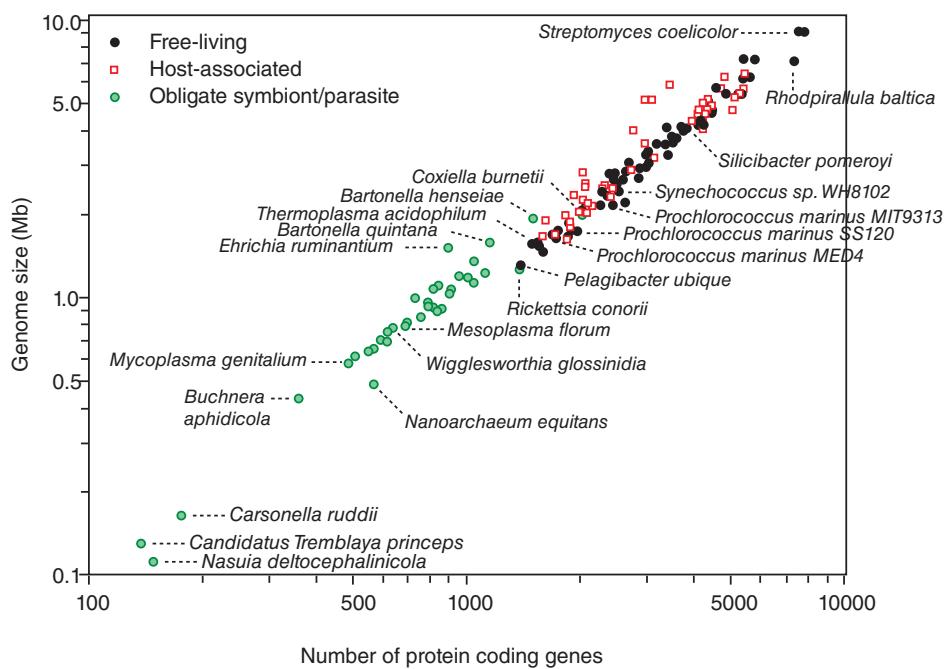
Species	Genome size (Mb)	Coding regions	GC content	Reference
<i>Sorangium cellulosum</i> So0157-2 (B)	14.8	10,400	72.1	Han <i>et al.</i> , 2013
<i>Sorangium cellulosum</i> So ce56 (B)	13.0	9,380	71.4	Schneiker <i>et al.</i> , 2007
<i>Solibacter usitatus</i> (B)	10	7,888	61.9	Unpublished; accession NC_008536
<i>Myxococcus xanthus</i> DK 1622 (B)	9.1	7,388	68.9	Goldman <i>et al.</i> , 2006
<i>Streptomyces coelicolor</i> (B)	8.67	7,825	72	Bentley <i>et al.</i> , 2002
<i>Methanosarcina acetivorans</i> C2A (A)	5.75	4,524	42.7	Galagan <i>et al.</i> , 2002
<i>Ureaplasma urealyticum</i> parvum biovar serovar 3 (B)	0.752	613	26	Glass <i>et al.</i> , 2000
<i>Mycoplasma pneumoniae</i> M129 (B)	0.816	677	40	Himmelreich <i>et al.</i> , 1996
<i>Mycoplasma genitalium</i> G-37 (B)	0.58	470	32	Fraser <i>et al.</i> , 1995
<i>Nanarchaeum equitans</i> (A)	0.49	552	31.6	Huber <i>et al.</i> , 2002; Waters <i>et al.</i> , 2003
<i>Buchnera aphidicola</i> (B)	0.42	362	20	Pérez-Brocal <i>et al.</i> , 2006
<i>Carsonella ruddii</i> (B)	0.16	182	16.5	Nakabachi <i>et al.</i> , 2006

genomes, protein-coding genes constitute about 85–95% of the genome. Intergenic and nongenic fractions are therefore small. An exception is the pathogen that causes leprosy, *Mycobacterium leprae*. Its genome underwent massive gene decay, and protein-coding genes constitute only 49.5% of the genome (Cole *et al.*, 2001; Singh and Cole, 2011). Another exception is the parasite *Rickettsia prowazekii*, described below, that has 24% noncoding DNA. The aphid symbiont *Serratia symbiotica* has a coding density of ~61% and 550 pseudogenes (Burke and Moran, 2011).

The density of genes in microbial genomes is consistently about one gene per kilobase. As an example, the genome of *Escherichia coli* K-12 substr. MG1655 (accession NC\_000913.3) is 4.64 Mb and encompasses 4497 genes (one gene per 1032 base pairs). Even in very small genomes such as *Mycoplasma genitalium*, reduced genome sizes are not associated with changes either in gene density or in the average size of genes (Fraser *et al.*, 1995). The genome sizes of selected large or small bacteria and archaea are shown in Table 17.5.

Examination of the sizes of several hundred bacterial and archaeal genomes in relation to the number of genes shows a linear relationship (Fig. 17.2). This figure (adapted from Giovannoni *et al.*, 2005) further distinguishes free-living, host-associated, and obligate symbiont organisms. The smallest bacterial genomes are from intracellular parasites or symbionts having an obligate relationship with a host. In general bacteria and archaea, having very small genome sizes, live in extremely stable environments in which the host provides reliable resources (e.g., nutrients) and homeostatic benefits (e.g., a constant pH). Organisms with small genomes evolved from ancestors with larger genomes. One of the smallest sequenced genomes of a free-living organism (and one of the first genomes to have been sequenced) is that of *Mycoplasma genitalium*, a urogenital pathogen. The *M. genitalium* has 580,070 bp encoding 470 protein-coding genes, 3 rRNA genes, and 33 tRNA genes (Fraser *et al.*, 1995). Mycoplasmas are bacteria of the class Mollicutes. They lack a cell wall and have a low GC content (32%) characteristic of this class.

Of the smallest bacterial genomes, *Buchnera aphidicola* has a genome of just 422,434 base pairs with 362 protein-coding genes (Pérez-Brocal *et al.*, 2006). The genome is organized in a circular chromosome and an additional 6 kb plasmid for leucine biosynthesis.



**FIGURE 17.2** Number of predicted protein-encoding genes versus genome size for 246 complete published genomes from bacteria and archaea. Giovannoni *et al.* (2005) reported that *P. ubique* has the smallest number of genes (1354 open reading frames) for any free-living organism that has been studied in the laboratory. Recent data from the smallest bacterial genomes are included. Adapted from Giovannoni *et al.* (2005) with permission from AAAS and S. Giovannoni.

Aphids are metazoans (animals) within the class Insecta. The *B. aphidicola* accession is NC\_008513.1.

There is an obligate endosymbiotic relationship between *B. aphidicola* and the cedar aphid *Cinara cedri*. The bacterium has lost most of its metabolic functions, depending on those provided by its host, while in turn it provides metabolites (the aphid diet is restricted to plant sap, so it needs essential amino acids and other nutrients). The relationship between host and bacterium is thought to have been established over 200 million years ago, with a continual reduction in the size of the bacterial genome such that it no longer possesses the capability to synthesize its own cell wall.

Another very small bacterial genome is that of another endosymbiont, *Carsonella ruddii* (indicated in Fig. 17.2). Its genome consists of a single circular chromosome of 159,662 base pairs with only 182 open reading frames (Nakabachi *et al.*, 2006). Both the small genome size and the low guanine plus cytosine content (GC content 16.5%) are exceptional. Half of the open reading frames encode proteins implicated in translation and amino acid metabolism. Like *B. aphidicola*, *C. ruddii* is an obligate endosymbiont of a sap-feeding insect, the psyllid *Pachypsylла venusta*.

Recently two even smaller genomes have been identified, again inside plant-feeding insects. *Candidatus Tremblaya princeps* has just under 139,000 base pairs and 121 protein-coding genes (see NC\_015736.1; Bennett and Moran, 2013). The *Nasuia deltocephalinicola* genome consists of 112,000 base pairs and 137 protein-coding genes. *Tremblaya* is a betaproteobacterium that resides inside a mealybug (*Planococcus citri*). Remarkably a gammaproteobacterium, *Candidatus Moranella endobia* lives inside *Tremblaya*. McCutcheon and von Dohlen (2011) describe this amazing case of nested symbiosis.

Among the archaea, the smallest genome is that of a hyperthermophilic organism that was cultured from a submarine hot vent, *Nanoarchaeum equitans* (Huber *et al.*, 2002). This archaeon appears to grow attached to another archeon, *Ignicoccus*. Because of its

small cell size (400 nm) and small genome size, Huber *et al.* (2002) suggested that *N. equitans* resembles an intermediate between the smallest living organisms (such as *M. genitalium*) and large viruses (such as the pox virus). Nonetheless, even parasitic intracellular bacteria and archaea originated as free-living organisms, so are classified as distinct from viruses.

By comparing small bacterial and archaeal genomes, it is possible to estimate the minimal number of genes required for life (Box 17.2). The *B. aphidicola* and *C. ruddii* genomes do not encode many proteins that serve transport functions, suggesting that their metabolites may freely diffuse to their hosts. Many required gene products could have been transferred to their hosts' nuclear genomes. Such a process has occurred in mitochondria, which depend on many proteins encoded by a eukaryotic nuclear genome.

See Andersson (2006) for a review of the *B. aphidicola* and *C. ruddii* genomes.

## Classification of Bacteria and Archaea Based on Lifestyle

In addition to the criteria of morphology and genome size and geometry, a third approach to classifying bacteria (and archaea) is based on their lifestyle. One main advantage of this approach is that it conveniently highlights the principle of extreme reduction in genome

### BOX 17.2 SMALL GENOME SIZES, MINIMAL GENOME SIZES, AND ESSENTIAL GENES

How many genes are required in the genome of the smallest living organism, that is, the smallest autonomous self-replicating organism? One approach is to identify the smallest genomes in nature. The *C. ruddii*, *N. deltocephalinicola*, and *Candidatus Tremblaya princeps* genomes encode only 182, 137, and 121 proteins, respectively. However, they are constrained to living within particular insect cells which support their survival. Bacteria of the genus *Mycoplasma* tend to have both small sizes and small genomes, and have therefore been studied in terms of minimal gene sets. At present, the genomes from 46 species of this genus have been sequenced. *M. genitalium* encodes 523–548 genes (depending on the strain) and has the smallest genome size of an autonomously replicating bacterium. The forces driving the evolution of small genome size include genome reduction from larger ancestral genomes in a process that may promote fitness of the organism. In thinking about a minimal genome size, we must always consider the ecological niche occupied by the organism; this will have an enormous influence on the particular genes of the endosymbiont as well as the mechanisms of reductive evolution.

A second approach involves comparative genomics by identifying the orthologs in common between several microbes. In the earliest days of complete genome sequencing, Mushegian and Koonin (1996) identified 239 genes in common between *Escherichia coli*, *H. influenzae*, and *M. genitalium*. This is considered one estimate of the minimal genome size. The functions of these 239 genes include several basic categories: translation, DNA replication, recombination and DNA repair, transcription, anaerobic metabolism, lipid and cofactor biosynthesis, and transmembrane transporters. Huang *et al.* (2013) described the overlap of the 517 *M. genitalium* genes with conserved core gene sets of Gram-negative and Gram-positive bacteria, reporting 151 common bacterial core genes (a total of 39 of these encode the 30S and 50S ribosomal subunits). A Database of Essential Genes (<http://www.essentialgene.org>) lists these genes (Luo *et al.*, 2014).

A third approach to determining the minimal number of genes required for life is experimental. Itaya (1995) randomly knocked out protein-coding genes in the bacterium *Bacillus subtilis*. Mutations in only 6 of 79 loci prevented growth of the bacteria and were indispensable. Extrapolating to the size of the complete *B. subtilis* genome, about 250 genes were estimated to be essential for life. Attempts are underway to create life forms from a specific set of genes. Pósfai *et al.* (2006) from the group of Frederick Blattner have experimentally reduced the genome size of *Escherichia coli* K-12 (by 20% to about 4 Mb), targeting the removal of insertion sequence elements and other mobile DNA elements as well as repeats that mediate structural changes (such as inversions, duplications, and deletions). Mizoguchi *et al.* (2007) further reduced the genome size to 3.6 Mb. For *M. tuberculosis*, random transposon mutagenesis has been employed to identify essential genes (Lamichhane *et al.*, 2003). This and related approaches can provide information on which genes and gene products are likely to be most useful as drug targets (Lamichhane and Bishai, 2007).

D'Elia *et al.* (2009) and Acevedo-Rocha *et al.* (2013) both note the importance of the context in defining a gene as essential. Many essential genes encode proteins having poorly characterized functions, and the physiological state of the bacterium likely influences the circumstances in which that gene acts.

Several groups have studied core sets of genes required for life, including Koonin (2003) and Gil *et al.* (2004). Koonin lists 63 genes that are present across all of ~100 genomes sequenced at the time. These include genes having functions in translation (e.g., ribosomal proteins and aminoacyl-transfer RNA synthetases and translation factors), transcription (RNA polymerase subunits), and replication and repair (DNA polymerase subunits, exonuclease, topoisomerase).

size that is associated with three lifestyles: extremophiles; and intracellular and epicellular bacteria and archaea.

- Extremophiles are microbes that live in extreme environments (Canganella and Wiegel, 2011). Archaea have been identified in hypersaline conditions (halophilic archaea), geothermal areas such as hot vents (hyperthermophilic archaea), and anoxic habitats (methanogens) (DeLong and Pace, 2001). One of the most extraordinary extremophiles is *Deinococcus radiodurans* that can survive dessication as well as massive doses of ionizing radiation (it thrives in nuclear waste). It achieves this feat by reassembling shattered chromosomes through a novel repair mechanism (Zahradka *et al.*, 2006).
- Intracellular bacteria invade eukaryotic cells; a well-known example is the  $\alpha$ -proteobacterium that is thought to have invaded eukaryotic cells and evolved into the present-day mitochondrion.
- Epicellular bacteria (and archaea) are parasites that live in close proximity to their hosts, but not inside host cells.

We may distinguish six basic lifestyles of bacteria and archaea (**Table 17.6**):

Worldwide, one-third of all people are infected with tuberculosis (and 9 million were sick with the disease in a single recent year) (see <http://www.cdc.gov/tb/>, WebLink 17.7). The *M. tuberculosis* genome was sequenced by Cole *et al.* (1998).

- Extracellular:** For example, *E. coli* commonly inhabits the human intestine without entering cells. Many free-living bacteria have relatively large genomes (as indicated in **Fig. 17.2**), such as the  $\delta$ -proteobacterium *Myxococcus xanthus* described above. Having a larger genome may provide a reservoir of genes that can be utilized to meet the needs of changing environments. As another example, the Gram-positive bacterium *Propionibacterium acnes* inhabits human skin and can cause acne. Its 2.5 Mb genome allows *P. acnes* the flexibility to grow under aerobic or anaerobic conditions and to utilize a variety of substrates available from skin cells (Brüggemann *et al.*, 2004).
- Facultatively intracellular bacteria** can enter host cells, but their behavior depends on environmental conditions. *Mycobacterium tuberculosis*, the cause of tuberculosis, can remain dormant within infected macrophages, only to activate and cause disease many decades later.
- Extrophilic microbes:** Initially, archaea were all identified in extreme environmental conditions. Some archaea have been found to grow at temperatures as high as 113°C, at pH 0, and in salt concentrations as high as 5 M sodium chloride. *Methanocaldococcus jannaschii*, the first archeal organism to have its genome completely sequenced (Bult *et al.*, 1996), grows at pressures over 200 atm and at an optimum temperature near 85°C. Archaea have subsequently been identified in less extreme habitats, including forest soil and ocean seawater (DeLong, 1998; Robertson *et al.*, 2005).
- Epicellular bacteria and archaea** grow outside of their hosts, but in association with them. *Mycoplasma pneumoniae*, a bacterium with a genome size of ~816,000 bp, is a major cause of respiratory infections. The bacterium is a surface parasite that attaches to the respiratory epithelium of its host. The genome was sequenced (Himmelreich *et al.*, 1996) and subsequently reannotated by Peer Bork and colleagues (Dandekar *et al.*, 2000).
- Obligately intracellular and symbiotic:** Tamas *et al.* (2002) compared the complete genome sequences of two bacteria, *Buchnera aphidicola* (Sg) and *Buchnera aphidicola* (Ap), that are endosymbionts of the aphids *Schizaphis graminum* (Sg) and *Acyrtosiphon pisum* (Ap). Each of these bacteria has a small genome size of about 640,000 bp. They have 564 and 545 genes, respectively, of which they share almost all (526). Remarkably, these bacteria diverged about 50 MYA yet they share complete conservation of genome architecture. There have been no inversions, translocations, duplications, or gene acquisitions in either bacterial genome since their divergence (Tamas *et al.*, 2002). This provides a dramatic example of genomic stasis. Although

**TABLE 17.6 Classification of bacteria and archaea based on ecological niche.**Adapted from <http://www.chlamydiae.com>.

Lifestyle	Bacterium	Genome size (Mb)	Reference
Extracellular	<i>Escherichia coli</i>	4.6	Blattner et al., 1997
	<i>Vibrio cholerae</i>	4.0	Heidelberg et al., 2000
	<i>Pseudomonas aeruginosa</i>	6.3	Stover et al., 2000
	<i>Bacillus subtilis</i>	4.2	Kunst et al., 1997
	<i>Clostridium acetobutylicum</i>	4.0	Nolling et al., 2001
	<i>Deinococcus radiodurans</i>	3.3	White et al., 1999
Facultatively intracellular	<i>Salmonella enterica</i>	4.8	Parkhill et al., 2001a
	<i>Yersinia pestis</i>	4.7	Parkhill et al., 2001b
	<i>Legionella pneumophila</i>	3.9	Bender et al., 1990
	<i>Mycobacterium tuberculosis</i>	4.4	Cole et al., 1998
	<i>Listeria monocytogenes</i>	2.9	Glaser et al., 2001
Extremophile	<i>Aeropyrum pernix</i>	1.7	Kawarabayasi et al., 1999
	<i>Methanococcus jannaschii</i>	1.7	Bult et al., 1996
	<i>Archeoglobus fulgidus</i>	2.2	Klenk et al., 1997
	<i>Thermotoga maritima</i>	1.9	Nelson et al., 1999
	<i>Aquifex aeolius</i>	1.6	Deckert et al., 1998
Epicellular	<i>Neisseria meningitidis</i>	2.2	Tettelin et al., 2000
	<i>Haemophilus influenzae</i>	1.8	Fleischmann et al., 1995
	<i>Mycoplasma genitalium</i>	0.6	Fraser et al., 1995
	<i>Mycoplasma pneumoniae</i>	0.8	Himmelreich et al., 1996
	<i>Ureaplasma urealyticum</i>	0.8	Glass et al., 2000
	<i>Mycoplasma pulmonis</i>	1.0	Chamraud et al., 2001
	<i>Borrelia burgdorferi</i>	0.9	Fraser et al., 1997; Casjens et al., 2000
	<i>Treponema pallidum</i>	1.1	Fraser et al., 1998
	<i>Helicobacter pylori</i>	1.7	Tomb et al., 1997; Alm et al., 1999
	<i>Pasteurella multocida</i>	2.3	May et al., 2001
Obligate intracellular, symbiotic	<i>Buchnera</i> sp.	0.6	Shigenobu et al., 2000
	<i>Wolbachia</i> spp.	1.1	Sun et al., 2001
	<i>Wigglesworthia glossinidia</i>	0.7	Akman et al., 2002
	<i>Sodalis glossinidius</i>	2.0	Akman et al., 2001
Obligate intracellular, parasitic	<i>Rickettsia prowazekii</i>	1.1	Andersson et al., 1998
	<i>Rickettsia conorii</i>	1.3	Ogata et al., 2001
	<i>Ehrlichia chaffeensis</i>	1.2	Hotopp et al., 2006
	<i>Cowdria ruminantium</i>	1.6	de Villiers et al., 2000
	<i>Chlamydia trachomatis</i>	1.1	Stephens et al., 1998; Read et al., 2000
	<i>Chlamydophila pneumoniae</i>	1.3	Kalman et al., 1999; Read et al., 2000; Shirai et al., 2000

it is extremely rare for obligate intracellular bacteria to share such genome conservation, it is common for endosymbionts to have relatively small genome sizes. This may reflect the dependence of these bacteria on nutrients derived from the host.

6. *Obligately intracellular and parasitic:* *Rickettsia prowazekii* is the bacterium that causes epidemic typhus. Its genome is relatively small, consisting of 1.1 Mb (Andersson et al., 1998). Like other *Rickettsia*, it is an  $\alpha$ -proteobacterium that infects eukaryotic cells selectively. It is also of interest because it is closely related to the

**TABLE 17.7 Vaccine-preventable bacterial diseases. Adapted from CDC-DPDx, <http://www.cdc.gov/vaccines/vpd-vac/vpd-list.htm> and <http://www.cdc.gov/DiseasesConditions/>.**

Disease	Species
Anthrax	<i>Bacillus anthracis</i>
Diarrheal disease (cholera)	<i>Vibrio cholerae</i>
Diphtheria	<i>Corynebacterium diphtheriae</i>
Community acquired pneumonia	<i>Haemophilus influenzae</i> type B, <i>Streptococcus pneumoniae</i>
Lyme disease	<i>Borrelia burgdorferi</i>
Meningitis	<i>Haemophilus influenzae</i> type B (HIB), <i>Streptococcus pneumoniae</i> , <i>Neisseria meningitidis</i>
Pertussis	<i>Bordetella pertussis</i>
Tetanus	<i>Clostridium tetani</i>
Tuberculosis	<i>Mycobacterium tuberculosis</i>
Typhoid	<i>Salmonella typhi</i>

mitochondrial genome. A closely related species, *Rickettsia conorii*, is an obligate intracellular parasite that causes Mediterranean spotted fever in humans. Its genome was sequenced by Ogata *et al.* (2001). Similar to the *Buchnera aphidicola* subspecies, the genome organization of the two *Rickettsia* parasites is well conserved.

Why are some bacterial genome sizes severely reduced? Intracellular parasites are subject to deleterious mutations and substitutions that cause gene loss, tending toward genome reduction (Andersson and Kurland, 1998; McCutcheon and Moran, 2011). A similar process occurred as a primordial  $\alpha$ -proteobacterium evolved into the modern mitochondrion, maintaining only a minuscule mitochondrial genome size (Chapter 15).

### Classification of Bacteria Based on Human Disease Relevance

Bacteria and eukaryotes have engaged in an ongoing war for millions of years. Bacteria occupy the nutritive environment of the human body in an effort to reproduce. Typical sites of bacterial colonization include the skin, respiratory tract, digestive tract (mouth, large intestine), urinary tract, and genital system (Eisenstein and Schaechter, 1999). It has been estimated that each human has more bacterial cells than human cells in the body. In the majority of cases, these bacteria are harmless to humans. However, many bacteria cause infections, often with devastating consequences.

In recent years, the widespread use of antibiotics has led to an increased prevalence of drug resistance among bacteria. It is therefore imperative to identify bacterial virulence factors and to develop strategies for vaccination (Bush *et al.*, 2011). One approach to this problem is to compare pathogenic and nonpathogenic strains of bacteria (see “Comparison of Bacterial Genomes” below). **Table 17.7** lists some of the bacterial diseases for which vaccinations are routinely administered. The worldwide disease burden caused by bacteria is enormous. For example, there are 690,000 new cases of leprosy reported annually worldwide; the causative agent is *Mycobacterium leprae*. There are millions of cases of salmonellosis each year, caused by *Salmonella enterica*. A pathogenic strain of *E. coli* (O157:H7) causes haemorrhagic colitis and infects 75,000 individuals in the United States each year. As mentioned above, *M. tuberculosis* infects billions of people and kills millions.

PATRIC is a bacterial bioinformatics resource center (Gillespie *et al.*, 2011). It centralizes information about large numbers of strains of pathogenic bacteria, including expert curation and analyses of metabolic pathways in those organisms. Annotations are performed using Rapid Annotation using Subsystem Technology (RAST; Overbeek *et al.*, 2014; see “Gene Annotation” below).

You can read about a variety of bacterial diseases at the Centers for Disease Control and Prevention website (<http://www.cdc.gov/DataStatistics/>, WebLink 17.8).

PATRIC is available online at <http://patricbrc.vbi.vt.edu/> (WebLink 17.4).

An emerging theme in the biology of bacteria and archaea is that, in addition to mutation, bacterial populations undergo recombination, causing genetic diversification (Fraser *et al.*, 2007). Species can be defined as clusters of genetically related strains, and the exchange of DNA by homologous recombination or other processes can complicate species definitions. Joyce *et al.* (2002) have reviewed recombination in the context of pathogenic bacteria such as *Helicobacter pylori* (a leading cause of gastric ulcers), *Streptococcus pneumoniae*, and *Salmonella enterica*. While eukaryotes achieve genetic diversity through sexual reproduction, bacteria and archaea also achieve tremendous genetic diversity through both recombination and lateral gene transfer (see section on “Lateral Gene Transfer” below).

We can consider bacteria that infect humans. Additionally, there are many pathogenic bacteria of plants which can cause human suffering by devastating crops. Mansfield *et al.* (2012) reported the results of a poll of the research community for the most economically and scientifically important plant pathogens. These included a group of *Pseudomonas syringae* pathovars in first place and three *Xanthomonas* species.

## Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences

A main way to describe the diversity of microbial life is by molecular phylogeny. Trees have been generated based on multiple sequence alignments of 16S rRNA and other small rRNAs from various species. Ribosomal RNA has excellent characteristics as a molecule of choice for phylogeny: it is distributed globally; it is highly conserved yet still exhibits enough variability to reveal informative differences; and it is only rarely transferred between species. An example of a rRNA-based tree is shown in Figure 17.1, and we saw a similar tree reconstruction in Figure 15.1. Through rRNA and other genome-based trees, bacterial and archaeal genomics are having a major impact on microbial systematics (Klenk and Göker, 2010; Zhi *et al.*, 2012).

A major conclusion of early rRNA studies by Carl Woese and colleagues (Woese and Fox, 1977; Fox *et al.*, 1980) is that bacteria and archaea are distinct groups. The deepest branching phyla are hyperthermophilic microbes, consistent with the hypothesis that the universal ancestor of life existed at hot temperatures (Achenbach-Richter *et al.*, 1987).

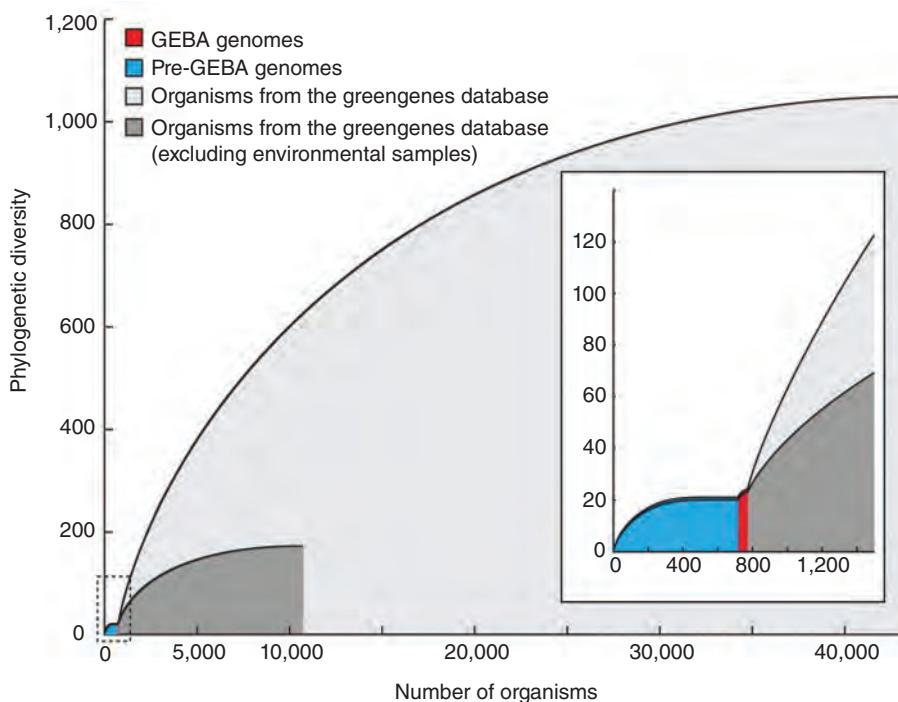
A great advance in our appreciation of microbial diversity came from the realization that the vast majority of bacteria and archaea are noncultivable (Hugenholtz *et al.*, 1998). It is straightforward to obtain microbes from natural sources and grow some of them in the presence of different kinds of culture medium. For the great majority of microbes however, perhaps >99%, culture conditions are not known. It is still possible to sample uncultivated (or uncultivable) microbes by extracting nucleic acids directly from naturally occurring habitats (DeLong and Pace, 2001). Norman Pace and colleagues pioneered the analysis of rRNA to characterize uncultivated species.

Because there is a sampling bias towards cultivatable microbes, just four bacterial phyla have been characterized most fully: Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes (Hugenholtz, 2002). These major groups account for over 90% of all known bacteria (discussed in Gupta and Griffiths, 2002). However, 35 bacterial and 18 archaeal phylum-level lineages are currently known (Hugenholtz, 2002). Analyses of uncultivated microbes will expand our view of bacterial and archaeal diversity.

In an approach that may be called diversity-driven phylogenomics, Jonathan Eisen and colleagues initiated the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project (Wu *et al.*, 2009). Its goal is to generate and analyze complete genome sequences from the tree of life based on phylogenetic diversity. They identified the most divergent lineages that lacked sequenced genomes, and selected cultivatable representatives for analysis. The first report included 56 complete genomes having ~16,800 protein families, ~1700 of which displayed no significant sequence similarity to any known proteins.

Reysenbach and Shock (2002) described a phylogenetic tree of extremophilic microbes based on 16S rRNA sequences. They used a software package designed for rRNA studies, called ARB (Chapter 10). You can obtain this software at <http://www.arb-home.de/> (WebLink 17.9).

The Joint Genomics Institute (JGI) offers a GEBA website at <http://genome.jgi.doe.gov/programs/bacteria-archaea/GEBA.jsf> (WebLink 17.10).



**FIGURE 17.3** Estimates of the phylogenetic diversity of bacteria and archaea from small subunit ribosomal RNA (SSU rRNA) genes. The plot is based on analysis of a phylogenetic tree of unique SSU rRNA sequences. Phylogenetic diversity (based on SSU rRNA sequences) was estimated from: (1) genome sequences before GEBA (blue); (2) 56 complete genomes contributed by the GEBA project (red); (3) all cultured organisms (gray); and (4) all available SSU rRNA genes (light gray).

Source: Wu *et al.* (2009). Reproduced with permission from Macmillan Publishers.

How much bacterial and archaeal diversity has yet to be characterized? The GEBA project has used small subunit rRNA gene sequences as a measure of organismal diversity (Fig. 17.3, y axis; Wu *et al.*, 2009). They estimated that half the genetic diversity of known, cultured bacteria and archaea could be captured by sequencing ~1500 isolates based on the criterion of phylogenetic diversity. The extent of diversity from uncultured species (as indicated by the analysis of rRNA sequences) is vastly greater. The authors estimated that sequencing ~9200 genomes from archaea and bacteria that are not cultured would encompass 50% of this additional diversity (Fig. 17.3).

In another example of diversity-driven phylogenomics, Shih *et al.* (2013) sequenced the genomes of 54 diverse strains of cyanobacteria. This study focused on a single phylum, and was called the CyanoGEBA dataset since it was inspired by the broader GEBA approach. About 21,000 of the identified proteins (out of a total of ~193,000 proteins) had no detectable homology to known proteins. The cyanobacteria are oxygenic photosynthetic organisms, and the genome sequences provided insight into the origins of plant plastids (photosynthetic organelles) that derive from cyanobacteria.

### Classification of Bacteria and Archaea Based on Other Molecular Sequences

In addition to rRNA, many other DNA, RNA, or protein sequences can be used for molecular phylogeny studies. One motivation to do this is that the analysis of 16S ribosomal RNA sequence occasionally yields conflicting results. For example, the  $\alpha$ -proteobacterium *Hyphomonas neptunium* is classified as a member of the order *Rhodobacterales* based on

16S rRNA but *Caulobacterales* based on 23S rRNA as well as according to ribosomal proteins HSP70 and EF-Tu (Badger *et al.*, 2005). This is potentially due to lateral gene transfer (see section on this topic below). In other instances, 16S rRNA of unusual composition has been identified (Baker *et al.*, 2006). Because of concerns about the properties of 16S rRNA for phylogenetic analysis, Teeling and Gloeckner (2006) introduced RibAlign, a database of ribosomal protein sequences. The HOGENOM database is another resource that is useful for phylogenetic studies. It includes large numbers of protein families across the tree of life.

The use of individual proteins (or genes) for such studies sometimes yields tree topologies that conflict with each other and with topologies obtained using rRNA sequences. These discrepancies are usually attributed either to lateral gene transfer (see below), which can confound phylogenetic reconstruction, or to the loss of phylogenetic signals due to saturating levels of substitutions in the gene or protein sequences. A strategy to circumvent this problem is to use combined gene or protein sets. Brown *et al.* (2001) aligned 23 orthologous proteins conserved across 45 species. Their trees supported thermophiles as the earliest evolved bacteria lineages (Fig. 17.4). Matrices of concatenated, conserved protein alignments are commonly used for phylogenomic studies. The Shih *et al.* (2013) study of cyanobacteria (see above) relied on 31 conserved proteins for phylogenetic reconstructions.

There are many other approaches to bacterial phylogeny. One is to identify conserved insertions and deletions in a large group of proteins. Such “signature sequences” can distinguish bacterial groups and form the basis of a tree (see Web Document 17.1; Gupta and Griffiths, 2002). This tree shows the relative branching order of bacterial species from completed genomes. In an early study, Eugene Koonin and colleagues (Wolf *et al.*, 2001) used five independent approaches to construct trees for 30 completely sequenced bacterial genomes and 10 sequenced archaeal genomes. Their approach included: (1) assessing genes that are present or absent in various categories of functional annotation; (2) assessing the conservation of local gene order (i.e., pairs of adjacent genes) among the genomes; (3) measuring the distribution of percent identity between likely orthologs; (4) aligning 32 ribosomal proteins into a multiple sequence alignment consisting of 4821 columns (characters) and then generating a tree using the maximum-likelihood approach; and (5) comparing multiple trees generated from a series of protein alignments. These approaches can produce complementary information about phylogenetic reconstructions.

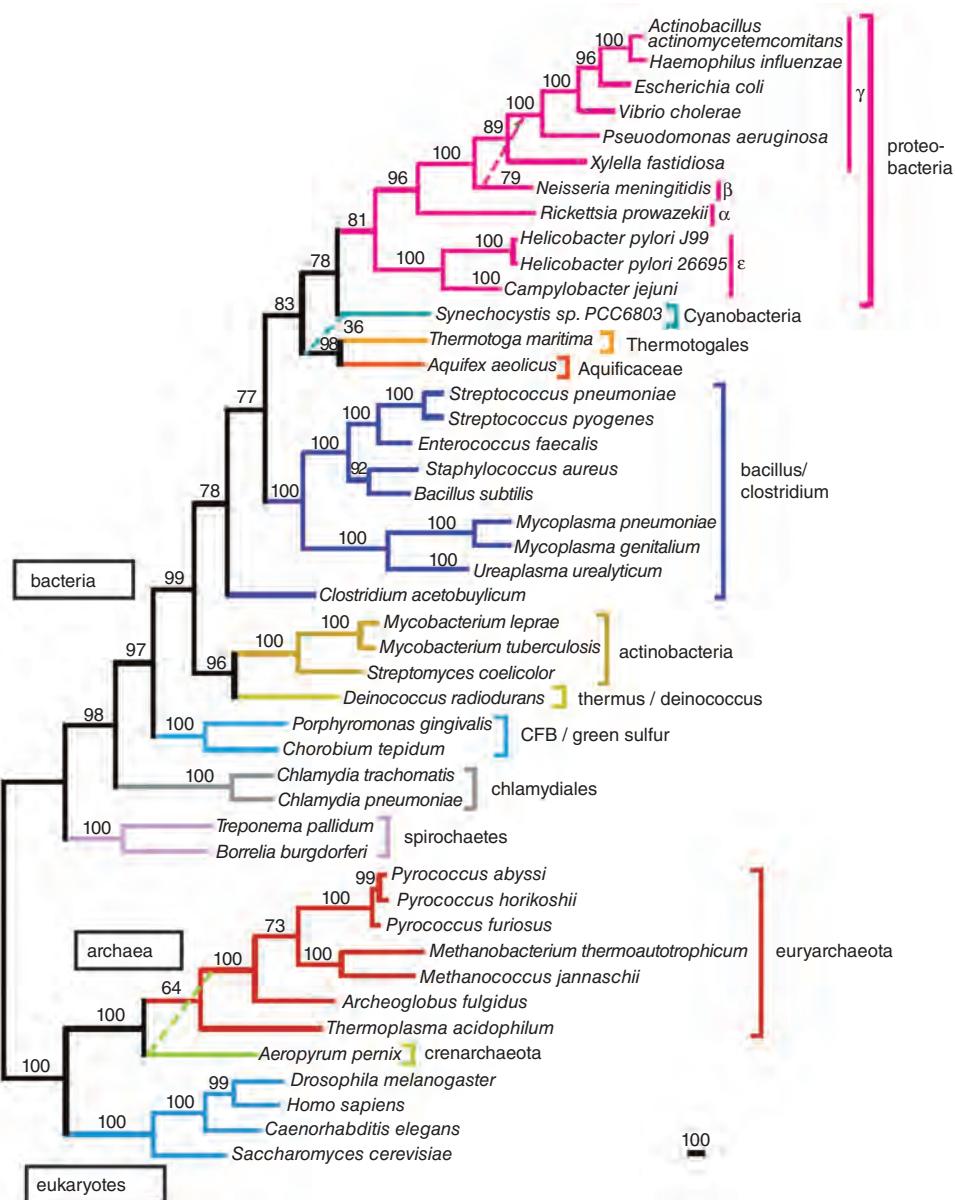
## THE HUMAN MICROBIOME

Most of us think of our bodies as consisting of mostly human cells, occasionally housing some bacteria in our mouths and guts. However, it has been estimated that there are ten times more bacterial cells than human cells in our bodies; citing earlier sources, Savage (1977) suggested a typical person has  $10^{13}$  animal cells and  $10^{14}$  bacterial cells. These bacteria, as well as some archaea, viruses, and eukaryotes, collectively may contain greater than two orders of magnitude more genes than are encoded by our human genome (Gill *et al.*, 2006). This collection of foreign genomes in our bodies is referred to as the human microbiome. Most are commensal, coexisting and helping to digest food and facilitate our metabolism; some are pathogenic. Together they weigh about 1.5 kg in a typical human gut.

Two large-scale projects have characterized our microbiome: the Human Microbiome Project (HMP) and the Metagenomics of the Human Intestinal Tract (MetaHIT). The HMP analyzed the microbiome of 242 healthy adults, sampling 15 or 18 body sites up to three times (Human Microbiome Project Consortium, 2012a). Their goal was to

RibAlign is available at <http://www.megx.net/ribalign> (WebLink 17.11). Its multiple sequence alignments of ribosomal proteins use MAFFT (Chapter 6). Homologous Sequences in Complete Genomes Database (HOGENOM) is available at <http://pbil.univ-lyon1.fr/databases/hogenom/home.php> (WebLink 17.12).

We study eukaryotes from the perspective of a tree that uses a combined protein dataset (Fig. 19.1).



**FIGURE 17.4** An unrooted tree of life redrawn from Brown *et al.* (2001) is based on an alignment of 23 proteins (spanning 6591 amino acid residues). These proteins are conserved across 45 species and include tRNA synthetases, elongation factors, and DNA polymerase III subunit. By combining these proteins, there are many phylogenetically informative sites. The tree consists of three major, monophyletic branches of life as described in Chapter 15. The tree was generated in PAUP by maximum parsimony (described in Chapter 7). Numbers along the branches show percentage of nodes in 1000 bootstrap replicates. Scale bar corresponds to 100 amino acid substitutions. Adapted from Brown *et al.* (2001).

The HMP website is <http://commonfund.nih.gov/hmp/index> (WebLink 17.13). Sponsored by the National Institutes of Health, it cost US\$ 170 million. The MetaHIT consortium was funded by the European Commission (for € 21.2 million), and its website is <http://www.metahit.eu/> (WebLink 17.14).

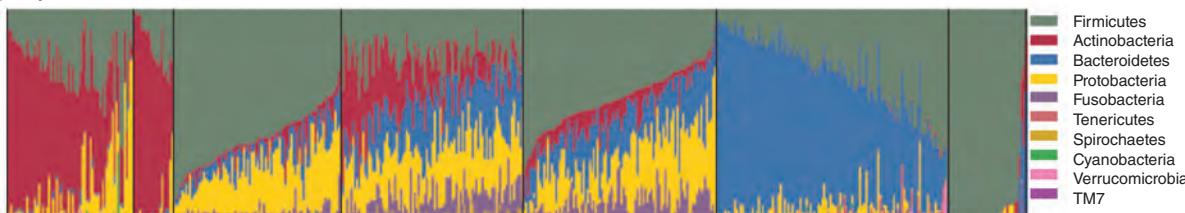
establish viral, bacterial, and eukaryotic reference genomes using 16S rRNA sequencing, whole-genome shotgun (WGS) sequencing, or metagenomic sequencing. The sampled sites spanned five body areas such as the oral cavity (from saliva to throat), nares (nostrils), skin specimens (from the retroauricular creases behind each ear and from the inner elbows), stool, and, for women, three vaginal sites (Human Microbiome Project Consortium, 2012a). The MetaHIT consortium focused on the gut microbiome by analyzing fecal samples from 124 European individuals (Qin *et al.*, 2010).

We may summarize some of the major findings of these consortia as follows (Pennisi, 2012; Morgan *et al.*, 2013; also see Dave *et al.*, 2012). Five leaders in this field provide an overview of findings and trends in the field (Blaser *et al.*, 2013).

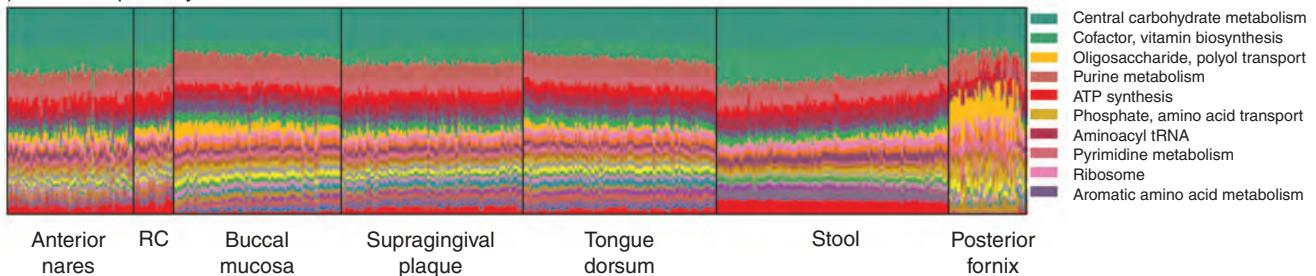
1. There are extraordinary bioinformatics challenges associated with these types of projects, which for HMP involved collecting over 3.5 terabases of DNA sequence by 2012 (Human Microbiome Project Consortium, 2012a). Weinstock (2012), Teeling and Glöckner (2012) as well as Rob Knight and colleagues (Kuczynski *et al.*, 2012) have reviewed some of the bioinformatics tools required for this research. IMG/M is an example of a set of software tools for metagenome analysis (Markowitz *et al.*, 2014).
2. Most of the microbiome is bacterial. The MetaHIT consortium reported that 0.14% of reads were human contamination (following standard efforts to remove human sequences), with additional sequences from other eukaryotes (accounting for 0.5% of reads), archaea (0.8%), and viruses (up to 5.8%) (Arumugam *et al.*, 2011).
3. There is no single reference microbiome because there is such enormous diversity of species within each individual and between individuals (Morgan *et al.*, 2013).
4. Each body region does have characteristic bacterial species within each individual, and these often occur in common between individuals. Despite the great diversity, bacterial species are therefore not randomly distributed. A plot from the HMP showing bacterial phyla across seven body regions shows some of the dominant phyla (**Fig. 17.5a**). For any given body region there tends to be a single major phylum (and often genus), although that phylum often differs between individuals. In feces, *Bacteroides* is the most abundant and the most variable species, and the amounts of these as well as *Prevotella* and *Ruminococcus* define three clusters or enterotypes of microbes (Arumugam *et al.*, 2011).
5. While bacterial phyla and genera vary greatly across body regions, the HMP made the remarkable discovery that most metabolic pathways are evenly distributed and evenly prevalent across body regions and between individuals (Human Microbiome Project Consortium, 2012a; **Fig. 17.5b**). Future efforts to alter the microbiome to promote health might therefore focus on understanding the status of functional

IMG is online at <http://img.jgi.doe.gov/> (WebLink 17.5).

(a) Phyla



(b) Metabolic pathways



**FIGURE 17.5** Characterization of bacterial taxa in human microbiomes. (a) Microbial phyla vary greatly across different body regions of human microbiomes. Regions include the retroauricular crease (RC; the skin behind the ears) and the posterior fornix of the vagina. (b) A series of 10 metabolic modules were characterized based on the functional characterization of the bacterial species. Most metabolic pathways are conserved across body regions and across individuals, in contrast to the particular phyla that are present.

Source: Human Microbiome Project Consortium (2012b). Reproduced with permission from Macmillan Publishers.

pathways then modifying them as needed, rather than trying to promote or eliminate particular species.

6. There has been great interest in the possible role of the microbiome in human disease including obesity, psoriasis, asthma, and bowel diseases (Cho and Blaser, 2012; Zhao, 2013). Turnbaugh *et al.* (2009) studied microbiomes from fecal samples of female monozygotic and dizygotic twin pairs concordant for leanness or obesity. They found that the gut microbial community is shared by family members, with reduced microbial diversity in obese individuals. The MetaHIT consortium reported consistent findings (Le Chatelier *et al.*, 2013), suggesting that even a few bacterial species can distinguish individuals who are lean versus obese (as well as distinguishing those with high versus low bacterial richness).

EcoCyc is online at <http://ecocyc.org/> (WebLink 17.15), Regulon is at <http://regulondb.ccg.unam.mx/> (WebLink 17.16), and EcoGene is available at <http://ecogene.org/> (WebLink 17.17). For each database try entering a query for the gene *BLC* and you will see a variety of data including its genomic context, links to structural genomics projects, and BLAST links. Julio Collado-Vides and colleagues have expertly curated the transcription initiation sites and operon organization of *E. coli* with an emphasis on elucidating the regulatory networks.

## ANALYSIS OF BACTERIAL AND ARCHAEL GENOMES

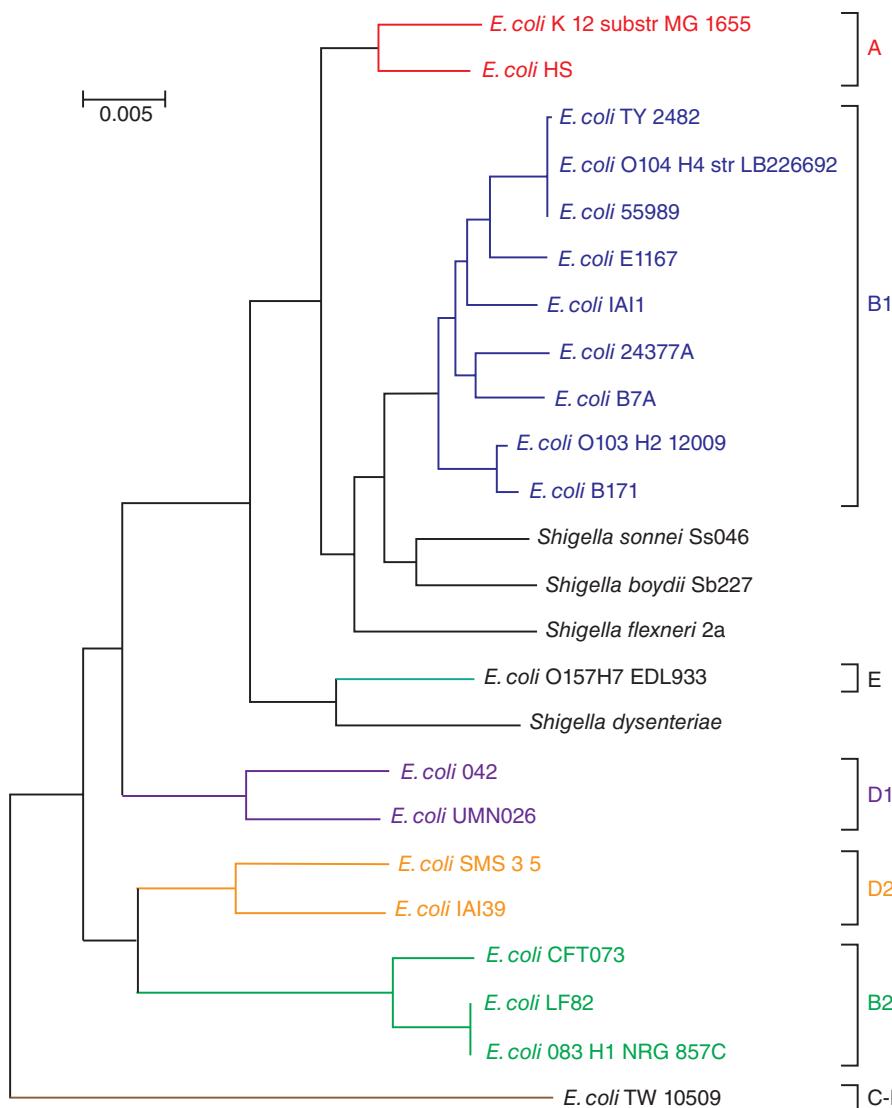
Some of the main attributes of a bacterial or archaeal genome are its genome size, nucleotide composition, gene content, extent of lateral gene transfer, and functional annotation. We can approach this subject by considering *Escherichia coli*, arguably the best-characterized bacterium.

We begin with a phylogenetic perspective (Chaudhuri and Henderson, 2012). The initial genome sequence analysis by Blattner *et al.* (1997) was of *E. coli* K-12 strain MG1655, and it continues to be annotated and used as a reference genome (Riley *et al.*, 2006). The annotation process includes an effort by the community to correct sequence errors, to update the boundaries for genes and transcripts (based for example on models for gene structures in related bacteria), and to assign functional descriptions for all genes (as described in Chapter 14). There are online resources that centralize information about *E. coli* such as EcoCyc (Keseler *et al.*, 2013), RegulonDB (Salgado *et al.*, 2006), and EcoGene (Rudd, 2000).

The next *E. coli* genomes to be sequenced were pathogenic EHEC O157:H7 strains Sakai (RIMD 0509952) and EDL933 (Fig. 17.6, clade B). The *E. coli* O157:H7 strain appears in contaminated food, causing disease such as hemorrhagic colitis. These strains diverged from *E. coli* K-12 MG1655 about 4.5 MYA (Reid *et al.*, 2000). Both genomes were sequenced and compared (Blattner *et al.*, 1997; Hayashi *et al.*, 2001; Perna *et al.*, 2001; reviewed in Eisen, 2001). *Escherichia coli* O157:H7 is about 859,000 bp larger than *E. coli* K-12. The two bacteria share a common backbone of about 4.1 Mb, while *E. coli* O157:H7 has an additional 1.4 Mb sequence comprised largely of genes acquired by lateral gene transfer.

The next *E. coli* genome to be sequenced was that of strain CFT073 (clade B2). Unexpectedly, only 10% of the CFT073-specific genes relative to MG 1655 were also shared by the O157:H7 genome. Chaudhuri and Henderson (2012) traced the efforts to characterize additional *E. coli* genomes. *Shigella*, identified in the late nineteenth century, was thought to belong to a distinct genus because of phenotypic differences (e.g., it is nonmotile in contrast to *E. coli*, and cannot ferment lactose). However, phylogenetic analyses clearly place *Shigella* spp. in the same genus as *E. coli*, as shown in Figure 17.6.

In May 2011 there was a large outbreak of Shiga toxin-producing *E. coli* O104:H4. There were over 4000 cases and 50 deaths; symptoms included diarrhea and a hemolytic-uremic syndrome. Several groups, including Rasko *et al.* (2011) promptly sequenced the German outbreak strain (see clade B1) as well as 12 additional *E. coli* genomes. They identified structural variation between O104:H4 and other enteropathogenic O104:H4 isolates. The outbreak strain included two lambda-like prophage elements, including one containing the genes for Shiga toxin. They concluded that this strain acquired its virulence by lateral transfer (see “Lateral Gene Transfer” below). This episode highlights the



**FIGURE 17.6** Phylogenetic relationships of *E. coli* strains. The tree was generated by aligning complete and draft genome sequences spanning 2.78 Mb (excluding positions with gaps), then using a maximum likelihood tree-building approach. Bootstrap replicates (not shown) were all 100. Note that group B1 includes two strains involved in a haemorrhagic uraemic syndrome outbreak in Germany in 2011 (TY 2482 and O104 H4 str LB226692) that are closely related to enteropathogenic *E. coli* (EAEC) strain 55989, a pathovar. Redrawn from Chaudhuri and Henderson (2012). Reproduced with permissions from Elsevier.

emerging role of next-generation sequencing in rapidly identifying disease-associated pathogens; the authors reported requiring five hours to sequence each isolate.

As we focus on *E. coli* K-12 MG1655, we can select from a series of extremely rich resources.

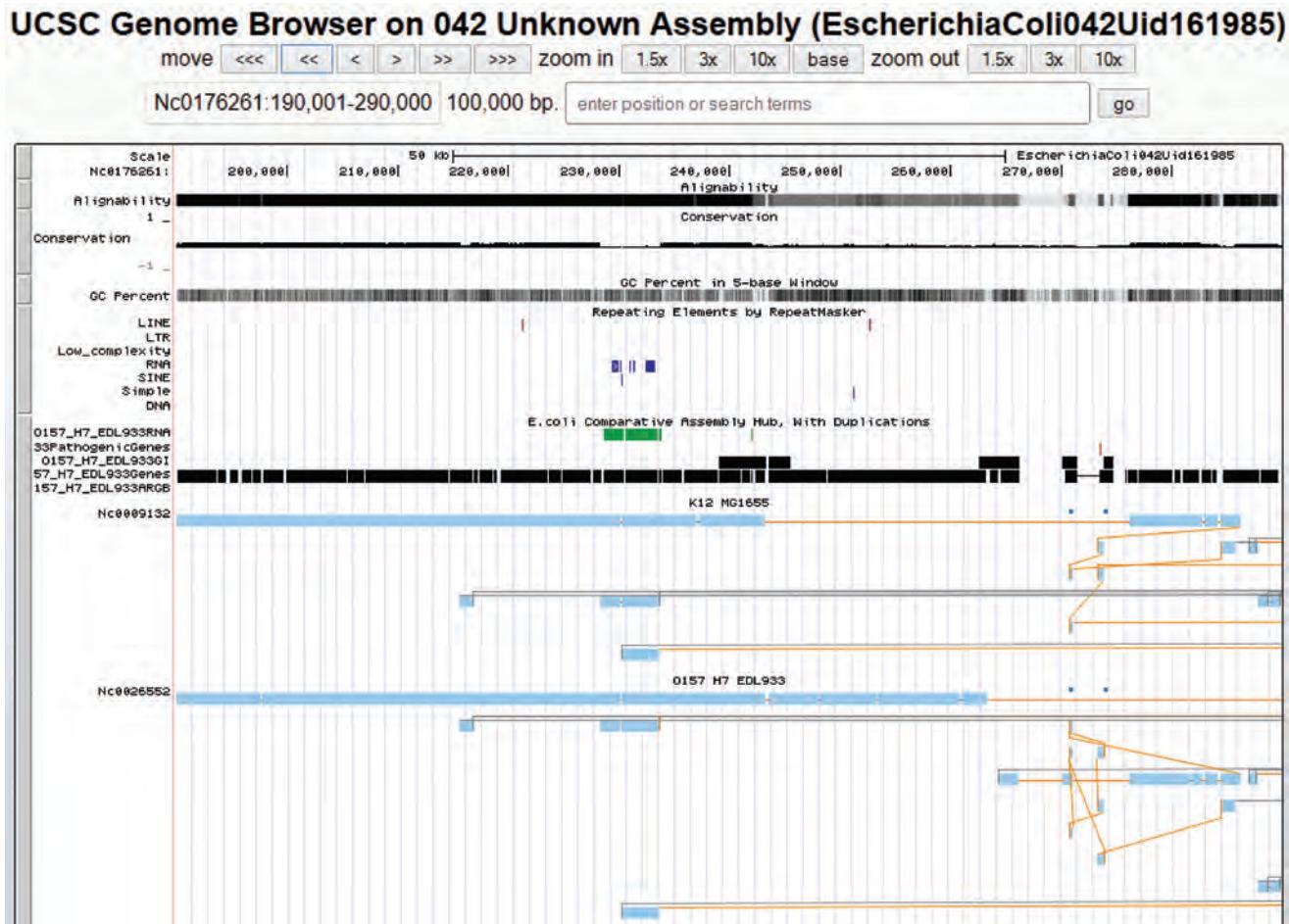
- A search of the Genomes Online Database (GOLD) for the organism *Escherichia coli* shows an interactive world map of where *E. coli* has been isolated, and also lists >2400 projects (62 complete and published genomes, ~1200 permanent drafts, and many incomplete projects). Following the link to *Escherichia coli* K-12, MG1655 (Goldstamp Gc00008), we find a page with a wealth of information on its genome including summaries of the DNA molecule (4640 kilobase pairs; 51% GC content; 4497 open reading frames) and external links.

	Number	% of Total
<b>DNA, total number of bases</b>	4639675	100.00%
DNA coding number of bases	3992744	86.06%
DNA G+C number of bases	2356477	50.79% <sup>1</sup>
<b>DNA scaffolds</b>	1	100.00%
CRISPR Count	2	
<b>Genes total number</b>	4497	100.00%
Protein coding genes	4321	96.09%
Pseudo Genes	178	3.96% <sup>2</sup>
RNA genes	176	3.91%
rRNA genes	22	0.49%
5S rRNA	8	0.18%
16S rRNA	7	0.16%
23S rRNA	7	0.16%
tRNA genes	89	1.98%
Other RNA genes	65	1.45%
Protein coding genes with function prediction	3906	86.86%
without function prediction	415	9.23%
Protein coding genes connected to SwissProt Protein Product	4264	94.82%
not connected to SwissProt Protein Product	57	1.27%
Protein coding genes with enzymes	1385	30.80%
Protein coding genes connected to Transporter Classification	739	16.43%
Protein coding genes connected to KEGG pathways <sup>3</sup>	1463	32.53%
not connected to KEGG pathways	2858	63.55%
Protein coding genes connected to KEGG Orthology (KO)	2933	65.22%
not connected to KEGG Orthology (KO)	1388	30.87%
Protein coding genes connected to MetaCyc pathways	1343	29.86%
not connected to MetaCyc pathways	2978	66.22%

**FIGURE 17.7** The Integrated Microbial Genomes (IMG) website offers data on bacterial genomes such as *E. coli* K-12 MG1655. IMG also offers extensive tools for metagenomics analyses.

Source: IMG.

- The NCBI Project for that organism provides access to raw DNA sequence (e.g., SRA files when available).
- The Integrated Microbial Genomes (IMG) site includes genomic data as shown in Figure 17.7 including a web browser, lists of putative laterally transferred genes, annotation data, and analyses of the phylogenetic distribution of genes.
- EcoCyc is a major resource for *E. coli* (Keseler *et al.*, 2013). EcoCyc is a part of BioCyc which encompasses ~3000 databases of pathways and organisms (Latendresse *et al.*, 2012).



**FIGURE 17.8** The UCSC Genome Browser offers an *E. coli* hub, currently with access to 72 *E. coli* genomes. Features include alignability, conservation, GC percent in windows, repeat elements from RepeatMasker (several of which are shown), and comparative assembly data (here with K-12 MG1655 and O157:H7 EDL933). UCSC also offers a microbial browser.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

- EnsemblBacteria includes large numbers of *E. coli* strains. For MG1655 it includes access to the sequence (in FASTA format or from the European Nucleotide Archive), comparative genomics tools such as gene trees, and data on the genome build.
- The UCSC Genome Browser include a microbial browser. There is also an annotation hub that can be used to compare genomic features for dozens of *E. coli* strains. An example comparing K-12 MG1655 and O157:H7 is shown in **Figure 17.8**.
- Galaxy includes access to the UCSC Archaea Table Browser as well as access to BoMart, UCSC, and EBI resources.

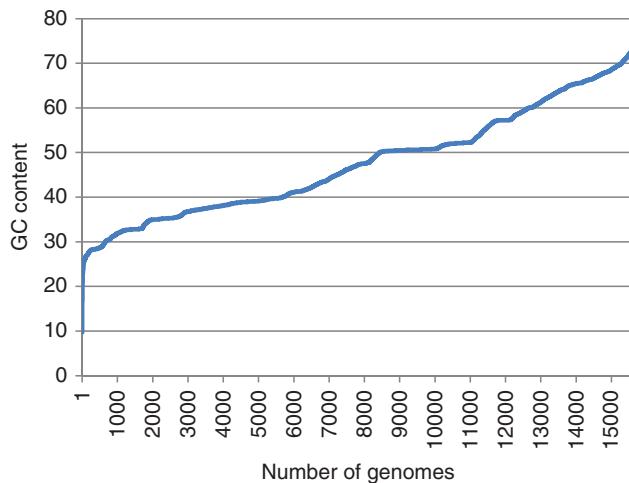
Visit the EnsemblBacteria at  
<http://bacteria.ensembl.org>  
 (WebLink 17.3).

Deciding which to use depends on your preferences and the nature of the project (e.g., IMG offers particularly strong metagenomics resources). Popular bioinformatics resources Ensembl, NCBI, and UCSC offer familiarity for those used to working with them.

## Nucleotide Composition

In the analysis of a completed genome, the nucleotide composition has characteristic properties. The GC content is the mean percentage of guanine and cytosine and, as first reported by Noboru Sueoka (1961) it typically varies from 25 to 75% in bacteria (**Fig. 17.9**). Eukaryotes almost always have a larger and more variable genome size than bacteria, but their GC content is very uniform (around 40–45%). Within each species, nucleotide composition tends to be uniform.

We showed the range of GC content in **Figure 15.13**.



**FIGURE 17.9** GC content for ~15,000 bacterial and archaeal genomes. Data from NCBI Genome.

Determine GC content with the Emboss program GEECEE (<http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::geecce,> WebLink 17.18) or with other programs such as GLIMMER (see the following section).

We modify an excellent online tutorial by Avril Coglan, available at <http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter2.html> (WebLink 17.19).

GC content varies within an individual genome. Regions having atypical GC content sometimes reflect invasions of foreign DNA (such as phage DNA incorporating into bacterial genomes). The GC content is highest (AT content lowest) in intergenic regions, possibly because of the requirements of transcription factor-binding sites (Mitchison, 2005). GC content is also related to the frequency of codon utilization; we explore this in a computer lab exercise (17.3) at the end of this chapter.

We analyze the GC content of an *E. coli* strain using an R package, `seqinr`. We can divide our task into three parts: (1) obtaining a genome sequence in the FASTA format; (2) determining the overall GC content; and (3) measuring the GC content in windows across the genome.

We can search for a standard strain of *E. coli* by entering the search term “escherichia coli K12” to the home page of NCBI. From there we find accession NC\_000913.3 corresponding to *Escherichia coli* str. K-12 substr. MG1655. The corresponding NCBI Nucleotide entry (at [http://www.ncbi.nlm.nih.gov/nuccore/NC\\_000913.3](http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3)) includes the option Send > Complete Record > Destination: File > Format: FASTA > Create File. This file is 4.7 MB; save it (or copy it) to your R working directory.

Next open the R program. You can install `seqinr` using the “Packages” pull-down menu from the RGui console, or RStudio offers a convenient option to select packages for installation.

```
> library("seqinr")
> ?seqinr # Explore features of this package
> ecoli <- read.fasta(file = "NC_000913.fasta")
```

We have created the object `ecoli` which includes the sequence. The command `str(ecoli)` will show us its of length 4.6 million. Next we place the sequence in a vector called `ecoliseq`. We can see this with the `length` command, and can also obtain the GC content:

```
> ecoliseq <- ecoli[[1]] # This puts the sequence in a vector
> ecoliseq[1:10] # This displays the first 10 nucleotides
[1] "a" "g" "c" "t" "t" "t" "t" "c" "a" "t"
> length(ecoliseq)
[1] 4641652
> GC(ecoliseq)
[1] 0.5079071
```

This strain of *E. coli* therefore has a GC content of about 50.8%. We may also want to evaluate the GC content across windows of fixed size. We begin with a size of 20,000 base pairs.

The plot of GC content is shown in **Figure 17.10**. Note that these windows are non-overlapping (for some applications it may be useful to work with overlapping windows).

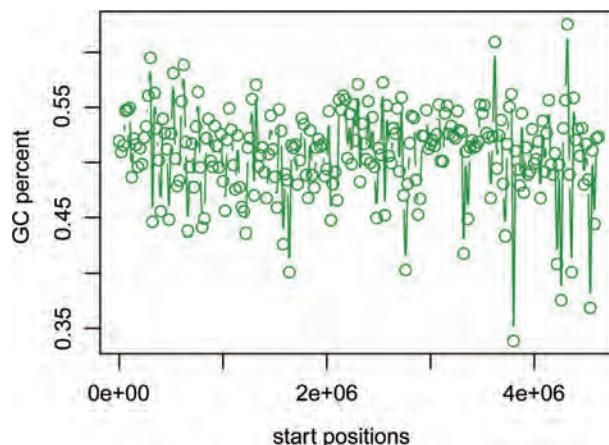
```
> starts <- seq(1, length(ecoliseq)-20000, by = 20000)
> n <- length(starts) # n is the length of the vector.
> n
[1] 232
> chunkGCs <- numeric(n)
# This creates a vector of the same length as starts.
> for (i in 1:n) {
  chunk <- ecoliseq[starts[i]:(starts[i]+1999)]
  chunkGC <- GC(chunk)
  print(chunkGC)
  chunkGCs[i] <- chunkGC
}
# This "for loop" iteratively determines the GC content in each window
> plot(starts,chunkGCs,type="b",xlab="start position",ylab="GC
percent",col=forestgreen) # the type "b" specifies a plot with the data
# points connected by lines. col specifies the color.
```

## Finding Genes

Bacteria and archaea are characterized by a high gene density (about one gene per kilobase), an absence of introns, and very little repetitive DNA. The problem of finding genes is therefore relatively simple in comparison to searching eukaryotic DNA (Chapter 8). Several programs are available for microbial gene identification, as listed in **Table 17.8**.

There are four main features of genomic DNA that are useful for gene recognition (Baytaluk *et al.*, 2002). These features apply to both bacterial and eukaryotic gene finding:

1. *Open reading frame (ORF) length.* An ORF is not necessarily a gene; for example, many short ORFs are not part of authentic genes (discussed further below). An ORF is defined by a start codon (i.e., ATG encoding a methionine) and a stop codon (TAA, TAG, TGA). In bacteria however, alternative start codons may be employed such as GTG or TTG, and there are rarely used alternative stop codons.



**FIGURE 17.10** GC content of *E. coli* strain K-12. The sequence of an *E. coli* strain was downloaded from NCBI, input to the R program `seqinr`, a `for` loop was used to calculate GC content in windows of 20,000 base pairs, and the data were plotted (see text for details).

**TABLE 17.8 Programs for gene finding in bacterial and archaeal genomes.**

Program	Description	URL
EasyGene	A web server from Anders Krogh and colleagues	<a href="http://www.cbs.dtu.dk/services/EasyGene/">http://www.cbs.dtu.dk/services/EasyGene/</a>
FrameD	Locates genes and frameshifts; optimized for GC-rich genomes	<a href="http://bioinfo.genopole-toulouse.prd.fr/apps/FrameD/FrameD.html">http://bioinfo.genopole-toulouse.prd.fr/apps/FrameD/FrameD.html</a>
GeneMarkP, GeneMarkS	Uses hidden Markov models	<a href="http://exon.gatech.edu/GeneMark/">http://exon.gatech.edu/GeneMark/</a>
GLIMMER	At Johns Hopkins University	<a href="http://ccb.jhu.edu/software.shtml">http://ccb.jhu.edu/software.shtml</a>

2. *Presence of a consensus sequence for ribosome binding in the immediate vicinity of the start codon.* In some cases, it is possible to identify two in-frame ATG codons, either of which could represent the start codon. Identifying a ribosome binding site can be an important indicator of which is the likely start site. In bacteria, the ribosome binding site is called a Shine–Dalgarno sequence. It is a purine-rich stretch of nucleotides that is complementary to the 3' end of 16S rRNA, extending from the –20 position (i.e., 5' to the initiation codon) to the +13 position (i.e., 13 nucleotides downstream in the 3' direction). Samuel Karlin and colleagues (Ma *et al.*, 2002) studied 30 prokaryotic genomes and correlated the features of the Shine–Dalgarno sequence with expression levels of genes based on codon usage bias (see below), type of codon, functional gene class, and type of start codon. They have shown a positive correlation between the presence of a strong Shine–Dalgarno sequence and high levels of gene expression.
3. *Presence of a pattern of codon usage that is consistent with genes.* Hidden Markov models (Chapter 6 and see below) have been particularly useful in defining the coding potential of putative protein-coding DNA sequences.
4. *Homology of the putative gene to other known genes.* Genomic DNA sequences, including putative genes, can be searched against protein databases using BLASTX (see Chapter 4). This approach is especially helpful in finding genes in eukaryotic organisms. For example, exons can be matched to expressed sequence tags (Chapter 8).

The first three of these features are studied using intrinsic approaches to gene finding. They are called intrinsic because the features do not necessarily depend on comparisons to gene sequences from other organisms. The fourth feature, relationship to other genes, is called an extrinsic approach. Bacterial gene-finding programs sometimes combine both intrinsic and extrinsic approaches.

Intrinsic approaches are also sometimes called *ab initio* approaches.

The GLIMMER system is one of the premier gene-finding algorithms, and identifies over 99% of all genes in a bacterial genome (Delcher *et al.*, 1999, 2007). The latest version has excellent sensitivity (determined based on comparisons to well-annotated bacterial genomes) and specificity (there are relatively few false positive results, i.e., gene predictions that do not correspond to authentic genes). The algorithm uses interpolated Markov models (IMMs). A Markov chain can describe the probability distribution for each nucleotide in a genomic DNA sequence. This probability can depend on the preceding  $k$  variables (nucleotides) in the sequence. A fixed-order Markov chain would describe the  $k$ -base context for each nucleotide position; for example, a fixed fifth-order Markov chain model describes  $4^5 = 1024$  probability distributions, one for each possible 5-mer. GLIMMER uses a fifth-order Markov chain because that corresponds to a model of two consecutive codons (six nucleotide positions). The  $k$ -mers are used as a training set to teach the algorithm the rules for which probability distributions are most likely to be relevant to this particular genomic sequence. Larger values for  $k$  are more informative but, since they occur more rarely, it is more difficult to sample enough data for a training set in order to model the probability of the next base in the sequence. IMMs are a specialization

of Markov models in which rare  $k$ -mers tend to be ignored and more common  $k$ -mers are weighted more heavily.

GLIMMER builds an IMM from a training set, then scans a genomic DNA sequence to predict genes. Criteria for gene-finding include the presence of an initiation codon and some particular minimal length for an open reading frame. GLIMMER further assigns functions to predicted genes through BLAST searches and HMM searches, and also searches for noncoding RNAs (e.g., using tRNAscan; Chapter 10), paralogs, and PROSITE motifs (Chapter 12).

A simplified form of GLIMMER is available online at the NCBI website. We can enter the accession number for *E. coli* str. K-12 substr. MG1655 (the complete genome of a well-known strain). We can then download the whole-genome sequence (4,641,652 base pairs) in the FASTA format and save it as a text file (Fig. 17.11a). Visiting the NCBI GLIMMER site, we can then upload the text file and perform analyses of open reading frames in this DNA (Fig. 17.11b).

GLIMMER was written by Owen White, Steven Salzberg and colleagues when at The Institute for Genomic Research. GLIMMER is an acronym for Gene Locator and Interpolated Markov Modeler.

(a) Obtaining the *E. coli* genome sequence in the FASTA format

**Escherichia coli str. K-12 substr. MG1655, complete genome**

NCBI Reference Sequence: NC\_000913.3

[FASTA](#) [Graphics](#)

Go to:

LOCUS	NC_000913	4641652 bp	DNA	circular	CON 03-N0
DEFINITION	Escherichia coli str. K-12 substr. MG1655, complete genome.				
ACCESSION	NC_000913				
VERSION	NC_000913.3	GI:556503834			
DBLINK	<a href="#">BioProject: PRJNA57779</a>				
KEYWORDS	RefSeq.				
SOURCE	Escherichia coli str. K-12 substr. MG1655				
ORGANISM	<a href="#">Escherichia coli str. K-12 substr. MG1655</a>				
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.					

(b) Output of the GLIMMER gene-finding program (web-based, NCBI)

GLIMMER (ver. 3.02; iterated) predictions:				
orfID	start	end	frame	score
<hr/>				
>gi 556503834 ref NC_000913.3	Escherichia coli str. K-12 substr. MG1655			
orf00001	337	2799	+1	12.01
orf00002	2801	3733	+2	11.36
orf00004	3734	5020	+2	14.41
orf00006	5537	5310	-3	2.45
orf00008	6459	5683	-1	12.91
orf00009	7911	6529	-1	7.93
orf00010	8238	9191	+3	13.60

**FIGURE 17.11** Identifying *E. coli* genes using the web-based GLIMMER3 program at NCBI. (a) Starting from the accession number of an *E. coli* strain (NC\_000913.3) the “send to” option is selected to download a text file with the nucleotide sequence in the FASTA format. (b) The first ten open reading frame predictions (of 4482 total) are shown.

Source: GLIMMER3, NCBI.

To find GLIMMER at NCBI, visit the Genome page at <http://www.ncbi.nlm.nih.gov/genome> (WebLink 17.20) then follow the link to microbes.

GLIMMER is available at <http://ccb.jhu.edu/software/glimmer/index.shtml> (WebLink 17.21).

GLIMMER was designed for command-line usage on Linux (or related) operating systems. We download GLIMMER from its website and transfer it to a folder called `glimmer`.

```
$ mkdir glimmer # this makes a new directory called glimmer
$ cd Downloads/ # change directory to the Downloads directory
$ cp glimmer302b.tar ~/glimmer/ # copy the downloaded glimmer program to
# the glimmer directory
$ cp sequence-6.fasta ~/glimmer/ # transfer (copy) the fasta file to the
# glimmer directory
$ mv sequence-6.fasta ecoliK12MG1655.fasta # we rename the downloaded
# sequence
$ tar xzf glimmer302b.tar.gz # uncompress the distribution file
```

A new directory called `glimmer3.02` is now created. Next, compile the program.

```
$ cd src/
$ make # the program is compiled. (If make command fails, see the
# documentation for help.)
```

We now run GLIMMER in two steps, as described in the following sections.

#### *Interpolated Context Model (ICM)*

First, we build an interpolated context model (ICM) which is a probability model of coding sequences. We select a set of *E. coli* nucleotide sequences in the FASTA format to train the model. We are doing this to demonstrate how GLIMMER3 works. In general, if you have a new genome for which you want to annotate genes, there are several options: you can identify a set of known genes from BLAST searches; you can select known genes from a closely related species; or, you can use the program `long-orfs` within GLIMMER3 to identify long open reading frames that represent candidate genes. To see its help documentation, type:

```
$ ./glimmer3.02/bin/long-orfs -h # Once you copy the executable to your
# home/bin directory you can invoke glimmer without needing the ./ prefix
# that specifies the location of the executable
```

From the NCBI Nucleotide page for *Escherichia coli* str. K-12 substr. MG1655 (<http://www.ncbi.nlm.nih.gov/nuccore/556503834?report=fasta>, NC\_000913.3), choose “Send” to send the coding sequences to a file as we described above. This file can be viewed in a text editor, renamed `Ecoli.fna.train` (where `fna` indicates a set of FASTA nucleotide sequences), and moved to our current directory:

```
$ cp ~/Downloads/Ecoli.fna.train . # the . symbol indicates that the file
# should be moved to the current directory which is ~/glimmer
```

We create a second file (`Ecoli2.fna.train`) that has about half the number of entries. To see how many are in each file we can use `grep`, a utility that searches our plain text document for a regular expression such as the `>` symbol that appears at the start of each nucleotide entry.

```
$ grep ">" Ecoli.fna.train | wc -l
4141
```

The regular expression we wish to grep is “`>`”. The pipe symbol `|` sends the output directly to the word count (`wc`) utility, and the `-l` modifier specifies that we want to know

how many lines are in the file (without that modifier we would see the numbers of lines, words, and characters).

```
$ grep ">" Ecoli2.fna.train | wc -l
2051
```

Let's look at the first ten lines of one of these files with the `head` command:

```
$ head Ecoli.fna.train
>lcl|NC_000913.3_cdsid_NP_414542.1 [gene=thrL] [protein=thr operon leader
peptide] [protein_id=NP_414542.1] [location=190..255]
ATGAAACGCATTAGCACCACCATTAACCAACCATCACCAATTACCAACAGGTAAACGGTGCGGGCTGA
>lcl|NC_000913.3_cdsid_NP_414543.1 [gene=thrA] [protein=fused
aspartokinase I and homoserine dehydrogenase I] [protein_id=NP_414543.1]
[location=337..2799]
ATGCGAGTGTGAAGTTCGGGGTACATCAGTGGCAAATGCAGAACGTTCTGCGTGTGCCGATATT
TGGAAAGCAATGCCAGGCAGGGCAGGTGCCACCGTCCTCTGCCCGCAAAATACCAACCCACT
GGTGGCGATGATTGAAAAAAACATTAGCGGCCAGGATGCTTACCCAATATCAGCGATGCCAACGTATT
TTTGCCTGAACTTTGACGGGACTCGCGCCGCCAGCCGGGTTCCCGCTGGCGCAATTGAAAACTTCG
TCGATCAGGAATTGCCAATAAAACATGCTCTGCATGGCATTAGTTGTTGGGCAGTGCCCCGATAG
CATCAACGCTGCCTGATTGCCGTGGCGAGAAAATGTCGATGCCATTATGCCGGCGTATTAGAACCG
CGCGGTACAACGTTACTGTTATCGATCCGGCGAAAACGTGCTGGCAGTGGGCATTACCTCGAATCTA
```

Next we can build our interpolated context model (ICM).

```
$ glimmer3.02/bin/build-icm --text my_icm.txt < Ecoli2.fna.train
```

This program takes just a few seconds to run. We used the `--text` option to produce a text-based version that we can look at (we redo the command without the `--text` option to create an ICM that GLIMMER3 can use). Let's use `head` and then `tail` to look at the top and bottom lines of this file.

```
$ head my_icm.txt
ver = 2.00 len = 12 depth = 7 periodicity = 3 nodes = 21845
0      ---|---|---|-*?  0.0519      0.183      0.265      0.288      0.263
1      ---|---|---|*a?  0.0944      0.315      0.212      0.195      0.278
2      ---|---|---|*c?  0.0350      0.193      0.280      0.353      0.174
3      ---|---|---|*g?  0.0828      0.081      0.403      0.185      0.331
4      ---|---|---|*t?  0.0803      0.111      0.222      0.390      0.277
5      ---|---|---|*aa?  0.0093      0.407      0.258      0.118      0.217
6      ---|---|---|*ca?  0.0297      0.235      0.139      0.430      0.196
7      ---|---|---|*ga?  0.0115      0.366      0.173      0.162      0.299
8      ---|---|---|*ta?  0.0103      0.067      0.385      0.007      0.541
$ tail my_icm.txt
21835 -|---*|cg|t|tt|t?  0.1115      0.259      0.308      0.132      0.301
21836 -|---*|ct|t|tt|t?  0.1780      0.247      0.327      0.126      0.300
21837 a|---*|g-t|tt|t?  0.1728      0.301      0.281      0.107      0.312
21838 c|---*|g-t|tt|t?  0.1276      0.303      0.297      0.093      0.307
21839 g|---*|g-t|tt|t?  0.0833      0.289      0.293      0.108      0.310
21840 t|---*|g-t|tt|t?  0.1093      0.273      0.288      0.114      0.325
21841 *|---|tat|tt|t?  0.1656      0.254      0.302      0.152      0.291
21842 -|---*|tct|tt|t?  0.3216      0.251      0.300      0.152      0.296
21843 -|---*|tgt|tt|t?  0.6490      0.256      0.298      0.153      0.293
21844 *|---|ttt|tt|t?  0.2363      0.260      0.304      0.161      0.275
```

There are seven columns. The first is an ID number. Second is a contextual pattern, starting with a single base and eventually including six bases in various patterns. Vertical lines demarcate the codons. The `?` symbol corresponds to the nucleotide that is being predicted, and the asterisk shows the position that has maximum mutual information with the predicted position. The third column displays mutual information, and columns 4–7 display the probabilities of A, C, G, and T.

**GLIMMER3**

Now we can run GLIMMER3.

```
$ glimmer3.02/bin/glimmer3 ecoliK12MG1655.fasta my_icm myoutput
```

`myoutput` is an example of a tag that you choose for naming the output files. This run is completed in several seconds. There are many options (e.g., linear versus circular genome, specifying ribosome binding sites, start and stop codons, minimum gene length, maximum overlap, and GC content). These options highlight the usefulness of command-line software relative to the simple web-based version of GLIMMER at NCBI (that offers no options). There are two output files. The first is `myoutput.detail`.

```
$ less myoutput.detail
Command: glimmer3.02/bin/glimmer3 ecoliK12MG1655.fasta my_icm myoutput
Sequence file = ecoliK12MG1655.fasta
Number of sequences = 1
ICM model file = my_icm
Excluded regions file = none
List of orfs file = none
Input is NOT separate orfs
Independent (noncoding) scores are used
Circular genome = true
Truncated orfs = false
Minimum gene length = 100 bp
Maximum overlap bases = 30
Threshold score = 30
Use first start codon = false
Start codons = atg,gtg,ttg
Start probs = 0.600,0.300,0.100
Stop codons = taa,tag,tga
GC percentage = 50.8%
Ignore score on orfs longer than 750
>gi|556503834|ref|NC_000913.3| Escherichia coli str.
K-12 substr. MG1655, complete genome
Sequence length = 4641652
-- Start --- - Length ----- Scores -----
ID Frame of Orf of Gene Stop of Orf of Gene Raw InFrm F1 F2 F3 R1 R2 R3 NC
+2 4641564 4641606 76 162 120 -7.17 0 - 0 0 - - - 99
-2 463 334 230 231 102 -4.25 1 - - - 0 1 0 98
+2 350 374 487 135 111 -3.57 2 0 2 0 0 - 0 97
-1 516 474 364 150 108 -16.90 0 0 1 0 0 - 0 98
-3 620 236 108 510 126 -8.51 0 - - - - 0 99
-1 747 654 517 228 135 -11.06 0 0 - 0 0 - - 99
-3 761 734 621 138 111 -11.40 0 0 - - 0 - 0 99
```

It shows the command used to run GLIMMER3; the list of parameters used by the program; and the FASTA header of the input file. It then shows a table with the following columns: ID (identifier for genes); Frame (+ for forward strand and – for reverse strand); start and stop positions of the ORF and the gene; the length of both the ORF and the gene (not including the bases in the stop codon); and then a series of scores for the six possible frames as well as NC (a normalized independent model score).

The second output file is `myoutput.predict`. Here are the first few lines:

```
$ less myoutput.predict
>gi|556503834|ref|NC_000913.3| Escherichia coli str. K-12 substr. MG1655,
complete genome
orf00001    337      2799      +1      2.98
orf00002    2801     3733      +2      2.95
orf00004    3734     5020      +2      2.96
orf00005    6459     5683      -1      2.93
```

orf00006	7959	6529	-1	2.96
orf00007	8175	9191	+3	2.88
orf00010	12163	14079	+1	2.97
orf00012	14138	15298	+2	2.90
orf00013	15445	16557	+1	2.95
orf00014	17489	18655	+2	2.90

This file includes the final gene predictions. The columns are (1) the identifier (matching those of the `.detail` file); (2) the start and (3) the end position of the gene; (4) the reading frame; and (5) the per-base raw score of the gene.

## Challenges of Bacterial and Archaeal Gene Prediction

There are several pitfalls associated with bacterial and archaeal gene prediction:

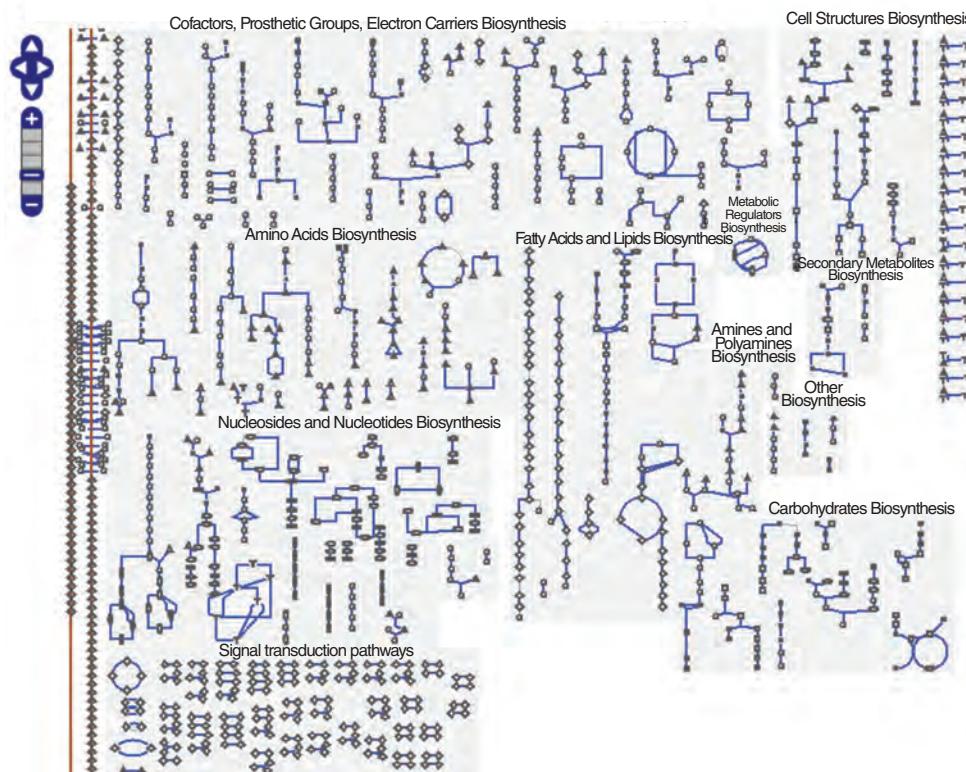
- There may be multiple genes that are encoded by one genomic DNA segment in an alternate reading frame on the same strand or opposite strand. GLIMMER includes features to address this situation.
- It is difficult to assess whether a short ORF is genuinely transcribed. According to Skovgaard *et al.* (2001), there are far too many short genes annotated in many genomes. For *E. coli*, they suggest that there are 3800 true protein-coding genes rather than the 4300 genes that have been annotated. Since stop codons (TAA, TAG, TGA) are AT rich, genomes that are GC rich tend to have fewer stop codons and more predicted long ORFs. For all predicted proteins in a genome, the proportion of hypothetical proteins (defined as predicted proteins for which there is no experimental evidence that they are expressed) rises greatly as sequence length is smaller.
- Frameshifts can occur, in which the genomic DNA is predicted to encode a gene with a stop codon in one frame but a continuing sequence in another frame on the same strand. A frameshift could be present because of a sequencing error or because of a mutation that leads to the formation of a pseudogene (a nonfunctional gene). GLIMMER extends gene prediction loci several hundred base pairs upstream and downstream to search for homology to known proteins, and is therefore designed to detect possible frameshifts.
- Some genes are part of operons that often have related functional roles in bacteria (or archaea). Operons have promoter and terminator sequence motifs, but these are not well characterized. Steven Salzberg and colleagues (Ermolaeva *et al.*, 2001) analyzed 7600 pairs of genes in 34 bacterial and archaeal genomes that are likely to belong to the same operon.
- Lateral gene transfer, also called horizontal gene transfer, commonly occurs in bacteria and archaea. We discuss this in the relevant section below.

An operon is a cluster of contiguous genes, transcribed from one promoter, that gives rise to a polycistronic mRNA.

## Gene Annotation

Gene annotation is used to assign functions to genes and, in some cases, to reconstruct metabolic pathways or other higher levels of gene function. Gene annotation pipelines seek to maximize accuracy, consistency, and completeness. An example of the functional groups assigned to *E. coli* genes by the EcoCyc database is shown in **Figure 17.12**.

The Rapid Annotations using Subsystems (RAST) server offers automated annotation of bacterial and archaeal genomes (Aziz *et al.*, 2008, 2012; Overbeek *et al.*, 2014). RAST annotation includes the following 16 steps. The input is a set of contigs in the FASTA format. (1) RAST identifies selenoproteins and other specialized proteins. (2) RAST estimates 30 closest phylogenetic neighbors using GLIMMER3. (3) It calls tRNA genes (using tRNAscan-SE; see Chapter 10) and rRNA genes (using BLASTN against a rRNA database). (4–7) Protein candidates are further evaluated, including iterative retraining of

Cellular overview of *Escherichia coli* K-12 substr. MG1655 (EcoCyc)

**FIGURE 17.12** The EcoCyc database includes a cellular overview of *E. coli* K-12 MG1655. This site organizes *E. coli* proteins according to function. Data may be explored based on biochemical pathways, reactions, genes, enzymes, or compounds.

Source: Adapted from SRI International (<http://ecocyc.org/>).

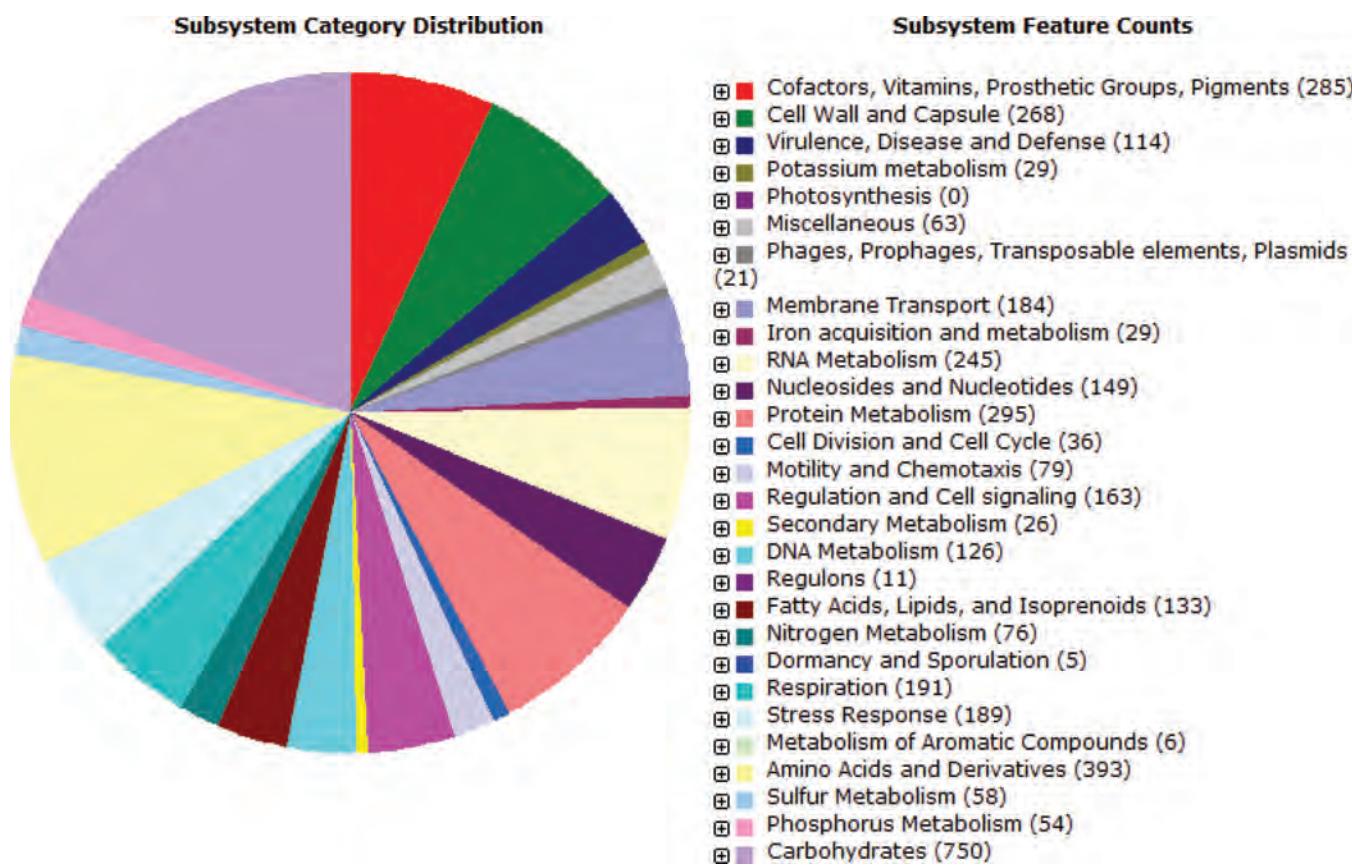
The RAST annotation server is at <http://rast.nmpdr.org> (WebLink 17.22). By early 2014 >12,000 users have annotated >60,000 genomes using RAST (Overbeek *et al.*, 2014). The SEED project, which is the underlying annotation database, is available at <http://pubseed.theseed.org> (WebLink 17.23).

GLIMMER3. (8) Gene candidates with frameshifts are evaluated and processed. (9) Loci with >1500 base pairs but lacking an annotated gene are assessed using BLASTX against the 30 nearest neighbors. (10–13) Assessments of gene function are made using BLASTP and other approaches. (14) Metabolic reconstructions are performed. (15) Comparisons to other annotated projects are made. (16) Genome annotations are exported in GenBank, GFF3, GTF, and other formats.

We can save the *E. coli* genome in the FASTA format as described above, then upload it to the RAST server. The output includes annotation spreadsheets as well as a metabolic model (Fig. 17.13).

Automated annotation pipelines are subject to many types of artifacts. Richardson and Watson (2013) provide the following examples.

- Pipelines rely heavily on homology to closely related species. However, often the genome of a new strain is sequenced because of its genetic or functional differences from its closest reference genomes. Such differences may correspond to loci that are absent from the reference and therefore not annotated in the new strain. While the leading annotation pipelines are automated, manual intervention is still necessary and direct comparisons of pipelines can lead to differing results (Kisand and Lettieri, 2013).
- Inconsistent annotation occurs such as genes and domains that are either split or fused. Richardson and Watson cite a block of genes in *E. coli* K-12 MG1655 and *E. coli* O157:H7 Sakai that share 97% nucleotide identity, yet have different gene names (e.g., *tbpA* versus *thiB*) or gene names corresponding to locus tags. Each annotated genome may have different types and amounts of information. It is therefore



**FIGURE 17.13** Automated annotation of bacterial and archaeal genomes is performed by services such as the RAST server. Raw nucleotide sequence is input, and the output includes functional annotation (as shown here) as well as tabular descriptions of functional assignments. Source: SEED/RAST. The Fellowship for the Interpretation of Genomes and Argonne National Laboratory.

important to select an optimal reference genome or even multiple reference genomes for annotation.

- Spelling mistakes occur, such as “syntase” instead of “synthase.” Such errors often propagate within and across databases. To see examples of this, search NCBI Nucleotide with the terms “syntase” or “psuedogene” (instead of “pseudogene”).
- The same gene name may be assigned different product names. The *int* gene has 12 different protein product names (e.g., integrase, putative phage integrase protein) in 17 *Salmonella* RefSeq entries.
- There are tens of thousands of “hypothetical proteins” that are predicted but lack homologs of known function. The naming and description of these proteins is variable between annotators. Some are artifacts that are propagated in databases.

## Lateral Gene Transfer

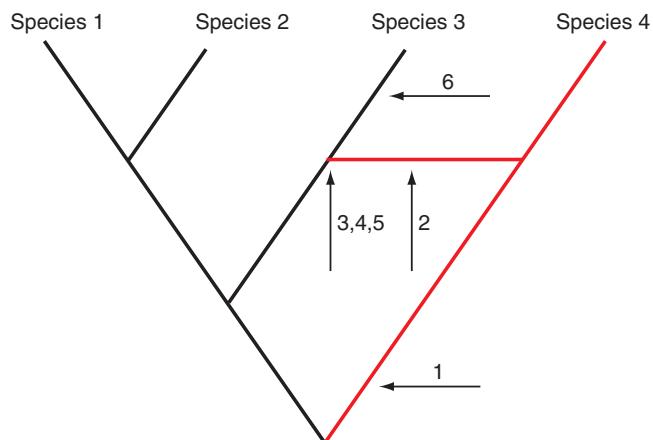
Lateral, or horizontal, gene transfer (LGT) is the phenomenon in which a genome acquires a gene from another organism directly, rather than by descent (Eisen, 2000; Koonin *et al.*, 2001; Boucher *et al.*, 2003). There are many situations in which examination of a genome shows that a particular gene is very closely related to orthologs in distantly related organisms. The simplest explanation for how a species acquired such a gene is through lateral gene transfer. This mechanism represents a major force in genome evolution. The gene transfer is unidirectional, rather than involving a reciprocal exchange of DNA, and it does not involve the usual pattern of inheritance from a parental lineage. Over 50% of archaeal

and a smaller percentage of bacterial species have one or more protein domains acquired by lateral gene transfer, in contrast to <10% of eukaryotic species (Choi and Kim, 2007; Andersson, 2009).

Lateral gene transfer is a significant phenomenon for several reasons:

1. This mechanism vastly differs from the normal mode of inheritance in which genes are transmitted from parent to offspring. Lateral gene transfer therefore represents a major shift in our conception of evolution.
2. This mechanism is very common, and many examples have also been described in eukaryotes. It has been observed within and between each of the three main branches of life, but is particularly prevalent in bacteria and archaea relative to eukaryotes (Choi and Kim, 2007).
3. Lateral gene transfer can greatly confound phylogenetic studies (Dagan, 2011). If a DNA, RNA, or protein is selected for phylogenetic analysis that has undergone lateral gene transfer, then the tree will not accurately represent the natural history of the species under consideration. An extreme interpretation of lateral gene transfer is that, if it is common enough, then it is impossible in principle to derive a single true tree of life. Daubin *et al.* (2003) and Choi and Kim (2007) have suggested that although lateral gene transfer is common it is not so prevalent that it greatly interferes with phylogenetic studies of organisms. Lateral gene transfer can be useful in phylogenetic studies to infer monophyletic groups and to elucidate the evolutionary history of both donor and recipient species (Huang and Gogarten, 2006). Dagan (2011) suggests that network-based rather than tree-based models of phylogeny are needed to reconstruct both vertical and lateral evolution.
4. Lateral gene transfer can profoundly affect the properties of basic biological processes, as reviewed extensively by Boucher *et al.* (2003). They describe its importance in a variety of processes such as photosynthesis, aerobic respiration, nitrogen fixation, sulfate reduction, and isoprenoid biosynthesis.

Lateral gene transfer occurs as a multistep process (Fig. 17.14; Eisen, 2000). A gene that evolves in one lineage (by the traditional Darwinian process of vertical descent) may



**FIGURE 17.14** Lateral gene transfer occurs in stages. In this hypothetical scenario, four species evolved from a common ancestor. Genes in each species descend in a vertical fashion over time (arrow 1). At some point in time, a gene transfers horizontally from the lineage of species 4 to the lineage of species 3 (arrow 2). Transferred genes must then be fixed in some individual genomes (arrow 3), maintained under strong positive selection (arrow 4), and spread through the population of species 3 (arrow 5). The laterally transferred gene then evolves as an integral part of the new genome (arrow 6). This gene may be distinguished from other genes in species 3 by having a nucleotide composition or codon usage profile that is characteristic of species 4. Adapted from Eisen (2000), with permission from Elsevier.

transfer to the lineage of a second species. This DNA transfer could be mediated by a viral vector or by a mechanism such as homologous recombination. Mobile genetic elements with transfer and recombination activities have key mechanistic roles (Toussaint and Chandler, 2012). Once a new gene is incorporated into the genome of individuals with a population (e.g., species 3 in Fig. 17.14), positive selection maintains its presence within those individuals. A transferred gene presumably must confer benefits to the new species in order to be maintained, propagated, and spread throughout the population of the new species. Finally, the new gene adapts to its new lineage, a process called “amelioration” (Eisen, 2000; Fig. 17.14, arrow 6).

How is lateral gene transfer identified? The main criteria are that a gene has an unusual nucleotide composition, codon usage, phylogenetic position, or intron features that distinguish it from most other genes in a genome. There are three principal methods by which lateral gene transfer may be inferred:

1. Phylogenetic trees of different genes may be compared. This is the favored approach (Eisen, 2000; Beiko and Ragan, 2008). If a tree based on a gene (or protein) has a topology different from that observed using ribosomal RNA, this discrepancy could be caused by lateral gene transfer.
2. Patterns of best matches for each gene in a genome may be used. A gene may have a highly unusual nucleotide composition or frequency of codon utilization, consistent with its origin in a distantly related genome.
3. The distribution pattern of genes across species can be assessed to search for genes that have undergone lateral gene transfer. If a gene is present in crenarcheota and a group of plants but not in other archaea, bacteria, or eukaryotes, this may be taken as evidence favoring a lateral gene transfer mechanism from crenarcheota to plants.

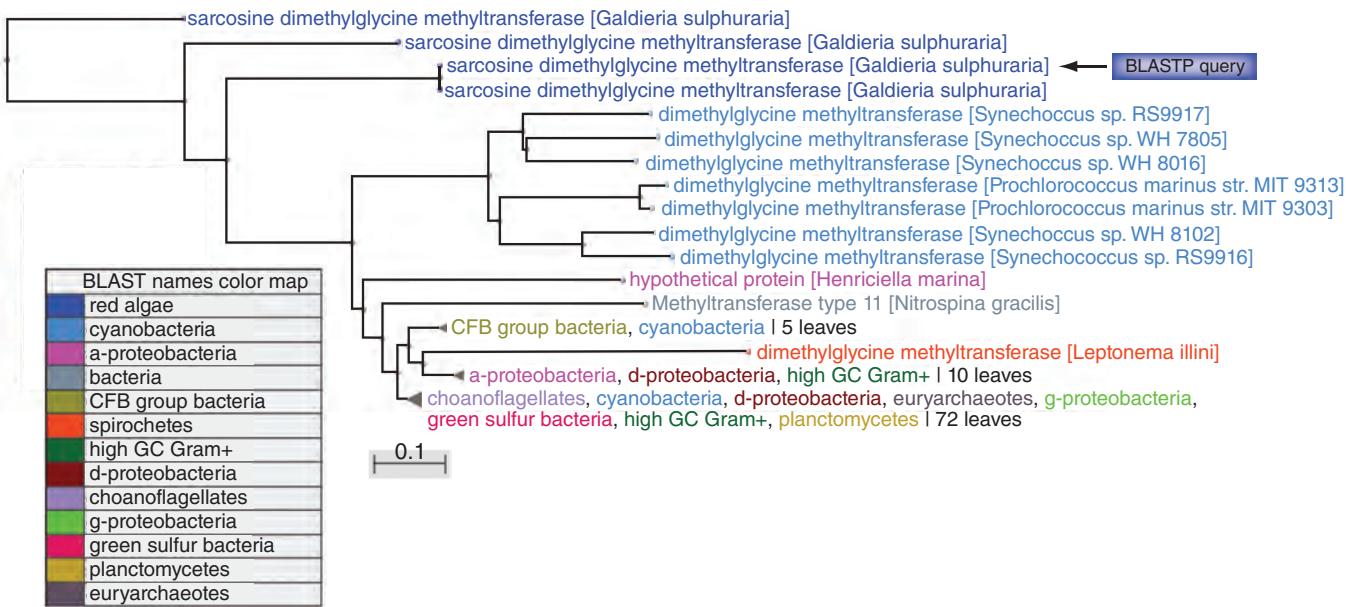
There are several reasons for caution in assigning a mechanism of lateral gene transfer. Consider the case of a gene widely distributed in bacteria that is observed in humans.

- If orthologs of the bacterial gene were present in an insect such as *Drosophila* or a plant, then the argument in favor of lateral gene transfer to humans would be considerably weakened. A concern in positing lateral gene transfer has been that the candidate gene might be present throughout the tree of life, but we might have insufficient sequence data to find it in other species; the recent flood of sequence data makes the possible lack of data less likely. Over time it will be progressively easier to assess evolutionary relationships.
- It is also possible that the gene in question has undergone rapid mutation, such that the phylogenetic signal is lost. This mechanism may lead to artifactual results (false positives) if gene loss or rapid mutation has occurred, but not lateral gene transfer.

The eukaryotic alga *Galdieria sulphuraria* provides a dramatic example of lateral gene transfer. This unicellular red alga lives in an extreme environment that is hot and acid (56°C, pH 0–4) such as volcanic hot sulfur springs. Schönknecht *et al.* (2013) sequenced its 13.7 Mb genome and identified 75 separate LGT events from bacteria and archaea. These transferred genes had fewer introns (an average of 0.8 versus 2.1), higher GC content (40.6% versus 39.9%), and differing dinucleotide usage. The laterally transferred genes include those that confer tolerance to high salt, heat, and otherwise toxic metals. As an example, the alga acquired a sarcosine dimethylglycine methyltransferase gene from halophilic cyanobacteria (i.e., those living in high salt concentrations). This is evident from a BLASTP search using a *G. sulphuraria* protein as a query (Fig. 17.15); there are close matches to many bacterial (and archaeal) species, but not to other eukaryotes.

Carl Woese (2002) suggested that in early evolution lateral gene transfer predominated to such an extent that primitive cellular evolution was a communal process, followed only later by vertical (Darwinian) evolution.

See the computer laboratory exercise (17.4) at the end of this chapter for another example of lateral gene transfer.



**FIGURE 17.15** Lateral gene transfer of a gene encoding a sarcosine dimethylglycine methyltransferase from cyanobacteria to the eukaryote *G. sulphuraria*. The *G. sulphuraria* protein (Gasu\_07590; XP\_005708533.1) was used as a query in a BLASTP search against the RefSeq database at the NCBI website. The resulting hits are viewed as a neighbor-joining tree using Kimura protein distances (redrawn from the BLASTP output). The scale bar represents 0.1 changes per site. Schöönknecht *et al.* (2013) reported a similar phylogenetic tree, including additional more distantly related orthologs from eukaryotes. This particular gene encodes an enzyme that is part of the S-adenosylmethionine-dependent methyltransferase (SAM) family.

Source: BLASTP, NCBI.

## COMPARISON OF BACTERIAL GENOMES

One of the most important lessons of whole-genome sequencing is that comparative analyses greatly enhance our understanding of genomes. It can be useful to compare genomes whether they are from closely or distantly related organisms. Some of the species that have had the genomes of closely related strains completely sequenced are indicated in **Table 17.9**. It will be significant to compare such genomes for several reasons:

- We may be able to discover why some strains are pathogenic.
- Eventually, we may be able to predict the clinical outcome of infections based on the genotype of the pathogen.
- We may develop strategies for vaccine development.

For an example of comparisons of bacterial genomes (and proteomes) we can consider *Chlamydiae*, obligate intracellular bacteria that are phylogenetically distinct from other bacterial divisions. *Chlamydia pneumoniae* infects humans, causing pneumonia and bronchitis. *Chlamydia trachomatis* causes trachoma (an ocular disease that leads to blindness) and sexually transmitted diseases. Why do these closely related bacteria affect different body regions and cause such distinct pathologies? Their genomes have been sequenced and compared (Stephens *et al.*, 1998; Kalman *et al.*, 1999; Read *et al.*, 2000). There are hundreds of genes present uniquely in each bacterium, including a family of outer membrane proteins that could be important in tissue tropism (Kalman *et al.*, 1999).

In the United States, 10% of all pneumonia cases and 5% of bronchitis cases are attributed to *C. pneumoniae*.

### TaxPlot

The NCBI offers a powerful tool for genome comparison that is easy to use. From the NCBI Genome page, select TaxPlot and you will be able to compare two proteomes (such

**TABLE 17.9** Bacterial and archaeal species for which genomes of at least two closely related strains have been determined.

Organism	Accession	Genome size (base pairs)
<i>Chlamydophila pneumoniae</i> AR39	NC_002179	1,229,858
<i>C. pneumoniae</i> CWL029	NC_000922	1,230,230
<i>C. pneumoniae</i> J138	NC_002491	1,226,565
<i>Escherichia coli</i> K-12	NC_000913	4,639,221
<i>E. coli</i> O157:H7	NC_002695	5,498,450
<i>E. coli</i> O157:H7 EDL933	NC_002655	5,528,445
<i>Helicobacter pylori</i> 26695	NC_000915	1,667,867
<i>H. pylori</i> J99	NC_000921	1,643,831
<i>Mycobacterium tuberculosis</i> CDC1551	NC_002755	4,403,836
<i>M. tuberculosis</i> H37Rv	NC_000962	4,411,529
<i>Neisseria meningitidis</i> MC58	NC_003112	2,272,351
<i>N. meningitidis</i> Z2491	NC_003116	2,184,406
<i>Staphylococcus aureus</i> aureus MW2	NC_003923	2,820,462
<i>S. aureus</i> aureus Mu50	NC_002758	2,878,040
<i>S. aureus</i> aureus N315	NC_002745	2,813,641
<i>Streptococcus agalactiae</i> 2603V/R	NC_004116	2,160,267
<i>S. agalactiae</i> NEM316	NC_004368	2,211,485
<i>S. pneumoniae</i> R6	NC_003098	2,038,615
<i>S. pneumoniae</i> TIGR4	NC_003028	2,160,837
<i>S. pyogenes</i> M1 GAS	NC_002737	1,852,441
<i>S. pyogenes</i> MGAS315	NC_004070	1,900,521
<i>S. pyogenes</i> MGAS8232	NC_003485	1,895,017

as *C. trachomatis* A/HAR-13 and *C. pneumoniae* AR39) against a reference proteome (the anthrax bacterium *B. anthracis* in the example of Fig. 17.16). In this plot, each point represents a protein from the reference genome. The *x* and *y* coordinates show the BLAST score for the closest match of each protein to the two *Chlamydia* proteomes being compared. Most proteins are found along a diagonal line, indicating that they have equal (or nearly equal) scores between the reference protein and either of the *Chlamydia* proteins. However, there are notable outliers which could represent genes important in the distinctive behavior of these two organisms. These points are clickable (see circled data point in Fig. 17.16, arrow 2), and the selected data point is highlighted (Fig. 17.16, arrow 3). This protein is identified as an arginine/ornithine antiporter in *B. anthracis* and *C. trachomatis*, and as an amino acid permease in *C. pneumoniae*. There are further links to the pairwise BLAST comparisons (not shown).

Another powerful application of TaxPlot is to select a genome for both reference and for one of the queries, then select a second genome for the second query. This is illustrated in Figure 17.17 for a *C. trachomatis* strain versus *C. pneumoniae*. All the data points fall on the diagonal (indicating that they share identity between the two species) or in the upper left section. No data points are in the lower right section because no *C. trachomatis* query protein can possibly be more related to *C. pneumoniae* than to its own protein sequence. The outliers, such as those indicated with arrows 1–4, are of particular interest because they are highly divergent between the two species, having high BLASTP scores in one but low scores in the other. All four of the arrows point to polymorphic outer membrane proteins. Several additional outlying data points correspond to proteins that are annotated as

There are several ways to access TaxPlot, including from the Tools link on the left sidebar of the main NCBI home page as well as from <http://www.ncbi.nlm.nih.gov/Genome> (WebLink 17.24).

### Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

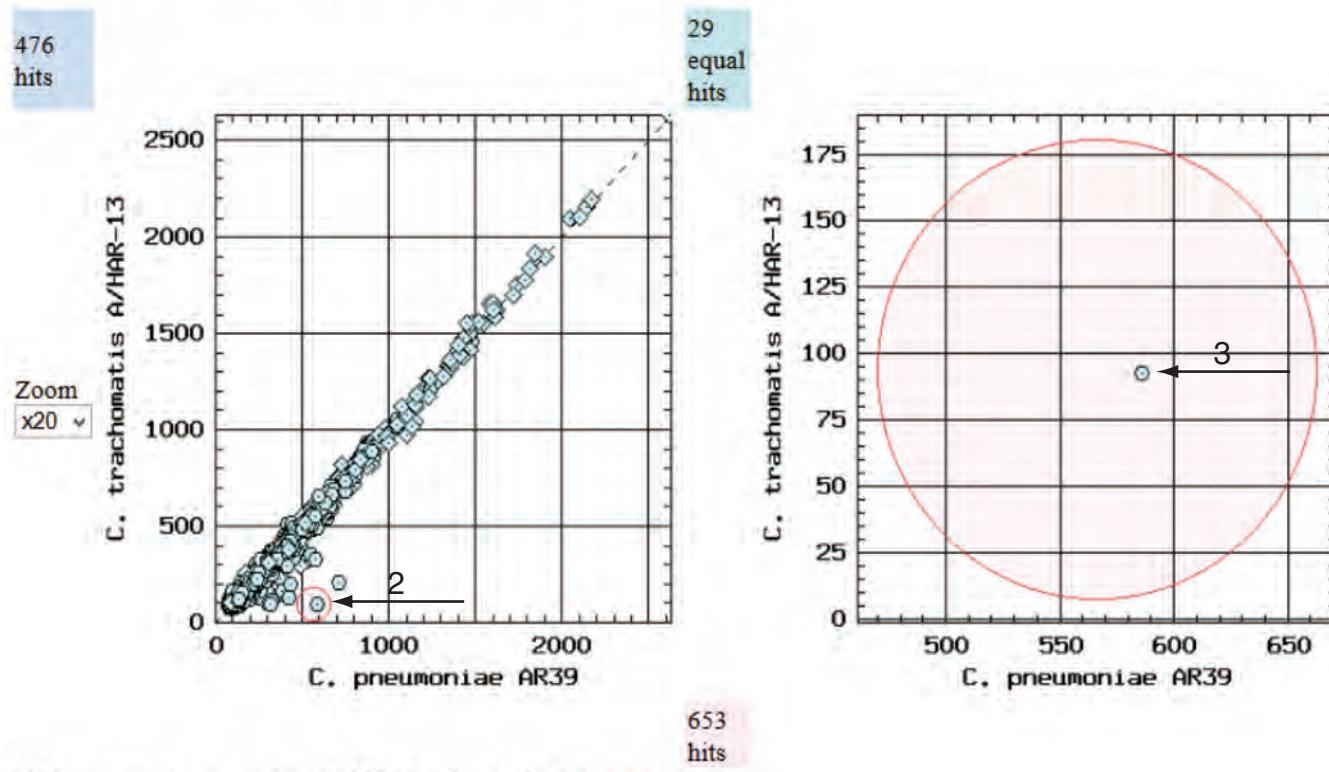
198094 B. anthracis str. Ames

Choose two species for comparison

315277 C. trachomatis A/HAR-13

115711 C. pneumoniae AR39

### Distribution of *B. anthracis* str. Ames homologs



5328 query proteins produced 1158 hits, from which 1 is selected.

**FIGURE 17.16** The TaxPlot tool at NCBI allows the comparison of two bacteria (*C. trachomatis* A/HAR-13 and *C. pneumoniae* AR39) to a reference genome (*B. anthracis* strain Ames in this case; Read *et al.*, 2002). The plot shows the distribution of BLASTP scores of each bacterial proteome against the reference proteome. Twenty-nine matches are identical, while 476 hits are at least marginally closer to *C. trachomatis* and 653 hits are closer to *C. pneumoniae*. The query genome (arrow 1) and the two species for comparison are selected using a pull-down menu. Most data points are placed along the unit diagonal line, indicating that the BLASTP score relative to the query (anthrax) proteome yields equivalent scores. A match of interest that has a higher pairwise BLASTP score in one proteome relative to the other comparison group can be clicked (arrow 2) leading to a zoom feature (arrow 3). The highlighted protein is identified in all three species (not shown) and there are links to the pairwise alignments from BLAST (Chapter 3). The significance of identifying outlier data points (such as that indicated by arrow 2) is that this protein has diverged greatly in one of the comparison species relative to the other, suggesting the possibility of functional differences.

Source: TaxPlot, Entrez, NCBI.

### Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

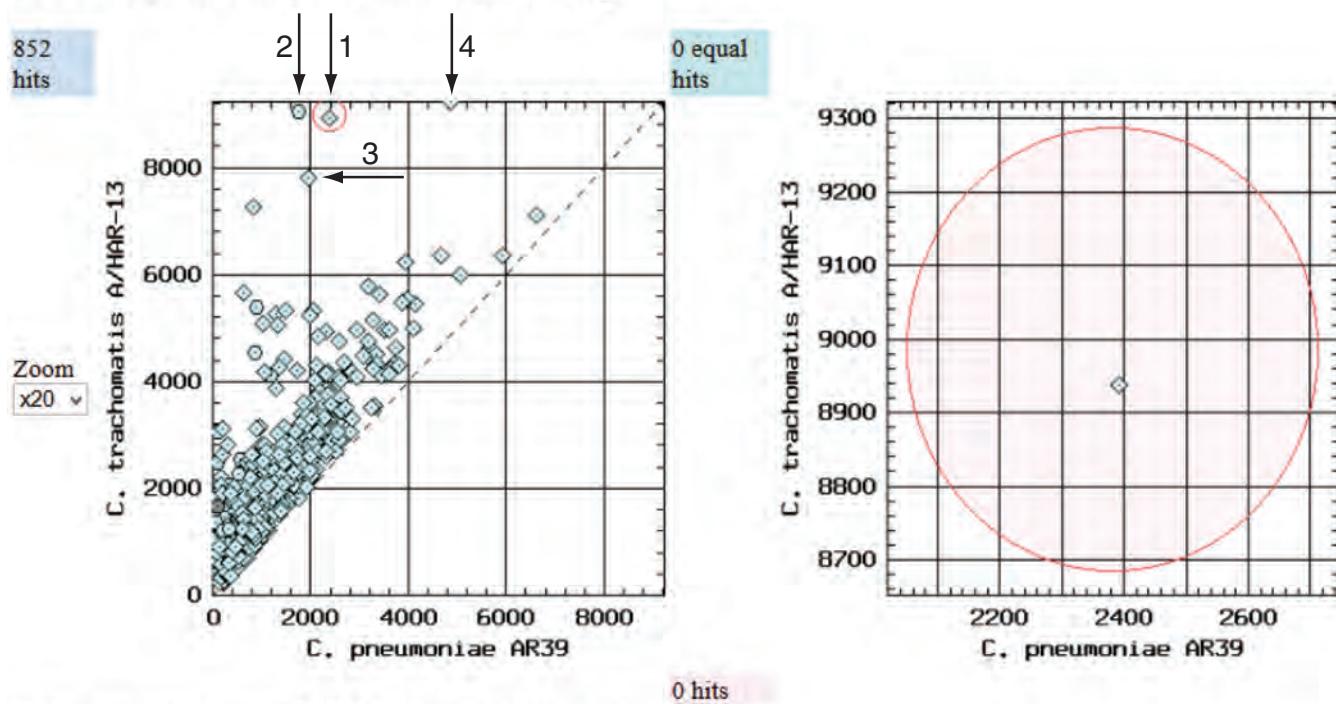
315277 C. trachomatis A/HAR-13

Choose two species for comparison

315277 C. trachomatis A/HAR-13

115711 C. pneumoniae AR39

#### Distribution of *C. trachomatis* A/HAR-13 homologs



919 query proteins produced 852 hits, from which 1 is selected.

**FIGURE 17.17** TaxPlot (NCBI) can be used with one proteome serving as both the reference and the first query (in this case, *C. trachomatis* A/HAR-13) while another proteome forms the second query (in this case, *C. pneumoniae* AR39). Points that fall off the diagonal line (e.g., see arrows 1–4) have a high BLASTP score in one proteome but a relatively low score in the other, indicating that they are relatively poorly conserved. Such proteins may be of great interest in explaining the particular physiology or behavior of a strain or species.

Source: TaxPlot, Entrez, NCBI.

hypothetical (therefore function has not been assigned). These are potentially important in distinguishing the functional differences between these two species.

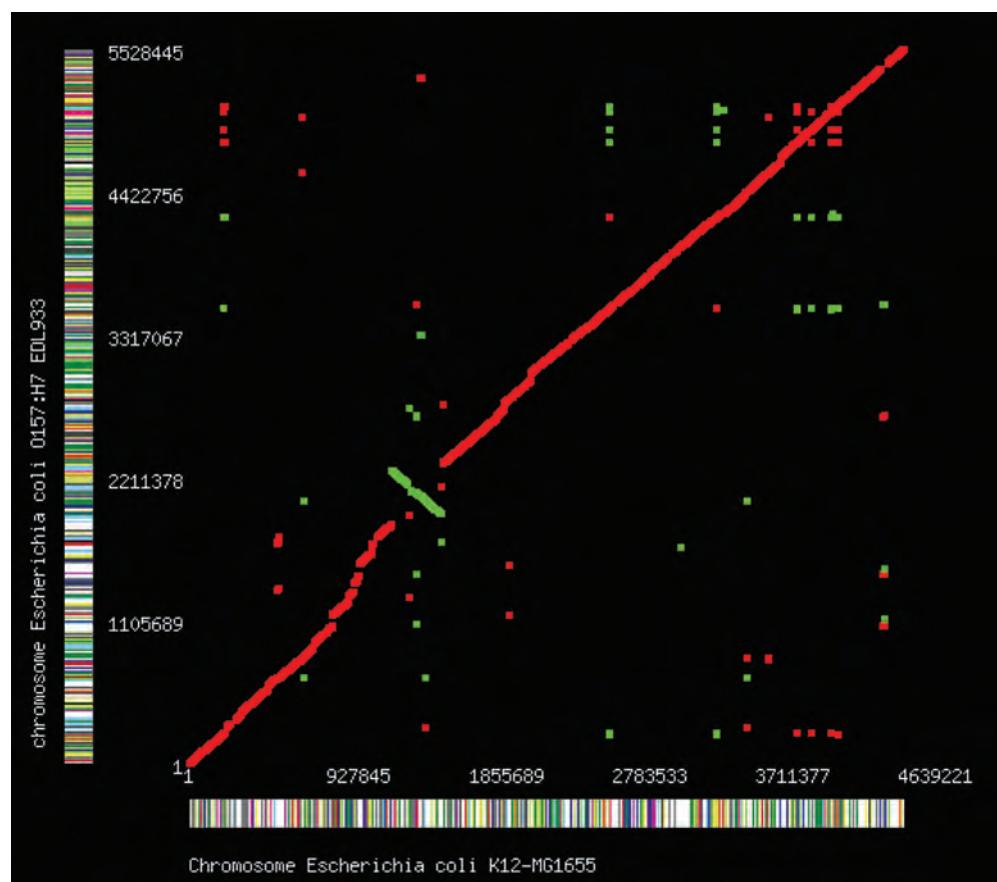
TaxPlot is therefore an easy way to identify proteins that are different in two microbial genomes of interest. The tool has also been extended to eukaryotes.

### MUMmer

We introduced MUMmer in Chapter 16 as a command-line tool to compare two segments of DNA such as bacterial genomes. Several web-based MUMmer applications are also available.

You can use MUMmer at IMG.

Two strains of *E. coli* are compared in the example of **Figure 17.18**: the harmless *E. coli* K-12 strain and *E. coli* O157:H7. The MUMmer output is useful to identify regions of the two genomes that are shared as well as regions in which the orientation is inverted. Eisen *et al.* (2000) used such analyses to describe symmetrical chromosomal inversions



**FIGURE 17.18** The MUMmer program allows you to select two microbial genomes of interest for comparison on a dotplot. The minimal alignment length can be adjusted. The MUMmer output consists of a dotplot that displays maximally unique matching subsequences (MUMs) between two genomes. This tool rapidly describes the relationship between two genomes, including information on the relative orientation of the genomic DNA and the presence of insertions or deletions. Here *E. coli* K-12 MG1655 is represented on the x axis, and the pathogenic strain *E. coli* 0157:H7 EDL933 is on the y axis. There is a major 45° line where the two closely related genomes align. A line segment near the center is oriented at a 90° angle. This represents an inversion in which the orientation of a genomic segment in one of the two strains is reversed relative to the other. Created using MUMmer.

near the origin of replication in comparisons of closely related species including *C. pneumoniae* versus *C. trachomatis*.

There are two further extensions of MUMmer. NUCmer (NUCleotide MUMmer) allows multiple reference and query sequences to be aligned. One application is to align a group of contigs. PROtein MUMmer (PROmer) is similar to NUCmer, but uses six-frame translations of each nucleotide sequence, offering superior sensitivity in aligning distantly related sequences.

## PERSPECTIVE

The recent and ongoing sequencing of thousands of bacterial and archaeal genomes has had a profound effect on virtually all aspects of microbiology. We can summarize the benefits of whole-genome sequencing of microbes as follows:

- Upon identifying the entire DNA sequence of a bacterial or archaeal genome, we obtain a comprehensive survey of all the genes and regulatory elements. This is

similar to obtaining a parts list of a machine, although we do not have the instruction manual.

- Through comparative genomics, we may learn the principles by which the “machine” is assembled and by which it functions.
- We can understand the diversity of microbial species through comparative genomics. We can therefore begin to uncover the principles of genome organization, and can compare pathogenic versus nonpathogenic strains. We can also appreciate the dramatic differences in genome properties between two strains from the same species.
- We are gaining insights into the evolution of both genes and species, and can now appreciate lateral gene transfer as one of the driving forces of microbial evolution. We can study gene duplication and gene loss. Having the complete genome available is important both to learn what genes comprise an organism as well as to learn what genes are absent.
- Complete genome sequences offer a starting point for biological investigations.

## PITFALLS

As complete bacterial and archaeal genomes are sequenced, two of the most important tasks are gene identification and genome annotation. Gene identification has become routine, but can be difficult for several reasons. It can be difficult to assess whether short ORFs correspond to transcripts that are actively transcribed. For example, in contrast to eukaryotes, bacteria and archaea do not always use AUG as a start codon.

Genome annotation is the critical process by which functions are assigned to predicted proteins. When genome sequences were first obtained in the 1990s, it was common for half of all predicted proteins to have no known homologs and their function was entirely obscure. Perhaps surprisingly this situation has persisted to a large extent, with many genes annotated as “hypothetical” or having unknown function.

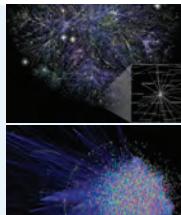
Gene annotation performed computationally should always be viewed as generating a hypothesis that needs to be experimentally tested. There are several kinds of common errors (Brenner, 1999; Peri *et al.*, 2001; Richardson and Watson, 2013). These include transitive catastrophes (inappropriately assigning a function to a gene based upon homology to another gene with a known function) and misidentification of small ORFs as authentic genes when they are not transcribed.

## ADVICE FOR STUDENTS

As with the advice in the previous chapter, choose a bacterial species (whether *E. coli*, *Y. pestis*, or any other) and: (1) read the primary genomics literature; and (2) download its genomic sequence and analyze it in depth. Try to annotate its genes using RAST, then repeat the annotation using different reference species. Select several genes that (according to the literature) were acquired by lateral gene transfer, then assess the evidence for lateral transfer by determining their GC content, dinucleotide frequencies, or phylogenetic positions.

## WEB RESOURCES

The Genomes Online Database (GOLD) provides an important starting point for any study of microbial genomes (<http://genomesonline.org/>; WebLink 17.2). IMG (<http://img.jgi.doe.gov/>; WebLink 17.5) offers useful analysis tools.



## Discussion Questions

**[17-1]** Anthrax strains vary in their pathogenicity. What bioinformatics approaches could you take to understand the basis of this difference? What specific proteins are involved in its pathogenicity?

**[17-2]** How can you assess whether bacterial genes have incorporated into the human genome through lateral gene transfer? What alternative explanations could there be for the presence of a human protein that is most closely related to a group of bacterial proteins, without having other eukaryotic orthologs?

**[17-3]** Consider the differences between *E. coli* K-12 and *E. coli* O157:H7 and other closely related pairs of bacteria. They undergo lateral gene transfer to different degrees, they have distinct patterns of pathogenicity, and these two strains even differ in genome size by over a million base pairs. What is the definition of a species? Is *E. coli* a species?

### PROBLEMS/COMPUTER LAB

**[17-1]** How many regions of *E. coli* K-12 are homologous to viruses? Visit the UCSC microbial browser (<http://microbes.ucsc.edu>), select the Table Browser, choose the Bacteria-Proteobacteria-Gamma clade and *Escherichia coli* K-12, set the group to Comparative Genomics and the track to BlastP Viruses. Output options include plain text, browser extensible data (BED), or hyperlinks to the genome browser.

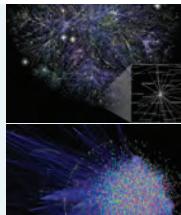
**[17-2]** Analyze a gene from *E. coli* in depth. First, find an *E. coli* gene that is known to have a homolog in eukaryotes (for example, DELTA-BLAST search human beta globin against the RefSeq database restricted to *E. coli*). Use the resources described in this chapter to characterize its paralogs, orthologs, and function.

**[17-3]** Explore the GC content and codon utilization of a bacterium. Search the genomes for *Yersinia pestis*, and select *Y. pestis* CO92. How many chromosomes and plasmids does it contain? Use the R tools described in this chapter (see Fig. 17.10) and compare the range of GC percent across the main chromosome and across plasmid pPCP1. Which has a higher GC content? Try using several window sizes.

**[17-4]** The bacterium *Wolbachia pipiensis* is an endosymbiont that lives inside insect and nematode hosts. A large fraction of its genome has transferred to the nuclear genome of some hosts (Hotopp *et al.*, 2007). Select a *Wolbachia* protein (e.g., NP\_965857.1) and provide evidence that an ortholog has been laterally transferred to a *Drosophila* species. As one strategy, first perform BLASTP with the protein as a query, restricting the output to bacteria and then restricting the output to eukaryotes. Try performing a TBLASTN search against the trace archives (a link is provided on the main NCBI blast web page). Try a TBLASTN search against the whole-genome shotgun read database restricted to the insects.

**[17-5]** Compare two completed genomes. Begin at NCBI Genome. Choose bacteria, then choose an organism such as *Rickettsia prowazekii*. Use TaxPlot to perform a three-way genome comparison. Repeat your analysis with MUMmer and Artemis at IMG. Identify the chromosomal segments that harbor outliers based on the TaxPlot analysis.

**[17-6]** We noted that the *Candidatus Carsonella ruddii* genome is extremely small (see accession NC\_008512.1). First note how many genes are annotated based on NCBI's Entrez database. Next, obtain the sequence (159,662 nucleotides) in FASTA format and input it to the GLIMMER program for gene prediction (either via the command line or via the NCBI genomes site). How many genes does the GLIMMER program annotate relative to the NCBI annotation?



## Self-Test Quiz

**[17-1]** A typical bacterial genome is composed of approximately how many base pairs of DNA?

- (a) 20,000 base pairs;
- (b) 200,000 base pairs;
- (c) 2,000,000 base pairs (2 Mb); or
- (d) 20,000,000 base pairs (20 Mb).

**[17-2]** *Myxococcus xanthus* has a relatively large genome size, even compared to other proteobacteria. One reason for this size may be:

- (a) *M. xanthus* acquired repetitive DNA sequences;
- (b) *M. xanthus* is a bacterium with a relatively large diameter size;

- (c) *M. xanthus* has a complex social lifestyle requiring large numbers of genes; or
- (d) *M. xanthus* acquired a large number of plasmids.

**[17-3]** The *E. coli* genome encodes about 4300 protein-coding genes. The total number of *E. coli* introns is approximately:

- (a) 10;
- (b) 430;
- (c) 4300; or
- (d) 43,000.

**[17-4]** The smallest bacterial genomes tend to be those of:

- (a) extremophiles;
- (b) viruses;
- (c) intracellular species; or
- (d) *Bacilli*.

**[17-5]** Which of the following constitutes strongest evidence that a gene became incorporated into the *E. coli* genome by lateral gene transfer?

- (a) the GC content of the gene varies greatly relative to other *E. coli* genes;
- (b) the frequency of codon utilization of the gene varies greatly relative to other *E. coli* genes;
- (c) phylogenetic analysis shows that proteobacteria closely related to *E. coli* lack this gene; or
- (d) any of the above.

**[17-6]** The pathogenic strain *E. coli* O157:H7 EDL933 is substantially larger than *E. coli* K-12 substr. MG1655. By using TaxPlot, MUMmer, or NCBI Genome, its approximate number of additional genes can be determined as:

- (a) 1000;
- (b) 2000;
- (c) 3000; or
- (d) 8000.

## SUGGESTED READING

There is a large literature on bacterial genomics. Important reviews are by Fraser-Liggett (2005), Ward and Fraser (2005), and Bentley and Parkhill (2004). Casjens (1998) provided an excellent introduction, and this was updated and expanded by Bentley and Parkhill on comparative genomics. Susannah Tringe and Edward Rubin (2005) and Riesenfeld *et al.* (2004) provide introductions to metagenomics. Five leaders in the field of the human microbiome (Martin Blaser, Peer Bork, Claire Fraser, Rob Knight, and Jun Wang) offer their opinions of key findings and trends (Blaser *et al.*, 2013).

David Edwards and Kathryn Holt (2013) offer an excellent guide to comparative bacterial genome analysis using next-generation sequence data. Their review discusses genome assembly, contigs, annotation, and comparative genomics. The supplement to their paper provides a detailed tutorial that introduces a wide range of bioinformatics software tools, and is highly recommended.

## REFERENCES

- Acevedo-Rocha, C.G., Fang, G., Schmidt, M., Ussery, D.W., Danchin, A. 2013. From essential to persistent genes: a functional approach to constructing synthetic life. *Trends in Genetics* **29**(5), 273–279. PMID: 23219343.
- Achenbach-Richter, L., Gupta, R., Stetter, K.O., Woese, C.R. 1987. Were the original eubacteria thermophiles? *Systematic and Applied Microbiology* **9**, 34–39.
- Akman, L., Rio, R.V., Beard, C.B., Aksoy, S. 2001. Genome size determination and coding capacity of *Sodalis glossinidius*, an enteric symbiont of tsetse flies, as revealed by hybridization to *Escherichia coli* gene arrays. *Journal of Bacteriology* **183**, 4517–4525.
- Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., Aksoy, S. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidiae*. *Nature Genetics* **32**, 402–407.

- Alm, R.A., Ling, L.S., Moir, D.T. *et al.* 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176–180. PMID: 9923682.
- Andersson, J.O. 2009. Gene transfer and diversification of microbial eukaryotes. *Annual Review of Microbiology* **63**, 177–193. PMID: 19575565.
- Andersson, S.G. 2006. The bacterial world gets smaller. *Science* **314**, 259–260.
- Andersson, S.G., Kurland, C.G. 1998. Reductive evolution of resident genomes. *Trends in Microbiology* **6**, 263–268.
- Andersson, S.G., Zomorodipour, A., Andersson, J.O. *et al.* 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998). PMID: 9823893.
- Arumugam, M., Raes, J., Pelletier, E. *et al.* 2011. Enterotypes of the human gut microbiome. *Nature* **473**(7346), 174–180. PMID: 21508958.
- Aziz, R.K., Bartels, D., Best, A.A. *et al.* 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75. PMID: 18261238.
- Aziz, R.K., Devoid, S., Disz, T. *et al.* 2012. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS One* **7**(10), e48053. PMID: 23110173.
- Badger, J.H., Eisen, J.A., Ward, N.L. 2005. Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders 'Rhodobacterales' and *Caulobacterales*. *International Journal of Systematic and Evolutionary Microbiology* **55**, 1021–1026.
- Baker, B.J., Tyson, G.W., Webb, R.I., Flanagan, J., Hugenholtz, P., Allen, E.E., Banfield, J.F. 2006. Lineages of acidophilic archaea revealed by community genomic analysis. *Science* **314**, 1933–1935.
- Baytaluk, M. V., Gelfand, M. S., Mironov, A. A. 2002. Exact mapping of prokaryotic gene starts. *Briefings in Bioinformatics* **3**, 181–194.
- Beiko, R.G., Ragan, M.A. 2008. Detecting lateral genetic transfer : a phylogenetic approach. *Methods in Molecular Biology* **452**, 457–469. PMID: 18566777.
- Bender, L., Ott, M., Marre, R., Hacker, J. 1990. Genome analysis of *Legionella* ssp. by orthogonal field alternation gel electrophoresis (OFAGE). *FEMS Microbiology Letters* **60**, 253–257.
- Bennett, G.M., Moran, N.A. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, doi: 10.1093/gbe/evt118. PMID: 23918810.
- Bentley, S.D., Parkhill, J. 2004. Comparative genomic structure of prokaryotes. *Annual Review of Genetics* **38**, 771–792.
- Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A. M. *et al.* 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147. PMID: 12000953.
- Binnewies, T.T., Motro, Y., Hallin, P.F. *et al.* 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional and Integrative Genomics* **6**, 165–185.
- Blaser, M., Bork, P., Fraser, C., Knight, R., Wang, J. 2013. The microbiome explored: recent insights and future challenges. *Nature Reviews Microbiology* **11**(3), 213–217. PMID: 23377500.
- Blattner, F. R., Plunkett, G. 3rd, Bloch, C. A. *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474. PMID: 9278503.
- Boucher, Y., Douady, C.J., Papke, R.T. *et al.* 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics* **37**, 283–328.
- Brenner, S. E. 1999. Errors in genome annotation. *Trends in Genetics* **15**, 132–133.
- Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., Stanhope, M. J. 2001. Universal trees based on large combined protein sequence data sets. *Nature Genetics* **28**, 281–285.
- Brüggemann, H., Henne, A., Hostetler, F. *et al.* 2004. The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin. *Science* **305**, 671–673.
- Bult, C. J., White, O., Olsen, G. J. *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073. PMID: 8688087.
- Burke G.R., Moran, N.A. 2011. Massive genomic decay in *Serratia symbiotica*, a recently evolved symbiont of aphids. *Genome Biology and Evolution* **3**, 195–208. PMID: 21266540.

- Bush, K., Courvalin, P., Dantas, G. *et al.* 2011. Tackling antibiotic resistance. *Nature Reviews Microbiology* **9**(12), 894–896. PMID: 22048738.
- Canganella, F., Wiegel, J. 2011. Extremophiles: from abyssal to terrestrial ecosystems and possibly beyond. *Naturwissenschaften* **98**(4), 253–279. PMID: 21394529.
- Casjens, S. 1998. The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics* **32**, 339–377.
- Casjens, S., Palmer, N., van Vugt, R. *et al.* 2000. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology* **35**, 490–516. PMID: 10672174.
- Chambaud, I., Heilig, R., Ferris, S. *et al.* 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Research* **29**, 2145–2153. PMID: 11353084.
- Charon, N. W., Goldstein, S. F. 2002. Genetics of motility and chemotaxis of a fascinating group of bacteria: The Spirochetes. *Annual Review of Genetics* **36**, 47–73.
- Chaudhuri, R.R., Henderson, I.R. 2012. The evolution of the *Escherichia coli* phylogeny. *Infection, Genetics and Evolution* **12**(2), 214–226. PMID: 22266241.
- Cho, I., Blaser, M.J. 2012. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* **13**(4), 260–270. PMID: 22411464.
- Choi, I.G., Kim, S.H. 2007. Global extent of horizontal gene transfer. *Proceedings of the National Academy of Sciences, USA* **104**, 4489–4494.
- Cole, S. T., Brosch, R., Parkhill, J. *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544. PMID: 9634230.
- Cole, S. T., Eiglmeier, K., Parkhill, J. *et al.* 2001. Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011. PMID: 11234002.
- D'Elia, M.A., Pereira, M.P., Brown, E.D. 2009. Are essential genes really essential? *Trends in Microbiology* **17**(10), 433–438. PMID: 19765999.
- Dagan, T. 2011. Phylogenomic networks. *Trends in Microbiology* **19**(10), 483–491. PMID: 21820313.
- Dandekar, T., Huynen, M., Regula, J. T. *et al.* 2000. Re-annotating the *Mycoplasma pneumoniae* genome sequence: Adding value, function and reading frames. *Nucleic Acids Research* **28**, 3278–3288. PMID: 10954595.
- Daubin, V., Moran, N.A., Ochman, H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832.
- Dave, M., Higgins, P.D., Middha, S., Rioux, K.P. 2012. The human gut microbiome: current knowledge, challenges, and future directions. *Translational Research* **160**(4), 246–257. PMID: 22683238.
- Deckert, G., Warren, P.V., Gaasterland, T. *et al.* 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358. PMID: 9537320.
- Delcher, A. L., Harmon, D., Kasif, S., White, O., Salzberg, S. L. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**, 4636–4641.
- Delcher, A.L., Bratke, K.A., Powers, E.C., Salzberg, S.L. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679.
- DeLong, E. F. 1998. Everything in moderation: Archaea as “non-extremophiles.” *Current Opinion in Genetics and Development* **8**, 649–654 (1998).
- DeLong, E. F., Pace, N. R. 2001. Environmental diversity of bacteria and archaea. *Systematic Biology* **50**, 470–478.
- de Villiers, E. P., Brayton, K. A., Zweygarth, E., Allsopp, B. A. 2000. Genome size and genetic map of *Cowdria ruminantium*. *Microbiology* **146**, 2627–2634.
- Dobell, C. 1932. *Antony van Leeuwenhoek and his “little animals”*. Harcourt, Brace and Company, New York.
- Edwards, D.J., Holt, K.E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation* **3**(1), 2. PMID: 23575213.
- Eisen, J. A. 2000. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Current Opinion in Genetics and Development* **10**, 606–611.

- Eisen, J. A. 2001. Gastrogenomics. *Nature* **409**, 463, 465–466.
- Eisen, J.A., Heidelberg, J.F., White, O., Salzberg, S.L. 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology* **1**, RESEARCH0011.
- Eisenstein, B. I., Schaechter, M. 1999. Normal microbial flora. In *Mechanisms of Microbial Disease* (eds M.Schaechter, N. C.Engleberg, B. I.Eisenstein, G.Medoff). Lippincott Williams and Wilkins, Baltimore, MD, Chapter 20.
- Ermolaeva, M. D., White, O., Salzberg, S. L. 2001. Prediction of operons in microbial genomes. *Nucleic Acids Research* **29**, 1216–1221.
- Fleischmann, R. D., Adams, M. D., White, O. *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512. PMID: 7542800.
- Fournier, P.E., Raoult, D. 2011. Prospects for the future using genomics and proteomics in clinical microbiology. *Annual Review of Microbiology* **65**, 169–188. PMID: 21639792.
- Fox, G. E., Stackebrandt, E., Hespell, R. B. *et al.* 1980. The phylogeny of prokaryotes. *Science* **209**, 457–463. PMID: 6771870.
- Fraser, C., Hanage, W.P., Spratt, B.G. 2007. Recombination and the nature of bacterial speciation. *Science* **315**, 476–480.
- Fraser, C. M., Gocayne, J. D., White, O. *et al.* 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403. PMID: 7569993.
- Fraser, C. M., Casjens, S., Huang, W. M. *et al.* 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580–586. PMID: 9403685.
- Fraser, C. M., Norris, S. J., Weinstock, G. M. *et al.* 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388. PMID: 9665876.
- Fraser, C. M., Eisen, J. A., Salzberg, S. L. 2000. Microbial genome sequencing. *Nature* **406**, 799–803.
- Fraser-Liggett, C.M. 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Research* **15**, 1603–1610.
- Galagan, J. E., Nusbaum, C., Roy, A. *et al.* 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Research* **12**, 532–542. PMID: 11932238.
- Gil, R., Silva, F.J., Pereto, J., Moya, A. 2004. Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Review* **68**, 518–537.
- Gill, S.R., Pop, M., Deboy, R.T. *et al.* 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359.
- Gillespie, J.J., Wattam, A.R., Cammer, S.A. *et al.* 2011. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and Immunity* **79**(11), 4286–4298. PMID: 21896772.
- Giovannoni, S.J., Tripp, H.J., Givan, S. *et al.* 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**, 1242–1245.
- Glaser, P., Frangeul, L., Buchrieser, C. *et al.* 2001. Comparative genomics of *Listeria* species. *Science* **294**, 849–852. PMID: 11679669.
- Glass, J. I., Lefkowitz, E. J., Glass, J. S. *et al.* 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407**, 757–762. PMID: 11048724.
- Goldman, B. S., Nierman, W. C., Kaiser, D. *et al.* 2006. Evolution of sensory complexity recorded in a myxobacterial genome. *Proceedings of the National Academy of Sciences, USA* **103**, 15200–15205. PMID: 17015832.
- Graur, D., Li, W.-H. 2000. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 2000.
- Gupta, R. S., Griffiths, E. 2002. Critical issues in bacterial phylogeny. *Theoretical Population Biology* **61**, 423–434.
- Han, K., Li, Z.F., Peng, R. *et al.* 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieus. *Scientific Reports* **3**, 2101. PMID: 23812535.
- Hayashi, T., Makino, K., Ohnishi, M. *et al.* 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Research* **8**, 11–22. PMID: 11258796.

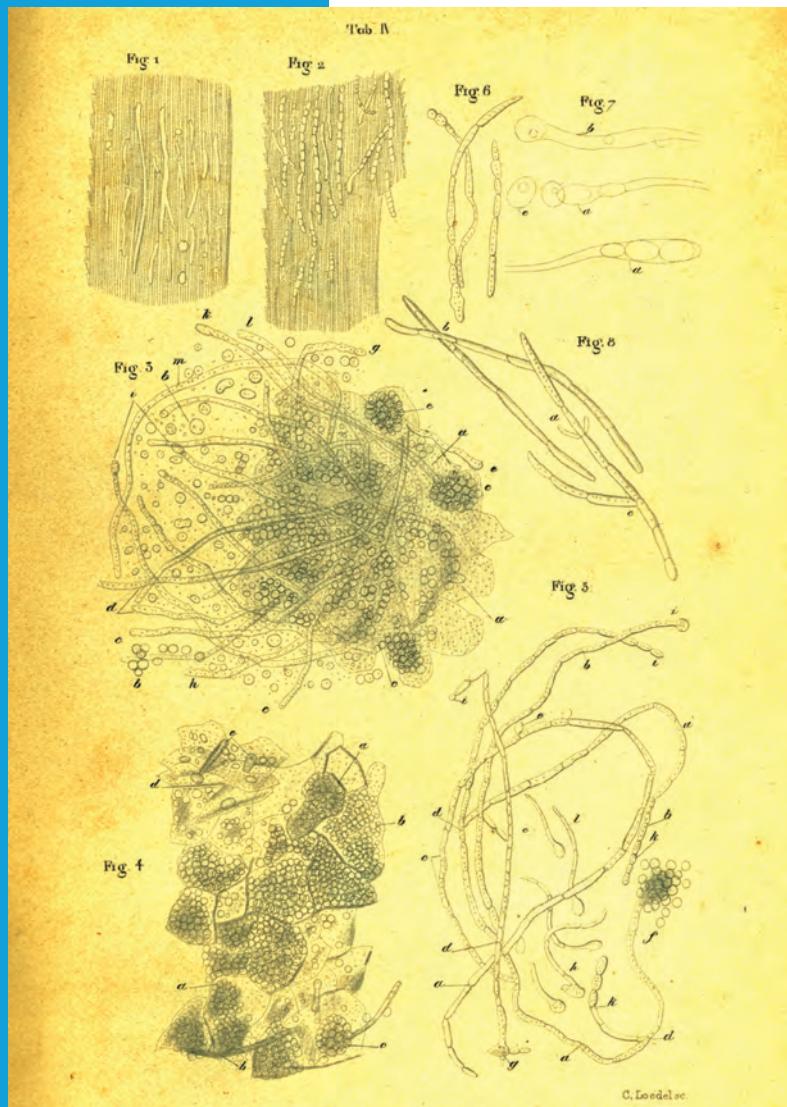
- Heidelberg, J. F., Eisen, J. A., Nelson, W. C. *et al.* 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483. PMID: 10952301.
- Himmelreich, R., Hilbert, H., Plagens, H. *et al.* 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research* **24**, 4420–4449. PMID: 8948633.
- Hotopp, J.C., Lin, M., Madupu, R. *et al.* 2006. Comparative genomics of emerging human ehrlichiosis agents. *PLoS Genetics* **2**, e21. PMID: 16482227.
- Hotopp, J.C., Clark, M.E., Oliveira, D.C. *et al.* 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756.
- Huang, C.H., Hsiang, T., Trevors, J.T. 2013. Comparative bacterial genomics: defining the minimal core genome. *Antonie Van Leeuwenhoek* **103**(2), 385–398. PMID: 23011009.
- Huang, J., Gogarten, J.P. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends in Genetics* **22**, 361–366.
- Huber, H., Hohn, M.J., Rachel, R. *et al.* 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67. PMID: 11986665.
- Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biology* **3**(2), doi: 10.1186/gb-2002-3-2-reviews0003.
- Hugenholtz, P., Goebel, B. M., Pace, N. R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* **180**, 4765–4774.
- Human Microbiome Project Consortium. 2012a. A framework for human microbiome research. *Nature* **486**(7402), 215–221. PMID: 22699610.
- Human Microbiome Project Consortium. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402), 207–214. PMID: 22699609.
- Itaya, M. 1995. An estimation of minimal genome size required for life. *FEBS Letters* **362**, 257–260.
- Joyce, E. A., Chan, K., Salama, N. R., Falkow, S. 2002. Redefining bacterial populations: A post-genomic reformation. *Nature Reviews Genetics* **3**, 462–473.
- Kaiser, D. 2013. Are Myxobacteria intelligent? *Frontiers in Microbiology* **4**, 335. PMID: 24273536.
- Kalman, S., Mitchell, W., Marathe, R. *et al.* 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genetics* **21**, 385–389. PMID: 10192388.
- Kawarabayasi, Y., Hino, Y., Horikawa, H. *et al.* 1999. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Research* **6**, 83–101, 145–152. PMID: 10382966.
- Kersey, P.J., Allen, J.E., Christensen, M. *et al.* 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research* **42**(1), D546–552. PMID: 24163254.
- Keseler, I.M., Mackie, A., Peralta-Gil, M. *et al.* 2013. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research* **41**(Database issue), D605–612. PMID: 23143106.
- Kisand, V., Lettieri, T. 2013. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics* **14**, 211. PMID: 23547799.
- Klenk, H. P., Clayton, R. A., Tomb, J. F. *et al.* 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364–370. PMID: 9389475.
- Klenk, H.P., Göker, M. 2010. En route to a genome-based classification of Archaea and Bacteria? *Systematic and Applied Microbiology* **33**(4), 175–182. PMID: 20409658.
- Koonin, E.V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology* **1**, 127–136.
- Koonin, E. V., Makarova, K. S., Aravind, L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annual Review of Microbiology* **55**, 709–742.
- Kuczynski, J., Lauber, C.L., Walters, W.A. *et al.* 2012. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics* **13**(1), 47–58. PMID: 22179717.
- Kunst, F., Ogasawara, N., Moszer, I. *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256. PMID: 9384377.
- Lamichhane, G., Bishai, W. 2007. Defining the ‘survivosome’ of *Mycobacterium tuberculosis*. *Nature Medicine* **13**, 280–282.

- Lamichhane, G., Zignol, M., Blades, N.J. *et al.* 2003. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences, USA* **100**(12), 7213–7218. PMID: 12775759.
- Latendresse, M., Paley, S., Karp, P.D. 2012. Browsing metabolic and regulatory networks with BioCyc. *Methods in Molecular Biology* **804**, 197–216. PMID: 22144155.
- Le Chatelier, E., Nielsen, T., Qin, J. *et al.* 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**(7464), 541–546. PMID: 23985870.
- Loman, N.J., Constantinidou, C., Chan, J.Z. *et al.* 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* **10**(9), 599–606. PMID: 22864262.
- Luo, H., Lin, Y., Gao, F., Zhang, C.T., Zhang, R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research* **42**(1), D574–580. PMID: 24243843.
- Ma, J., Campbell, A., Karlin, S. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *Journal of Bacteriology* **184**, 5733–5745.
- Mansfield, J., Genin, S., Magori, S. *et al.* 2012. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology* **13**(6), 614–629. PMID: 22672649.
- Markowitz, V.M., Chen, I.M., Chu, K. *et al.* 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research* **42**(1), D568–573. PMID: 24136997.
- Martin, W., Koonin, E.V. 2006. A positive definition of prokaryotes. *Nature* **442**, 868.
- Mavromatis, K., Land, M.L., Brettin, T.S. *et al.* 2012. The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. *PLoS One* **7**(12), e48837. PMID: 23251337.
- May, B. J., Zhang, Q., Li, L. L. *et al.* 2001. Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proceedings of the National Academy of Sciences USA* **98**, 3460–3465. PMID: 11248100.
- McCutcheon, J.P., Moran, N.A. 2011. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology* **10**(1), 13–26. PMID: 22064560.
- McCutcheon, J.P., von Dohlen, C.D. 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology* **21**(16), 1366–1372. PMID: 21835622.
- Medini, D., Serruto, D., Parkhill, J. *et al.* 2008. Microbiology in the post-genomic era. *Nature Reviews Microbiology* **6**(6), 419–430. PMID: 18475305.
- Mitchison, G. 2005. The regional rule for bacterial base composition. *Trends in Genetics* **21**, 440–443.
- Mizoguchi, H., Mori, H., Fujio, T. 2007. *Escherichia coli* minimum genome factory. *Biotechnology and Applied Biochemistry* **46**(Pt 3), 157–167. PMID: 17300222.
- Morgan, X.C., Segata, N., Huttenhower, C. 2013. Biodiversity and functional genomics in the human microbiome. *Trends in Genetics* **29**(1), 51–58. PMID: 23140990.
- Mushegian, A. R., Koonin, E. V. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences USA* **93**, 10268–10273.
- Nakabachi, A., Yamashita, A., Toh, H. *et al.* 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* **314**, 267.
- NCBI Resource Coordinators. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **42**(1), D7–D17. PMID: 24259429.
- Nelson, K. E., Clayton, R. A., Gill, S. R. *et al.* 1999. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329. PMID: 10360571.
- Nolling, J., Breton, G., Omelchenko, M. V. *et al.* 2001. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *Journal of Bacteriology* **183**, 4823–4838. PMID: 11466286.
- Ogata, H., Audic, S., Renesto-Audiffren, P. *et al.* 2001. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* **293**, 2093–2098. PMID: 11557893.
- Overbeek, R., Olson, R., Pusch, G.D. *et al.* 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**(1), D206–214 PMID: 24293654.

- Owen, R. 1843. *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. Longman, Brown, Green, and Longmans, London.
- Pace, N.R. 2009. Problems with “prokaryote”. *Journal of Bacteriology* **191**(7), 2008–2010. PMID: 19168605.
- Pagani, I., Liolios, K., Jansson, J. et al. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research* **40**(Database issue), D571–579. PMID: 22135293.
- Parkhill, J., Dougan, G., James, K.D. et al. 2001a. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *typhi* CT18. *Nature* **413**, 848–852. PMID: 11677608.
- Parkhill, J., Wren, B.W., Thomson, N.R. et al. 2001b. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**, 523–527. PMID: 11586360.
- Pennisi, E. 2012. Microbiology. Microbial survey of human body reveals extensive variation. *Science* **336**(6087), 1369–1371. PMID: 22700898.
- Pérez-Brocal, V., Gil, R., Ramos, S. et al. 2006. A small microbial genome: the end of a long symbiotic relationship? *Science* **314**, 312–313.
- Peri, S., Ibarrola, N., Blagoev, B., Mann, M., Pandey, A. 2001. Common pitfalls in bioinformatics-based analyses: Look before you leap. *Trends in Genetics* **17**, 541–545.
- Perna, N. T., Plunkett, G. 3rd, Burland, V. et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533. PMID: 11206551.
- Pósfaí, G., Plunkett, G. 3rd, Fehér, T. et al. 2006. Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046.
- Qin, J., Li, R., Raes, J. et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**(7285), 59–65. PMID: 20203603.
- Rasko, D.A., Webster, D.R., Sahl, J.W. et al. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. *New England Journal of Medicine* **365**(8), 709–717. PMID: 21793740.
- Read, T.D., Brunham, R.C., Shen, C. et al. 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research* **28**, 1397–1406. PMID: 10684935.
- Read, T. D., Salzberg, S. L., Pop, M. et al. 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296**, 2028–2033. PMID: 12004073.
- Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., Whittam, T. S. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67.
- Rendulic, S., Jagtap, P., Rosinus, A. et al. 2004. A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**, 689–692.
- Reysenbach, A. L., Shock, E. 2002. Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296**, 1077–1082.
- Richardson, E.J., Watson, M. 2013. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics* **14**(1), 1–12. PMID: 22408191.
- Riesenfeld, C.S., Schloss, P.D., Handelsman, J. 2004. Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* **38**, 525–552.
- Riley, M., Abe, T., Arnaud, M. B. et al. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot: 2005. *Nucleic Acids Research* **34**, 1–9. PMID: 16397293.
- Robertson, C.E., Harris, J.K., Spear, J.R., Pace, N.R. 2005. Phylogenetic diversity and ecology of environmental Archaea. *Current Opinion in Microbiology* **8**, 638–642.
- Rudd, K.E. 2000. EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Research* **28**, 60–64.
- Salgado, H., Santos-Zavaleta, A., Gama-Castro, S. et al. 2006. The comprehensive updated regulatory network of *Escherichia coli* K-12. *BMC Bioinformatics* **7**, 5.
- Savage, D.C. 1977. Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology* **31**, 107–133. PMID: 334036.

- Schaechter, M. 1999. Introduction to the pathogenic bacteria. In *Mechanisms of Microbial Disease* (eds M.Schaechter, N. C.Engleberg, B. I.Eisenstein, G.Medoff). Lippincott Williams and Wilkins, Baltimore, MD, Chapter 10.
- Schloss, P.D., Handelsman, J. 2004. Status of the microbial census. *Microbiology and Molecular Biology Reviews* **68**, 686–691.
- Schneiker, S., Perlova, O., Kaiser, O. *et al.* 2007. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nature Biotechnology* **25**(11), 1281–1289. PMID: 17965706.
- Schönknecht, G., Chen, W.H., Ternes, C.M. *et al.* 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**(6124), 1207–1210. PMID: 23471408.
- Sherratt, D. 2001. Divide and rule: The bacterial chromosome. *Trends in Genetics* **17**, 312–313.
- Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H. 2000. Genome sequence of the endo-cellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86.
- Shih, P.M., Wu, D., Latifi, A. *et al.* 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences, USA* **110**(3), 1053–1058. PMID: 23277585.
- Shirai, M., Hirakawa, H., Kimoto, M. *et al.* 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Research* **28**, 2311–2314. PMID: 10871362.
- Singh, P., Cole, S.T. 2011. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. *Future Microbiology* **6**(1), 57–71. PMID: 21162636.
- Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., Krogh, A. 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends in Genetics* **17**, 425–428.
- Stepanauskas, R. 2012. Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology* **15**(5), 613–620. PMID: 23026140.
- Stephens, R. S., Kalman, S., Lammel, C. *et al.* 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759. PMID: 9784136.
- Stover, C.K., Pham, X.Q., Erwin, A.L. *et al.* 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964. PMID: 10984043.
- Sueoka, N. 1961. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences, USA* **47**, 1141–1149. PMID: 16590864.
- Sun, L.V., Foster, J.M., Tzertzinis, G. *et al.* 2001. Determination of *Wolbachia* genome size by pulsed-field gel electrophoresis. *Journal of Bacteriology* **183**, 2219–2225. PMID: 11244060.
- Tamas, I., Klasson, L., Canbäck, B. *et al.* 2002. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376–2379. PMID: 12089438.
- Teeling, H., Gloeckner, F.O. 2006. RibAlign: a software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits. *BMC Bioinformatics* **7**, 66.
- Teeling, H., Glöckner, F.O. 2012. Current opportunities and challenges in microbial metagenome analysis: a bioinformatic perspective. *Briefings in Bioinformatics* **13**(6), 728–742. PMID: 22966151.
- Tettelin, H., Saunders, N.J., Heidelberg, J. *et al.* 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**, 1809–1815. PMID: 10710307.
- Tomb, J.F., White, O., Kerlavage, A.R. *et al.* 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539–547. PMID: 9252185.
- Toussaint, A., Chandler, M. 2012. Prokaryote genome fluidity: toward a system approach of the mobileome. *Methods in Molecular Biology* **804**, 57–80. PMID: 22144148.
- Tringe, S.G., Rubin, E.M. 2005. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* **6**, 805–814.
- Turnbaugh, P.J., Hamady, M., Yatsunenko, T. *et al.* 2009. A core gut microbiome in obese and lean twins. *Nature* **457**(7228), 480–484. PMID: 19043404.
- Ward, N., Fraser, C.M. 2005. How genomics has affected the concept of microbiology. *Current Opinion in Microbiology* **8**, 564–571.

- Waters, E., Hohn, M.J., Ahel, I. *et al.* 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences, USA* **100**, 12984–12988. PMID: 14566062.
- Weinstock, G.M. 2012. Genomic approaches to studying the human microbiota. *Nature* **489**(7415), 250–256. PMID: 22972298.
- White, O., Eisen, J.A., Heidelberg, J.F. *et al.* 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571–1577. PMID: 10567266.
- Woese, C.R. 2002. On the evolution of cells. *Proceedings of the National Academy of Sciences, USA* **99**, 8742–8747.
- Woese, C.R., Fox, G.E. 1977. The concept of cellular evolution. *Journal of Molecular Evolution* **10**, 1–6.
- Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., Koonin, E. V. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology* **1**, 8.
- Wu, D., Hugenholtz, P., Mavromatis, K. *et al.* 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**(7276), 1056–1060. PMID: 20033048.
- Zahradka, K., Slade, D., Bailone, A., Sommer, S., Averbeck, D., Petranovic, M., Lindner, A.B., Radman, M. 2006. Reassembly of shattered chromosomes in *Deinococcus radiodurans*. *Nature* **443**, 569–573.
- Zhao, L. 2013. The gut microbiota and obesity: from correlation to causality. *Nature Reviews Microbiology* **11**(9), 639–647. PMID: 23912213.
- Zhi, X.Y., Zhao, W., Li, W.J., Zhao, G.P. 2012. Prokaryotic systematics in the genomics era. *Antonie Van Leeuwenhoek* **101**(1), 21–34. PMID: 22116211.



Some 200 fungal species are known to be pathogenic for humans, distressing millions of people. From the times of Greek and Roman antiquity to the middle of the nineteenth century, only two fungal diseases were known: ringworm (tinea) and thrush (oral candidiasis) (Ainsworth, 1993). Ringworm is caused by fungi of the genera *Microsporum*, *Trichophyton*, and *Epidermophyton*. Candidiasis (including thrush) is caused by *Candida albicans* and other *Candida* species. This image from Kuchenmeister (1857, plate IV) shows the thrush fungus, at that time called *Oidium albicans* (fig. 3 to 8).

Source: Kuchenmeister (1857).

# Eukaryotic Genomes: Fungi

# CHAPTER 18

*The chromosome III sequence has revealed 145 novel protein-encoding genes and a start has been made on their functional analysis. The results so far indicate that there are vast areas of yeast genetics of which we are completely ignorant and emphasize the need for molecular genetics and physiological studies to proceed hand-in-hand. The data also call for a radical reappraisal of our view of the yeast genetic map. The availability of the sequence establishes unequivocally the locations of the different genes on the chromosome. In consequence, the genetic map acquires a different emphasis; it becomes much more of a tool with which to study recombination and the dynamics of chromosome evolution. The goal of sequencing the entire yeast genome is achievable with present technology and this sequence will prove at least as important to the future development of eukaryotic molecular biology as the classical *S. cerevisiae* genetic map has in the past. The complete sequence of the yeast genome will open up new areas of molecular genetics and establish a foundation for the interpretation of sequence data from higher organisms.*

—Stephen Oliver et al. (1992) reporting the complete sequence of *S. cerevisiae* chromosome III, the first eukaryotic chromosome to be sequenced.

## LEARNING OBJECTIVES

After reading this chapter you should be able to:

- provide an overview of how fungi are classified;
- describe the features of the *Saccharomyces cerevisiae* genome;
- discuss genome duplication in *S. cerevisiae*;
- describe comparative genomics of the genus *Saccharomyces*; and
- describe comparative genomics of other fungal phyla.

## INTRODUCTION

According to the classification system of Whittaker (1969), there are five kingdoms of life: monera (prokaryotes); protists; animals; fungi; and plants. We have examined the bacteria and archaea in Chapter 17, and introduced the eukaryotic chromosome in Chapter 8. In this chapter we begin our exploration of eukaryotes by studying the kingdom of fungi. This diverse and interesting group of organisms last shared a common ancestor with plants and animals 1.5 billion years ago (BYA) (Wang *et al.*, 1999, discussed in Chapter 19). Some may think of fungi as organisms such as mushrooms that might be studied by botanists. Surprisingly, fungi are far more closely related to animals than to

Mycology (from the Greek word μύκης meaning “fungus”) is the study of fungi. Mycosis is a disease or ailment caused by fungi. The suffix *-mycota* refers to fungi: The kingdom fungi is also called the kingdom Eumycota (“true fungi”).

plants. In Chapter 19 we extend our study to the entire kingdom of eukaryotes, including animals, plants, and a variety of protozoa. We then discuss humans (Chapter 20).

The first eukaryotic genome to be fully sequenced was the 13-million-base-pair (Mb) genome of a fungus, the budding yeast *Saccharomyces cerevisiae*. Its genome is very small compared with that of humans (3 billion base pairs or gigabase pairs Gb), and its size is only severalfold larger than a typical bacterial genome. This yeast has served as a model eukaryotic organism for genetics studies because it grows rapidly, it can easily be genetically modified, and many of its cellular functions are conserved with metazoans and other eukaryotes. More recently, it became a model organism for functional genomics studies (Chapter 14). Every one of its approximately 6000 genes has been deleted, over-expressed, and characterized functionally using a variety of assays.

Now, as whole-genome sequencing has become routine, the sequencing of yeasts and other fungi has progressed at an accelerated pace. While the fungi are eukaryotes and share many properties in common with the metazoans (animals), most have relatively small genome sizes. Through comparative analysis we are gaining new insights into many basic properties of genome structure and evolution, including whole-genome duplications and the fate of duplicated genes (Dujon, 2006).

This chapter begins with an overview of the fungi. We then describe bioinformatic approaches to analyzing the *S. cerevisiae* genome. Finally, we describe the sequencing of other fungal genomes and the early lessons of comparative genomics in fungi.

## DESCRIPTION AND CLASSIFICATION OF FUNGI

Morphologically, fungi are characterized by hyphae (filaments) that grow and may branch. The Museum of Paleontology at the University of California, Berkeley, offers an introduction to fungi, including photographs of many species, at  
 <http://www.ucmp.berkeley.edu/fungi/fungi.html>  
 (WebLink 18.1).

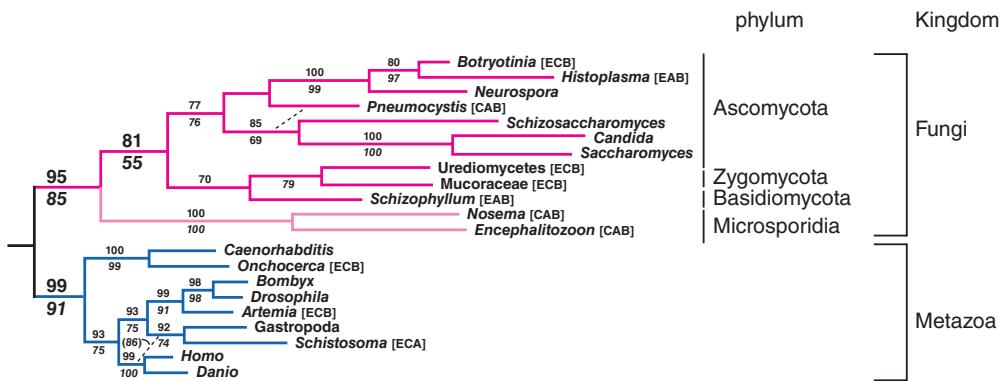
We explore this comprehensive tree in detail in Chapter 19 (Fig. 19.1).

Fungi are grown on food products such as Camembert and Brie cheeses to provide flavor. Fungi are used to produce soy sauce and many other foods and medicines.

Fungi are eukaryotic organisms that can be filamentous (as in the case of molds) or unicellular (as in the case of yeasts such as *S. cerevisiae*). The main criteria for classifying fungi are based on morphology (e.g., ultrastructure), biochemistry (e.g., growth properties or cell wall composition), and molecular sequence data (DNA, RNA, and protein sequences). Most fungi are aerobic, and all are heterotrophs that absorb their food. Fungi are typically very hardy, forming spores composed of chitin. They have a major role in the ecosystem in degrading organic waste material. Fungi are important causative agents of disease to humans, other animals, and plants (van de Wouw and Howlett, 2011). Fungi also have key roles in fermentation; the fungal mold *Rhizopus nigricans* is used in the manufacture of steroids such as cortisone, and *Penicillium chrysogenum* produces the antibiotic penicillin.

The relationships of many species throughout the tree of life have been described in phylogenetic analyses based on small-subunit ribosomal RNA (Fig. 15.1). In a complementary approach, W. F. Doolittle and colleagues defined a phylogeny of the eukaryotes based upon the concatenated amino acid sequences from four proteins: elongation factor-1 $\alpha$ , actin,  $\alpha$ -tubulin, and  $\beta$ -tubulin (Baldauf *et al.*, 2000). A portion of the tree shows that fungi form a monophyletic clade that is a sister group to animals (metazoan; Fig. 18.1). This close relationship between fungi and animals has been somewhat surprising given the apparently simple, unicellular nature of many fungi. However, fungi and animals share many similarities. Chitin is the main component of the fungal cell wall, and it is also a constituent of the arthropod exoskeleton. (Plant cell walls use cellulose.) Many of the fundamental processes of yeast, such as cell cycle control, DNA repair, and intracellular vesicle trafficking, are closely conserved with mammalian cells.

Advances in genomics have enabled continued progress in taxonomy, including sequence-based phylogenies (Casaregola *et al.*, 2011). According to the phylogenetic classification of Hibbett *et al.* (2007), the kingdom Fungi has seven phyla (see Box 18.1 for a discussion of fungal taxonomy). Of these phyla the subkingdom Dikarya includes the *Ascomycota* (including *Saccharomyces cerevisiae*) and *Basidiomycota*. The Hibbett *et al.* classification was consistent with a sampling of nearly 200 fungal species of every



**FIGURE 18.1** Phylogenetic analysis of the fungi reveals that they form a sister group with the metazoa (animals). This tree is a detailed view of a broad analysis of the eukaryotes (see Fig. 19.1) by Baldauf *et al.* (2000). The tree was generated using a multiple sequence alignment of four concatenated protein sequences: elongation factor 1 $\alpha$  (EF-1 $\alpha$ ; abbreviated E in tree), actin (C),  $\alpha$ -tubulin (A), and  $\beta$ -tubulin (B). Microsporidia were formerly classified as deep-branching eukaryotes but are now grouped with fungi. The fungal phylum Chytridiomycota is not shown in this tree.

Source: Baldauf *et al.* (2000). Reproduced with permission from AAAS.

major clade of *Fungi* by James *et al.* (2006). Phylogenetic analysis relied on a set of six genes (Box 18.1). Figure 18.2 depicts a phylogenetic tree based on James *et al.*

For web resources that address fungal taxonomy, visit the Index Fungorum (<http://www.indexfungorum.org/>, WebLink 18.2), MycoBank (<http://www.mycobank.org/>, WebLink 18.3), and the Global Biodiversity Information Facility (<http://www.gbif.org/>, WebLink 18.4).

From the time of Anton van Leeuwenhoek (1632–1723), yeast were thought to be chemical substances that are not living. Theodor Schwann (1810–1882) and Baron Charles Cagniard-Latour (1777–1859) independently discovered in 1836–1837 that yeast are composed of living cells. Schwann studied fermenting yeast and called them *Zuckerpilz* (sugar fungus), from which the term *Saccharomyces* is derived (Bulloch, 1938).

## INTRODUCTION TO BUDDING YEAST *SACCHAROMYCES CEREVISIAE*

The budding yeast *S. cerevisiae* was the first species domesticated by humans at least 10,000 years ago. It is commonly called brewer's yeast or baker's yeast, and it ferments glucose to ethanol and carbon dioxide. For almost 200 years, researchers have exploited this organism for biochemical, genetic, molecular, and cell biological studies. Because many of its characteristics are also conserved in human cells, yeast has emerged as a powerful instrument for basic research.

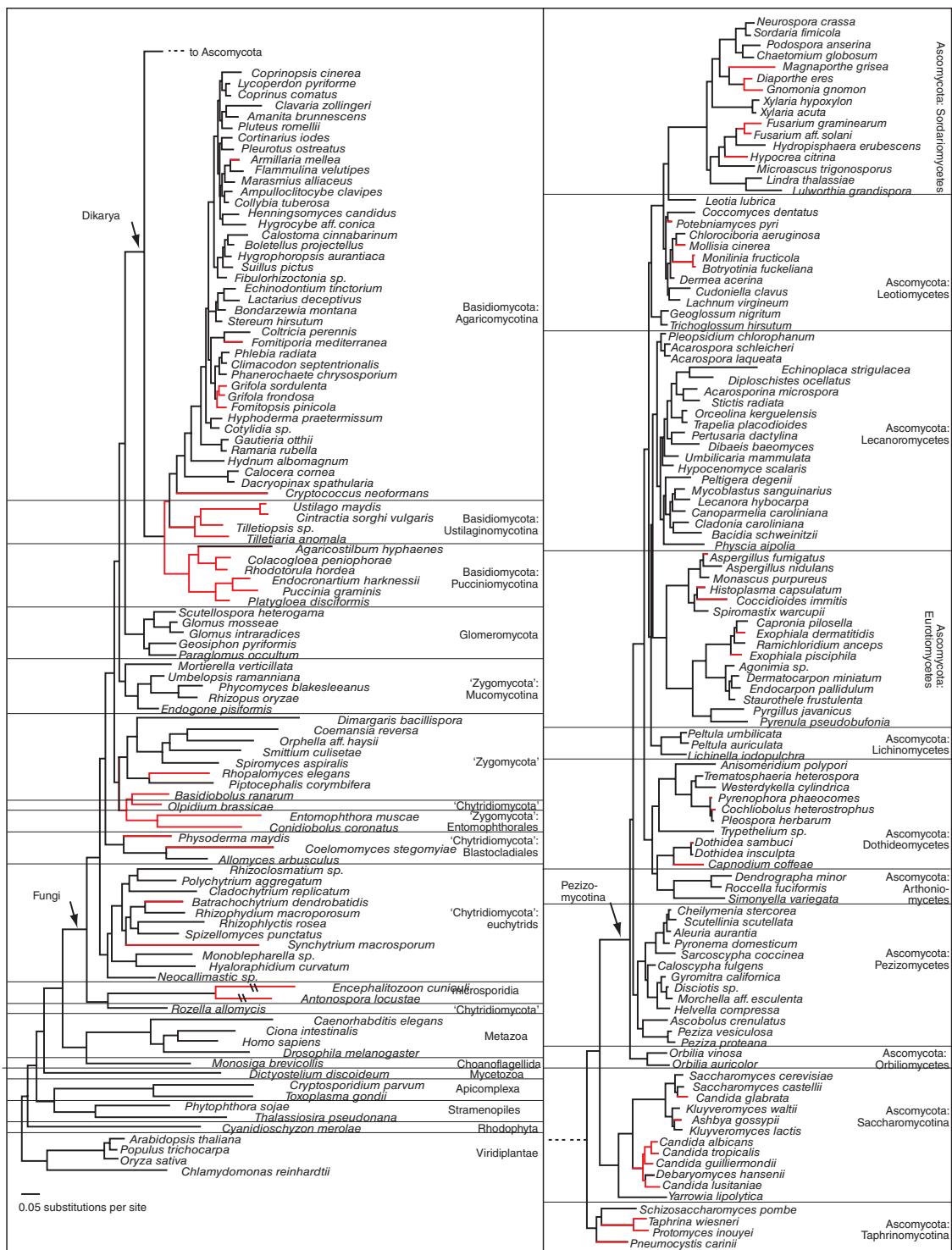
### BOX 18.1 FUNGAL TAXONOMY

Approximately 70,000 fungal species were described in 1995, although the total number of species is estimated to be at least 1.5 million. These fungi were classified in four phyla: Ascomycota, Basidiomycota, Chytridiomycota, and Zygomycota (described and illustrated by Guarro *et al.*, 1999). (1) Ascomycota includes yeasts, blue-green molds, truffles, and lichens; about 30,000 species are known, including the genera *Aspergillus*, *Candida*, *Cryptosporium*, *Histoplasma*, *Neurospora*, and *Saccharomyces*. (2) Basidiomycota includes rusts, smuts, and mushrooms; they are distinguished by club-shaped reproductive structures called basidia. (3) The phylum Chytridiomycota, sometimes classified in the kingdom Protista (Margulis and Schwartz, 1998), includes the genera *Allomyces* and *Polyphagus*. (4) Finally, fungi of the phylum Zygomycota lack septa (cross walls), typically feed on decaying vegetation and include the genera *Glomus*, *Mucor*, and *Rhizopus*.

The phylum Ascomycota is of particular interest because it includes the yeasts. The phylum is further divided into four classes: Hemiascomycetidae (e.g., *S. cerevisiae*), Euascomycetidae (e.g., *Neurospora crassa*), Loculoascomycetidae (e.g., *Elsinoe proteae*), and Laboulbeniomycetidae (parasites of insects).

Hibbett *et al.* (2007), in a paper with 67 authors, proposed a reclassification of the *Fungi* into one kingdom (*Fungi*), one subkingdom (*Dikarya*, encompassing the clade containing *Ascomycota* and *Basidiomycota*), seven phyla, 35 classes, and 129 orders. The seven phyla are the *Chytridiomycota*, *Neocallimastigomycota*, *Blastocladiomycota*, *Microsporidia*, *Glomeromycota*, *Ascomycota*, and *Basidiomycota*.

The *Dikarya* encompass about 98% of all known fungal species. The Hibbett *et al.* classification is consistent with the phylogeny of James *et al.* (2006) who analyzed sequence data from six genes in 199 taxa: 18S rRNA, 28S rRNA, 5.8S rRNA, elongation factor 1- $\alpha$ , and the two RNA polymerase II subunits *RPB1* and *RPB2*.



**FIGURE 18.2** Fungal phylogeny. Nearly 200 fungal species were sampled, and six molecules were analyzed (see Box 18.1). The majority of known fungal species are from the phyla Ascomycota and Basidiomycota of the subkingdom Dikarya. Adapted from James *et al.* (2006) with permission from Macmillan Publishers.

## Sequencing Yeast Genome

Currently, genomes are sequenced using next-generation sequencing (Chapter 9). In contrast, the yeast genome was sequenced in the early to mid-1990s by chromosome. This was accomplished by a worldwide consortium of over 600 researchers (Mewes *et al.*, 1997). The work proceeded in several phases. First, a crude physical map of its 16 chromosomes was constructed using rare-cutter restriction enzymes. Second, a library of ~10-kilobase genomic DNA inserts was constructed in phage lambda, and the inserts were fingerprinted using restriction enzymes. Clones with overlapping inserts were identified and assembled into 16 large contigs. A set of clones covering the genome with minimal overlap was selected and parsed out to individual laboratories for sequencing followed by assembly and annotation using a standardized nomenclature. (The final error rate was less than 3 per 10,000 bases, or 0.03%; Mewes *et al.*, 1997.) Today, this approach would be considered arduous, inefficient, and expensive. However, the collaboration worked extremely well.

*Saccharomyces cerevisiae* is often called a “budding yeast” to distinguish it from a “fission yeast,” *Schizosaccharomyces pombe*, the second fungal genome to be sequenced (see “Fission Yeast *Schizosaccharomyces pombe*” below). *Saccharomyces cerevisiae* is a single-celled organism that “buds” off in the process of replication.

## Features of Budding Yeast Genome

The *S. cerevisiae* genome consists of about 13 Mb of DNA in 16 chromosomes. With the complete sequencing of the genome, the physical map (determined directly from DNA sequencing) was unified with the genetic map (determined by tetrad analysis to derive genetic distances between genes; Cherry *et al.*, 1997). The final sequence was assembled from 300,000 independent sequence reads (Mewes *et al.*, 1997). Some of the features of the *S. cerevisiae* sequence are listed in **Table 18.1**, based on the initial annotation of the genome (Goffeau *et al.*, 1996) as well as recent updates.

In the nearly two decades since the initial sequence analysis, the annotation has been regularly updated as models of genes are corrected and additional information (e.g., based on comparative analyses with other fungal genomes) allows a more accurate assessment of genome features. In 2010 the reference genome sequence of the major strain *S. cerevisiae* strain S288C was updated (and called S288C 2010; Engel *et al.*, 2013).

**TABLE 18.1 Features of *S. cerevisiae* genome. ORF: open reading frame; snoRNA: small nucleolar RNA; tRNA: transfer RNA; Ty: retrotransposons; UTR: untranslated region.**

Feature	Amount
Sequenced length*	12,157,105 base pairs
Length of repeats	1321 kb
Total length	13,389 kb
Total ORFs*	6,607 ORFs
Verified ORFs*	5,072 ORFs
Uncharacterized ORFs*	748 ORFs
Dubious ORFs*	787 ORFs
Introns in ORFs	220 introns
Introns in UTRs	15 introns
Pseudogenes*	19 pseudogenes
Autonomously replicating sequence	337 sequences
Intact Ty elements*	50 elements
tRNA genes*	299 genes
snRNA genes *	6 genes
snoRNA genes*	77 genes
noncoding RNA*	9 genes

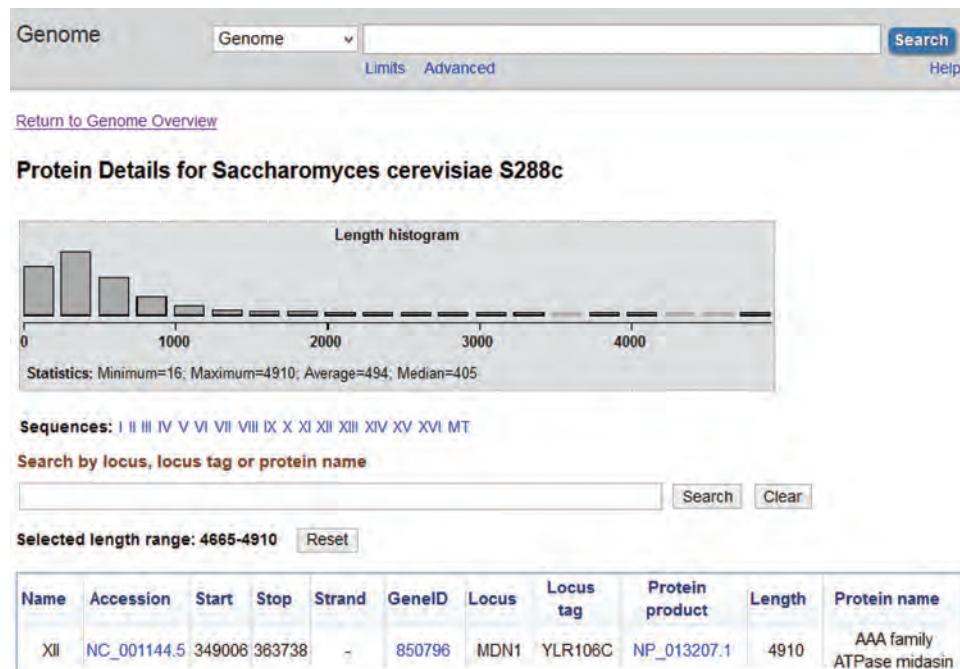
*Source:* Adapted from Goffeau *et al.* (1996) and (\*) Saccharomyces Genome Database, 2014 (<http://www.yeastgenome.org>).

By definition, all ORFs begin with a start codon (typically AUG encoding methionine) and end with a stop codon (usually UAG, UAA, or UGA).

A notable feature of the yeast genome is its high gene density (about one gene every 2 kb). While bacteria have a density of about one gene per kilobase, most higher eukaryotes have a much sparser density of genes. At the time the genomic sequence was initially annotated, there were 6275 predicted open reading frames (ORFs). An ORF was defined as  $\geq 100$  codons (300 nucleotides) in length, thus specifying a protein of at least  $\sim 11,500$  daltons. Of these, 390 were listed as questionable (Table 18.1) because they were short and unlikely to encode proteins (Dujon *et al.*, 1994). Questionable ORFs display an unlikely preference for codon usage based on a “codon adaptation index” of  $< 0.11$ .

Is it possible that short ORFs encode authentic proteins? These questions are fundamental to our understanding of any eukaryotic genome. In annotating the yeast genome, there are false positives (identified ORFs that do not encode an authentic gene) and false negatives (true genes with short ORFs that are not annotated). The *Saccharomyces* Genome Database (introduced below) lists categories of verified ORFs, uncharacterized ORFs, and dubious ORFs. There are 40,000 ORFs longer than 20 codons (Mackiewicz *et al.*, 2002). Below the arbitrary cutoff of 100 codons, there are many ORFs that meet the criteria of having a codon adaptation index of  $> 0.11$  and which do not overlap a longer ORF (Harrison *et al.*, 2002). The main criteria for deciding whether they are protein-coding genes are: (1) evidence of conservation in other organisms; and/or (2) experimental evidence of gene expression and/or expression of the corresponding protein by mass spectrometry.

The NCBI Genome entry for *S. cerevisiae* S288c lists 5906 proteins with a range of 16 to 4910 amino acid residues (average 494, median 405 residues; Fig. 18.3). A total of 69 of these proteins have a length of 16–50 amino acid residues. Are these authentic? Two are ribosomal 60S subunit proteins, and most are hypothetical proteins. For example, YJR151W-A (NP\_878108.1) is a 16 amino acid peptide annotated at NCBI as a “hypothetical protein; identified by fungal homology and RT-PCR; predicted to have



**FIGURE 18.3** Proteins in *S. cerevisiae* 288c. The NCBI Genome entry for this yeast species includes a genome annotation providing a histogram of proteins based on size. By clicking on the right-most portion of the histogram the entry for the largest protein is shown (AAA family ATPase midasin having a length of 4910 residues). In the case of small predicted proteins (e.g.,  $< 100$  codons) it is important to confirm that the gene is transcribed and translated *in vivo* and does not represent a fortuitous open reading frame that is not biologically meaningful.

Source: NCBI Genome, NCBI.

**TABLE 18.2** Ten most common protein domains in *S. cerevisiae* from InterPro.

ID	InterPro name	Number of genes	Number of Ensembl hits
IPR011009	Protein kinase-like domain	130	131
IPR000719	Protein kinase, catalytic domain	117	236
IPR011046	WD40 repeat-like-containing domain	110	116
IPR008271	Serine/threonine-protein kinase, active site	108	108
IPR016024	Armadillo-type fold	104	119
IPR001680	WD40 repeat	100	1038
IPR017441	Protein kinase, ATP binding site	87	87
IPR003593	ATPase, AAA+ type, core	86	120
IPR016196	Major facilitator superfamily domain, general substrate transporter	85	89
IPR017986	WD40-repeat-containing domain	81	89

Source: Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

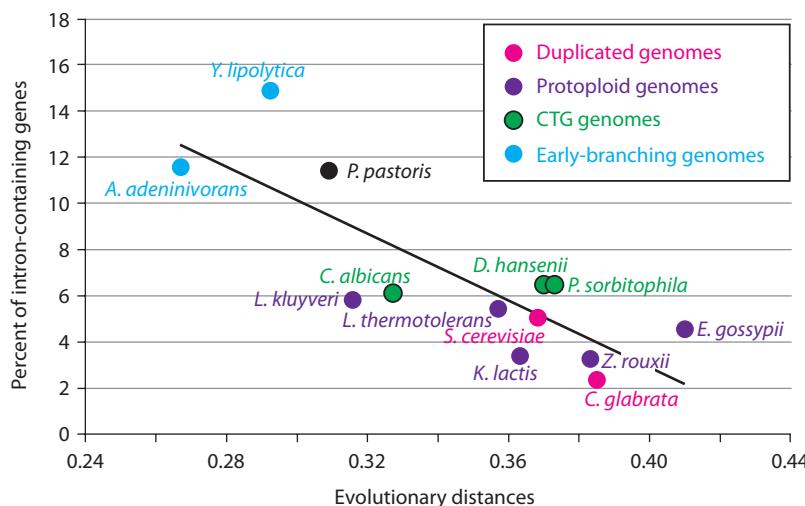
a role in transcription based on computational ‘guilt by association’ analysis.” Strong evidence therefore supports the biological relevance of this and other very small ORFs.

The largest gene in the *S. cerevisiae* genome, *YLR106c*, is assigned to chromosome XII. This gene encodes a protein with 4910 amino acids (Mdn1p; accession NP\_013207.1; Garbarino and Gibbons, 2002). Midasin, a human ortholog, is 5596 amino acids long (over 600 kDa; RefSeq accession NP\_055426.1).

The most common protein families in *S. cerevisiae* are listed in Table 18.2.

Introns are a basic feature of almost all eukaryotic protein-coding genes; in the human genome there are ~8 introns per gene. *S. cerevisiae* offers a stark exception: only 4% of its genes are interrupted by introns. In the fission yeast *S. pombe* (introduced below), 40% of the genes have introns. The lack of introns makes *S. cerevisiae* an attractive model organism for the identification of genes from genomic DNA. Neuvéglise *et al.* (2011) characterized the intron content of 13 hemiascomycetous yeast genomes. They found that the more rapidly evolving species tend to lose their introns (Fig. 18.4). Kelkar and Ochman (2012) analyze intron frequency in terms of genetic drift: in general, genome expansions in fungi are associated with decreases in gene density and increases in intron frequency.

Cécile Neuvéglise introduced Génosplicing, a website describing spliceosomal introns of hemiascomycetous yeasts (<http://genome.jouy.inra.fr/genosplicing/>, WebLink 18.5). According to the SacCer\_ Apr2011-Primary assembly there are 279 RefSeq coding introns and 60 RefSeq noncoding introns (<http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=4932&build=3&ver=1>, WebLink 18.6).



**FIGURE 18.4** *S. cerevisiae* has very few introns. The percent of intron-containing genes (y axis) in 13 yeast genomes is plotted versus evolutionary distances (x axis) based on a phylogenetic analysis using the alignment of 84 proteins. Genomes with the fewest introns in their genes tend to have the largest evolutionary distances (the correlation coefficient  $r^2 = 0.662$ ). Adapted from Neuvéglise *et al.* (2011) with permission from Elsevier and C. Neuvéglise.

In addition to protein-coding genes, there are many transcribed genes that encode functional RNA molecules but are not subsequently translated into protein. In addition to the 299 tRNA genes shown in **Table 18.1**, there are 140 tandemly repeated copies of rRNA genes as well as small nucleolar (snoRNA; Lowe and Eddy, 1999) and other RNA species. In common with the human, mouse, and nematode genomes, the yeast genome is pervasively transcribed, likely controlling gene expression and/or chromatin domain formation (Tisseur *et al.*, 2011; Wu *et al.*, 2012).

Almost half the human genome is composed of transposable elements; we explore them in more detail in Chapter 20.

The *S. cerevisiae* genome encodes 50 intact retrotransposons (called Ty1, Ty2, Ty3, Ty4, and Ty5). These are endogenous retrovirus-like elements that mediate transposition (i.e., insertion into a new genomic location; Roth, 2000). They are flanked by long-terminal repeats (LTRs) that function in integration of the retrotransposon into a new genomic site. Retrotransposons have shaped the genomic landscape of all eukaryotic genomes.

We previously introduced lateral gene transfer (Chapter 17). Extensive lateral gene transfer has occurred from bacteria, plants, and other fungi into fungal genomes (Fitzpatrick, 2012). This includes biotin synthesis pathway genes into *S. cerevisiae*.

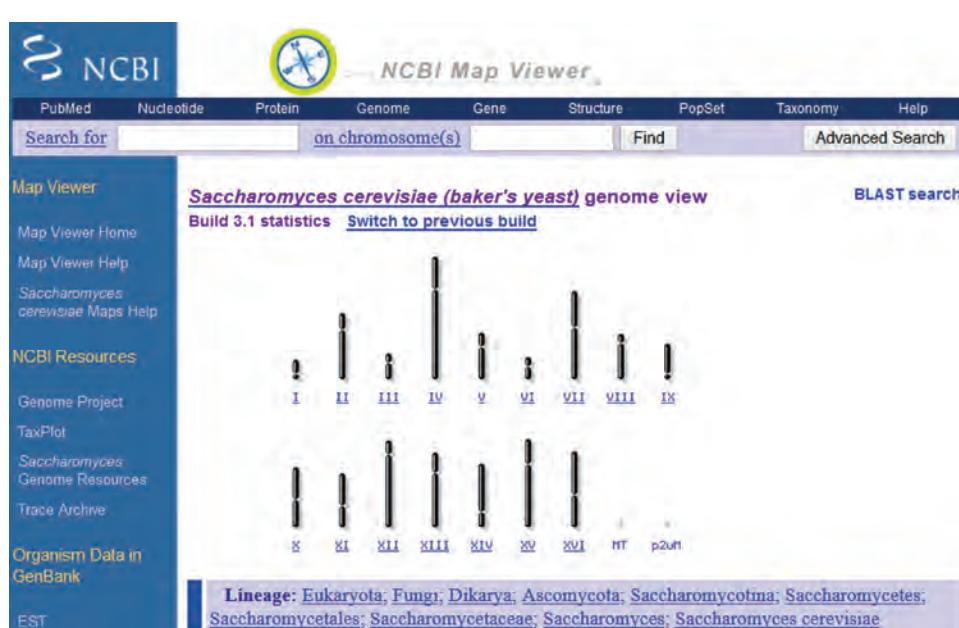
## Exploring Typical Yeast Chromosome

We select chromosome XII (Johnston *et al.*, 1997) to explore the features of a typical yeast chromosome.

### *Web Resources for Analyzing a Chromosome*

You can access the DNA sequence of any *S. cerevisiae* chromosome through several websites, along with assorted annotation. These sites include the following.

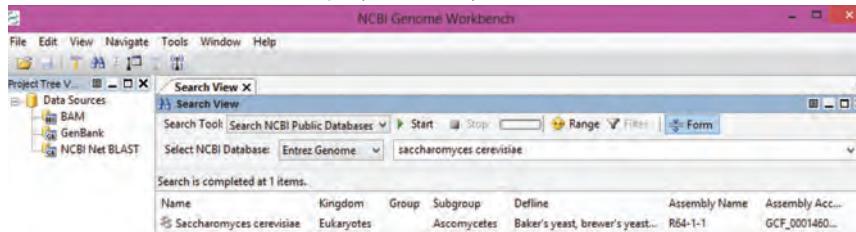
NCBI Genome resources on fungi are at <http://www.ncbi.nlm.nih.gov/genome?term=saccharomyces%20cerevisiae> (WebLink 18.7).



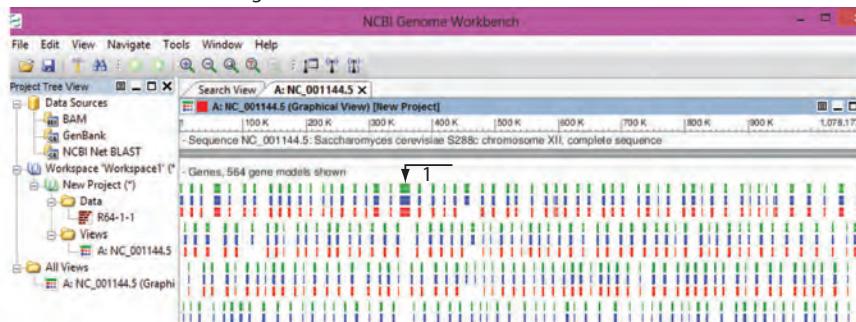
**FIGURE 18.5** The NCBI Map Viewer site includes this page on *S. cerevisiae*. Each of the 16 chromosomes can be explored separately. The left sidebar includes links to resources for *S. cerevisiae* and other fungi.  
Source: NCBI Map Viewer, NCBI.

The NCBI Genome Workbench offers convenient access to genomic data. We can view chromosome XII (Fig. 18.6) or zoom to see a typical gene on that chromosome, *VPS33/YLR396C* encoding Vps33p (Fig. 18.7). (For the nomenclature system used for *S. cerevisiae* genes and proteins, see Box 18.2.) There are links to tools such as BLAST, MUSCLE, genomic aligners, and phylogenetics tools.

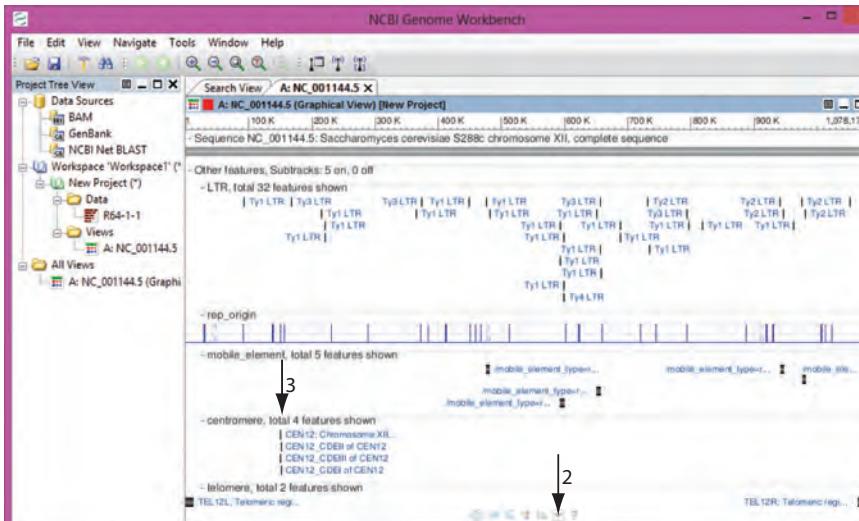
(a) Search view of Genome Workbench: query *Saccharomyces cerevisiae*



(b) Genome Workbench view of genes on chromosome XII

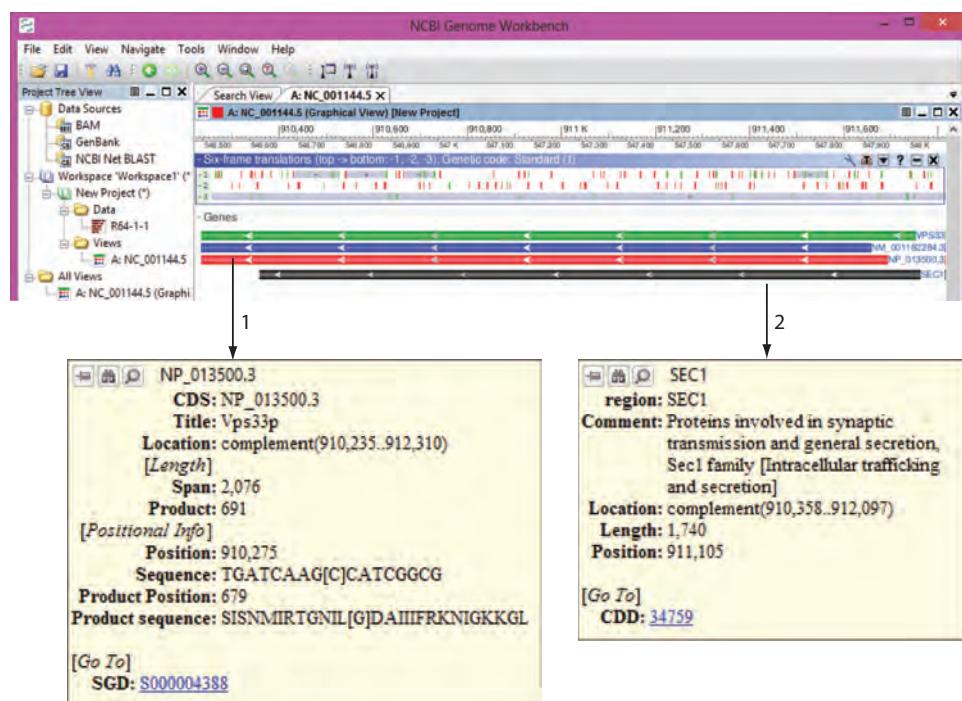


(c) Additional tracks available on graphical view



**FIGURE 18.6** The Genome Workbench at NCBI can be used to analyze sequences of interest. (a) Select “Entrez Genome” in the search view and enter “*saccharomyces cerevisiae*” (without quotes), and press enter to load that genome into a Data subfolder on the left sidebar. (b) The assembly name (R64-1-1) is given to the dataset. Right-click (on a PC) R64-1-1 on the left sidebar and select “open new view.” Choose “graphical view.” There is a list of the 16 chromosomes plus the mitochondrial chromosome; select chrXII (NC\_001144.5). The global view of the chromosome is shown; bars in green, blue, and red correspond to data on the gene, mRNA, and protein. The largest yeast protein, midasin, is evident in the top row of gene models at ~350,000 base pairs (arrow 1). (c) Additional tracks can be shown via a menu (arrow 2). These include long terminal repeats, replication origins, mobile elements, centromeric elements (arrow 3), and telomeres.

Source: Genome Workbench, NCBI.



**FIGURE 18.7** Genome Workbench display for a single gene on chromosome XII (*VPS33/YLR396C*). Tracks for the gene, mRNA, and protein are available. Mousing over the protein track reveals information (arrow 1) including a link to the *Saccharomyces* Genome Database entry. The related gene *SEC1* is shown (arrow 2), and mousing over also shows relevant data including a link to the Conserved Domain Database. *Source:* Genome Workbench, NCBI.

For SGD, visit <http://www.yeastgenome.org/> (WebLink 18.8).

We can visit Ensembl at <http://www.ensembl.org> (WebLink 18.9) then browse to *S. cerevisiae*.

- The *Saccharomyces* Genome Database (SGD) (Engel and Cherry, 2013) is one of the pre-eminent model organism databases and is arguably the most central resource for information about *S. cerevisiae* (see Chapter 14 for more information).
- The UCSC Genome Browser includes a *S. cerevisiae* browser. It includes tracks such as genes, mRNAs and expressed sequence tags, regulatory regions, and conservation.
- Ensembl includes a page for *S. cerevisiae* (Fig. 18.8). This includes the familiar format of the Ensembl browser.

Information in these databases is often cross-referenced; for example, the UCSC tracks include SGD data. Each of these various browsers offers unique feature (Box 18.3).

## BOX 18.2 GENE NOMENCLATURE IN *SACCHAROMYCES CEREVISIAE*

All ORFs that are  $\geq 100$  codons were assigned unique names consisting of three letters followed by a numeral and a subscript to describe its genomic position. For example, the gene name *YKL159c* refers to the ORF number 159 (from the centromere) on the left arm (L) of chromosome XI (K is the 11th letter of the alphabet) of yeast (Y). The designations *c* or *w* ("Crick" or "Watson") reflect the orientation of the gene on the chromosome. Once a gene has been characterized and assigned some kind of function, the investigators may assign a new name that reflects the function (in this case *RCN1* for "regulator of calcineurin"). Dominant alleles (typically the wildtype allele) are listed with three uppercase letters while recessive alleles (typically knockout mutations or loss of function alleles) are listed with three lowercase letters. The protein product of the gene is designated without italics and with only the first letter in uppercase, and with "p" appended to designate protein. Many genes have multiple names (synonyms) because investigators have identified them in independent functional screens. Some examples of nomenclature are as follows.

Wildtype allele	Protein product	Mutant alleles
<i>CNA1</i>	Cna1p	cna1 $\Delta$
<i>RCN1</i>	Rcn1p	rcn1, rcn1::URA3
<i>YKL159c</i>	Ykl159cp	ykl159c

The screenshot shows the Ensembl homepage for *Saccharomyces cerevisiae*. At the top, there's a navigation bar with links for BLAST/BLAST+, BioMart, Tools, Downloads, More, and a search bar. Below the header, a banner for 'What's New in *Saccharomyces cerevisiae* release 74' lists Ensembl 74 main databases, FASTA & GTF dumps, and External reference projection, with a 'More news...' link.

**Genome assembly: EF4**

- More information and statistics
- Download DNA sequence (FASTA)
- Display your data in Ensembl

Other assemblies: EF3 (Ensembl release 64) Go

**Gene annotation**

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download genes, cDNAs, ncRNA, proteins (FASTA)
- Update your old Ensembl IDs

**Pax6 IRS BRC2 DMD ssh Example gene**

**Example transcript**

**Comparative genomics**

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

**Variation**

What can I find? Short sequence variants.

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

**ATCGAGCT ATCCAGCT ATCGAGAT Example variant**

**Regulation**

What can I find? Microarray annotations.

- More about the Ensembl microarray annotation strategy

**FIGURE 18.8** Ensembl includes resources for *S. cerevisiae* (as for many other organisms) including genome assembly data, comparative genomics, regulation, annotation, and variation.

Source: Ensembl Release 73; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

### Exploring Variation in a Chromosome with Command-Line Tools

We now explore the Ensembl resource in more detail. To explore genomic variants on chromosome XII of *S. cerevisiae*, we can use a genome browser. Alternatively, we can explore the variants in a command-line environment. Follow the link to “Download all variants (GVF).” On a Unix platform, we can copy the link location and use the `wget` utility. On a Macintosh we can click to send the variation files to a download directory, and then use the `mkdir` command to create a directory called `yeast` and the `cp` utility to copy them there. There is a README file we can view using `cat README`, telling us that the files include all germline variations in the current Ensembl release. The file format is Genome Variation Format (GVF) (Reese *et al.*, 2010).

Since the variation files are compressed, we can unpack them:

```
$ gunzip Saccharomyces_cerevisiae.gvf.gz
```

The GVF format is described at  
<http://www.sequenceontology.org/gvf.html> (WebLink 18.10). We referred to it in Chapter 9 since (along with VCF files) it serves as input to VAAST.

### BOX 18.3 MULTIPLE YEAST GENOME BROWSERS

Prominent yeast genome browsers include those at NCBI, MIPS, SGD, and UCSC. Each offers different advantages, and there is no single best resource. The SGD is arguably the central web resource for the yeast genomics community. The strength of NCBI is its critical role in the bioinformatics community. The UCSC Genome Browser is an increasingly essential resource for the visualization, annotation, and analysis of vertebrate genomes, although its application to fungi is currently limited. MIPS offers expert curation. Notably, its web browser is based on the Generic Model Organism Database project (GMOD; <http://www.gmod.org/>, O’Connor *et al.*, 2008). GMOD is a set of interconnected applications and databases including the Generic Genome Browser (GBrowse). The research communities involved in a variety of organisms have contributed to GMOD (including the SGD and model organism projects described in Chapter 19 such as FlyBase, WormBase, and TAIR). Recently toolkits have been developed to facilitate the development of model organism websites, including using the Drupal content management system (Papanicolaou and Heckel, 2010; Ficklin *et al.*, 2011; Sanderson *et al.*, 2013).

Here `gunzip` is a utility to unpack a zipped (compressed) file. We can then inspect the file:

```
$ wc -l Saccharomyces_cerevisiae.gvf
263033 Saccharomyces_cerevisiae.gvf
$ less Saccharomyces_cerevisiae.gvf
##gff-version 3
##gvf-version 1.07
##file-date 2013-12-01
##genome-build ensembl EF4
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=4932
##feature-ontology http://song.cvs.sourceforge.net/viewvc/song/ontology/
so.obo?revision=1.283
##data-source Source=ensembl;version=74;url=http://e74.ensembl.org/
Saccharomyces_cerevisiae
##file-version 74
##sequence-region I 1 230218
##sequence-region II 1 813184
... # these dots indicate that we omit a series of additional comment lines
I SGRP SNV 84 84 . + . ID=1;Variant_seq=A;Dbxref=SGRP:s01-84;Reference_seq=G
I SGRP SNV 109 109 . + . ID=2;Variant_seq=C;Dbxref=SGRP:s01-109;Reference_seq=G
I SGRP SNV 111 111 . + . ID=3;Variant_seq=T;Dbxref=SGRP:s01-111;Reference_seq=C
I SGRP SNV 114 114 . + . ID=4;Variant_seq=C;Dbxref=SGRP:s01-114;Reference_seq=T
I SGRP SNV 115 115 . + . ID=5;Variant_seq=G;Dbxref=SGRP:s01-115;Reference_seq=C
```

The `wc -l` result tells us that this file has ~263,000 rows. The `less` command shows us the beginning of the file, with a series of header lines (each beginning with the `##` symbols) at the start of each line. We then see entries listing single-nucleotide variants (SNVs) in the genome. For example, the first variant that is described is an A residue at position 84 of chromosome I where the reference nucleotide is a G residue.

Which variants are assigned to chromosome XII? We can use the `grep` utility to select the rows having the expression “XII”. However this will include entries from chromosomes XII and also XIII, so we can selectively exclude XIII with the `grep -v` command. To learn how to use any utility such as `grep`, simply try a web browser search engine where you will find questions similar to yours and answers from experts. On a Linux platform, be sure to type `man grep` to read the manual.

```
$ grep "XII" Saccharomyces_cerevisiae.gvf | grep -v "XIII" | wc -l
22336
```

There are therefore ~22,000 variants on chromosome XII, and adding the modifier `grep -v "SNV"` confirms that all are single-nucleotide variants. We can end our previous command with `>` to send the output to a text file:

```
$ grep "XII" Saccharomyces_cerevisiae.gvf | grep -v "XIII"
> yeast_chrXII_SNVs.gvf
```

A GVF-formatted file is useful in many ways, including: you can upload a GVF to the UCSC Genome Browser as a custom track; in Chapter 9 we introduced BEDtools which can be used to analyze the relation of the nucleotide variants to a variety of other features; and you can upload a GVF file to Galaxy.

### Finding Genes in a Chromosome with Command-Line Tools

As another example of using command-line tools, we can return to the Ensembl page for *S. cerevisiae* and select “Download genes, cDNAs, ncRNA, proteins (FASTA).” This provides files listing cDNA, peptides, coding sequences (CDS), DNA, or noncoding DNA. Once downloaded you can transfer the files to a directory (such as `yeast`) that you create (with the `mkdir` command) and unpack them (e.g., `gunzip` the file). We focus on

The file `yeast_chrXII_SNVs.gvf` is available as Web Document 18.1 at <http://bioinfbook.org> (WebLink 18.11). You can view it in a text editor such as NotePad (PC) orTextEdit (Mac) or `vim`, `emacs`, or `nano` (Linux).

a file having entries in the FASTA format (note the extension `.fa`) containing noncoding RNAs (ncrna).

```
$ wc -l Saccharomyces_cerevisiae.EF4.74.ncrna.fa
1977 Saccharomyces_cerevisiae.EF4.74.ncrna.fa
```

This tells us that the file has 1977 rows. But how many entries does it have? Each entry is preceded by the `>` symbol, so we can use `grep`.

```
$ grep ">" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | wc -l
413
```

There are therefore 413 entries. We can look at the contents of the file, one page at a time, using `less`:

```
$ less Saccharomyces_cerevisiae.EF4.74.ncrna.fa
```

This tells us that there are several different types of noncoding RNA (rRNA, tRNA, snRNA, snoRNA; for descriptions see Chapter 10). We can determine how many of each type there are in the file. For example:

```
$ grep "biotype:snRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | wc -l
6
$ grep "biotype:tRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | wc -l
299
$ grep "biotype:rRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | wc -l
16
$ grep "biotype:ncRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | wc -l
15
```

We can further restrict our output to chromosome XII entries, count them, send them to a file, view them with `less`, or (as shown next) we can view the first few lines.

```
$ grep "snRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | grep "XII" |
grep -v "XIII" | less
>snR6 sgd:snRNA chromosome:EF4:XII:366235:366346:1 gene:snR6 gene_
biotype:snRNA transcript_biotype:snRNA
$ grep "snoRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | grep "XII" |
grep -v "XIII" | head -3
>snR30 sgd:snoRNA chromosome:EF4:XII:198784:199389:1 gene:snR30 gene_
biotype:snoRNA transcript_biotype:snoRNA
>snR34 sgd:snoRNA chromosome:EF4:XII:899180:899382:1 gene:snR34 gene_
biotype:snoRNA transcript_biotype:snoRNA
>snR44 sgd:snoRNA chromosome:EF4:XII:856710:856920:1 gene:snR44 gene_
biotype:snoRNA transcript_biotype:snoRNA
$ grep "tRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | grep "XII" | grep
-v "XIII" | head -3
>tR(ACG)L sgd:tRNA chromosome:EF4:XII:374355:374427:1 gene:tR(ACG)L gene_
biotype:tRNA transcript_biotype:tRNA
>tL(UAA)L sgd:tRNA chromosome:EF4:XII:962972:963055:-1 gene:tL(UAA)L gene_
biotype:tRNA transcript_biotype:tRNA
>tP(UGG)L sgd:tRNA chromosome:EF4:XII:92548:92650:1 gene:tP(UGG)L gene_
biotype:tRNA transcript_biotype:tRNA
$ grep "rRNA" Saccharomyces_cerevisiae.EF4.74.ncrna.fa | grep "XII" | grep
-v "XIII" | head -3
>RDN25-1 sgd:rRNA chromosome:EF4:XII:451786:455181:-1 gene:RDN25-1 gene_
biotype:rRNA transcript_biotype:rRNA
>RDN18-2 sgd:rRNA chromosome:EF4:XII:465070:466869:-1 gene:RDN18-2 gene_
biotype:rRNA transcript_biotype:rRNA
>RDN5-4 sgd:rRNA chromosome:EF4:XII:482045:482163:1 gene:RDN5-4 gene_
biotype:rRNA transcript_biotype:rRNA
```

Chromosome XII (accession number NC\_001144.5) has 1,078,177 bp.

The centromere is the site at which chromosomes attach to the mitotic or meiotic spindle. In yeast, the centromere divides each chromosome into the left and right arm; in humans, it divides each chromosome into a short (or p) arm and a long (or q) arm.

The telomere is the terminal region of each chromosome arm (Chapter 8). These arms are important in the maintenance of chromosome structure. They have been implicated in processes ranging from aging to intellectual disability.

Such command-line queries are flexible, powerful, and allow you to perform queries with a broad range of other command-line tools.

### *Properties of Yeast Chromosome XII*

We can now turn to some of the properties of Chromosome XII.

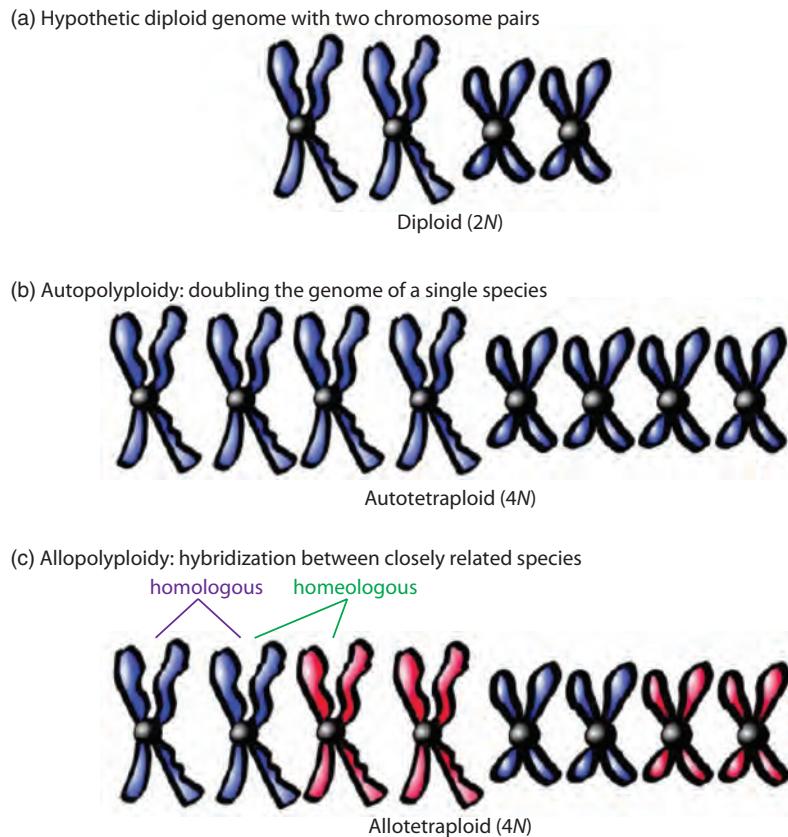
- The overall GC content of chromosome XII is 38%. The GC content tends to be highest in localized regions corresponding to a high density of protein-coding genes. There are three regions of particularly low GC content (below 37%); one of these corresponds to the centromere. This feature is typical of all eukaryotic centromeres. Yeast centromeres in particular contain structural elements called CDEI, CDEII, and CDEIII that are required for assembly, and they are evident in the graphical view of chromosome XII in **Figure 18.6c** (arrow 3) at nucleotide position ~150,000.
- Overall, there is very little repetitive DNA throughout the *S. cerevisiae* genome. The rDNA repeats are all on chromosome XII (encoding rRNAs). This region of the chromosome also has the highest GC content (approximately 42%). In addition, *S. cerevisiae* chromosomes have telomeric and subtelomeric repetitive DNA elements. This feature is typical of essentially all eukaryotic chromosomes.
- There are few spliceosomal introns (~235 total). These are probably due to homologous recombination of cDNAs produced by reverse transcription of spliced mRNAs. On chromosome XII, 17 ORFs (3.2% of the total) contain introns; half of these genes encode ribosomal proteins. The extreme lack of introns contrasts with other fungi such as *Cryptococcus neoformans* (see “*Cryptococcus neoformans*” below) which averages 6.3 exons and 5.3 introns for its 6572 predicted protein-coding genes (Loftus *et al.*, 2005).
- There are six transposable elements (Ty elements) on chromosome XII. Additionally, there are hundreds of fragments of transposable elements.
- The density of ORFs is extremely high. Seventy-two percent of chromosome XII contains protein-coding genes, a fraction that is typical of the other yeast chromosomes. There are 534 ORFs of 100 or more codons on chromosome XII, with an average codon size of 485 codons.

## GENE DUPLICATION AND GENOME DUPLICATION OF *S. CEREVIAE*

As the genome sequence of *S. cerevisiae* was analyzed, it became apparent that there are many duplications of DNA sequence involving both ORFs and larger genomic regions. In many cases, the gene order and orientation (top or bottom strand) is preserved between the duplicated regions. The duplications are both intrachromosomal (occurring within a chromosome) and interchromosomal (occurring between chromosomes).

These changes in genetic material are fundamental in explaining the evolution of species in yeast or in any branch of life. We will see that in the human genome and a variety of other eukaryotic genomes, as many as 25% of the genes are duplicated (Chapters 19 and 20). There are two compelling questions (Conant and Wolfe, 2008): by what mechanisms does duplication occur, and how does selection optimize the novelty of newly duplicated DNA? We first address the origin of new, duplicate genes. There are two main mechanisms.

1. *Segments of a genome can duplicate.* We discuss segmental duplication of the human genome in Chapter 20; it is sometimes defined as consisting of two loci sharing 90% or more identity over a length of 1000 base pairs or more. About 5% of the human genome is segmentally duplicated.
2. *An entire genome can duplicate,* a process called polyploidy (**Fig. 18.9**; Hufton and Panopoulou, 2009). In the case of *S. cerevisiae*, this is a tetraploidization. If this



**FIGURE 18.9** Whole-genome duplication. (a) A hypothetical diploid genome has two chromosome pairs (large, small). (b) Genome duplication within an organism generates an autotetraploid. (c) Hybridization between two closely related species generates an allotetraploid, preserving the full genome content of both parent species. Redrawn from Hufton and Panopoulou (2009) with permission from Elsevier.

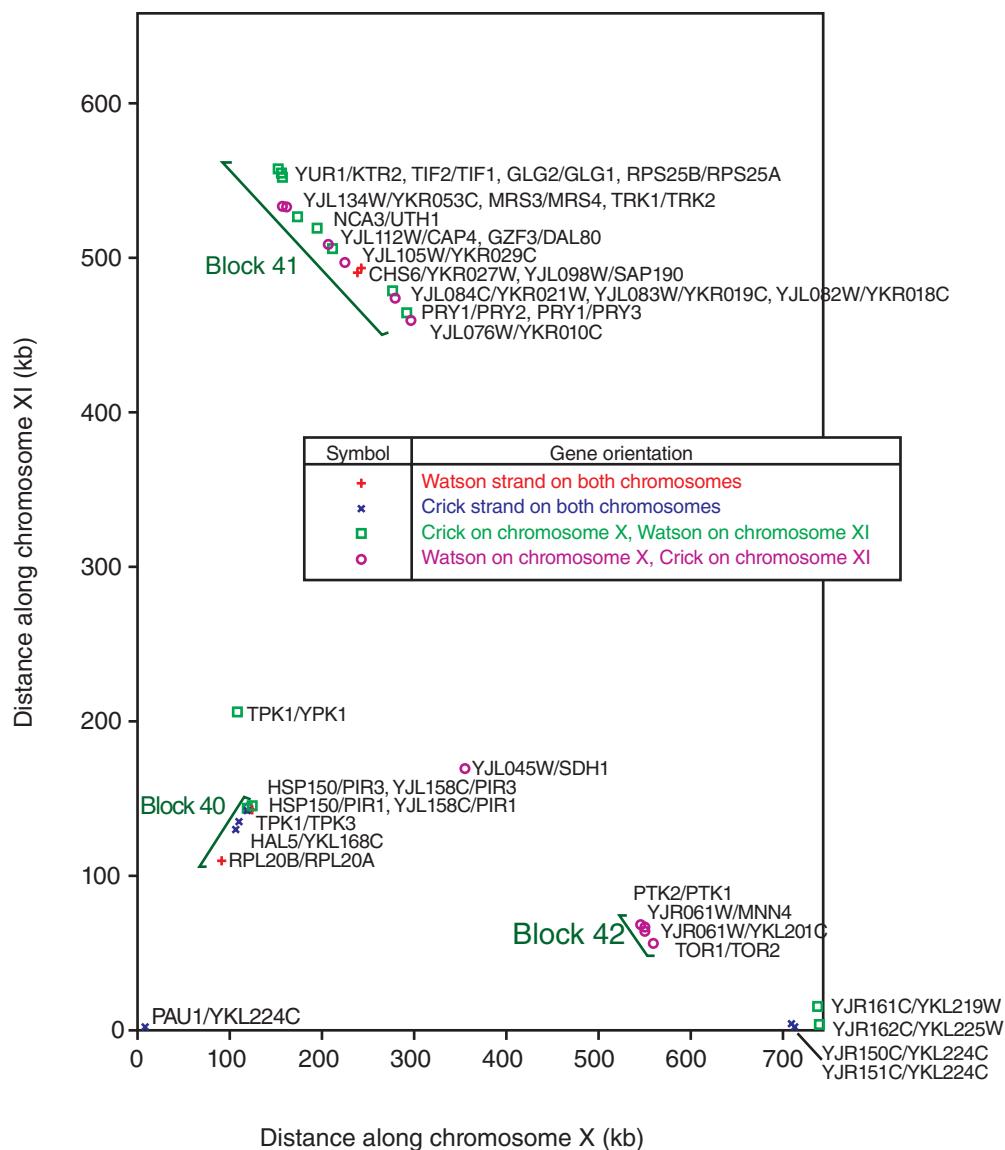
resulted from the combining of two genomes from one species it is called autopolyploidy; if two distinct species fuse it is allopolyploidy.

Other mechanisms of creating new genes may be important, but occur less commonly and are therefore less relevant. New genes can arise by gene conversion. In this process, genes are transferred nonreciprocally from one genomic region to another. (This occurs between repetitive regions of the human Y chromosome; Rozen *et al.*, 2003.) Genes can be introduced into a genome by lateral (horizontal) gene transfer, as mentioned above (Fitzpatrick, 2012). We also describe lateral gene transfer for *Encephalitozoon* (see “Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*” below). In eukaryotes lateral gene transfer may introduce some functionally important genes, but it does not account for a large quantity of new genes.

In 1970, Susumu Ohno published the brilliant book *Evolution by Gene Duplication*. He proposed that vertebrate genomes evolved by two rounds of whole-genome duplication. According to this hypothesis, these duplication events occurred early in vertebrate evolution and allowed the development of a variety of cellular functions. Ohno (1970) wrote:

Had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged. The creation of metazoans, vertebrates, and finally mammals from unicellular organisms would have been quite impossible, for such big leaps in evolution required the creation of new gene loci with previously nonexistent function. Only the cistron that became redundant was able to escape from the relentless pressure of natural selection. By escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus.

Tetraploidy is the presence of four haploid sets of chromosomes in the nucleus.



**FIGURE 18.10** Wolfe and Shields (1997) performed BLASTP searches of proteins from *S. cerevisiae* and found 55 blocks of duplicate regions, providing strong evidence that the entire genome underwent an ancient duplication. This figure (redrawn from the original) depicts the result of BLASTP searches of proteins encoded by genes on chromosomes X and XI. Matches with scores >200 are shown, arranged in several blocks of genes. Redrawn from Wolfe and Shields (1997). Reproduced with permission from Macmillan Publishers.

Which mechanism of gene duplication might have occurred in *S. cerevisiae*? Smith (1987) examined duplicate histone genes (histones H3-H4 and H2A-H2B) and suggested the possibility of an early whole-genome duplication event. Soon after the complete sequence of the genome became available, Wolfe and Shields (1997) provided strong support for Ohno's whole-genome duplication paradigm. They assessed the duplicated regions of the yeast genome by performing systematic BLASTP searches of all yeast proteins against each other and plotting the matches on dot matrices. Duplicate regions were observed as diagonal lines, such as the three duplicated regions seen in a comparison of proteins derived from chromosomes X and XI (Fig. 18.10). In the whole genome, they identified 55 duplicated regions and 376 pairs of homologous genes. In subsequent studies, they employed the more sensitive Smith-Waterman algorithm and identified a few

additional regions of duplication (Seoighe and Wolfe, 1999). Based on these results, they proposed a single, ancient duplication of the *S. cerevisiae* genome, approximately 100 million years ago (Wolfe and Shields, 1997). Subsequent to this duplication event, many duplicated genes were deleted. Other genes were rearranged by reciprocal translocation.

There are two main explanations for the presence of so many duplicated regions. There could have been whole-genome duplication (tetraploidy) followed by translocations as well as gene loss, or alternatively there could have been a series of independent duplications. Wolfe and Shields (1997) favored the tetraploidy model for two reasons:

1. For 50 of the 55 duplicate regions, the orientation of the entire block was preserved with respect to the centromere. If each block were generated independently, a random orientation would be expected.
2. Fifty-five successive, independent duplications of blocks would be expected to result in about seven triplicated regions, but only zero (or possibly one) such triplicated region was observed.

In sum, polyploidy has occurred in *S. cerevisiae*, as well as other fungi (Dujon, 2010; Kelkar and Ochman, 2012; Albertin and Marullo, 2012). Whole-genome duplication also occurred in many eukaryotes from plants to fish to protozoans (Chapter 19).

What is the fate of genes after duplication? The presence of extra copies of genes is usually deleterious to an organism. In the model of Wolfe and colleagues, the genome of an ancestral yeast doubled (from the diploid number of about 5000 to the tetraploid number of 10,000 genes) then lost the majority of its duplicated genes, yielding the present-day number of about 6200 ORFs. Overall, between 50 and 92% of duplicated genes are eventually lost (Wagner, 2001). For eukaryotes, the half-life of duplicated genes is only a few million years (Lynch and Conery, 2000). There are four main possibilities:

1. Both copies can persist, maintaining the function of the original gene. In the scenario of a local duplication, there is a gene dosage effect because of the extra copy of the gene. In whole-genome duplication, the stoichiometry of the genes (and gene products) may be maintained as in the original state.
2. One copy could be completely deleted. This is the most common fate of duplicated genes, as confirmed by recent whole-genome studies (described in “Gene Duplication and Genome Duplication of *S. cerevisiae*” below). A rationale for this fate is that, since the duplicated genes share identical functions initially, either one of them may be subject to loss-of-function mutations (Wagner, 2001).
3. One copy can accumulate mutations and evolve into a pseudogene (a gene that does not encode a functional gene product). This represents a loss of gene function, although it occurs without the complete deletion of the duplicate copy. Over time, the pseudogene may be lost entirely.
4. One or both copies of the gene could diverge functionally. According to this hypothesis, gene duplications (regardless of mechanism) provide an organism with the raw material needed to expand its repertoire of functions. Furthermore, loss of either gene having overlapping functions might not be tolerated. The functionally diverged genes would therefore both be positively selected.

After a gene duplicates, why does one of the members of the newly formed gene pair often become inactivated? At first glance, it might seem highly advantageous to have two copies because one may functionally diverge (driving the process of evolution to allow a cell to perform new functions) or one may be present in an extra copy in case the other undergoes mutation. However, gene duplication instead appears to be generally deleterious, leading to the loss of duplicated genes. The logic is that some mutations in a gene are *forbidden* rather than *tolerable* (these terms were used by Ohno (1970) in describing gene duplication). Forbidden mutations severely affect the function of a gene product, for

Wolfe and Shields (1997) used BLASTP rather than BLASTN to study duplicated regions of chromosomes. This is because protein sequence data are more informative than DNA for the detection of distantly related sequences. See Chapter 3.

In humans, an extra copy of chromosome 21 (i.e., trisomy 21) causes Down syndrome. Trisomies 13 and 18 are also sometimes compatible with life, but other autosomal trisomies are not. Duplications of even limited regions of the genome cause intellectual disability and other diseases (see Chapter 21 on human disease). This highlights the deleterious nature of duplications at the level of individual organisms.

instance by altering the properties of the active site of an enzyme. (A tolerable mutation causes a change that remains compatible with the function of the gene product.) Natural selection can eliminate forbidden mutations, because the individual is less fit to reproduce. After a gene duplicates, a deleterious mutation in one copy of a gene might then be tolerated because the second gene can assume its function. A second reason that duplicated genes may be deleterious is that, in their presence, the crossing-over of homologous chromosomes during meiosis may be mismatched, causing unequal crossing-over.

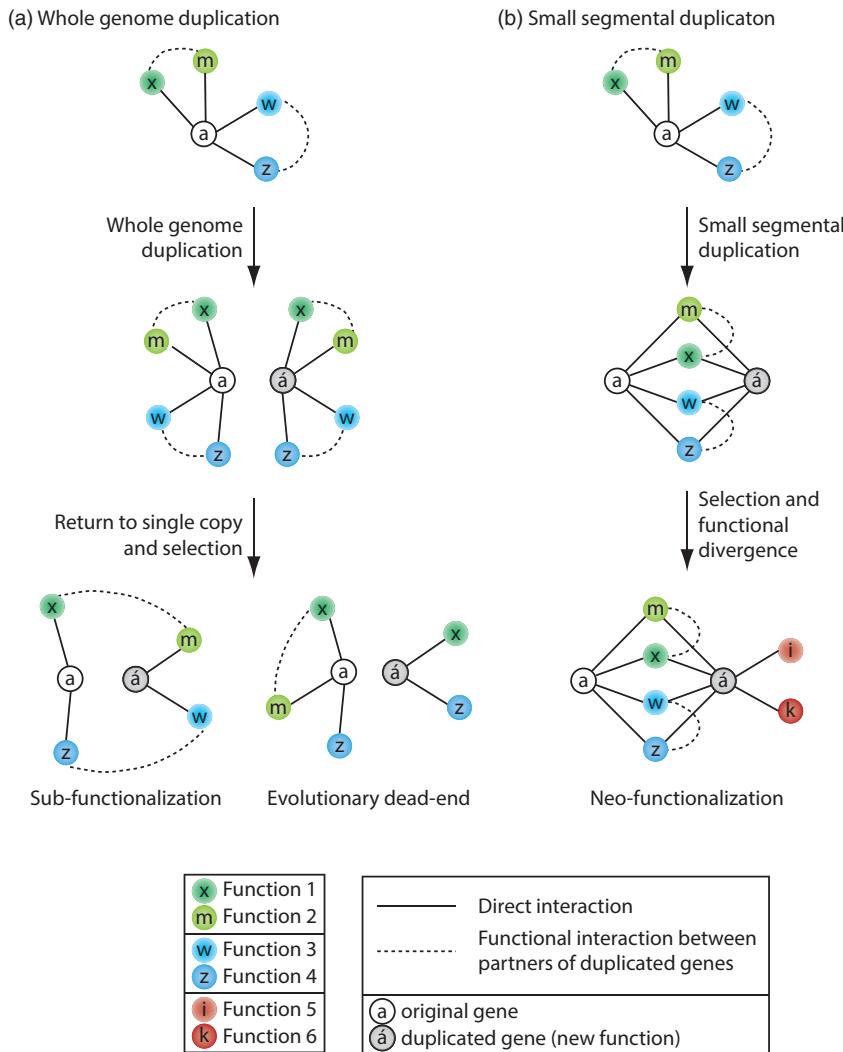
We can consider the possible fates of duplicated genes with the specific example of genes encoding proteins that are essential for vesicle trafficking. We introduced *SSO1* in Chapter 14 (Fig. 14.4). In yeast and all other eukaryotes, spherical intracellular vesicles transport various cargo to destinations within the cell. These vesicles traffic cargo to the appropriate target membrane through the binding of vesicle proteins (e.g., Snc1p in yeast or VAMP/synaptobrevin in mammals) to target membrane proteins (e.g., Sso1p in yeast or syntaxin in mammals) (Aalto *et al.*, 1993; Protopopov *et al.*, 1993). In *S. cerevisiae*, genome duplications presumably caused the appearance of two paralogous genes in each case: *SNC1* and *SNC2* as well as *SSO1* and *SSO2*. The *SNC1* and *SNC2* genes are on corresponding regions of chromosomes I and XV, while the *SSO1* and *SSO2* genes are on chromosomes XVI and XIII, respectively.

What could the consequences of genome duplication have been? The two pairs of syntaxin-like and VAMP/synaptobrevin-like yeast proteins might have maintained the same function of the original proteins (before genome duplication). A search for *SSO1* at the SGD website shows that the gene is nonessential (the null mutant is viable), but the double knockout is lethal (see Fig. 14.4). It is therefore likely that these paralogs offer functional redundancy for the organism; in the event a gene is lost (e.g., through mutation), the organism can survive because of the presence of the other gene. Similarly, the *SNC1* null mutant is viable, but the double knockout of *SNC1* and *SNC2* is deficient in secretion.

As an alternative explanation of the duplication of these genes, it is possible that whole-genome duplication provided the new genetic materials with which the intracellular secretion machinery could be diversified. Syntaxin and VAMP/synaptobrevin proteins function at a variety of intracellular trafficking steps, and these gene families diversified throughout eukaryotic evolution (Dacks and Doolittle, 2002). Particular combinations of these proteins interact to confer specificity to vesicular trafficking events (Pevsner and Scheller, 1994).

There are several models for the consequences of duplication. Andreas Wagner (2000) addressed the question of how *S. cerevisiae* protects itself against mutations by one of two mechanisms: (1) having genes with overlapping functions (such as paralogs that maintain related functions); or (2) through the interactions of nonhomologous genes in regulatory networks. He found that genes whose loss of function caused mild rather than severe effects on fitness did not tend to have closely related paralogs. This is consistent with a model in which gene duplication does not provide robustness against mutations.

The fate of gene duplicates is likely to differ between those generated by whole-genome duplication versus small-scale duplication (Conant and Wolfe, 2008). Fares *et al.* (2013) present a model for the emergence of new functions by whole-genome duplication versus small-scale duplications (Fig. 18.11). After whole-genome duplication, duplicated genes are in dosage balance and tend to maintain their functions. An ultimate outcome may be sub-functionalization in which each of the two gene copies performs a subset of the ancestral (pre-duplication) gene function. Both gene copies therefore undergo comparable selective pressure. In contrast, small-scale duplications (involving one or a few genes) can produce genetic robustness in which there is selective pressure to maintain both copies, and one copy may diverge to acquire new functions (neo-functionalization).



**FIGURE 18.11** Model of evolution after gene duplication by (a) whole-genome duplication (WGD) or (b) small-scale duplication. After WGD, the stoichiometry of gene products is maintained, and duplicated genes maintain their genetic interactions and their functions. Interaction partners (solid lines) are maintained, and partners of duplicated genes interact functionally (dashed lines). There are relaxed selective constraints on the duplicated genes, so one may be lost. Redrawn from Fares *et al.* (2013). Licensed under the Creative Commons Attribution License 3.0.

## COMPARATIVE ANALYSES OF HEMIASCOMYCETES

Analysis of *S. cerevisiae* has elucidated many fundamental principles concerning genome structure, function, and evolution. Comparison of phylogenetically related genomes has opened an entirely new dimension on genome analysis. Some of the first genomes selected were hemiascomycetes phylogenetically close to *S. cerevisiae* such as *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, and *Yarrowia lipolytica* (Dujon *et al.*, 2004; Souciet, 2011). In all, hundreds of fungal genomes are currently being sequenced, facilitating comparative genomics. In parallel to this, population genomics studies allow the measurement of genetic diversity within species such as *S. cerevisiae* and *S. paradoxus* (Liti and Schacherer, 2011) or within a genus such as *Saccharomyces* (Dequin and Casaregola, 2011; Hittinger, 2013).

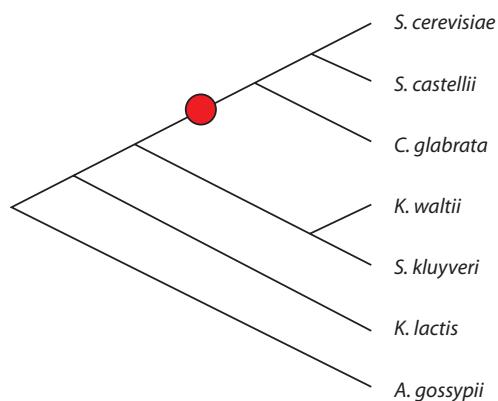
## Comparative Analyses of Whole-Genome Duplication

The hypothesis that yeast underwent a whole-genome duplication event has been tested by analyzing whole-genome sequences. By becoming polyploid, an organism doubles its complement of chromosomes (and therefore genes). This might appear to be an appealing mechanism to increase the repertoire of genes available for adaptation to new environments. However, polyploidy leads to genome instability, partly because of difficulties for the cell to perform proper chromosome segregation.

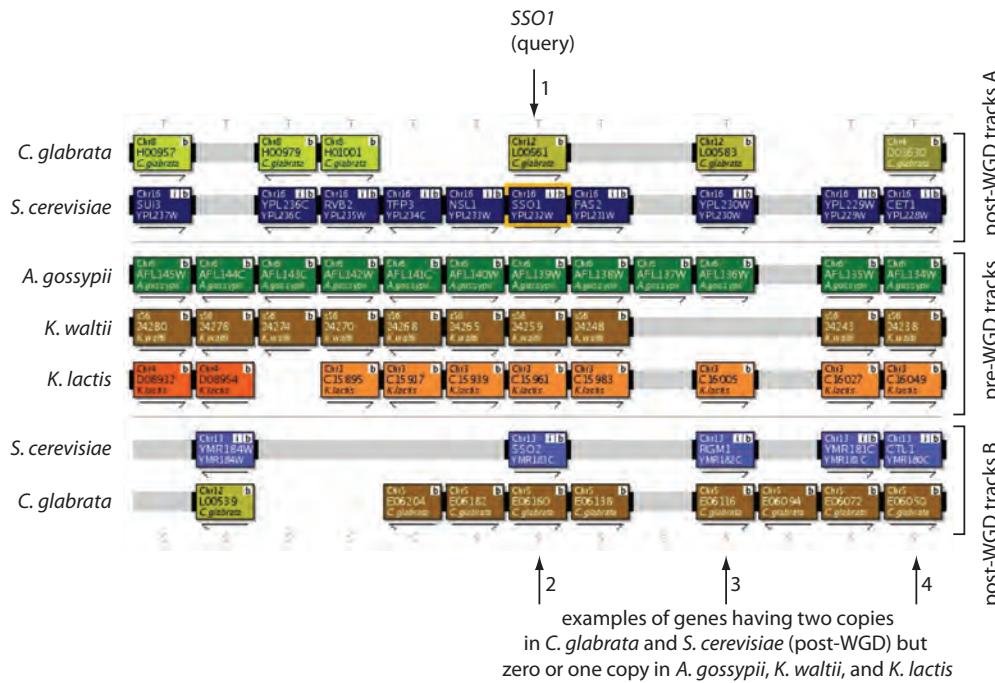
To understand the whole-genome duplication of *S. cerevisiae*, Kellis *et al.* (2003) sequenced the genome of *Kluyveromyces waltii*, a related yeast that diverged before the whole-genome duplication event (Fig. 18.12). They sequenced the eight chromosomes of *K. waltii*, and annotated 5230 putative protein-coding genes. They identified blocks of conserved synteny (loci containing orthologous genes in the same order between the two species). Most regions of *K. waltii* mapped to two separate regions of *S. cerevisiae*. However, these regions of *S. cerevisiae* show evidence of massive gene loss (with 12% of the paralogous gene pairs retained, and 88% of paralogous genes deleted to leave one copy remaining).

Kellis *et al.* considered the rate of evolution of 457 gene pairs in *S. cerevisiae* that arose by whole-genome duplication. Seventy-six of these gene pairs displayed accelerated evolution (based on amino acid substitution rates in the *S. cerevisiae* lineage relative to *K. waltii*). Remarkably, in 95% of these cases, the accelerated evolution was restricted to just one of the two paralogs. This supports Ohno's suggestion that, after duplication, one copy of a gene can preserve the original function while the other may diverge to acquire a novel function.

With the continuing production of new genome-sequencing data, Scannell *et al.* (2006) considered six yeast species: three that descended from a common ancestor that is thought to have undergone a whole-genome duplication (*S. cerevisiae*, *Saccharomyces castellii* and *Candida glabrata*), as well as three additional yeasts that diverged before the whole-genome duplication event (*Kluyveromyces waltii*, *Kluyveromyces lactis*, and *Ashbya gossypii*). They used the Yeast Gene Order Browser to compare the six species. This browser is available online (Byrne and Wolfe, 2005, 2006). An example is shown in Figure 18.13 for the query *SSO1* as well as six adjacent upstream and downstream genes. There are seven horizontal tracks in this example. Three in the center show the genes in the reference species that diverged before the whole-genome duplication event



**FIGURE 18.12** Phylogeny of several yeasts after Kurtzman and Robnett (2003). A red circle indicates the likely place at which a whole-genome duplication (WGD) occurred. Adapted from <http://wolfe.gen.tcd.ie/ygob/> with permission from K. H. Wolfe.



**FIGURE 18.13** The Yeast Gene Order Browser of Kevin Byrne (in the group of Kenneth Wolfe) provides evidence supporting whole-genome duplication events. Upon entering the query (*SSO1*; top arrow) and selecting the species to display, the query and varying numbers of adjacent genes are displayed. Each box represents a gene, and boxes are color-coded to correspond to particular chromosomes. Solid bars connect genes that are immediately adjacent. Here, the first and seventh rows correspond to *C. glabrata*, and the second and sixth rows correspond to *S. cerevisiae* (chromosome 16, including *SSO1* gene, on row 2; chromosome 13, including the paralog *SSO2*, on row six). In this view there are three genes that have two copies in *C. glabrata* and *S. cerevisiae* that may have resulted from whole-genome duplication. For yeast lineages that are hypothesized to have not undergone whole-genome duplication (*A. gossypii*, *K. waltii*, and *K. lactis*) there tends to be only one copy of these genes. For all species, occasional gene losses are evident (e.g., *K. waltii*, the gene indicated by arrow 3). Yeast Gene Order Browser includes additional features such as links to the raw sequences and to phylogenetic reconstructions of each gene family. Adapted from <http://wolfe.gen.tcd.ie/ygob/> with permission from K.H. Wolfe.

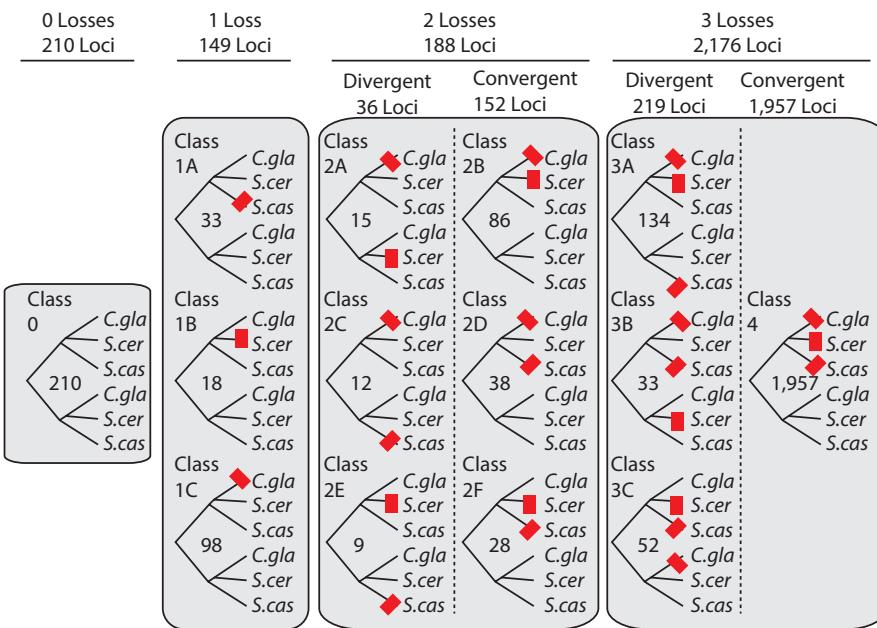
(*A. gossypii*, *K. waltii*, and *K. lactis*). For *S. cerevisiae* and *C. glabrata* there are pairs of tracks, both above and below the reference species. For genes such as *SSO1*, *YPL230W*, and *WPL228W* (Fig. 18.13, arrows 2–4) there are two copies in both *S. cerevisiae* and *C. glabrata* but only one copy in the reference genomes. These two copies occur in adjacent positions along separate chromosomes.

A variety of patterns of loss can occur. Scannell *et al.* (2006) described 14 patterns by which gene loss can occur (outlined in Fig. 18.14). Out of 2723 ancestral loci that aligned appropriately, in only 210 cases was there no gene loss among the three genomes that underwent whole-genome duplication. In the great majority of cases (1957 instances or 72% of the total), all three species lost one of the two copies of a given duplicated gene, and most commonly all three species lost the same copy of the gene. Genes involved in highly conserved biological processes such as ribosome function were especially likely to experience gene loss.

Wolfe and colleagues have extended YGOB to discover previously unannotated genes in various yeast species (ÓÉigearaigh *et al.*, 2011). They also developed an automated Yeast Genome Annotation Pipeline that relies on YGOB (Proux-Wéra *et al.*, 2012).

The Yeast Gene Order Browser is online at the website of Kenneth Wolfe, <http://wolfe.gen.tcd.ie/ygob/> (WebLink 18.12).

*Saccharomyces cerevisiae* can live under anaerobic conditions, while *K. lactis* cannot. It is possible that the *S. cerevisiae* genome duplication resulted in physiological changes that allowed this organism to acquire the new growth phenotype (Piskur, 2001).



**FIGURE 18.14** Patterns of gene loss after whole-genome duplication in three species. For three species that underwent whole-genome duplication (*C. glabrata*, *S. cerevisiae*, and *S. castellii*) there are 14 possible fates including loss of no genes (class 0), loss of one gene from any one of the three lineages (class 1A, 1B, 1C), loss of two genes (class 2), loss of three genes from different loci (class 3), or loss of three genes in a convergent manner (class 4; loss of duplicated orthologs). Class 4 represents the most common fate of duplicated genes. Redrawn from Scannell *et al.* (2006). Reproduced with permission from Macmillan Publishers.

## Identification of Functional Elements

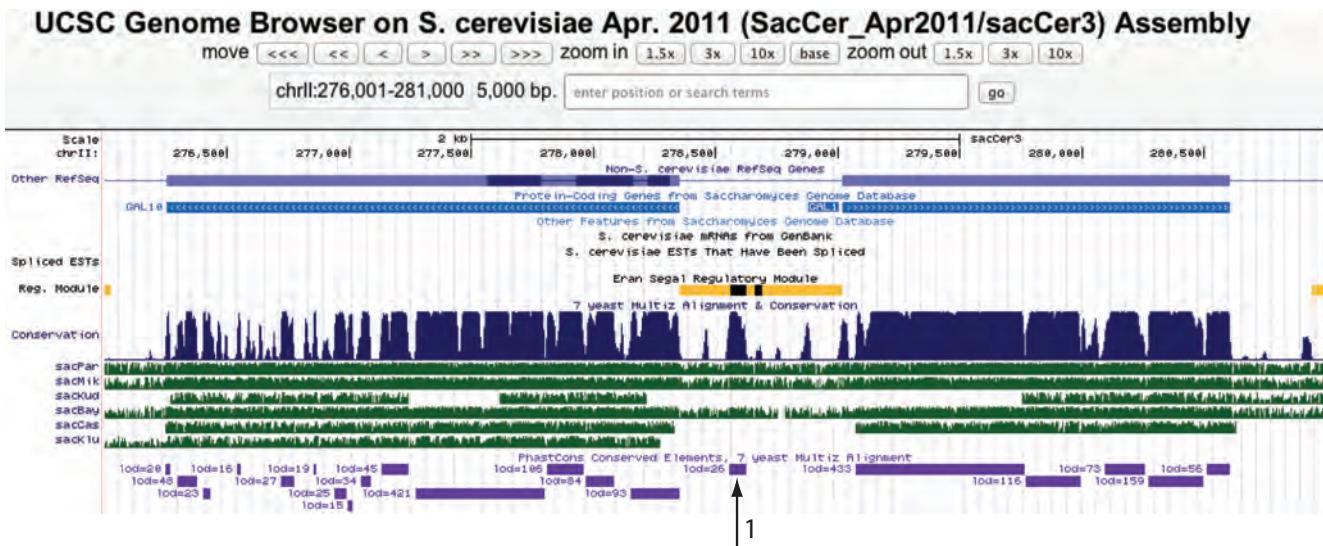
It is extraordinarily difficult to identify genes and gene regulatory regions (such as promoters) from genomic sequence data alone. Matching expressed sequence tags (ESTs; Chapter 10) to genomic DNA is one useful approach to defining protein-coding genes. Comparative analyses between genomic sequences also provide a powerful approach to identifying functionally important elements.

Kellis *et al.* (2003) obtained the draft sequences of *Saccharomyces paradoxus*, *S. mikatae*, and *S. bayanus* which diverged from *S. cerevisiae* some 5–20 million years ago. Almost all of the 6235 open reading frames (ORFs) in the SGD annotation of *S. cerevisiae* had clear orthologous matches in each of the other three species. A noticeable exception is at all 32 telomeres (i.e., both ends of the 16 chromosomes), where matches are often ambiguous. Genes assigned to subtelomeric regions are often present in different number, order, or orientation, and these regions have undergone multiple reciprocal translocations. Kellis *et al.* refer to changes in the telomeric regions as “genomic churning.” For all ORFs in the four *Saccharomyces* genomes, Kellis *et al.* introduced a reading frame conservation test to classify each ORF as authentic (if conserved) or spurious (if not well conserved). As a result of their analysis, Kellis *et al.* proposed revising the entire *S. cerevisiae* gene catalog to 5538 ORFs of  $\geq 100$  amino acids. Their analyses further revised the count of introns (predicting 58 new introns beyond the 240 previously predicted).

Another aspect of the comparison of four *Saccharomyces* genome sequences is the opportunity to identify regulatory elements. Gal4 is one of the best-characterized transcription factors. It regulates genes involved in galactose metabolism including the *GAL1* and *GAL10* genes. These two genes can be viewed at the UCSC Genome Browser (Fig. 18.15). They are transcribed from a short intergenic region that includes the Gal4 binding

*S. cerevisiae* and *S. bayanus* share 62% nucleotide identity in conserved regions; for comparison, human and mouse share 66% nucleotide identity in conserved regions.

Dramatic genomic changes that occur in subtelomeric regions have also been observed in the malaria parasite *Plasmodium falciparum* (see Chapter 19), and in humans subtelomeric deletions are a major cause of intellectual disability.



**FIGURE 18.15** View of the transcription factor Gal4 binding site region between the *GAL1* and *GAL10* genes of *S. cerevisiae*. A 5000-base-pair view of yeast chromosome 2 is shown (from the UCSC Genome Browser coordinates chrII:276,001–281,000). The short intergenic region (arrow 1) includes regions defined as having regulatory properties by several databases (see annotation tracks). The conservation track shows that some of the intergenic region is highly conserved among four *Saccharomyces* species, and contains binding sites for Gal4.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

motif CGGn<sub>(11)</sub>CCG where n<sub>(11)</sub> refers to any 11 nucleotides. By clicking the conservation track, you can see several copies of this motif in a multiple sequence alignment of DNA from the four *Saccharomyces* species. Kellis *et al.* (2003, 2004) studied both previously known and predicted motifs, and predicted 52 new motifs. Other groups such as Cliften *et al.* (2003) and Harbison *et al.* (2004) also identified yeast functional elements through comparative genomics.

## ANALYSIS OF FUNGAL GENOMES

In addition to *S. cerevisiae*, the genomes of many other fungi are now being sequenced including *Ascomycetes* (Table 18.3), *Basidiomycetes* (Table 18.4), and others (Table 18.5).

**TABLE 18.3** Fungal genome projects: representative examples of the Ascomycetes. ID refers to the NCBI Genome Project identifier; by entering this into the search box at the home page of NCBI you can link to information on this genome project.

Organism	Chromosomes	Genome		Comment	ID	GC %
		size (Mb)				
<i>Ajellomyces capsulatus</i> G186AR	7	30.5		Causes histoplasmosis, an infection of the lungs	12635	44.5
<i>Aspergillus fumigatus</i> Af293	8	29.4		Most frequent fungal infection worldwide	131	49.8
<i>Candida albicans</i> SC5314	8	27.6		Diploid fungal pathogen	10701	33.4
<i>Coccidioides immitis</i> RS	4	29.0		Causes the disease coccidioidomycosis (valley fever)	12883	46
<i>Kluyveromyces lactis</i> NRRL Y-1140	6	10.7		Related to <i>S. cerevisiae</i>	13835	38.7
<i>Magnaporthe grisea</i> 70-15	7	41.0		Rice blast fungus	13840	51.6
<i>Pneumocystis carinii</i>	15	7.7		Opportunistic pathogen; causes pneumonia in rats	125	31.1
<i>Saccharomyces cerevisiae</i> S288c	16	12.2		Baker's yeast	13838	38.2
<i>Schizosaccharomyces pombe</i> 972h-	3	12.6		Fission yeast	13836	36
<i>Yarrowia lipolytica</i> CLIB122	6	20.6		Nonpathogenic yeast, distantly related to other yeasts	13837	49

Source: NCBI Genome, NCBI.

**TABLE 18.4 Fungal genome projects: representative examples of the Basidiomycetes.**

Organism	Chromosomes	Genome size (Mb)	Comment	ID	GC %
<i>Coprinopsis cinerea</i> okayama7#130	13	37.5	Multicellular basidiomycete, undergoes complete sexual cycle	1447	51.6
<i>Cryptococcus neoformans</i> var. <i>neoformans</i> JEC21	14	19.1	Pathogenic fungus, causes cryptococcosis	13856	48.5
<i>Lentinula edodes</i> L-54	8	33	Edible shiitake mushroom	17581	30.7
<i>Phanerochaete chrysosporium</i> RP-78	10	30.0	Wood-decaying white rot fungus	135	57
<i>Puccinia graminis</i> f. sp. <i>tritici</i> CRL 75-36-700-3	18	88.7	Pathogenic fungus causes stem rust in cereal crops	18535	43.3
<i>Ustilago maydis</i> 521	23	19.8	Causes corn smut disease	1446	53.7

Source: NCBI Genome, NCBI.

We discuss some of these fascinating projects – *Aspergillus*, *Candida albicans*, *Cryptococcus neoformans*, the microsporidial parasite *Encephalitozoon cuniculi*, *Neurospora crassa*, the Basidiomycete *Phanerochaete chrysosporium*, and the fission yeast *Schizosaccharomyces pombe* (the second fungal genome to be completely sequenced) – in the following sections. All these projects highlight the remarkable diversity of fungal life. In Chapter 19 we will describe comparative genomics projects on more familiar organisms such as humans and fish (which diverged ~450 MYA), the fruit fly and mosquito (estimated to have diverged ~250 MYA), as well as closely related species that diverged more recently. The fungi offer an opportunity to analyze highly divergent species (e.g., *S. cerevisiae* and *S. pombe* diverged ~400 MYA), as well as closely related species.

One of the leading bioinformatics resources for fungal genome research is the MycoCosm portal (Grigoriev *et al.*, 2014). Projects such as this are important to centralize information about fungal genomes and also to help promote consistent annotation across projects.

The MycoCosm fungal genomics portal is at <http://jgi.doe.gov/fungi> (WebLink 18.13).

### Fungi in the Human Microbiome

We encountered the diversity of bacteria living in various regions of the human body in Chapter 17. Not surprisingly, human skin is an inviting habitat that also harbors diverse fungi. Findley *et al.* (2013) cultured fungi from 14 body regions in 10 healthy individuals, sequencing 18S rRNA. Across 11 body and arm regions they identified both Ascomycetes and Basidiomycetes (in particular, of the genus *Malassezia*). The greatest fungal diversity occurred in the foot, including the plantar heel (median richness of ~80 genera), toe web, and toenail. Studies such as this reveal the complexity of the skin ecosystem and can help us to learn more about the roles of fungi in health and disease.

**TABLE 18.5 Fungal genome projects: representative examples of fungi other than Ascomycetes and Basidiomycetes.** ND: not determined.

Organism	Chromosomes	Genome size (Mb)	Comment	ID	GC %
<i>Allomyces macrogyrus</i>	ND	57.1	Filamentous chytrid fungus	20563	61.6
<i>Antonospora locustae</i>	ND	2.9	Intracellular microsporidian parasite	186881	–
<i>Batrachochytrium dendrobatidis</i> JEL423	20	23.9	Aquatic chytrid fungus kills amphibians	13653	39.3
<i>Encephalitozoon cuniculi</i> GB-M1	11	2.5	Intracellular parasite, infects mammals	13833	47.3
<i>Rhizopus oryzae</i> RA 99-880	ND	46.2	Opportunistic pathogen causes mucormycosis	13066	35.6

Source: NCBI Genome, NCBI.

## Aspergillus

The genus *Aspergillus* consists of filamentous Ascomycetes. Of the 250 known species of *Aspergillus*, over two dozen are human pathogens. Fourteen genomes have now been sequenced, and dozens more are in progress. All *Aspergillus* genomes that have been sequenced have eight chromosomes and a genome size of 28–40 Mb, but the species harbor as much sequence diversity as species of our phylum, the vertebrates (Gibbons and Rokas, 2013). Information about these fungi is centralized at the *Aspergillus* Genome Database (Cerdeira *et al.*, 2014). This resource promotes consistent annotation, and provides access to data (including RNA-seq) and tools.

We introduce three prominent species. (1) *Aspergillus nidulans* has had a long-standing role as a model organism in genetics; its genome was sequenced by Galagan *et al.* (2005). (2) *Aspergillus fumigatus* is the most common mold that causes infection worldwide. It is an opportunistic pathogen to which immunocompromised individuals are particularly susceptible. Nierman *et al.* (2005) sequenced its genome and identified candidate pathogenicity genes as well as genes that may facilitate its unusual lifestyle (e.g., thriving at temperatures up to 70°C). One of the many unique features of this genome is the presence of *A. fumigatus*-specific proteins that are closely related to a class of arsenite reductases previously seen only in bacteria. (3) *Aspergillus oryzae* is a fungus from which sake, miso, and soy sauce are prepared. Like *A. nidulans* and *A. fumigatus*, its genome is organized into eight chromosomes but the total genome size is 7–9 megabases larger (29–34% larger; Machida *et al.*, 2005). This is due to blocks of sequence that are dispersed throughout the *A. oryzae* genome.

Comparative analyses revealed the presence of conserved noncoding DNA elements (Galagan *et al.*, 2005), analogous to the studies of *Saccharomyces* described above. Of the three *Aspergilli*, *A. fumigatus* and *A. oryzae* reproduce through asexual mitotic spores while *A. nidulans* has a sexual cycle. Comparative analysis of the three genomes suggested that, surprisingly, *A. fumigatus* and *A. oryzae* have the necessary genes for a sexual cycle (reviewed in Scazzochio, 2006). Another surprising aspect of the comparative analyses is that peroxisomes in *Aspergilli* (organelles responsible for fatty acid β-oxidation) resemble those of mammalian cells more than yeasts because: (1) β-oxidation occurs in both peroxisomes and mitochondria, and both *Aspergilli* and mammals have two sets of the necessary genes; and (2) both *Aspergilli* and mammalian genomes encode peroxisomal acyl-CoA dehydrogenases. The yeasts have served as important model systems for the study of human peroxisomal disorders such as adrenoleukodystrophy.

Next-generation sequencing has transformed comparative genomics and, in addition to comparing species, it is becoming routine to sequence the genomes of strains. Umemura *et al.* (2012) sequenced the genome of an industrial isolate of *A. oryzae*, comparing its sequence to that of the wildtype isolate characterized in 2005. They found frequent mutations at loci that lacked conserved synteny among *A. oryzae*, *A. fumigatus*, and *A. nidulans*.

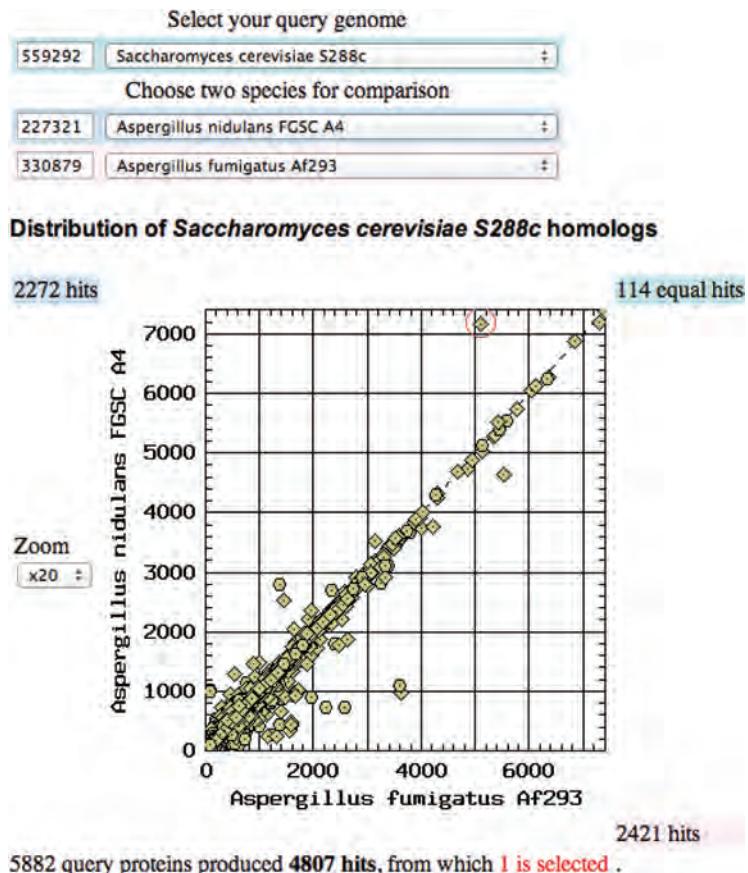
A comparison of *A. nidulans* and *A. fumigatus* using TaxPlot at NCBI (Chapter 17), with *S. cerevisiae* as a reference, shows that many proteins are conserved between those three species (Fig. 18.16). Of those that differ a notable example is midasin (circled), the giant protein from *S. cerevisiae* chromosome XII.

## Candida albicans

*Candida albicans* is a diploid sexual fungus that frequently causes opportunistic infections in humans (Kim and Sudbery, 2011). The skin, nails, and mucosal surfaces are typical targets, but deep tissues can also be infected. The genome size is approximately 14.8 Mb (which is typical for many fungi), but the chromosomal arrangement is unusual: the genome has eight chromosome pairs, seven of which are constant and one of which is variable (ranging from about 3 to 4 Mb). Another unusual feature is that it has no known

The *Aspergillus* Genome

Database is at <http://www.aspgd.org/> (WebLink 18.14).



**FIGURE 18.16** The TaxPlot tool at NCBI shows proteins from *A. nidulans* and *A. fumigatus* in relation to a reference proteome of *S. cerevisiae*. TaxPlot can help to identify organism-specific innovations that may underlie the distinct physiologies of these *Aspergilli*. A midasin homolog that is more closely related to *S. cerevisiae* in *A. nidulans* is circled.

Source: TaxPlot, Entrez, NCBI.

The *C. albicans* genome was sequenced by Ron Davis and colleagues at Stanford University.

CandidaDB is available online at  
<http://www.candidagenome.org/> (WebLink 18.15).

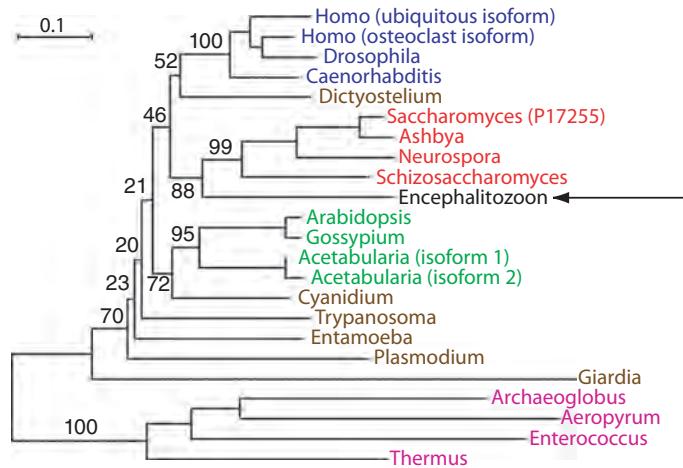
haploid state; the diploid genome was therefore sequenced (Jones *et al.*, 2004; reviewed by Odds *et al.*, 2004). This was challenging because heterozygosity commonly occurs at many alleles, making it difficult to assign a sequence to one heterozygous locus rather than two independent loci. On average there is one polymorphism every 237 bases, a considerably higher frequency than occurs in human (Chapter 20).

Information on the *Candida* genome is centralized at the CandidaDB database (Rossignol *et al.*, 2008). The reference genome initially contained 7677 ORFs (of size 100 amino acids or greater) although, as is routine for any genome project, the annotation process is ongoing. About half the predicted proteins match human, *S. cerevisiae*, and *Schizosaccharomyces pombe*, and only 22% of the ORFs did not match any of those three genomes.

A specialized feature of *C. albicans* (shared by *Debaryomyces hansenii*; Dujon *et al.*, 2004) is that the codon CUG is translated as serine rather than the usual product, leucine. Bezerra *et al.* (2013) engineered *C. albicans* strains that misincorporate varying levels of Leu at CUG sites. They concluded that this organism uses ambiguity in the genetic code to shape gene evolution, increasing phenotypic variation.

### *Cryptococcus neoformans*: model fungal pathogen

*C. neoformans* is a soil-dwelling fungus that causes cryptococcosis, one of the most life-threatening infections in AIDS patients. Its genome of 20 megabases is organized into 14 chromosomes as well as a mitochondrial genome. Loftus *et al.* (2005) sequenced two



**FIGURE 18.17** Phylogenetic analysis of vacuolar ATPase subunit A from animals, plants, fungi, protists, bacteria, and archaea supports a fungal origin for the microsporidial parasite *Encephalitozoon cuniculi* (arrow). This tree was generated using a neighbor-joining method, and values are bootstrap percentages (see Chapter 7). Redrawn from Katinka *et al.* (2001) with permission from Macmillan Publishers.

separate strains. Transposons constitute about 5% of the genome and are dispersed among all 14 chromosomes. In contrast to *S. cerevisiae*, there is no evidence of a whole-genome duplication. Another difference between the two fungi is that *C. neoformans* gene organization is more complex. Its 5672 predicted protein-coding genes are characterized by introns (an average of 5.3 per gene of 67 base pairs), alternatively spliced transcripts, and endogenous antisense transcripts.

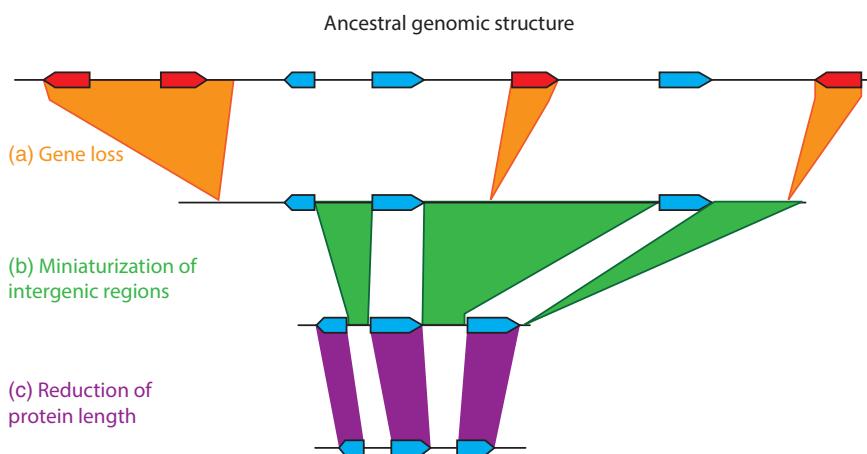
### Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*

Microsporidia are single-celled eukaryotes that lack mitochondria and peroxisomes. These organisms infect animals (including humans) as obligate intracellular parasites. The complete genome of the microsporidium *E. cuniculi* was determined by several research groups in France (Katinka *et al.*, 2001). The genome is highly compacted, having about 2000 protein-coding genes in 2.9 Mb. Analogous to parasitic bacteria (Chapter 17), these pathogens have therefore undergone a reduction in genome size. Phylogenetic analyses using several *E. cuniculi* proteins suggest that these parasites are atypical fungi that once possessed, but subsequently lost, their mitochondria (Fig. 18.17; Katinka *et al.*, 2001).

Many other microsporidia have undergone genome reduction (Corradi and Slámovits, 2011). This can occur by gene loss, by reducing the size of intergenic regions, and by reducing the lengths of proteins and introns (Fig. 18.18). In some cases, microsporidia have gained genes by lateral gene transfer. *Encephalitozoon hellem* and *E. romaleae*, each having genomes between 2 and 3 Mb, acquired genes responsible for folate and purine metabolism by lateral transfer from different eukaryotic and bacterial donors (Pombert *et al.*, 2012). In *E. hellem* these transferred genes are functional, while in *E. romaleae* multiple frameshift mutations created pseudogenes involving one particular functional pathway (that of *de novo* synthesis of folate). Pombert *et al.* speculate on mechanisms and reasons for this specific loss, possibly involving the metabolic environment provided by each organism's host.

### *Neurospora crassa*

The orange bread mold *Neurospora* has served as a beautiful and simple model organism for genetic and biochemical studies since Beadle and Tatum used it to establish the



**FIGURE 18.18** Mechanisms of reduction of genome size in microsporidia. An ancestral genome is shown schematically with seven genes (blue, red) and large intergenic regions (black line). (a) Gene losses (shaded orange, leading to loss of red genes) reduce the genome size. (b) Reduction of intergenic region sizes leads to increased gene density. (c) Shortening of gene regions encoding proteins reduces the genome size. These three types of events may occur in any order. Adapted from Corradi and Slamovits (2011) with permission from Oxford University Press.

*Neurospora crassa* genome database websites are available at the Broad Institute (<http://www.broadinstitute.org/annotation/genome/neurospora/MultiHome.html>, WebLink 18.16) and at MIPS (<http://mips.helmholtz-muenchen.de/genre/proj/ncrassa/>, WebLink 18.17). Ensembl also offers a *N. crassa* resource ([http://fungi.ensembl.org/Neurospora\\_crassa/info/Index](http://fungi.ensembl.org/Neurospora_crassa/info/Index), WebLink 18.18).

George Beadle and Edward Tatum shared a Nobel Prize in 1958 (with Joshua Lederberg) "for their discovery that genes act by regulating definite chemical events" (<http://www.nobel.se/medicine/laureates/1958/>, WebLink 18.19). They irradiated *N. crassa* with X-rays to study gene function.

one-gene–one-enzyme model in the 1940s. *Neurospora* is the best characterized of the filamentous fungi, a group of organisms critically important to agriculture, medicine, and the environment (Perkins and Davis, 2000). The developmental complexity of *Neurospora* contrasts with other unicellular yeasts (Casselton and Zolan, 2002). *Neurospora* is widespread in nature and, like the fly *Drosophila*, is exceptionally suited as a subject for population studies.

As for *S. cerevisiae*, *Neurospora* is an ascomycete and therefore shares the advantage of this group of organisms in yielding complete tetrads for genetic analyses. However, it is more similar to animals than yeasts in many important ways. For example, unlike yeast but like mammals, it contains complex I in its respiratory chain, it has a clearly discernable circadian rhythm, and it methylates DNA to control gene expression. The seven decades of intensive studies on the genetics, biochemistry, and cell biology of *Neurospora* establish this organism as an important source of biological knowledge.

Galagan *et al.* (2003) reported the complete genome sequence of *Neurospora*. They sequenced about 39 Mb of DNA on 7 chromosomes, and identified 10,082 protein-coding genes (9200 longer than 100 amino acids). Of these proteins, 41% have no similarity to known sequences, and 57% do not have identifiable orthologs in *S. cerevisiae* or *S. pombe*.

The *Neurospora* genome has only 10% repetitive DNA, including ~185 copies of rDNA genes (Krumlauf and Marzluf, 1980). Other repeated DNA is dispersed and tends to be short and/or diverged, presumably because of the phenomenon of "RIP" (repeat-induced point mutation). RIP is a mechanism by which the genome is scanned for duplicated (repeated) sequences in haploid nuclei of special premeiotic cells. The RIP machinery efficiently finds them, and then litters them with numerous GC-to-AT mutations (Selker, 1990). Apparently RIP serves as a genome defense system for *Neurospora*, inactivating transposons and resisting genome expansion (Kinsey *et al.*, 1994). Galagan *et al.* (2003) found relatively few *Neurospora* genes that are in multigene families, and a mere eight pairs of duplicated genes that encode proteins >100 amino acids. Also, 81% of the repetitive DNA sequences were mutated by RIP. RIP has therefore suppressed the creation of new genes through duplication in *Neurospora* (Perkins *et al.*, 2001; Galagan *et al.*, 2003).

## First Basidiomycete: *Phanerochaete chrysosporium*

*Phanerochaete chrysosporium* is the first fungus of the phylum Basidiomycota to have its genome completely sequenced. This is a white rot fungus that degrades many biomaterials, including pollutants and also lignin (a polymer that provides strength to wood, among other roles). Fungi appeared about 1–1.5 billion years ago, and the Basidiomycota diverged from the better-characterized Ascomycota over 500 million years ago. There were therefore relatively little sequence data available from closely related organisms, and annotation of this genome was particularly difficult. The genome consists of about 30 Mb of DNA arranged in 10 chromosomes. Martinez *et al.* (2004) predicted 11,777 genes, of which three-quarters had significant matches to previously known proteins. White rot fungi are able to degrade the major components of plant cell walls, including cellulose and lignins, using a series of oxidases and peroxidases. The genome encodes hundreds of enzymes that are able to cleave carbohydrates. An updated annotation of the genome reveals additional gene models for secreted proteins (vanden Wymelenberg *et al.*, 2006).

Wood is notably resistant to decay. White rot fungi (such as *P. chrysosporium*) as well as some brown rot are the only organisms able to decompose lignin and cellulose in wood. To understand the evolutionary origin of this process, Floudas *et al.* (2012) performed comparative analyses of 31 fungal genomes (including 12 that they sequenced), identifying oxidoreductases, carbohydrate-active enzymes, and peroxidases implicated in wood decay. Their phylogenetic analyses suggested the emergence of white rot with wood-decaying capabilities about 295 MYA. While *P. chrysosporium* degrades lignin and cellulose, its close relative *Ceriporiopsis subvermispora* degrades lignin but not cellulose. Fernandez-Fueyo *et al.* (2012) sequenced the *C. subvermispora* genome and compared the inventories of genes encoding peroxidases and other enzymes. These studies highlight the rapid impact of genomics on the study of physiological processes.

The *P. chrysosporium* genome-sequencing project was undertaken by the US Department of Energy (<http://genome.jgi-psf.org/Phchr1/Phchr1.info.html>, WebLink 18.20).

## Fission Yeast *Schizosaccharomyces pombe*

The fission yeast *S. pombe* has a genome size of 13.8 Mb. The complete sequencing of this genome was reported by a large European consortium (Wood *et al.*, 2002). The genome is divided into three chromosomes (Table 18.6).

Notably, there are 4940 predicted protein-coding genes (including 11 mitochondrial genes) and 33 pseudogenes. This is substantially fewer genes than is found in *S. cerevisiae* and is among the smallest number of protein-coding genes observed for any eukaryote. Some bacterial genomes encode more proteins, such as *Mesorhizobium loti* (6752 predicted genes) and *Streptomyces coelicolor* (7825 predicted genes).

The gene density in *S. pombe* is about one gene per 2400 bp, which is slightly less dense than is seen for *S. cerevisiae*. The intergenic regions are longer, and about 4730 introns were predicted. In *S. cerevisiae*, only 4% of the genes have introns.

*Schizosaccharomyces pombe* and *S. cerevisiae* diverged between 330 and 420 MYA. Some gene and protein sequences are equally divergent between these two fungi as they

For extensive information on *S. pombe* genome sequence analysis, see PomBase (<http://www.pombase.org/>, WebLink 18.21).

Leland Hartwell, Timothy Hunt, and Sir Paul Nurse won the Nobel Prize in Physiology or Medicine in 2001 for their work on cell cycle control. Nurse's studies employed *S. pombe*, while Hartwell studied *S. cerevisiae* and Hunt studied sea urchins and other organisms. See (<http://www.nobel.se/medicine/laureates/2001/>) (WebLink 18.22).

TABLE 18.6 Features of *S. pombe* genome.

Chromosome number	Length (Mb)	Number of genes	Mean gene length (bp)	Coding (%)
1	5.599	2255	1446	58.6
2	4.398	1790	1411	57.5
3	2.466	884	1407	54.5
Whole-genome	12.462	4929	1426	57.5

Source: Wood *et al.* (2002). Reproduced with permission from Macmillan Publishing Ltd.

are between fungi and their vertebrate (e.g., human) orthologs. To identify such genes, use the TaxPlot tool on the NCBI Genome website. Comparative analyses are likely to elucidate the genetic basis for differences in the biology of these fungi, such as the propensity of *S. pombe* to divide by binary fission and the relatively fewer number of transposable elements in *S. pombe*.

### Other Fungal Genomes

In addition to those described in this chapter, many other fungal genomes have been sequenced and characterized. These include *Fusarium* (Ma *et al.*, 2010); *Pichia pastoris* (used for industrial production of proteins and metabolites; Gonçalves *et al.*, 2013); *Pseudomonas* (including plant, insect, and human pathogens; Silby *et al.*, 2011); *Trichoderma* (Druzhinina *et al.*, 2011); *Tuber melanosporum* (the Périgord black truffle; Martin *et al.*, 2010); and *Yarrowia lipolytica* (Nicaud *et al.*, 2012).

### Ten Leading Fungal Plant Pathogens

A survey of experts suggested ten fungal pathogens of greatest scientific and economic importance (Dean *et al.*, 2012): (1) *Magnaporthe oryzae* is a filamentous ascomycete that afflicts rice and wheat; (2) *Botrytis cinerea* or gray mould can infect 200 plant species; (3) *Puccinia* spp.; (4) *Fusarium graminearum*; (5) *Fusarium oxysporum*; (6) *Blumeria graminis*; (7) *Mycosphaerella graminicola*; (8) *Colletotrichum* spp.; (9) *Ustilago maydis*; and (10) *Melampsora lini*.

Similar survey results were presented for viruses in Chapter 16 and for bacteria in Chapter 17.

### PERSPECTIVE

The budding yeast *S. cerevisiae* is one of the most significant organisms in biology for several reasons:

- It represents the first eukaryotic genome to have been sequenced. It was selected because of its compact genome size and structure.
- As a single-celled eukaryotic organism, its biology is simple relative to humans and other metazoans.
- The biology community has acquired a deep knowledge of yeast genetics and has collected a variety of molecular tools that are useful to elucidate the function of yeast genes. Functional genomics approaches based on genome-wide analysis of gene function have been implemented (Chapter 14). For example, each of its >6000 genes has been knocked out and tagged with molecular barcodes, allowing massive, parallel studies of gene function.

Many additional fungal genomes are now being sequenced. In each branch of biology, we are learning that comparative genomic analyses are essential in helping to identify protein-coding genes (by homology searching), in studies of functional elements in noncoding DNA, in evolutionary studies such as analyses of genome duplications, and in helping us to uncover biochemical pathways that allow cells to survive.

We can consider the nature of genomes and the forces that shape them (Conant and Wolfe, 2008). (1) What is the mechanism by which portions of a genome are increased or streamlined? Fungi afford many examples of whole-genome duplication, as well as segmental duplication and in some cases lateral gene transfer, that introduce new genetic material to a genome. Some fungi such as *Encephalitozoon* provide examples of genome reduction. (2) How are newly arisen genomic features acted on by natural selection and other forces that modify genome structure and function? The fungi provide important organisms to address these questions.

## PITFALLS

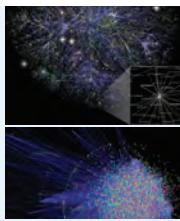
Although *S. cerevisiae* serves as an important model organism, it is essential to realize the scope of our ignorance. How does the genotype of a single-gene knockout lead to a particular phenotype? We urgently need to answer this question for gene mutations in humans that cause disease, but even in a so-called simple model organism such as yeast we do not understand the full repertoire of protein–protein interactions that underlie cell function. If we think of the genome as a blueprint of a machine, we now have a “parts list” in the form of a list of the gene products. We must next figure out how the parts fit together to allow the machine to function in a variety of contexts. Gene annotation in yeast databases such as SGD, including the results of broad functional genomics screens, provide an excellent starting point for functional analyses.

## ADVICE FOR STUDENTS

It has been suggested that *Saccharomyces cerevisiae* is the best-characterized eukaryote, if not the best-understood organism across the entire tree of life. You can use bioinformatics resources to probe this deep understanding. Choose a single protein and explore its properties in detail, from binding partners to chromosomal context to paralogs to gene expression changes. Choose a single biological process and explore how studies of yeast have enabled us to understand fundamental principles. Studies of secretion (for which Randy Scheckman recently received a Nobel Prize; see Fig. 14.4) provide an example: a functional screen led to the discovery of a few dozen secretory (*sec*) mutants, and these gene products were shown to interact in biochemical pathways involving vesicular transport. Such studies are further relevant to the function of all human cells. Given the distant divergence time between humans and fungi (who last shared a common ancestry 1.5 BYA), this highlights the remarkable conservation of this particular pathway.

## WEB RESOURCES

The SGD (<http://www.yeastgenome.org/>, WebLink 18.9) lists a series of yeast resources. Another useful gateway is the SGD Wiki (<http://wiki.yeastgenome.org/>, WebLink 18.23). The Fungal Genomics Program of the Joint Genome Institute provides a useful starting point for diverse fungal species (<http://genome.jgi.doe.gov/programs/fungi/>, WebLink 18.24).



## Discussion Questions

**[18-1]** The budding yeast *Saccharomyces cerevisiae* is sometimes described as a simple organism because it is unicellular, its genome encodes a relatively small number of genes (about 6000), and it has served as a model organism for genetics studies. Still, we understand the function of only about half its genes. Many functional genomics tools are now available, such as a complete collection of yeast knockout strains (i.e., null alleles of each gene). How would you use such functional genomics tools to further our knowledge of gene function in yeast?

**[18-2]** The fungi are a sister group to the metazoans (animals) (Fig. 19.1). Do you expect the principles of genome

evolution, gene function, and comparative genomics that are elucidated by studies of fungi to be closely applicable to metazoans such as humans, worms, and flies? For example, we discussed the whole-genome duplication of some fungi; how would you test the hypothesis that the human genome also underwent a similar duplication? In comparative genomics, do you expect fungi to be far more similar to each other in their biological properties than metazoans are to each other?

**[18-3]** In *C. albicans*, the CUG codon is sometimes read as serine (rather than the usual leucine). This may have the positive effect of diversifying the proteome. It could have a deleterious effect, however. If this phenomenon occurred in humans, how often would it be lethal?

## PROBLEMS/COMPUTER LAB

**[18-1]** This problem primarily uses the UCSC Genome Browser and the Yeast Genome Order Browser (YGOB) to study yeast. Visit the UCSC Genome Browser and navigate to the *S. cerevisiae* genome. Enter chrXII, the chromosome we explored in this chapter. Set the track for PhastCons conserved elements (Chapter 6) to full, and limit it to scores of at least 900. This shows a cluster of highly conserved, neighboring genes. For this exercise explore them in more depth. How many conserved genes are there? What happens as you raise or lower the PhastCons score threshold? Do the highly conserved genes share functional properties? Next, explore their conservation in the YGOB. Are the genes that have paralogs due to whole-genome duplication? Are these essential genes? You can determine whether they are essential at the *Saccharomyces* Genome Database (SGD).

**[18-2]** How many genes are on each *Saccharomyces cerevisiae* chromosome? Use EDirect. This problem is adapted from <http://www.ncbi.nlm.nih.gov/books/NBK179288/>. Use the following code (in blue; you can copy and paste relevant code from the EDirect website). Compare your result to that given here (in black).

```
for chr in I II III IV V VI VII VIII IX X XI XII
XIII XIV XV XVI MT
do
  esearch -db gene -query " Saccharomyces
cerevisiae [ORGN] AND $chr [CHR]" |
  efilter -query "alive [PROP] AND genotype
protein coding [PROP]" |
  efetch -format docsum |
  extract -pattern DocumentSummary -NAME Name \
-block GenomicInfoType -match "ChrLoc:$chr" \
-tab "\n" -element ChrLoc,"&NAME" |
  grep '.' | sort | uniq | cut -f 1 |
  sort-uniq-count-rank
done
94 I
408 II
161 III
755 IV
280 V
127 VI
530 VII
282 VIII
211 IX
359 X
313 XI
508 XII
461 XIII
398 XIV
537 XV
464 XVI
19 MT
```

**[18-3]** Explore NCBI Genome resources for *S. cerevisiae*. Visit <http://www.ncbi.nlm.nih.gov/genome/15> (WebLink 18.25). According the Genome Projects report, how many strains have been sequenced? What is their range of sizes and GC content?

**[18-4]** Use the *Saccharomyces* Genome Database:

- Go to the SGD site (<http://www.yeastgenome.org/>).
- Pick an uncharacterized ORF. To find one, use the Gene/Seq Resources (one of the analysis tools), pick a chromosome (e.g., XII), then select Chromosomal Features Table. The first hypothetical ORF listed is *YLL067C*.
- Explore what its function might be. For some uncharacterized ORFs there will be relatively little information available; for others you may find a lot of information. From the Chromosomal Features Table click “Info” to view a page similar to that shown in Chapter 14.
- What are the physical properties of the protein (e.g., molecular weight, isoelectric point)?
- Does the protein have known domains?
- Have interactions been characterized between this and other proteins?
- Is the gene either induced or repressed in various physiological states, such as stress response or during sporulation?
- In what other organisms is this gene present? Compare the usefulness of exploring SGD versus YGOD versus performing your own BLAST searches to answer this question. Which is best?

**[18-5]** Visit SGD > Analyze > Gene Lists to access YeastMine (or visit <http://yeastmine.yeastgenome.org>). Explore the many resources offered here, such as lists of centromeres (and accompanying descriptions), a broad range of queries (e.g., feature types), and analyses of a list of queries (e.g., try Sso1p).

**[18-6]** ABC transporters constitute a large family of trans-membrane-spanning proteins that hydrolyze ATP and drive the transport of ligands such as chloride across a membrane. How many ABC transporters are there in yeast?

**[18-7]** Create a phylogenetic tree of the fungi using 18S ribosomal RNA sequences. Align them, and create a tree using MEGA or related software (Chapter 7). Does the tree agree with those shown in this chapter? If not, why not?

## Self-Test Quiz

**[18-1]** The *Saccharomyces cerevisiae* genome is characterized by the following properties except:

- (a) very high gene density (2000 base pairs per gene);
- (b) very low number of introns;
- (c) high degree of polymorphism; or
- (d) 16 chromosomes.

**[18-2]** The yeast *Saccharomyces cerevisiae* is an attractive model organism for many reasons. Which one of the following is NOT a useful feature of yeast?

- (a) The genome size is relatively small.
- (b) Gene knockouts by homologous recombination are possible.
- (c) Large repetitive DNA sequences serve as a good model for higher eukaryotes.
- (d) There is high open reading frame (ORF) density.

**[18-3]** The *Saccharomyces cerevisiae* genome is small (it encodes about 6000 genes). It is thought that, about 100 MYA:

- (a) The entire genome duplicated, followed by tetraploidization.
- (b) The genome underwent many segmental duplications, followed by gene loss.
- (c) The entire genome duplicated, followed by gene loss.
- (d) The genome duplicated, followed by gene conversion.

**[18-4]** After gene duplication, the most common outcome is the loss of the duplicated gene. A reasonable explanation of why this might occur is that this second copy:

- (a) is superfluous;
- (b) may acquire forbidden mutations that are deleterious to the fitness of the organism;
- (c) is under intense negative selection; or
- (d) is a substrate for nonallelic homologous recombination.

**[18-5]** Comparative analyses of *S. cerevisiae* and two closely related species (*S. castelli*, *C. glabrata*) allow a description of the patterns of gene retention and gene loss in multiple organisms following whole-genome duplication. Across thousands of gene loci in three genomes that underwent genome duplication, which of the following occurred?

- (a) For about three-quarters of the loci, all three species lost one of the two copies of a duplicated gene.
- (b) For about half of the loci, no gene loss occurred.
- (c) For about half of the loci, there was partial loss of both copies of a duplicated gene.
- (d) For about three-quarters of all loci, all three loci lost both copies of the duplicated gene.

**[18-6]** Features of the *Candida albicans* genome include:

- (a) an accessory plasmid;
- (b) one of its chromosomes has a highly variable length;
- (c) the DNA is characterized by an extraordinarily high amount of polymorphism; or
- (d) the CTG codon that encodes leucine in most organisms encodes serine in *C. albicans*.

**[18-7]** The filamentous fungus *Neurospora crassa* has an extremely low amount of repetitive DNA (spanning only 10% of its 39 megabase genome). This is because it uses:

- (a) chromatin diminution;
- (b) repetitive DNA inversion;
- (c) repeat-induced point mutations, a phenomenon in which repeats are inactivated; or
- (d) repeat-induced synchronization to inactivate repeats.

**[18-8]** One of the most remarkable features of the *Schizosaccharomyces pombe* genome is that:

- (a) It is predicted to encode fewer than 5000 proteins, making its genome (and proteome) smaller than even some bacterial genomes.
- (b) The number of predicted introns is about the same as the number of predicted ORFs.
- (c) It has as many genes that are homologous to bacterial genes as it has genes that are homologous to *S. cerevisiae* genes.
- (d) Its genome size is approximately the same as that of *S. cerevisiae*, even though these species diverged hundreds of millions of years ago.

**[18-9]** Yeast is the only major research organism approved by the US Food and Drug Administration (FDA) for human consumption:

- (a) true; or
- (b) false.

## SUGGESTED READING

A superb overview of fungal taxonomy is provided by Guarro *et al.* (1999), while important papers are by Hibbett *et al.* (2007) and James *et al.* (2006). Bernard Dujon (2010) reviews yeast genomics in relation to eukaryotic genome evolution. For overviews of whole-genome duplication and factors affecting genome size in yeast, see Kelkar and Ochman (2012). On the topic of gene duplication, see Conant and Wolfe (2008). I also strongly recommend Susumu Ohno's 1970 book *Evolution by Gene Duplication*. Gibbons and Rokas (2013) provide an excellent overview of the *Aspergillus* genome.

## REFERENCES

- Aalto, M.K., Ronne, H., Keranen, S. 1993. Yeast syntaxins Sso1p and Sso2p belong to a family of related membrane proteins that function in vesicular transport. *EMBO Journal* **12**, 4095–4104.
- Ainsworth, G.C. 1993. Fungus infections (mycoses). In *The Cambridge World History of Human Disease* (ed. K. F.Kiple). Cambridge University Press, New York, pp. 730–736.
- Albertin, W., Marullo, P. 2012. Polyploidy in fungi: evolution after whole-genome duplication. *Proceedings of the Royal Society B: Biological Sciences* **279**(1738), 2497–2509. PMID: 22492065.
- Baldauf, S.L., Roger, A. J., Wenk-Siefert, I., Doolittle, W. F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972–977.
- Bezerra, A.R., Simões, J., Lee, W. *et al.* 2013. Reversion of a fungal genetic code alteration links proteome instability with genomic and phenotypic diversification. *Proceedings of the National Academy of Sciences, USA* **110**(27), 11079–11084. PMID: 23776239.
- Bullock, W. 1938. *The History of Bacteriology*. Oxford University Press, New York.
- Byrne, K.P., Wolfe, K.H. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15**, 1456–1461.
- Byrne, K.P., Wolfe, K.H. 2006. Visualizing synteny relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Research* **34**, D452–455.
- Casaregola, S., Weiss, S., Morel, G. 2011. New perspectives in hemiascomycetous yeast taxonomy. *Comptes Rendus Biologies* **334**(8–9), 590–598. PMID: 21819939.
- Casselton, L., Zolan, M. 2002. The art and design of genetic screens: Filamentous fungi. *Nature Reviews Genetics* **3**, 683–697.
- Cerdeira, G.C., Arnaud, M.B., Inglis, D.O. *et al.* 2014. The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Research* **42**(1), D705–710. PMID: 24194595.
- Cherry, J. M., Ball, C., Weng, S. *et al.* 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67–73. PMID: 9169866.
- Cliften, P., Sudarsanam, P., Desikan, A. *et al.* 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76.
- Conant, G.C., Wolfe, K.H. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* **9**(12), 938–950. PMID: 19015656.
- Corradi, N., Slamovits, C.H. 2011. The intriguing nature of microsporidian genomes. *Briefings in Functional Genomics* **10**(3), 115–124. PMID: 21177329.
- Dacks, J. B., Doolittle, W. F. 2002. Novel syntaxin gene sequences from *Giardia*, *Trypanosoma* and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *Journal of Cell Science* **115**, 1635–1642.
- Dean, R., Van Kan, J.A., Pretorius, Z.A. *et al.* 2012. The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology* **13**(4), 414–430. PMID: 22471698.
- Dequin, S., Casaregola, S. 2011. The genomes of fermentative *Saccharomyces*. *Comptes Rendus Biologies* **334**(8–9), 687–693. PMID: 21819951.

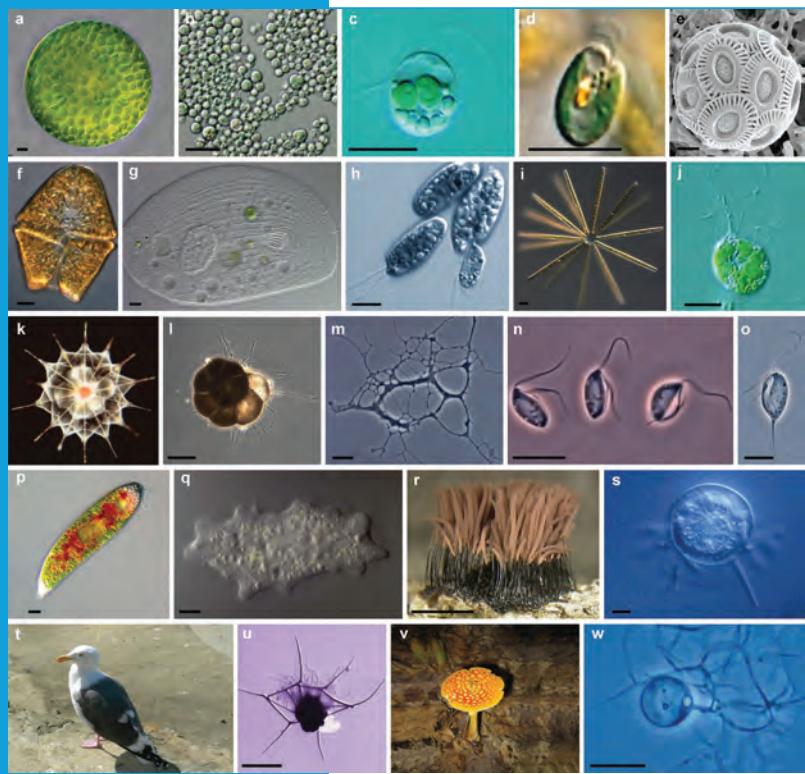
- Druzhinina, I.S., Seidl-Seiboth, V., Herrera-Estrella, A. *et al.* 2011. *Trichoderma*: the genomics of opportunistic success. *Nature Reviews Microbiology* **9**(10), 749–759. PMID: 21921934.
- Dujon, B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends in Genetics* **22**, 375–387.
- Dujon, B. 2010. Yeast evolutionary genomics. *Nature Reviews Genetics* **11**(7), 512–524. PMID: 20559329.
- Dujon, B., Alexandraki, D., André, B. *et al.* 1994. Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371–378. PMID: 8196765.
- Dujon, B., Sherman, D., Fischer, G. *et al.* 2004. Genome evolution in yeasts. *Nature* **430**, 35–44. PMID: 15229592.
- Engel, S.R., Cherry, J.M. 2013. The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database (Oxford) **2013**, bat012. PMID: 23487186.
- Engel, S.R., Dietrich, F.S., Fisk, D.G. *et al.* 2013. The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. *G3 (Bethesda)* **pii**, g3.113.008995v1. PMID: 24374639.
- Fares, M.A., Keane, O.M., Toft, C., Carretero-Paulet, L., Jones, G.W. 2013. The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. *PLoS Genetics* **9**(1), e1003176. PMID: 23300483.
- Fernandez-Fueyo, E., Ruiz-Dueñas, F.J., Ferreira, P. *et al.* 2012. Comparative genomics of *Ceriporiopsis subvermispora* and *Phanerochaete chrysosporium* provide insight into selective ligninolysis. *Proceedings of the National Academy of Sciences, USA* **109**(14), 5458–5463. PMID: 22434909.
- Ficklin, S.P., Sanderson, L.A., Cheng, C.H. *et al.* 2011. Tripal: a construction toolkit for online genome databases. *Database (Oxford)* **2011**, bar044. PMID: 21959868.
- Findley, K., Oh, J., Yang, J. *et al.* 2013. Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**(7454), 367–370. PMID: 23698366.
- Fitzpatrick, D.A. 2012. Horizontal gene transfer in fungi. *FEMS Microbiology Letters* **329**(1), 1–8. PMID: 22112233.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Floudas, D., Binder, M., Riley, R. *et al.* 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336**(6089), 1715–1719. PMID: 22745431.
- Galagan, J.E., Calvo, S.E., Borkovich, K.A. *et al.* 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**(6934), 859–868 3). PMID: 12712197.
- Galagan, J.E., Calvo, S.E., Cuomo, C. *et al.* 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**(7071), 1105–1115. PMID: 16372000.
- Garbarino, J. E., Gibbons, I. R. 2002. Expression and genomic analysis of midasin, a novel and highly conserved AAA protein distantly related to dynein. *BMC Genomics* **3**, 18.
- Gibbons, J.G., Rokas, A. 2013. The function and evolution of the *Aspergillus* genome. *Trends in Microbiology* **21**(1), 14–22. PMID: 23084572.
- Goffeau, A., Barrell, B.G., Bussey, H. *et al.* 1996. Life with 6000 genes. *Science* **274**, 546, 563–567. PMID: 8849441.
- Gonçalves, A.M., Pedro, A.Q., Maia, C. *et al.* 2013. *Pichia pastoris*: a recombinant microfactory for antibodies and human membrane proteins. *Journal of Microbiology and Biotechnology* **23**(5), 587–601. PMID: 23648847.
- Grigoriev, I.V., Nikitin, R., Haridas, S. *et al.* 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research* **42**(1), D699–704. PMID: 24297253.
- Guarro, J., Gene J., Stchigel, A. M. 1999. Developments in fungal taxonomy. *Clinical Microbiology Reviews* **12**, 454–500.
- Harbison, C.T., Gordon, D.B., Lee, T.I. *et al.* 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104. PMID: 15343339.

- Harrison, P. M., Kumar, A., Lang, N., Snyder, M., Gerstein, M. 2002. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Research* **30**, 1083–1090.
- Hibbett, D.S., Binder, M., Bischoff, J.F. *et al.* 2007. A higher-level phylogenetic classification of the Fungi. *Mycological Research* **111**, 509–547. PMID: 17572334.
- Hittinger, C.T. 2013. *Saccharomyces* diversity and evolution: a budding model genus. *Trends in Genetics* **29**(5), 309–317. PMID: 23395329.
- Hufton, A.L., Panopoulou, G. 2009. Polyploidy and genome restructuring: a variety of outcomes. *Current Opinion in Genetics and Development* **19**(6), 600–606. PMID: 19900800.
- James, T.Y., Kauff, F., Schoch, C.L. *et al.* 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818–822. PMID: 17051209.
- Johnston, M., Hillier, L., Riles, L. *et al.* 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature* **387**, 87–90. PMID: 9169871.
- Jones, T., Federspiel, N.A., Chibana, H. *et al.* 2004. The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences, USA* **101**, 7329–7334.
- Katinka, M. D., Duprat, S., Cornillot, E. *et al.* 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414**, 450–453. PMID: 11719806.
- Kelkar, Y.D., Ochman, H. 2012. Causes and consequences of genome expansion in fungi. *Genome Biology and Evolution* **4**(1), 13–23. PMID: 22117086.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.
- Kellis, M., Birren, B.W., Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624.
- Kim, J., Sudbery, P. 2011. *Candida albicans*, a major human fungal pathogen. *Journal of Microbiology* **49**(2), 171–177. PMID: 21538235.
- Kinsey, J. A., Garrett-Engele, P. W., Cambareri, E. B., Selker, E. U. 1994. The *Neurospora* transposon Tad is sensitive to repeat-induced point mutation (RIP). *Genetics* **138**, 657–664.
- Krumlauf, R., Marzluf, G. A. 1980. Genome organization and characterization of the repetitive and inverted repeat DNA sequences in *Neurospora crassa*. *Journal of Biological Chemistry* **255**, 1138–1145.
- Kuchenmeister, F. 1857. *On Animal and Vegetable Parasites of the Human Body, a Manual of their Natural History, Diagnosis, and Treatment*. Sydenham Society, London.
- Kurtzman, C.P., Robnett, C.J. 2003. Phylogenetic relationships among yeasts of the ‘*Saccharomyces* complex’ determined from multigene sequence analyses. *FEMS Yeast Research* **3**, 417–432.
- Liti, G., Schacherer, J. 2011. The rise of yeast population genomics. *Comptes Rendus Biologies* **334**(8–9), 612–619. PMID: 21819942.
- Loftus, B.J., Fung, E., Roncaglia, P. *et al.* 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324. PMID: 15653466.
- Lowe, T. M., Eddy, S. R. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–1171.
- Lynch, M., Conery, J. S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- Ma, L.J., van der Does, H.C., Borkovich, K.A. *et al.* 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **464**(7287), 367–373. PMID: 20237561.
- Machida, M., Asai, K., Sano, M. *et al.* 2005. Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* **438**, 1157–1161. PMID: 16372010.
- Mackiewicz, P., Kowalcuk, M., Mackiewicz, D. *et al.* 2002. How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? *Yeast* **19**, 619–629. PMID: 11967832.
- Margulis, L., Schwartz, K. V. 1998. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth*. W. H. Freeman and Company, New York.
- Martin, F., Kohler, A., Murat, C. *et al.* 2010. Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* **464**(7291), 1033–1038. PMID: 20348908.

- Martinez, D., Larrondo, L.F., Putnam, N. *et al.* 2004. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology* **22**, 695–700. PMID: 15122302.
- Mewes, H. W., Albermann, K., Bähr, M. *et al.* 1997. Overview of the yeast genome. *Nature* **387**, 7–65. PMID: 9169865.
- Neuvéglise, C., Marck, C., Gaillardin, C. 2011. The intronome of budding yeasts. *Comptes Rendus Biologies* **334**(8–9), 662–670. PMID: 21819948.
- Nicaud, J.M. 2012. *Yarrowia lipolytica*. *Yeast* **29**(10), 409–418. PMID: 23038056.
- Nierman, W.C., Pain, A., Anderson, M.J. *et al.* 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* **438**, 1151–1156. PMID: 16372009.
- O'Connor, B.D., Day, A., Cain, S. *et al.* 2008. GMODWeb: a web framework for the Generic Model Organism Database. *Genome Biology* **9**(6), R102. PMID: 18570664.
- Odds, F.C., Brown, A.J., Gow, N.A. 2004. *Candida albicans* genome sequence: a platform for genomics in the absence of genetics. *Genome Biology* **5**, 230.
- ÓhÉigearthaigh, S.S., Armisen, D., Byrne, K.P., Wolfe, K.H. 2011. Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments. *BMC Genomics* **12**, 377. PMID: 21791067.
- Ohno, S. 1970. *Evolution by Gene Duplication*. SpringerVerlag, Berlin.
- Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L. *et al.* 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**(6373), 38–46. PMID: 1574125.
- Papanicolaou, A., Heckel, D.G. 2010. The GMOD Drupal bioinformatic server framework. *Bioinformatics* **26**(24), 3119–3124. PMID: 20971988.
- Perkins, D.D., Davis, R.H. 2000. *Neurospora* at the millennium. *Fungal Genetics and Biology* **31**(3), 153–167. PMID: 11273678.
- Perkins, D. D., Radford, A., Sachs, M. S. 2001. *The Neurospora Compendium: Chromosomal loci*. Academic Press, San Diego, CA.
- Pevsner, J., Scheller, R.H. 1994. Mechanisms of vesicle docking and fusion: insights from the nervous system. *Current Opinion in Cell Biology* **6**(4), 555–560. PMID: 7986533.
- Piskur, J. 2001. Origin of the duplicated regions in the yeast genomes. *Trends in Genetics* **17**, 302–303.
- Pombert, J.F., Selman, M., Burki, F. *et al.* 2012. Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. *Proceedings of the National Academy of Sciences, USA* **109**(31), 12638–12643. PMID: 22802648.
- Protopopov, V., Govindan, B., Novick, P., Gerst, J. E. 1993. Homologs of the synaptobrevin/VAMP family of synaptic vesicle proteins function on the late secretory pathway in *S. cerevisiae*. *Cell* **74**, 855–861.
- Proux-Wéra, E., Armisen, D., Byrne, K.P., Wolfe, K.H. 2012. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* **13**, 237. PMID: 22984983.
- Reese, M.G., Moore, B., Batchelor, C. *et al.* 2010. A standard variation file format for human genome sequences. *Genome Biology* **11**(8), R88. PMID: 20796305.
- Robbertse, B., Tatusova, T. 2011. Fungal genome resources at NCBI. *Mycology* **2**(3), 142–160. PMID: 22737589.
- Rossignol, T., Lechat, P., Cuomo, C. *et al.* 2008. CandidaDB: a multi-genome database for *Candida* species and related *Saccharomycotina*. *Nucleic Acids Research* **36**(Database issue), D557–561. PMID: 18039716.
- Roth, J. F. 2000. The yeast Ty virus-like particles. *Yeast* **16**, 785–795.
- Rozen, S., Skaletsky, H., Marszalek, J.D. *et al.* 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876. PMID: 12815433.
- Sanderson, L.A., Ficklin, S.P., Cheng, C.H. *et al.* 2013. Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database (Oxford)* **2013**, bat075. PMID: 24163125.

- Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., Wolfe, K.H. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345.
- Scazzocchio, C. 2006. *Aspergillus* genomes: secret sex and the secrets of sex. *Trends in Genetics* **22**, 521–525.
- Selker, E. U. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annual Review of Genetics* **24**, 579–613.
- Silby, M.W., Winstanley, C., Godfrey, S.A., Levy, S.B., Jackson, R.W. 2011. *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiology Review* **35**(4), 652–680. PMID: 21361996.
- Seoighe, C., Wolfe, K. H. 1999. Updated map of duplicated regions in the yeast genome. *Gene* **238**, 253–261.
- Smith, M. M. 1987. Molecular evolution of the *Saccharomyces cerevisiae* histone gene loci. *Journal of Molecular Evolution* **24**(3), 252–259. PMID: 3106640.
- Souciet, J.L. 2011. Génolevures Consortium GDR CNRS 2354. Ten years of the Génolevures Consortium: a brief history. *Comptes Rendus Biologies* **334**(8–9), 580–584. PMID: 21819937.
- Tisseur, M., Kwapisz, M., Morillon, A. 2011. Pervasive transcription: Lessons from yeast. *Biochimie* **93**(11), 1889–1896. PMID: 21771634.
- Umemura, M., Koike, H., Yamane, N. *et al.* 2012. Comparative genome analysis between *Aspergillus oryzae* strains reveals close relationship between sites of mutation localization and regions of highly divergent genes among *Aspergillus* species. *DNA Research* **19**(5), 375–382. PMID: 22912434.
- van de Wouw, A.P., Howlett, B.J. 2011. Fungal pathogenicity genes in the age of ‘omics’. *Molecular Plant Pathology* **12**(5), 507–514. PMID: 21535355.
- vanden Wymelenberg, A., Minges, P., Sabat, G. *et al.* 2006. Computational analysis of the *Phanerochaete chrysosporium* v2.0 genome database and mass spectrometry identification of peptides in ligninolytic cultures reveal complex mixtures of secreted proteins. *Fungal Genetics and Biology* **43**(5), 343–356. PMID: 16524749.
- Wagner, A. 2000. Robustness against mutations in genetic networks of yeast. *Nature Genetics* **24**, 355–361.
- Wagner, A. 2001. Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends in Genetics* **17**, 237–239.
- Wang, D. Y., Kumar, S., Hedges, S. B. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings of the Royal Society of London, B: Biological Sciences* **266**, 163–171.
- Whittaker, R.H. 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science* **163**, 150–160.
- Wolfe, K.H., Shields, D. C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713.
- Wood, V., Gwilliam, R., Rajandream, M.A. *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880. PMID: 11859360.
- Wu, J., Delneri, D., O’Keefe, R.T. 2012. Non-coding RNAs in *Saccharomyces cerevisiae*: what is the function? *Biochemical Society Transactions* **40**(4), 907–911. PMID: 22817757.





Representative eukaryotic lineages from six putative supergroups. (a–c) 'Plantae': (a) *Eremosphaera viridis*, a green alga; (b) *Cyanidium* species, red algae; (c) *Cyanophora* species, a glaucophyte. (d–i) 'Chromalveolata': (d) *Chroomonas* species, a cryptomonad; (e) *Emiliania huxleyi*, a haptophyte; (f) *Akashiwo sanguinea*, a dinoflagellate; (g) *Trithigmostoma cucullulus*, a ciliate; (h) *Colpodella perforans*, an apicomplexan; (i) *Thalassionema* species, colonial diatom (Stramenopile). (j–m) 'Rhizaria': (j) *Chlorarachnion reptans*, an autotrophic amoeba (Cercozoa); (k) *Acantharea* species, a radiolarian; (l) *Ammonia beccarii*, a calcareous foraminifera; (m) *Corallomyxa tenera*, a reticulate amoeba. (n–p) 'Excavata': (n) *Jakoba* species, a jakobid with two flagella; (o) *Chilomastix cuspidata*, a flagellate retortamonad; (p) *Euglena sanguinea*, an autotrophic Euglenozoa. (q–s) 'Amoebozoa': (q) *Trichosphaerium* species, an amoeba; (r) *Stemonitis axifera*, an acellular slime mold; (s) *Arcella hemisphaerica*, a testate amoeba. (t–w) Opisthokonta: (t) *Larus occidentalis*, a bird; (u) *Campyloacantha* species, a choanoflagellate; (v) *Amanita flavoconia*, a basidiomycete fungus; (w) *Chytriomyces* species, a chytrid. All scale bars = 10  $\mu\text{m}$ , except (b, l) 100  $\mu\text{m}$  and (r) 5 mm.

Source: Tekle *et al.* (2009). Reproduced with permission from Oxford University Press.

# Eukaryotic Genomes: From Parasites to Primates

# CHAPTER 19

*Since the middle Miocene – an epoch of abundance and diversity for apes throughout Eurasia and Africa – the prevailing pattern of ape evolution has been one of fragmentation and extinction. The present-day distribution of nonhuman great apes, existing only as endangered and subdivided populations in equatorial forest refugia, is a legacy of that process. Even humans, now spread around the world and occupying habitats previously inaccessible to any primate, bear the genetic legacy of past population crises. All other branches of the genus *Homo* have passed into extinction. It may be that in the condition of Gorilla, Pan and Pongo we see some echo of our own ancestors before the last 100,000 years, and perhaps a condition experienced many times over several million years of evolution. It is notable that species within at least three of these genera continued to exchange genetic material long after separation, a disposition that may have aided their survival in the face of diminishing numbers. As well as teaching us about human evolution, the study of the great apes connects us to a time when our existence was more tenuous, and in doing so, highlights the importance of protecting and conserving these remarkable species.*

—Scally et al. (2012, p. 174)

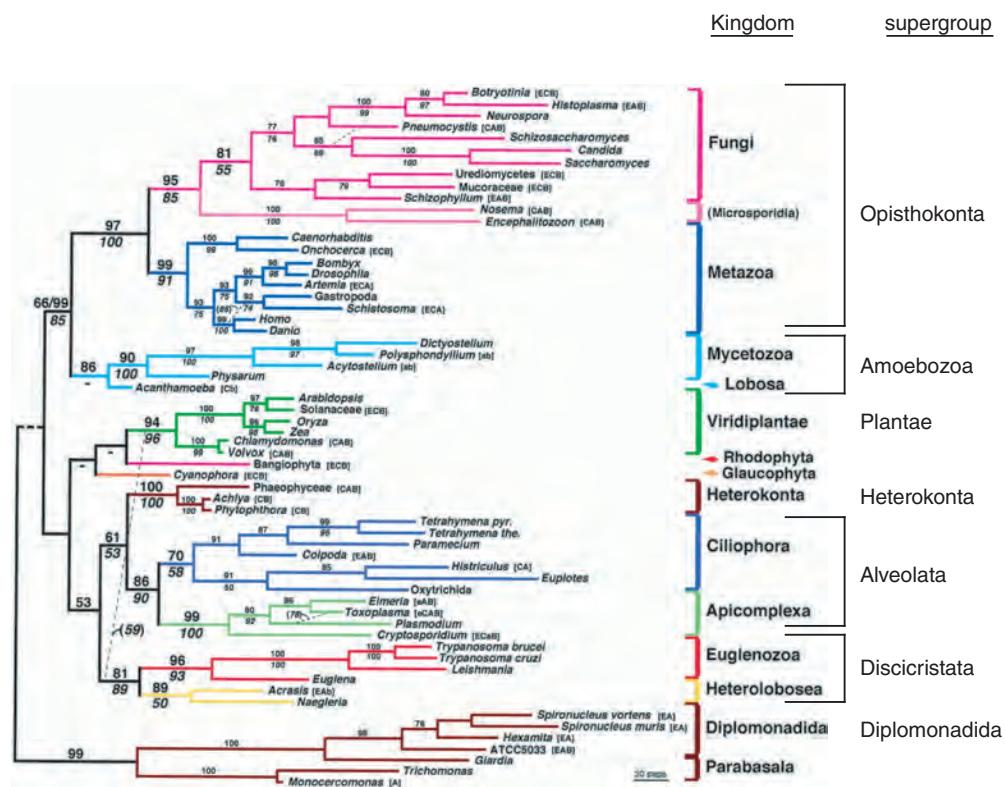
## LEARNING OBJECTIVES

Upon completing this chapter you should be able to:

- list the major groups of eukaryotes;
- define the key genomic features of selected eukaryotes including genome size and number of genes;
- provide examples of whole-genome duplication in the eukaryotes and discuss its significance; and
- provide a general time-line for the last common ancestor between humans and a range of animals from insects to primates.

## INTRODUCTION

In this chapter we explore individual eukaryotic genomes, from parasites to primates. We refer to a phylogenetic tree of the eukaryotes that was produced by Baldauf *et al.* (2000; Fig. 19.1). This tree was created by parsimony analysis using four concatenated protein sequences: elongation factor 1a (EF-1 $\alpha$ ); actin;  $\alpha$ -tubulin; and  $\beta$ -tubulin. We discussed



**FIGURE 19.1** A phylogeny of eukaryotes based on parsimony analysis of concatenated protein sequences. The proteins analyzed were EF-1 $\alpha$  (abbreviated E in tree), actin (C),  $\alpha$ -tubulin (A), and  $\beta$ -tubulin (B). This tree may be compared to the eukaryotic portion of the global tree of life based upon small-subunit ribosomal RNA sequences (Fig. 15.1). In this tree, 14 kingdoms are indicated as well as seven supergroups. One of the supergroups, Opisthokonta, includes fungi and microsporidia (Chapter 18) and metazoa (vertebrate and invertebrate animals). The tree was constructed by maximum parsimony (with bootstrap values indicated above the horizontal branches) and by maximum-likelihood analysis of second-codon-position nucleotides. For taxa with missing data, the sequences used are indicated in brackets, for example, [EAB]. Adapted from Baldauf *et al.* (2000), with permission from AAAS and S. Baldauf.

fungal genomes in Chapter 18; they are represented in a group that is adjacent to the metazoa (animals). We examine representative organisms in this tree, moving from the bottom up. This includes the diplomonad *Giardia lamblia* and other protozoans, such as the malaria parasite, *Plasmodium falciparum*; the plants, including the first sequenced plant genome (that of the thale cress, *Arabidopsis thaliana*) and rice (*Oryza sativa*); and the metazoans, from worms and insects to fish and mammals. We address the human genome in Chapter 20.

Following the outline introduced in Chapter 15, we consider five aspects of various genomes.

1. *Cataloguing information* includes describing the complete sequence of each chromosome, annotating the DNA to identify and characterize noncoding DNA, and identifying protein-coding genes and other noncoding genes. We survey chromosome number and structure (such as regions of duplication or deletion). This chapter provides a large amount of information about genome sizes. In many cases the exact size of a genome in megabases (Mb) or the exact number of genes are unknown; in

some cases, even the number of chromosomes is unknown. A goal of this chapter is to provide a survey of currently available information about eukaryotic genomes that orients you to the scales of genome sizes.

2. *Comparative genomics* is an essential part of any genome analysis. The availability of draft (or finished) genome sequences from closely related species permits a series of questions to be addressed about recent evolutionary changes such as lineage-specific expansions or contractions of gene families. The availability of distantly related species (such as fish genome sequences that last shared a common ancestor with humans over 400 million years ago or MYA) permits different kinds of questions to be addressed, such as the presence of conserved gene structures and regulatory elements.
3. *Biological principles* can be explored through genome sequences. For example, the genome of an underwater sea urchin unexpectedly encodes receptors that in other animals facilitate hearing and chemoreception, suggesting unsuspected sensory abilities of these animals. In general, genome sequence analysis can be used in an attempt to relate the genomic sequence to the phenotype of the organism. This phenotype includes an organism's strategies for adaptation to its environment, evolution, metabolism, growth, development, maintenance of homeostasis, and reproduction.
4. *Analysis of genomic sequences* offers a unique perspective on human disease (and diseases afflicting other organisms). In the case of many eukaryotes – from the protozoans such as *Plasmodium* to pathogenic fungi and parasitic worms – we want to understand the genetic basis of how the organism causes disease and how we can counterattack. At present, there are almost no vaccines available to prevent diseases caused by any eukaryotic parasites that infect humans, including protozoans (such as trypanosomes) and helminths (parasitic nematodes). The availability of whole-genome sequences may provide clues as to which antigens are promising targets for vaccine development and pharmacological intervention. For example, predicted secreted surface proteins can be expressed in bacteria and used to immunize mice in order to develop potential vaccines (Fraser *et al.*, 2000).
5. *Bioinformatics approaches* are constantly evolving, such as techniques for whole-genome sequencing and assembly as well as analytic tools. Analysis of genomes involves the use of next-generation sequencing as well as many of the tools we introduced in Chapters 2–7, including BLAST and molecular phylogeny. In Part I we discussed many of the complexities of multiple sequence alignment and phylogeny, and showed that the same raw data can be used to generate many alternative results. As you read about various genomes in this chapter, accession numbers (for genome projects and/or genes and proteins) are provided that will allow you to independently analyze many sequence analysis problems.

A phylogenetic description of the eukaryotes is essential for our understanding of both evolutionary processes that shaped the development of species and the diversity of life today. Evolutionary reconstructions that are based on molecular sequence data typically use small-subunit ribosomal RNA because it has many sites that are phylogenetically informative across all life forms (Van de Peer *et al.*, 2000). We saw an example of such a tree in **Figure 15.1**. However, there is no uniform consensus on the optimal approach to making a tree (Box 19.1; Chapter 7). For other phylogenetic trees of the eukaryotes, differing in some details from **Figure 19.1**, see Keeling (2007) in an introduction to the *Giardia lamblia* genome project and Embley and Martin (2006).

The word protozoan derives from the Greek *proto* ("early") and *zōion* ("animal"). This contrasts with the word metazoan (animal) from the Greek *meta* ("after"; at a later stage of development) and *zōion*.

## BOX 19.1 INCONSISTENT PHYLOGENIES

It is important to note that many phylogenetic reconstructions are inconsistent with each other. There are three main sources of conflicting results (Philippe and Laurent, 1998):

1. Gene duplication followed by random gene loss can cause artifacts in tree reconstruction. This occurred at the whole-genome level in yeast (Chapter 18) and other eukaryotes such as plants and fish.
2. Lateral gene transfer can confuse phylogenetic interpretation (Chapter 17).
3. The technical artifact of long branch chain attraction can confuse phylogenetic analyses. This is a phenomenon where the longest branches of a tree are grouped together, regardless of the true tree topology (Fig. 7.27). It is essential to account for differences in substitution rates among sites within a molecule. Reyes *et al.* (2000) consider this problem in their phylogeny of the order Rodentia.

Researchers often overcome these potential problems by concatenating multiple protein (or nucleic acid) sequences. For example, the tree in **Figure 19.1** is based on four concatenated proteins. With the advent of whole-genome sequencing, it has become common to identify thousands of 1:1 orthologs among multiple species for phylogenetic analysis.

## PROTOZOANS AT BASE OF TREE LACKING MITOCHONDRIA

The eukaryotes include deep-branching protozoan species from the parabasala (e.g., *Trichomonas*), diplomonadida (such as *Giardia*), discicristata (e.g., *Euglena*, *Leishmania*, and *Trypanosoma*), alveolata (e.g., *Toxoplasma* and *Plasmodium*), and heterokonta (Fig. 19.1). We begin at the bottom of the tree of **Figure 19.1** by describing *Trichomonas* and *Giardia*.

There is strong evidence that mitochondrial genes, present in most eukaryotes, are derived from an  $\alpha$ -proteobacterium (see Chapter 15). Previously, it was hypothesized that deep-branching organisms such as *Giardia* and *Trichomonas* lack mitochondria. They were thought to have evolved from other eukaryotes prior to the symbiotic invasion of an  $\alpha$ -proteobacterium. However, analyses of *Giardia* and *Trichomonas* suggest the presence of mitochondrial genes (Embley and Hirt, 1998; Lloyd and Harris, 2002; Williams *et al.*, 2002). Some protozoans (including trichomonads and ciliates) lack typical mitochondria but have a derived organelle called the hydrogenosome. This double-stranded structure produces adenosine triphosphate (ATP) and molecular hydrogen via fermentation.

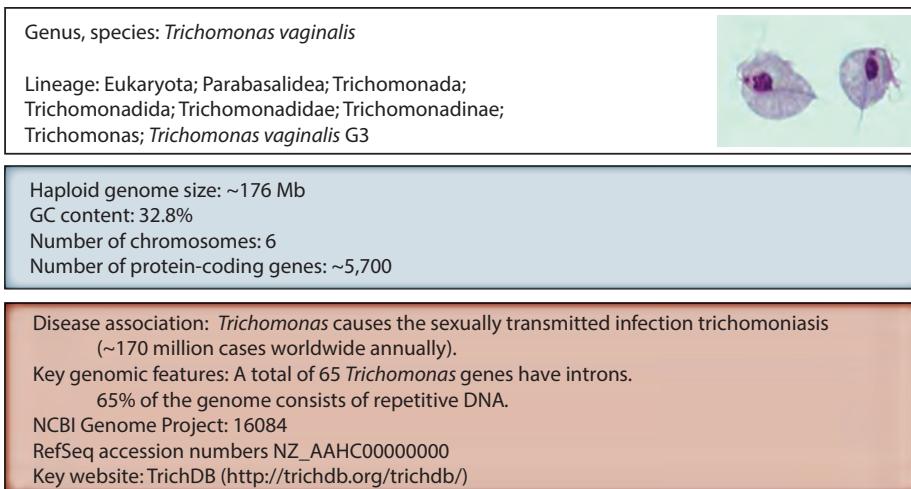
### *Trichomonas*

*Trichomonas vaginalis*, a flagellated protist and member of the parabasalids, is a sexually transmitted pathogen (Fig. 19.2). The World Health Organization estimates that there are ~170 million cases annually worldwide. *Trichomonas* is a single-celled organism that resides in the genitourinary tract where it phagocytoses vaginal epithelial cells, erythrocytes, and bacteria. Its genome of ~176 Mb has several remarkable features (Carlton *et al.*, 2007; reviewed in Conrad *et al.*, 2013): 62% of the genome consists of repetitive DNA, confounding efforts to characterize the genome architecture. Many of these repeats are of viral, transposon, or retrotransposon origin. There are 60,000 predicted protein-coding genes, one of the highest numbers among all life forms. Several gene families have undergone massive expansion such as protein kinases ( $n = 927$ ), the *BspA-like* gene family ( $n = 658$ ), and small GTPases ( $n = 328$ ). The *BspA-like* proteins are surface antigens that participate in host cell adherence and aggregation. *T. vaginalis* has apparently acquired 152 genes by lateral gene transfer from bacteria that thrive in the intestinal flora; most of these genes encode metabolic enzymes.

Analysis of the genome sequence by Carlton *et al.* suggests mechanisms by which *T. vaginalis* obtains its energy, functions as a parasite adhering to and invading host cells, and degrades proteins (via a complex degradome; Carlton *et al.*, 2007; Hirt *et al.*, 2011).

The microsporidia such as *Encephalitozoon* used to be classified as deep-branching eukaryotes. Subsequent analysis of the complete *E. cuniculi* genome revealed that this microsporidial parasite is closely related to the fungi, as described in Chapter 18.

The *Trichomonas* Genomics Resource is online at <http://trichdb.org/trichdb/> (WebLink 19.1).



**FIGURE 19.2** The Parabasala (see Fig. 19.1) are protozoans including *Trichomonas vaginalis*. Photograph from the Centers for Disease Control (CDC) Parasite Image Library ([http://www.dpd.cdc.gov/dpdx/HTML/Image\\_Library.htm](http://www.dpd.cdc.gov/dpdx/HTML/Image_Library.htm)) shows two trophozoites obtained from in vitro culture. Reproduced with permission from CDC-DPDx.

Tools to analyze this genome are available at the *Trichomonas* Genomics Resource, TrichDB (Aurrecoechea *et al.*, 2009a).

### *Giardia lamblia*: A Human Intestinal Parasite

*Giardia lamblia* (also called *Giardia intestinalis*) is a protozoan, water-borne parasite that lives in the intestines of mammals and birds (Adam, 2001). It is the cause of giardiasis, the most frequent source of nonbacterial diarrhea in North America. Like some other unicellular protozoans, *Giardia* lack not only mitochondria but also peroxisomes (responsible for fatty acid oxidation) and nucleoli. The genome of *Giardia* could therefore reflect the adaptations that led to the early emergence of eukaryotic cells.

The *Giardia* haploid genome is ~11.7 Mb (Morrison *et al.*, 2007; Upcroft *et al.*, 2010; Fig. 19.3). Each cell has two morphologically identical nuclei, each nucleus having five chromosomes ranging from 0.7 to over 3 Mb. A total of 6470 open reading frames (ORFs) were identified, spanning 77% of the genome, with 1800 overlapping genes and an additional 1500 ORFs spaced within 100 nucleotides of an adjacent ORF.

As we consider the genomes of various eukaryotes a consistent theme is that transposable elements are extremely abundant, occupying half the entire human genome (Chapter 20) and causing massive genomic rearrangements. In order to understand their origins and their function, it is therefore of interest to find eukaryotes that lack these elements. *Giardia* provides such an example. Arkhipova and Meselson (2000) examined 24 eukaryotic species for the presence of two major classes of transposable elements (retrotransposon reverse transcriptases and DNA transposons). They found them present in all species except bdelloid rotifers, an asexual animal. Deleterious transposable elements thrive in sexual species, but they are unlikely to propagate in asexual species because of strong selective pressure against having active elements. Further inspection of the asexual *Giardia* by Arkhipova and Morrison (2001) revealed just three retrotransposon families. One of these is inactive and the other two are telomeric. This location could provide a buffer between protein-coding genes and the telomeres, and these elements could contribute to the ability of *Giardia* to vary the length of its chromosomes in response to environmental pressures; for example, chromosome 1 can expand from 1.1 to 1.9 Mb (Pardue *et al.*, 2001).

Organisms that lack peroxisomes could provide us insight into fatty acid metabolism or other metabolic processes. This in turn could prove helpful to our understanding of human diseases that affect such organelles. The most common human genetic disorder affecting peroxisomes is adrenoleukodystrophy, caused by mutations in the *ABCD1* gene (RefSeq accession NM\_000033.3). Does *Giardia* have an ortholog of this gene?

*Giardia* was the first parasitic protozoan of humans observed with a microscope by Antony van Leeuwenhoek (in 1681). The diplomonadida are also called diplomonads. This group includes the family Hexamitidae, which further includes the genus *Giardia*. Information on *Giardia* is available at the US FDA (<http://www.fda.gov/Food/FoodborneIllnessContaminants/CausesOfIllnessBadBugBook/ucm070716.htm>, WebLink 19.2) and CDC websites (<http://www.cdc.gov/healthywater/swimming/rwi/illnesses/giardia.html>, WebLink 19.3).

The *Giardia* genome project website is at <http://www.giardiadb.org/giardiadb/> (WebLink 19.4); see also Aurrecoechea *et al.* (2009a).

Genus, species: *Giardia lamblia*

Lineage: Eukaryota; Diplomonadida; Hexamitidae; Giardiinae; Giardia; *Giardia lamblia* ATCC 50803

Haploid genome size: 12 Mb GC content: 49% Number of chromosomes: 5 Number of protein-coding genes: 6,470 Number of genes per kilobase: 0.58	
--	---

Disease association: *Giardia* causes ~100 million infections annually, and is the most prevalent parasitic protist in the United States.

Key genomic features: *Giardia* lacks mitochondria, hydrogenosomes, and peroxisomes. The organism has two similar, active, diploid nuclei. The genome encodes simplified machinery for DNA replication, transcription and RNA processing. There are no Krebs cycle proteins and few genes encoding proteins involved in amino acid metabolism.

NCBI Genome Project: 1439  
Project accession number: AACB02000000  
Key website: GiardiaDB (<http://www.giardiadb.org>)

**FIGURE 19.3** The Diplomonadida (see Fig. 19.1) are protozoans including *Giardia*. Image shows three trophozoites stained with Giemsa (from the CDC Parasite Image Library, [http://www.dpd.cdc.gov/dpdx/HTML/Image\\_Library.htm](http://www.dpd.cdc.gov/dpdx/HTML/Image_Library.htm)). Each protist has two prominent nuclei. Reproduced with permission from CDC-DPDx.

Tsetse flies are insects that feed on vertebrate blood. To obtain additional nutrients beyond what is available in blood, tsetse flies harbor two obligate intracellular bacteria: *Wigglesworthia glossinidia* and *Sodalis glossinidius*. The *W. glossinidia* genome (RefSeq accession NC\_004344.2) has a reduced genome size of only 700,000 base pairs (Akman *et al.*, 2002). For information on sleeping sickness, including the *T. brucei* lifecycle, see the Centers for Disease Control and Prevention (CDC) website at <http://www.cdc.gov/parasites/sleepingsickness/> (WebLink 19.5). A *Trypanosoma cruzi* Genome Initiative Information Server from the Oswaldo Cruz Institute, Brazil is available at <http://www.dbbm.fiocruz.br/TcruziDB> (WebLink 19.6). The Wellcome Trust Sanger Institute *T. brucei* website is <http://www.sanger.ac.uk/resources/downloads/protozoa/trypanosoma-brucei.html> (WebLink 19.7).

Another basic question about eukaryotic genomes is the origin of introns. Spliceosomal introns occur commonly in the “crown group” of eukaryotes (the kingdoms Animalia, Plantae, and Fungi). However, their presence in the earliest branching protzoa has been disputed (Johnson, 2002), and introns have not been detected in parabasalids such as *Trichomonas*. Nixon *et al.* (2002) identified a 35-bp intron in a gene encoding a putative [2Fe-2S] ferredoxin, and analysis of the draft genome sequence by Morrison *et al.* (2007) identified three more. Simpson *et al.* (2002) also identified several introns in *Carpediemonas membranifera*, a eukaryote thought to be a close relative of *Giardia*. These findings suggest that if introns were a eukaryotic adaptation, they arrived early in evolution and possibly in the last common eukaryotic ancestor.

We introduce nucleomorph genomes below (see section entitled “Nucleomorphs”); they are the functional, remnant nuclei of algal endosymbionts in several eukaryotic lineages. The first four sequenced nucleomorph genomes, all of which have undergone severe reduction in size to under a megabase, have 0, 2, 17, and 24 introns (Moore *et al.*, 2012).

## GENOMES OF UNICELLULAR PATHOGENS: TRYpanosomes AND LEISHMANIA

### Trypanosomes

There are about 20 species in the protozoan genus *Trypanosoma* (reviewed in Donelson, 1996). Two of these are pathogenic in humans (Cox, 2002). *Trypanosoma brucei* subspecies cause several forms of sleeping sickness, a fatal disease that infects hundreds of thousands of people in Africa (Fig. 19.4). *Trypanosoma cruzi* causes Chagas’ disease, prevalent in South and Central America. The adverse impact of these trypanosomes is even greater because they also afflict livestock. Tsetse flies or other insects transmit the trypanosomes to humans.

Genus, species: *Trypanosoma brucei*  
*Trypanosoma cruzi*  
*Leishmania major* (Friedlin strain)

Lineage: Eukaryota; Euglenozoa; Kinetoplastida (order);  
 Trypanosomatidae (family); Trypanosoma



	<i>T. brucei</i>	<i>T. cruzi</i>	<i>L. major</i>
Haploid genome size:	35 Mb	60 Mb	32.8 Mb
GC content	46.4%	51%	59.7%
Number of chromosomes:	11*	~28 (variable)	36
Number of genes (incl. pseudogenes)	9,068	~12,000	8,311

\* includes ~100 mini- and intermediate size chromosomes

Disease association: *T. brucei* causes trypanosomiasis (sleeping sickness). The incidence is 300,000 to 500,000 cases per year. *T. cruzi* causes Chagas disease in humans; 16–18 million people are infected with 21,000 deaths per year. Leishmaniasis is an infectious disease with 2 million new cases annually and 350 million people at risk; 20 *Leishmania* species infect humans. No vaccines and few drugs are available.

Key genomic features: These three species share a conserved core proteome of ~6,200 proteins.

NCBI Genome project identifiers: 11756 (*T. brucei*), 11755 (*T. cruzi*), 10724 (*L. major*).

Key websites: <http://www.genedb.org/Homepage/Tbruceibrucei927>  
<http://www.sanger.ac.uk/resources/downloads/protozoa/trypanosoma-brucei.html>

**FIGURE 19.4** The Euglenozoa (see Fig. 19.1) include the kinetoplastid parasitic protozoa *Trypanosoma brucei*, *T. cruzi*, and *Leishmania major*. The image (from the CDC Parasite Image Library) shows a *T. brucei* from a blood smear in the trypanostigote stage. There is a centrally located nucleus, a small kinetoplast at the posterior end (upper right), and an undulating membrane with a flagellum exiting the body at the anterior end. Length in the range 14–33 µm. Reproduced with permission from CDC-DPDx.

Berriman *et al.* (2005) reported the genome sequence of *T. brucei*. The genome is 26 Mb, although its size varies by up to 25% in different isolates (reviewed in El-Sayed *et al.*, 2000). There are at least 11 pairs of large, diploid, nuclear chromosomes (ranging in size from about 1 Mb to >6 Mb). Additionally, there are variable numbers of intermediate chromosomes (200–900 kb), and there are about 100 linear minichromosomal DNA molecules (50–150 kb). Some of these minichromosomes contain a 177 base pair repeat that comprises more 90% of the total sequence (El-Sayed *et al.*, 2000). The genome includes 9068 predicted genes, of which about 900 are pseudogenes and ~1700 are specific to *T. brucei*.

Another remarkable feature of trypanosomes is the presence of a massive network of circular rings of mitochondrial DNA, termed kinetoplast DNA. Thousands of rings of kinetoplast DNA interlock in a shape resembling medieval armor (Shapiro and Englund, 1995). Kinetoplast DNA occurs as maxicircles (present in several dozen copies) and minicircles (present in thousands of copies). These include a universal minicircle sequence of 12 nucleotides that serves as a replication origin (Morris *et al.*, 2001).

For a major portion of their life-cycles, trypanosomes thrive in the bloodstreams of their hosts. They evade assault from the immune system by densely coating their exteriors with variant surface glycoprotein (VSG) homodimers. There are over 1000 VSG genes and pseudogenes encoded in the *T. brucei* genome, of which only one is expressed at a time (Berriman *et al.*, 2005; reviewed by Taylor and Rudenko, 2006). Remarkably, fewer than 7% of these encode functional proteins, while 66% encode full-length pseudogenes and the remainder are gene fragments or otherwise atypical. Most of the VSG genes are located in subtelomeric arrays of 3–250 copies. Taylor and Rudenko suggest that the pseudogenes could be advantageous in the generation of antigenic diversity during chronic infections of the bloodstream. The limited number of intact VSG genes could

The accession number of a typical VSG protein from *T. brucei* is XP\_822273.1. Try a DELTA-BLAST search using it as a query.

The *Trypanosoma brucei* GeneDB is available at <http://www.genedb.org/Homepage/Tbruceibrucei927> (WebLink 19.8). It is part of the GeneDB pathogen database that includes resources for trypanosomes, *Leishmania*, Apicomplexans, as well as helminths and parasite vectors (Logan-Klumpler *et al.*, 2012).

See Problem (19.3) for an exercise on a trypanosome universal minicircle binding protein. For an example of a maxicircle sequence and the genes it encodes, see GenBank accession M94286.1.

be used, but also segmental gene conversion of pseudogenes could create novel, intact, mosaic, VSG genes.

*T. cruzi* infects 16–18 million people and is the cause of 21,000 deaths per year from Chagas disease. El-Sayed *et al.* (2005b) reported the diploid genome sequence of two different haplotypes that averaged 5.4% sequence divergence. The diploid genome size is ~106–111 Mb and is predicted to contain 22,570 genes, while the haploid genome contains ~12,000 genes. There is a notable large family of 1377 copies of mucin-associated surface protein (*masp*) genes, which may be involved in immune system evasion.

### *Leishmania*

*Leishmania major* is another deadly protozoan parasite in the Euglenozoa (Fig. 19.1). Twenty different species of *Leishmania* cause the disease leishmaniasis, for which there is no effective vaccine and limited pharmacological intervention available. A total of 150 million people are afflicted, with a complex host-pathogen interface (Kaye and Scott, 2011). The various *Leishmania* species have 34–36 chromosomes (Myler *et al.*, 2000). While the Old World groups *L. major* and *L. donovani* have 36 chromosome pairs (in the range 0.28–2.8 Mb), New World groups *L. mexicana* and *L. braziliensis* have undergone chromosomal fusions (Chapter 8) and have 34 or 35 chromosomal pairs.

The *Leishmania major* genome is about 34 Mb with 36 chromosomes (0.3–2.5 Mb). The nucleotide sequence was determined for chromosome 1 (the smallest chromosome) and was found to have a remarkable genomic organization (Myler *et al.*, 1999). The first 29 genes (from the left telomere) are all transcribed from the same DNA strand, while the remaining 50 genes are all transcribed from the opposite strand. This polarity is unprecedented in eukaryotes and resembles bacterial-like operons. It has a 257-kb region that is filled with 79 protein-coding genes (~1 gene per 3200 base pairs). Ivens *et al.* (2005) reported the *L. major* genome sequence (Fig. 19.4). There are 8272 predicted protein-coding genes including ~3000 that cluster into 662 different families of paralogs. These families arose principally by tandem gene duplication. The *L. major* genome encodes relatively few proteins involved in transcriptional control, and gene duplication may be a mechanism for increasing expression levels.

In addition to *L. major* (32.8 Mb), Peacock *et al.* (2007) sequenced the genomes of *Leishmania infantum* (32.0 Mb) and *L. braziliensis* (32.0 Mb). The three genomes share a comparable GC percentage and number of predicted genes. *L. major* and *L. braziliensis* diverged between 20 and 100 million years ago; this broad range reflects uncertainty as to whether the *Leishmania* genus speciated due to migration events or to the breakup of the supercontinent Gondwana (Fig. 15.3). *L. braziliensis* has 35 chromosomes rather than 36 because of the fusion of chromosomes 20 and 34. There is conserved synteny for more than 99% of the genes across the three genomes, and the average nucleotide and amino acid identities are high (e.g., 92% amino acid identity between *L. major* and *L. infantum*). While many pathogenic protozoans have large gene families involved in immune evasion localized to subtelomeric regions, such families are not evident in the *Leishmania* species. Peacock *et al.* identified only 5 genes that are specific to *L. major*, 26 to *L. infantum*-specific genes, and ~47 to *L. braziliensis*.

With the further sequencing of *Leishmania mexicana* and *Leishmania donovani* genomes, comparative genomics continues to advance rapidly. Each of the five sequenced *Leishmania* genomes has a similar size (34–36 chromosomes, 30–33 Mb, >8000 protein coding genes). The availability of these sequences allows correction of gene models and discovery of previously unannotated genes (see Nirujogi *et al.*, 2014). Comparisons of the three trypanosomatid genomes of *L. major*, *T. brucei*, and *T. cruzi* have revealed a shared core of 6200 genes (El-Sayed *et al.*, 2005a). Some protein domains are specific

The World Health Organization offers information on leishmaniasis at <http://www.who.int/mediacentre/factsheets/fs375/en/> (WebLink 19.9). The *Leishmania major* Friedlin Genome Project at the Wellcome Trust Sanger Institute is available at <http://www.sanger.ac.uk/resources/downloads/protozoa/leishmania-major.html> (WebLink 19.10).

to just one group, such as the variant surface glycoprotein (VSG) expression site-associated domains (Pfam families PF03238 and PF00913) in *T. brucei*. Some domains appear to have expanded or contracted selectively and insertions, deletions, and substitutions occurred. However, there is a notable high overall gene conservation between the three species.

## THE CHROMALVEOLATES

The Chromalveolates are a supergroup of unicellular eukaryotes, distinct from the Excavates (such as *Giardia*). Many of them have cryptic mitochondria (e.g., hydrogenosomes rather than traditional mitochondria). They include six groups or phyla (Keeling, 2007): (1) the *Apicomplexa* consist of protozoan pathogens that invade host cells using a specialized apical complex (they are typically transmitted by an invertebrate vector such as mosquitoes or flies, and this phylum includes parasites such as *P. falciparum* and *Toxoplasma gondii*); (2) dinoflagellates include a cause of paralytic shellfish poisoning, *Alexandrium*; (3) ciliates include *Paramecium* and *Tetrahymena thermophila*; (4) heterokonts; (5) haptophytes; and (6) cryptomonads. In the tree of Figure 19.1, these groups are organized as the Apicomplexa, Ciliophora, and Heterokonta. In the following sections of this chapter we will turn to the Viridiplantae (plants), the Mycetozoa, and the Metazoa (animals).

### Malaria Parasite *Plasmodium falciparum*

Malaria kills over a million people each year (mostly children in Africa) and almost 500 million people are newly infected each year. Worldwide malaria deaths were 995,000 in 1980; 1,817,000 in 2004; and 1,238,000 in 2010 (Murray *et al.*, 2012). It is caused by the apicomplexan parasite *Plasmodium falciparum*. While there are 120 species of *Plasmodium*, only four typically infect humans: *P. falciparum* (most responsible for mortality); *P. vivax* (most responsible for morbidity); *P. ovale*; and *P. malariae*. The main vector for malaria in Africa is the mosquito, *Anopheles gambiae*.

*Plasmodium falciparum* has a complex lifestyle, contributing to the challenge of developing a successful vaccine (Cowman and Crabb, 2002; Long and Hoffman, 2002; Winzeler, 2008). *Plasmodium* resides in the salivary glands and gut of the mosquito *A. gambiae*. When a mosquito bites a human, it introduces the parasite in the sporozoite form that infects the liver. *Plasmodium* then matures to the merozoite form, which attaches to and invades human erythrocytes through host cell receptors. Within erythrocytes, trophozoites form. Some merozoites transform into gametocytes, which are captured when mosquitoes feed on infected individuals. A goal of sequencing the *P. falciparum* genome is to find gene products that function at selective stages of the parasite life cycle, offering targets for drug therapy or vaccine development.

Historically, in much of the twentieth century, malaria was treated with the inexpensive drugs chloroquine and pyrimethamine-sulphadoxine. *Plasmodium* has become broadly resistant and currently the artemisinins are the only effective class of antimalarial drug. In one drug screen Gamo *et al.* (2010) tested 2 million compounds for inhibition of *P. falciparum*, identifying several thousands of new candidates. The mechanism of action of these drugs can be understood in the context of the genome, providing a motivation to sequence complete genomes. Furthermore, it is possible to combine chemical screens with genome-wide association studies (GWAS; Chapter 21). For example, Yuan *et al.* (2011) performed high-throughput chemical screening to identify candidates, and performed GWAS to identify specific gene mutations associated with responsiveness of *Plasmodium* to drugs. This strategy also allowed Cheeseman *et al.* (2012) to identify loci conferring resistance to artemisinin.

The name Apicomplexa derives from a characteristic apical complex of microtubules. Read more about apicomplexans at <http://www.ucmp.berkeley.edu/protista/apicomplexa.html> (WebLink 19.11) or <http://www.tulane.edu/~wiser/protozoology/notes/api.html> (WebLink 19.12). For online facts on malaria, see <http://malaria.wellcome.ac.uk/> (WebLink 19.13) and <http://www.who.int/mediacentre/factsheets/fs094/en/> (WebLink 19.14).

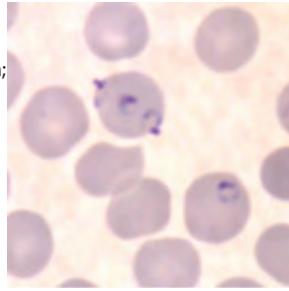
Charles Louis Alphonse Laveran won a Nobel Prize in 1907 for his work on malaria-causing parasites ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1907/](http://nobelprize.org/nobel_prizes/medicine/laureates/1907/), WebLink 19.15). Earlier, Ronald Ross was awarded a Nobel Prize for his studies of malaria ([http://nobelprize.org/nobel\\_prizes/medicine/laureates/1902/](http://nobelprize.org/nobel_prizes/medicine/laureates/1902/), WebLink 19.16).

Genus, species: *Plasmodium falciparum*

Selected lineages: Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Haemosporida; Plasmodium; Plasmodium (Laverania); *Plasmodium falciparum*

Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Piroplasmida; Theileriidae; Theileria; *Theileria annulata*

Eukaryota; Alveolata; Apicomplexa; Coccidia; Eucoccidiorida; Eimeriorina; Sarcocystidae; Toxoplasma; *Toxoplasma gondii* RH



	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>P. falciparum</i> 3D7	22.8 Mb	19.4%	14	5,268	13173
<i>P. yoelli yoelli</i>	23.1 Mb	22.6%	14	5,878	1436
<i>Babesia bovis</i>	8.2 Mb	41.8%	4	3,671	18731
<i>Cryptosporidium hominis</i>	9.2 Mb	31.7%	8	3,956	13200
<i>Cryptosporidium parvum</i>	9.1 Mb	30.3%	8	3,886	144
<i>Theileria annulata</i>	8.4 Mb	32.5%	4	3,792	153
<i>Theileria parva</i>	8.3 Mb	34.1%	4	4,035	16138
<i>Toxoplasma gondii</i>	65 Mb	52.3%	9	8,032	16727

The *P. falciparum* genome was sequenced by a consortium including the Wellcome Trust Sanger Institute, The Institute for Genomic Research, the US Naval Medical Research Center (Maryland), and Stanford University. The genome of the slime mold *Dictyostelium discoideum* also has a high AT content (see “Social Slime Mold *Dictyostelium discoideum*” below).

A plastid is any photosynthetic organelle. The most well-known plastid is the chloroplast, found in green algae and land plants (Gilson and McFadden, 2001). See the section on “Plant Genomes” below.

PlasmoDB is at <http://www.plasmodb.org/> (WebLink 19.17). GeneDB includes a *P. falciparum* resource at <http://www.genedb.org/Homepage/Pfalciparum> (WebLink 19.18). ProtozoaDB is at <http://protozoadb.biowebdb.org/> (WebLink 19.19).

Selected divergence dates: the Apicomplexa lineage originated less than 1,000 million years ago. Disease association: Each of these apicomplexans is a parasite that causes disease in mammals.

*B. babesii* causes babesiosis, a tick-borne disease that threatens half the cattle in the world. *P. falciparum* causes malaria. *T. gondii* causes toxoplasmosis.

Key genomic features: *Theileria* parasites are the only eukaryotes that transform lymphocytes (and thus induce lymphoma).

Key websites: ApiDB for apicomplexans (<http://www.apidb.org/apidb/>); PlasmoDB for *Plasmodium* (<http://plasmodb.org/>).

**FIGURE 19.5** The Apicomplexa (see Fig. 19.1) include the malaria parasite *Plasmodium falciparum*. This image shows multiply infected red blood cells in thin blood smears (from the CDC Parasite Image Library). Reproduced with permission from CDC-DPDx.

The complete genome sequence of *P. falciparum* was reported by an international consortium (Gardner *et al.*, 2002; Fig. 19.5). The sequencing was extraordinarily challenging because the AT (adenine and thymine) content of the genome is 80.6% overall, which is among the highest for any eukaryotic genome. In intergenic regions and introns, the AT content reached 90% in some cases. A whole-chromosome (rather than a whole-genome) shotgun sequencing strategy was employed. With this approach, chromosomes were separated on pulsed-field gels, DNA was extracted, and shotgun libraries containing 1–3 kb of DNA were constructed and sequenced. The genome is 22.8 Mb, with 14 chromosomes in the range 0.6–3.3 Mb.

Gardner *et al.* (2002) identified 5268 protein-coding genes in *P. falciparum*. This is the same number as that predicted for *Schizosaccharomyces pombe* (Chapter 18), although the genome size is twice as large. There is one gene approximately every 4300 base pairs overall. Gene Ontology Consortium terms (Chapter 12) were assigned to about 40% of the gene products (~2100). However, about 60% of the predicted proteins had no detectable homology to proteins in other eukaryotes. These proteins are potential targets for drug therapies. For example, some are essential for the function of the apicoplast. This is a plastid, unique to Apicomplexa and homologous to the chloroplast, that functions in fatty acid and isoprenoid biosynthesis.

PlasmoDB (Aurrecoechea *et al.*, 2009b) is the centralized resource for *P. falciparum* genomic data. There are many complementary resources such as ProtozoaDB (Dávila *et al.*, 2008).

In addition to the initial *P. falciparum* genome project, the genomes of 9 *Plasmodium* species and over 30 strains have been sequenced as of 2014. A consortium sequenced the genome of the rodent malaria parasite, *Plasmodium yoelii yoelii* (Carlton *et al.*, 2002) and Hall *et al.* (2005) sequenced the genomes of the rodent malaria parasites *Plasmodium berghei* and *P. chabaudi*. Further projects addressed the genomes of *P. vivax* (Carlton *et al.*, 2008a), the simian and human parasite *P. knowlesi* (Pain *et al.*, 2008), and the chimpanzee parasite *P. reichenowi* (Jeffares *et al.*, 2007).

What is the significance of sequencing additional *Plasmodium* genomes? In the case of *P. yoelii yoelii*, *Plasmodium berghei*, and *P. chabaudi* this is an extremely important accomplishment because the complete life cycle of *P. falciparum* cannot be maintained *in vitro* while the rodent parasites can. The *P. yoelii yoelii* genome is 23.1 Mb and has 14 chromosomes, as does *P. falciparum*; the AT content is comparably high (77.4%). The genomes are also predicted to encode a comparable number of genes. When the full set of predicted *P. falciparum* proteins (5268) were searched against the predicted *P. yoelii yoelii* proteins (5878 proteins) by BLASTP searching (with an *E* value cutoff of  $10^{-15}$ ), 3310 orthologs were identified. These include vaccine antigen candidates known to elicit immune responses in exposed humans (Carlton *et al.*, 2002).

Having the genome sequences of *P. falciparum* and several rodent parasites available, how can bioinformatics and genomics approaches be used to understand the basic biology of these organisms? Data are now available on thousands of previously unknown genes, offering many new potential strategies to combat malaria (Hoffman *et al.*, 2002; Florent *et al.*, 2010).

- Sites of positive selection can be inferred through comparative genomics (Carlton *et al.*, 2008b). These are likely to be involved in host-parasite interactions. Further comparative studies have established a far higher genetic diversity in *P. vivax* than in *P. falciparum*, possibly impacting intervention strategies (Neafsey *et al.*, 2012).
- The apicoplast is a potential drug target. Zuegge *et al.* (2001) analyzed the amino terminal sequences of 84 proteins targeted to apicoplasts and 102 nonapicoplast (e.g., cytoplasmic, secretory, or mitochondrial) sequences. They used principal components analysis, neural networks, and self-organizing maps (Chapter 11) to build a predictive model for apicoplast targeting signals.
- Comparative genomics approaches yield important insight into the genome structure, gene content, and other genomic features of closely related species (Carlton *et al.*, 2008b). Carlton *et al.* (2001) compared ESTs and genome survey sequences (see Chapter 2) from *P. falciparum*, *P. vivax*, and *P. berghei*. As part of this analysis, they identified the most highly expressed genes such as the *rif* gene family of *P. falciparum* that is implicated in antigenic variation.
- Hall *et al.* (2005) measured synonymous versus nonsynonymous substitution rates in genes from three rodent *Plasmodium* species in comparison to *P. falciparum*. They measured gene expression, categorizing transcripts according to the four categories of: housekeeping; host-related; invasion, replication, and development-related; or stage-specific.
- A map of conserved synteny regions between *P. yoelii yoelii* and *P. falciparum*, covering over 16 Mb overall, provides insight into the evolution of these parasites. Carlton *et al.* (2002) used the MUMmer program (Chapters 16 and 17) to align protein-coding regions. The conserved synteny map reveals regions of conserved gene order, allows analysis of chromosomal break points, and confirms the absence of some genes (such as *var* and *rif* in *P. yoelii yoelii*).
- Genes that function in antigenic variation and immune system evasion can be investigated. In *P. vivax*, there are as many as 1000 copies of *vir*, a gene family localized to subtelomeric regions. *Plasmodium yoelii yoelii* has 838 copies of a related gene, *yir* (Carlton *et al.*, 2002).

The Prediction of Apicoplast Targeted Sequences (PATS) database is available at <http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php> (WebLink 19.20).

We encountered *vir* in Chapter 5 (Problem (5.2)) where we used BLASTP and DELTA-BLAST to evaluate the family.

Isoprenes are five-carbon chemical molecules that combine to form many thousands of natural compounds, including steroids, retinol, and odorants. (RBP and OBP are lipocalins that transport isoprenoids.)

- Several groups applied proteomics approaches to analyze the proteins of *P. falciparum* at four stages of the life cycle (sporozoites, merozoites, trophozoites, and gametocytes). Florens *et al.* (2002) identified 2415 expressed proteins, about half of which are annotated as hypothetical. An unexpected finding was that the *var* and *rif* genes – thought to be involved in immune system invasion – were abundantly present in the sporozoite stage. Together, these studies define stage-specific expression of proteins, suggesting possible protein functions. Proteomics approaches also validate the gene-finding approaches from genomic DNA. Lasonder *et al.* (2002) identified some protein sequences by mass spectrometry that were not initially predicted using gene-finding algorithms to analyze genomic DNA.
- It is possible to identify *Plasmodium* metabolic pathways as therapeutic targets (Gardner *et al.*, 2002; Hoffman *et al.*, 2002). All organisms studied to date synthesize isoprenoids using isopentyl diphosphate as a building block. An atypical pathway employed by some plants and bacteria involves 1-deoxy-D-xylulose 5-phosphate (DOXP). This DOXP pathway is absent in mammals. Jomaa *et al.* (1999) used TBLASTN (with a bacterial DOXP reductoisomerase protein as a query against a *Plasmodium* genomic DNA database) and found an orthologous *Plasmodium* gene. They showed that this protein is likely localized to the apicoplast and that *P. falciparum* survival is sensitive to low levels of two inhibitors of the enzyme. They further showed that these drugs have antimalarial activity in mice infected with *Plasmodium vinckeii*. This type of bioinformatics-based approach holds great promise in the search for additional antimalarial drugs.

## More Apicomplexans

There are 5000 species in the phylum Apicomplexa, causing a wide range of diseases by mechanisms that are now being elucidated through genome sequence analysis (reviewed in Roos, 2005). Other apicomplexan genomes that have been sequenced include the following (summarized in Fig. 19.5):

- *Babesia bovis*, the cause of tick fever in cattle, threatens livestock globally. Brayton *et al.* (2007) reported its 8.2 Mb genome sequence. It has extremely limited metabolic potential, lacking genes encoding proteins that are required for gluconeogenesis, the urea cycle, fatty acid oxidation, and heme, nucleotide, and amino acid biosynthesis. It therefore relies on its host for many nutrients, and the *B. bovis* genome encodes many transporters. Analogous to *Plasmodium falciparum*, its genome encodes about 150 copies of a polymorphic variant erythrocyte surface antigen protein (*ves1* gene) family.
- *Theileria annulata* and *T. parva* are tick-borne parasites that cause tropical theileriosis and East Coast fever, respectively, in cattle. Pain *et al.* (2005) and Gardner *et al.* (2005) reported their 8.4 Mb genome sequences. *T. parva* reversibly, malignantly transforms its host cell, the bovine lymphocyte, causing lymphoma; *T. annulata* transforms macrophages. The *T. parva* genome encodes about 20% fewer genes than *P. falciparum*, but it has a higher density of genes.
- *Cryptosporidium hominis* causes diarrhea and acute gastroenteritis. Unlike other Apicomplexans that are transmitted via an invertebrate host, *C. hominis* is transmitted by ingestion of oocysts in water. Xu *et al.* (2004) sequenced the 8.8 Mb *C. hominis* genome, while Abrahamsen *et al.* (2004) sequenced the 9.1 Mb genome of the related parasite *C. parvum* that infects humans and other mammals. Like *B. bovis* and many other parasites, these genomes have very limited metabolic capabilities and rely on host cells for nutrients.
- *Toxoplasma gondii* causes toxoplasmosis. The Centers for Disease Control estimates that 60 million people in the United States are infected, although most are asymptomatic.

After infection, oocysts and tissue cysts transform into tachyzoites and localize in neural and muscle tissue. Complete genome data have been deposited in NCBI Genome for three *T. gondii* strains, each with a genome size of 63–65 Mb, ~8000 genes, and 52% GC content. Yang *et al.* (2013) identified candidate variants that may underlie phenotypic differences among these strains.

The *T. gondii* database ToxoDB is available at <http://toxodb.org/toxo/> (WebLink 19.21) (Gajria *et al.*, 2008).

- *Hammondia hammondi* is an avirulent, close relative to *T. gondii*, sharing greater than 95% conserved synteny. Walzer *et al.* (2013) sequenced this genome and used comparative analyses to suggest which virulence factors are most relevant to *T. gondii*.

### Astonishing Ciliophora: *Paramecium* and *Tetrahymena*

Ciliates are unicellular eukaryotes that are part of the monophyletic alveolate clade that includes the Apicomplexans (see Fig. 19.1). Ciliates share two properties: they use vibrating cilia for locomotion and food capture; and they have two nuclei with separate germline and somatic functions (nuclear dimorphism). One nucleus is a diploid germinal micronucleus that undergoes meiosis and is therefore responsible for transmitting genetic information to the progeny (but is otherwise silent). The other is a polyhaploid somatic macronucleus that is responsible for gene expression. This macronucleus is lost with each generation and is replenished following meiosis and development of the micronuclear lineage. The reason these protozoa break apart and reassemble their somatic genomes at every sexual generation is to eliminate parasitic transposons or other mobile elements (Coyne *et al.*, 2012).

*Paramecium tetraurelia* is a ciliate that lives in freshwater environments. It has long served as a model organism for many aspects of eukaryotic biology. *Paramecium* has an unknown number of micronuclear chromosomes (>50 ; Fig. 19.6). As the macronuclear

Genus, species: <i>Paramecium tetraurelia</i> <i>Tetrahymena thermophila</i> <i>Sterkiella histriomuscorum</i> (also called <i>Oxytricha trifallax</i> )	<i>Sterkiella histriomuscorum</i> ( <i>Oxytricha trifallax</i> ) 
Lineage: Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Oligohymenophorea; Peniculida; Parameciidae; Paramecium; <i>Paramecium tetraurelia</i>	
Lineage: Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Oligohymenophorea; Hymenostomatida; Tetrahymenina; Tetrahymenidae; <i>Tetrahymena</i> ; <i>Tetrahymena thermophila</i>	

	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Paramecium tetraurelia</i> (macronuclear genome)	~72 Mb	28%	~200	39,642	18363
<i>Tetrahymena thermophila</i> (macronuclear genome)	~104 Mb	22%	~225	27,424	12564
<i>Sterkiella histriomuscorum</i> (macronuclear genome)	~50 Mb	not avail.	~24,500	~26,800	12857

Selected divergence dates: the ciliates diverged from other eukaryotes ~1,000 million years ago.

Key genomic features: *Paramecium* has a macronuclear nucleus (with somatic functions) and a diploid micronuclear nucleus (with germline functions). The gene content is extraordinarily high, and the genome underwent at least three whole genome duplications.

Key websites: <http://www.ciliate.org>; <http://paramecium.cgm.cnrs-gif.fr/>

**FIGURE 19.6** The Ciliophora (see Fig. 19.1) include *Paramecium* and *Tetrahymena*. In some classifications, the Apicomplexa and Ciliophora are grouped together to form the Alveolata.

Source: <http://www.k12summerinstitute.org/workshops/asset.html>. Courtesy of A. Bell.

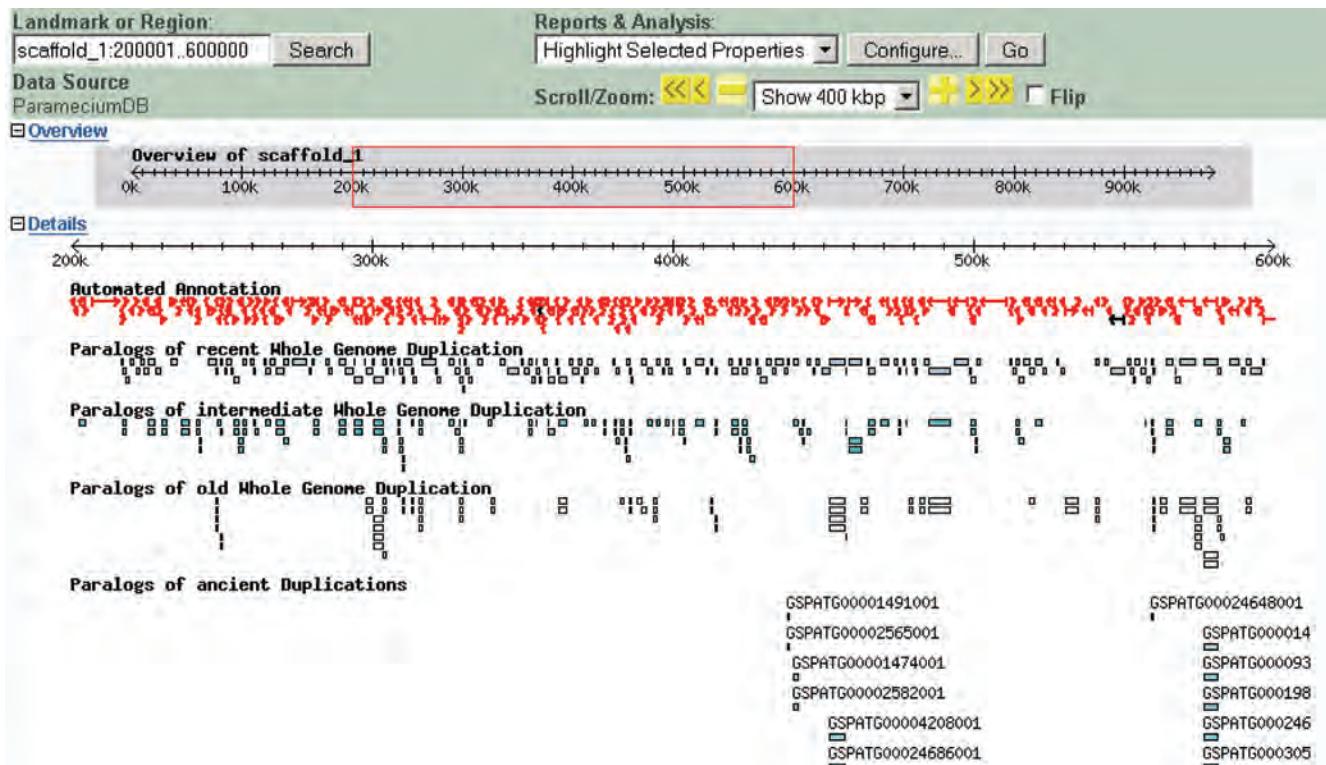
Chromatin diminution also occurs in nematodes (Chapter 8).

The *Paramecium* genome project website, including the ParameciumDB genome browser, can be viewed at <http://paramecium.cgm.cnrs-gif.fr/> (WebLink 19.22).

chromosome develops, it is amplified to ~800 copies and is extensively rearranged through a process of DNA elimination. Tens of thousands of unique copy elements are removed and, in a separate process, transposable elements and other repeats are deleted. This leads to a fragmented set of about 200 acentric chromosomes, ranging in size from ~50 kilobases to 981 kilobases. Aury *et al.* (2006) sequenced the *Paramecium* macronuclear genome which, although fragmented, is genetically homogenous because of the sexual process of autogamy by which it arose. The total coverage was 72 Mb, and most of the 188 largest scaffolds likely represent macronuclear chromosomes because they contain telomeric repeats.

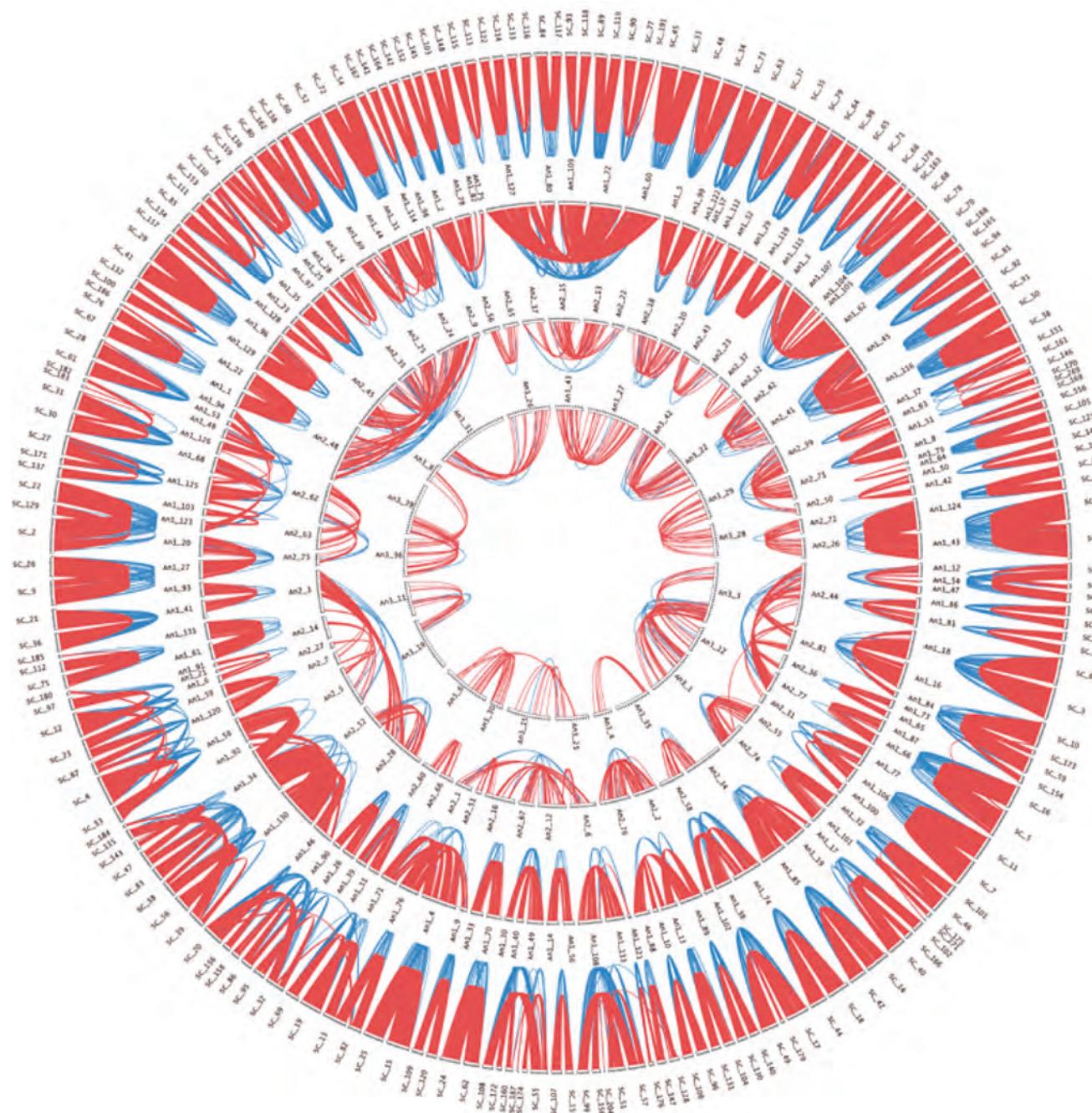
While the presence of two nuclei and the process of DNA rearrangement and elimination are extraordinary, another startling finding is that *Paramecium* encodes about 40,000 protein-coding genes (a far greater number than is found in animals or fungi). The genome sequencing process resulted in the creation of several hundred scaffolds. As we view the scaffold 1, corresponding to the longest chromosome observed by pulsed-field gel electrophoresis, we can see the compact nature of the coding portion of the genome (Fig. 19.7). Across the genome, 78% of the nucleotides contain genes and the intergenic regions average 352 bases.

Yet another surprising finding is the series of three whole-genome duplications that Aury *et al.* (2006) inferred (Fig. 19.8). All proteins were searched against each other using the Smith–Waterman algorithm (Chapter 3). Two-thirds of the predicted proteins occur in paralog pairs, maintaining conserved synteny across large portions of the chromosomes. The remaining third of the proteins presumably lost their duplicates after the whole-genome duplication event(s). The situation contrasts with the fungi (Chapter 18) and plant and fish genomes (see “*Arabidopsis thaliana* Genome” and “450 MYA: Vertebrate Genomes



**FIGURE 19.7** The *Paramecium* genome is proposed to have undergone at least three whole-genome duplications. The longest chromosome (scaffold 1 of the genome assembly) is viewed in the genome browser of ParameciumDB. A region of 400,000 base pairs is displayed, and the annotation tracks show the conservation to many paralogs reflecting recent, intermediate, and old whole-genome duplications.

Source: <http://paramecium.cgm.cnrs-gif.fr/>



**FIGURE 19.8** Whole-genome duplication in the ciliate *Paramecium tetraurelia* is inferred by analysis of protein paralogs. The outer circle displays all chromosome-sized scaffolds from the genome sequencing project. Lines link pairs of genes with a “best reciprocal hit” match. The three interior circles show the reconstructed ancestral sequences obtained by combining the paired sequences from each previous step. The inner circles are progressively smaller and reflect fewer conserved genes with a smaller average similarity.

Source: Aury et al. (2006). Reproduced with permission from Macmillan Publishers.

of Fish” below) in which whole-genome duplication events are followed by rapid gene loss and large-scale chromosomal rearrangements. By inferring ancestral blocks and then iteratively repeating the within-proteome alignments to search for conserved blocks sharing progressively less conservation, Aury *et al.* inferred the occurrence of three whole-genome duplications (Fig. 19.8). For a discussion of the software used to make the figure, see Box 19.2.

## BOX 19.2 GRAPHICALLY REPRESENTING WHOLE-GENOME DUPLICATIONS

We introduced the ideogram as a representation of a karyotype in Chapter 8. Traditionally, linear eukaryotic chromosomes are depicted as straight bars. However, when depicting the relationships between genes (or proteins or other elements) on multiple chromosomes, the patterns of relationships can be so complex that the visual presentation is confusing. Circular plots offer a concise method of viewing relationships between chromosomal elements. **Figure 19.8** shows an example of *Paramecium* chromosomes made by Aury *et al.* (2006) using Circos software developed by Martin Krzywinski (available as free software at <http://circos.ca/?home>). This website also offers a tutorial and a gallery of visually stunning samples.

Chromowheel is a related tool, developed by Ekdahl and Sonnhammer (2004) and available at Karolinska Institutet as a web service (<http://chromowheel.sbc.su.se/>). The user can submit a generic data definition format file which is then converted into an image in the Scalable Vector Graphics (SVG) format. Other software (such as the Circular Genome Viewer CGView, <http://wishart.biology.ualberta.ca/cgview/>; Stothard and Wishart, 2005) allow representation of circular genomes such as those of bacteria or mitochondria.

A primary *Tetrahymena* genome database is at <http://www.ciliate.org/> (WebLink 19.23), while a *Tetrahymena* genome sequencing website is at <http://lifesci.ucsb.edu/~genome/Tetrahymena/> (WebLink 19.24). The *Tetrahymena* functional genomics database is online at <http://fgd.ihb.ac.cn> (WebLink 19.25; Xiong *et al.*, 2013).

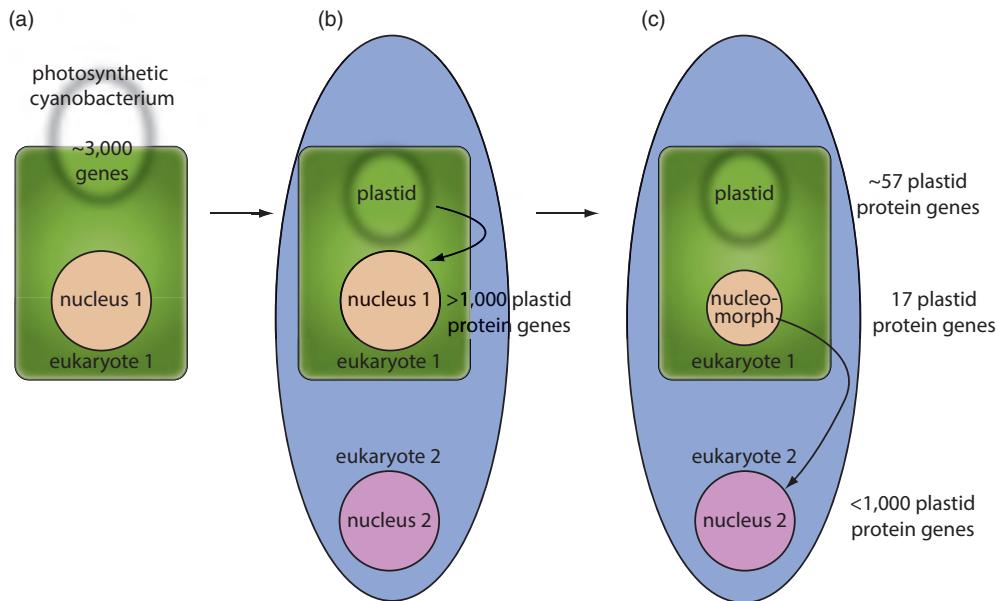
*Tetrahymena thermophila* is another ciliate that has long served as a model organism for biological research (Collins and Gorovsky, 2005). Discoveries made using *Tetrahymena* include catalytic RNA, telomeric repeats, telomerase, and the function of histone acetylation. Eisen *et al.* (2006) reported the sequence of its macronuclear genome which is 104 Mb and composed of about 225 chromosomes with a ploidy of ~45. In marked contrast to *Paramecium*, they did not find evidence for either segmental or whole-genome duplications. The relatively high gene count is explained by extensive tandem duplication of genes. The availability of the macronuclear genome will facilitate future sequencing of the micronuclear genome, which contains substantially more repetitive DNA. Such studies may elucidate the fascinating relationship between macro- and micronuclear chromosomes in the ciliates. This in turn may reveal fundamental mechanisms by which genome-wide rearrangement occurs. Additional *Tetrahymena* macronuclear genomes have now been sequenced (*T. malaccensis*, *T. elliotii*, and *T. borealis*).

A third ciliate genome is that of *Sterkiella histriomuscorum*, formerly called *Oxytricha trifallax* and indicated in the *Oxytrichida* group in **Figure 19.1**. *Sterkiella histriomuscorum* is of the class Spirotrichea. This macronuclear genome fragments into an astonishing number of about 24,500 minichromosomes (called nanochromosomes). Doak *et al.* (2003) described its genome project, including evidence of a ploidy of ~1000 per macronuclear genome. Swart *et al.* (2013) sequenced >16,000 nanochromosomes. They have a mean length of just 3.2 kilobases, each typically encoding a single gene. What biological problem does this organism solve with this system? Could our ~20,300 human protein-coding genes be packaged into a similar number of chromosomes?

### Nucleomorphs

The chloroplast is a plastid (photosynthetic organelle) in plants that contains the green pigment chlorophyll. Chloroplasts convert light to energy. A major hypothesis about their origin is that a eukaryotic cell acquired a cyanobacterium soon after the divergence of plants from animals and fungi (see “Plant Genomes” below). A radically different mechanism is also common, however. A eukaryote can ingest an alga (i.e., another eukaryote) that already has a chloroplast (Gilson and McFadden, 2002; Archibald and Lane, 2009; Moore and Archibald, 2009). This process, called endosymbiosis, may have occurred independently in at least seven separate eukaryotic groups: apicomplexa (discussed above), chlorarachniophytes, cryptomonads, dinoflagellates, euglenophytes, heterokonts, and haptophytes (reviewed in Gilson and McFadden, 2002).

Most chloroplast-containing plants and some algae have three genomes in each cell: a nuclear genome, a mitochondrial genome, and a chloroplast genome. In cryptomonads (such as *Guillardia theta*) and chlorarachniophytes (such as *Bigelowiella natans*), there is



**FIGURE 19.9** Sequential endosymbioses result in a eukaryote with four genomes. (a) In a primary endosymbiotic event, a eukaryotic host (eukaryote 1) acquires a photosynthetic bacterium such as a cyanobacterium. (b) Over time, the nuclear genome of eukaryote 1 acquires over 1000 plastid protein-coding genes. The plastid is the engulfed bacterial genome, that is, the chloroplast. Secondary endosymbiosis occurs when another nonphotosynthetic organism (eukaryote 2) engulfs and retains eukaryote 1 and so acquires photosynthetic capability. (c) Over time plastid protein genes are transferred to the nuclear genome of organism 2, resulting in the emergence of a severely reduced nucleomorph genome. The numbers of genes in the figure are for the chlorarachniophyte *Bigelowia natans*, whose nucleomorph genome is among the smallest known of all eukaryotes. Adapted from Gilson *et al.* (2006) with permission from the National Academy of Sciences.

an additional, fourth distinct genome: the vestigial nuclear genome of the engulfed alga. This second nucleus is called a nucleomorph. The process of sequential endosymbioses is outlined in **Figure 19.9**.

Just as the genome of intracellular bacteria is highly reduced, the nucleomorph genome is extremely small. Douglas *et al.* (2001) sequenced the nucleomorph genome of *G. theta* consisting of only 551,264 base pairs. The gene density is extraordinarily high, with one gene per 977 base pairs. The noncoding regions are extremely short, and there is only one pseudogene. Some otherwise essential genes, such as those encoding DNA polymerases, are absent. The gene product must be imported to the plastid across four separate membranes.

Another small nucleomorph genome is of the chlorarachniophyte *Bigelowia natans*. Its size is 373,000 base pairs, containing 331 genes on three chromosomes (Gilson *et al.*, 2006). Its nature is clearly eukaryotic including the presence of 852 introns, although these “pygmy introns” are the smallest known, having lengths of 18–21 nucleotides. Although *G. theta* and *B. natans* are phylogenetically distinct, Patron *et al.* (2006) compared their highly reduced nucleomorph genomes relative to their corresponding nuclear and plastid genomes. They concluded that *B. natans* nucleomorph genes are evolving at a rapid rate, while *G. theta* has stabilized. Additional nucleomorph genomes were sequenced and have very similar properties (Tanifuji *et al.*, 2011; Moore *et al.*, 2012). These are *Chroomonas mesostigmatica*, *Cryptomonas paramecium*, and *Hemiselmis andersenii*.

The lineage of *G. theta* is Eukaryota; Cryptophyta; Cryptomonadaceae; Guillardia. The lineage of *B. natans* is Eukaryota; Cercozoa; Chlorarachniophyceae; Bigelowia.

General principles are emerging (Moore and Archibald, 2009).

- The range of known nucleomorph genome sizes is ~380 kb to ~845 kb, encoding ~500 protein-coding genes (and having 0–24 spliceosomal introns). They are therefore models for extreme genome reduction (**Fig. 19.9**). In an unexpected example of convergent evolution, both cryptophyte and chlorarachniophyte algae nucleomorph genomes have three chromosomes and subtelomeric rRNA operons.
- Nucleomorph genomes are reduced in size because they have transferred genes to the nuclear genomes of their hosts.
- These small genomes tend to have very low GC content (~25%).
- These genomes maintain very high levels of transcription. Tanifuji *et al.* (2014) performed RNA-seq and mapped expression data for >99% of each of four nucleomorph genomes. About 10–12% of each genome is noncoding, indicating that even noncoding regions are transcribed (cf. the ENCODE project, described in Chapters 8 and 10, indicating that >80% of the human genome is at least occasionally transcribed). This suggests that all nucleomorph genes are transcribed, and mRNA synthesis levels are higher in the nucleomorph than nuclear genomes.

The circular plastid DNA of *G. theta* is also very compacted. Douglas and Penny (1999) sequenced this genome of 121,524 base pairs and found that 90% of the DNA is coding, with no pseudogenes or introns. (In contrast, only 68% of the rice plastid genome is coding.) You can explore the *G. theta* plastid genome at NCBI (accession NC\_000926.1) and compare it with the plastid genome of the red alga *Pophyra purea*, a rhodophyte (accession NC\_000925.1). These two genomes show a high degree of conserved synteny. You can also compare the *G. theta* plastid genome to that of the diatom *Odontella sinensis* (accession NC\_001713.1). This is a related alga that also acquired its plastid by secondary endosymbiosis but lacks a nucleomorph.

What of the nuclear genomes? Both the cryptophyte *G. theta* and the chlorarachniophyte *B. natans* genomes encode >21,000 proteins (Curtis *et al.*, 2012). Mitochondrial genes continue to transfer to the nucleus, but genes from the plastid and the nucleomorph do not. Protein-based phylogenetic analyses suggest that ~6–7% of the nuclear genes have an algal endosymbiont origin in each of the nuclear genomes.

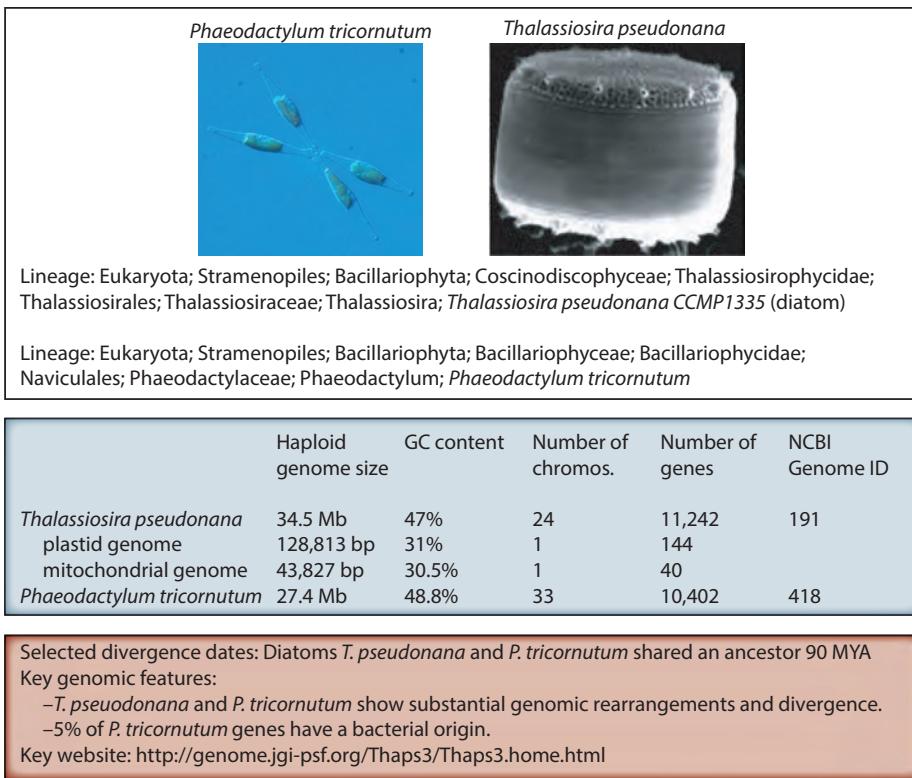
## Kingdom Stramenopila

The kingdom Stramenopila includes a wide range of fascinating organisms such as the oömycetes (e.g., the *Phytophthora* plant pathogens) and photosynthetic algae (e.g., diatoms, brown algae such as kelp, and the golden-brown algae). The Stramenopila group is represented in **Figure 19.1** as part of the Heterokonta, and we summarize several genomes in **Figures 19.10** and **19.11**.

A principal website for *Thalassiosira pseudonana* is at <http://genome.jgi-psf.org/Thaps3/Thaps3.home.html> (WebLink 19.26, from the Joint Genome Institute).

Diatoms are single-celled algae that occupy vast expanses of the oceans and are responsible for ~20% of global carbon fixation (Bowler *et al.*, 2010). They have an intricately patterned silicified (glass) cell wall called the frustule that displays beautiful, species-specific patterns as seen for example in **Figure 19.10**. Armbrust *et al.* (2004) determined the sequences of the three genomes of the diatom *Thalassiosira pseudonana*: a diploid nuclear genome of 34.5 Mb organized in 24 pairs, a plastid genome acquired by secondary endosymbiosis perhaps 1300 million years ago, and a mitochondrial genome. The plastid was acquired when a nonphotosynthetic, eukaryotic diatom ancestor engulfed a photosynthetic eukaryote (probably a red algal endosymbiont), a remarkable process described above (**Fig. 19.9**).

The second diatom genome to be sequenced was *Phaeodactylum tricornutum* (Bowler *et al.*, 2008; **Fig. 19.10**). While these organisms last shared a common ancestor 90 MYA (about the time mouse and human last shared a common ancestor), they share only ~60%



**FIGURE 19.10** The Heterokonta (see Fig. 19.1) include the diatoms. Photographs are from the NCBI Genomes website (*Thalassiosira pseudonana* by DOE-Genomes to Life, US Department of Energy Genomic Science program) and Kiene (2008), reproduced with permission from Macmillan Publishers.

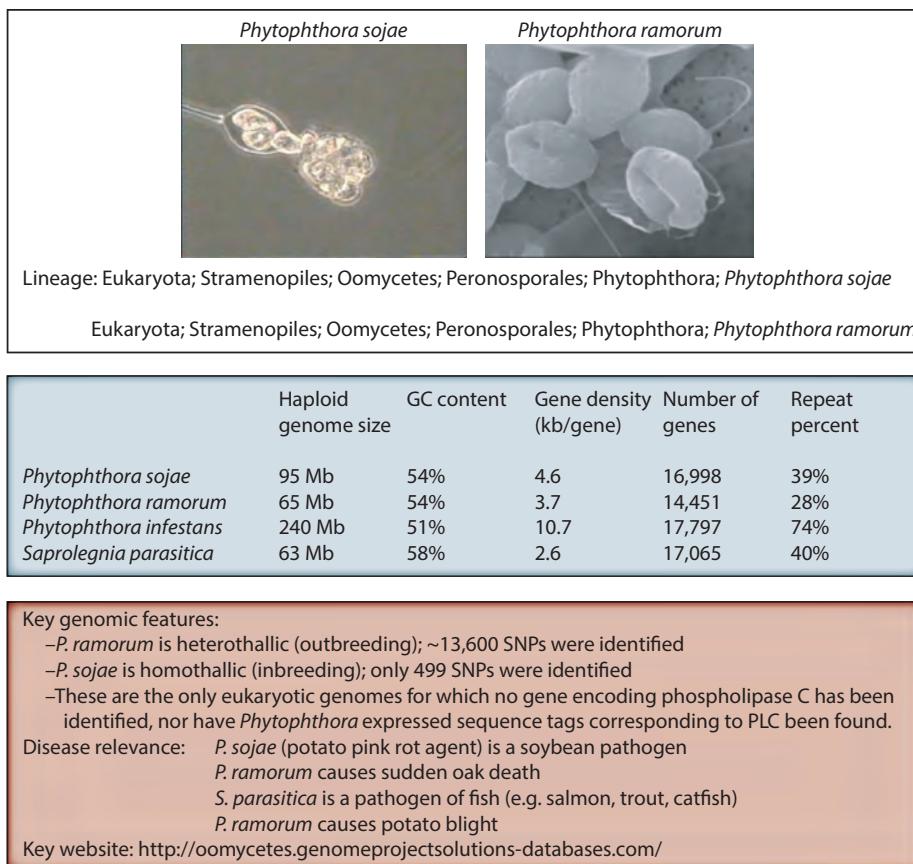
of their genes. Humans and fish shared a common ancestor ~450 MYA and protein orthologs share 61.4% identity on average; yet for *P. tricornutum* and *T. pseudonana*, orthologs share only 54.9% identity on average (Bowler *et al.*, 2010). Diatom genomes have therefore undergone rapid rates of diversification, including lineage-specific gene family expansions and also the acquisition of hundreds of bacterial genes by lateral transfer.

Oomycetes (also called water molds) are members of the kingdom Stramenopila but only distantly related to the diatoms. They include the soybean pathogen *Phytophthora sojae* and the sudden oak death pathogen *Phytophthora ramorum*. There are 59 known species of the genus *Phytophthora*, and together these cost tens of billions of dollars per year because of their destruction of plant species including crops. Tyler *et al.* (2006) reported draft genome sequences for both these plant pathogens (summarized in Fig. 19.11). The two genomes encode comparable numbers of genes including about 9700 pairs of orthologs (with extensive colinearity of orthologs spanning up to several megabases per block). While neither organism is photosynthetic, both contain many hundreds of genes that are derived from a red alga or cyanobacterium, suggesting that there was a photosynthetic ancestor.

Additional oomycete genomes have been sequenced (Jiang *et al.*, 2013). It is becoming clear that different pathogenic oomycetes have unique host interactions that lead to distinct patterns of gene loss and gene expansion. As an example, Jiang *et al.* found that the fish pathogen *Saprolegnia parasitica* genome encodes 270 proteases and 543 kinases, many of which are induced upon infection.

Such studies are likely to lead to a deeper understanding of the evolution of these organisms and possible strategies to reduce their ability to kill vast numbers of fish, insects, amphibians, and crustaceans worldwide.

The Department of Energy Joint Genome Institute (DOE JGI) website for *P. ramorum* is [http://genome.jgi-psf.org/Phyra1\\_1/Phyra1\\_1.home.html](http://genome.jgi-psf.org/Phyra1_1/Phyra1_1.home.html) (WebLink 19.27).



**FIGURE 19.11** The Heterokonta (see Fig. 19.1) include oomycetes such as *Phytophthora*. Photographs are from the NCBI Genomes website (*Phytophthora sojae* by Edward Braun, Iowa State University; *Phytophthora ramorum* by Edwin R. Florance, Lewis & Clark College).

The *Epifagus virginiana* chloroplast genome has been sequenced (NC\_001568.1; Wolfe *et al.*, 1992). *Epifagus* is parasitic on the roots of beech trees. The original major function of its chloroplast genome – photosynthesis – has become obsolete. It lacks six ribosomal protein and 13 tRNA genes that are present in the chloroplast genomes of photosynthetic plants (Wolfe *et al.*, 1992).

Brown algae (Phaeophyceae) represent another group among the Stramenopiles. They are seaweeds, and are one of five eukaryotic lineages that independently evolved multicellularity (along with metazoans (animals), fungi, and two plant groups: red algae and green algae/plants). Cock *et al.* (2010) sequenced the 214 Mb genome of the seaweed *Ectocarpus siliculosus*. Of note, they found genes likely to participate in the organism's complex photosynthetic system, allowing it to adapt to variable light conditions in harsh tidal environments.

## PLANT GENOMES

### Overview

Hundreds of thousands of plant species occupy the planet. Molecular phylogeny shows us that plants form a distinct clade within the eukaryotes (see Viridiplantae, Fig. 19.1). These include algae and the familiar green plants. All plants (other than algae) are multicellular because they develop from embryos, which are multicellular structures enclosed in maternal tissue (Margulis and Schwartz, 1998). Most plants have the capacity to perform photosynthesis, although some (such as the beech drop, *Epifagus*) do not.

The analysis of plant genomes allows us to address the molecular genetic basis of characteristics that distinguish plants from animals such as the presence of specialized

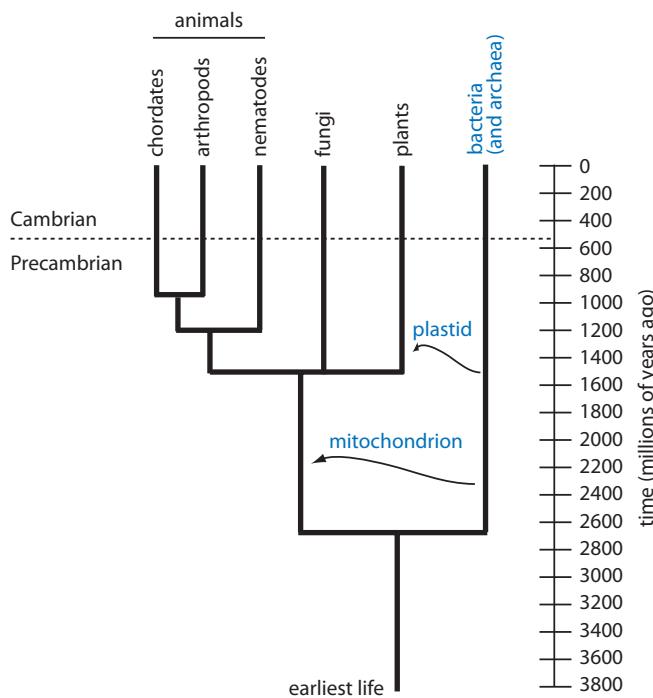
cell walls, vacuoles, plastids, and cytoskeleton. Plants are sessile and depend on photosynthesis. The sequencing of plant genomes is likely to lead to explanations for many of these basic features.

When did the lineages leading to today's plants diverge from animals, fungi, and other organisms? The earliest evidence of life is from about 3.8 billion years ago (BYA), while eukaryotic fossils have been dated to 2.7 BYA. These events are depicted in the schematic tree of **Figure 19.12**, based on separate studies by Meyerowitz (2002) and Wang *et al.* (1999). There are no very early plant fossils extant from earlier than 750 MYA, so it is difficult to assess the dates that species diverged from each other. Various researchers have used molecular clocks based on protein, DNA (nuclear or mitochondrial), or RNA data. A study by Wang *et al.* (1999) used a combined analysis of 75 nuclear genes to estimate the divergence times of plants, fungi, and several animal phyla. Their estimates of divergence time were calibrated based on evidence from the fossil record that birds and mammals diverged around 310 MYA. They found that animals and plants diverged around 1547 MYA, at almost exactly the same time that animals and fungi diverged (1538 MYA; **Fig. 19.12**).

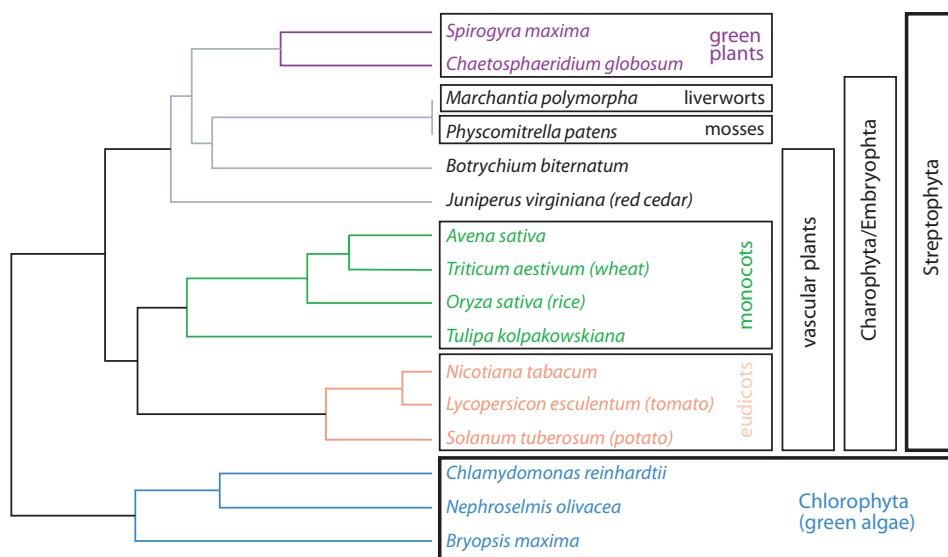
The early appearance of plants, animals, and fungi may have occurred with the divergence of a unicellular progenitor. A comparison of plants and animals therefore allows us to see how plants and animals independently evolved into multicellular forms (Meyerowitz, 2002; Niklas and Newman, 2013). The mitochondrial genes of plants and animals are homologous, indicating that their common ancestor was invaded by an  $\alpha$ -proteobacterium

Plants and animals differ greatly in their content of selected genes. For example, plants lack intermediate filaments and the genes that encode intermediate filament proteins such as cytokeratin and vimentin.

The use of 18S RNA has suggested an animal-fungi clade (**Fig. 19.1**), consistent with **Figure 19.12**.



**FIGURE 19.12** The evolution of plants, animals, and fungi. The estimated time of divergence of plants, fungi, and animals is 1.5 BYA according to a phylogenetic study (adapted from Wang *et al.*, 1999). Prior to this divergence event, a single-celled eukaryotic organism acquired an  $\alpha$ -proteobacterium (the modern mitochondrion, present today in animals, fungi, and plants). After the divergence of plants from animals and fungi about 1.5 BYA, the plant lineage acquired a plastid (the chloroplast). According to this model, metazoans diverged about 400 million years earlier than predicted by the fossil record. Nematodes (e.g., *C. elegans*) diverged earlier than chordates (e.g., vertebrates) and arthropods (e.g., insects). Adapted from Wang *et al.* (1999) with permission from Royal Society. Additional data from Meyerowitz (2002).



**FIGURE 19.13** Phylogenetic tree of the plants. A neighbor-joining tree of the plants using rubisco protein.

The earliest known plant fossils date from the Silurian period (430–408 MYA; Margulis and Schwartz, 1998).

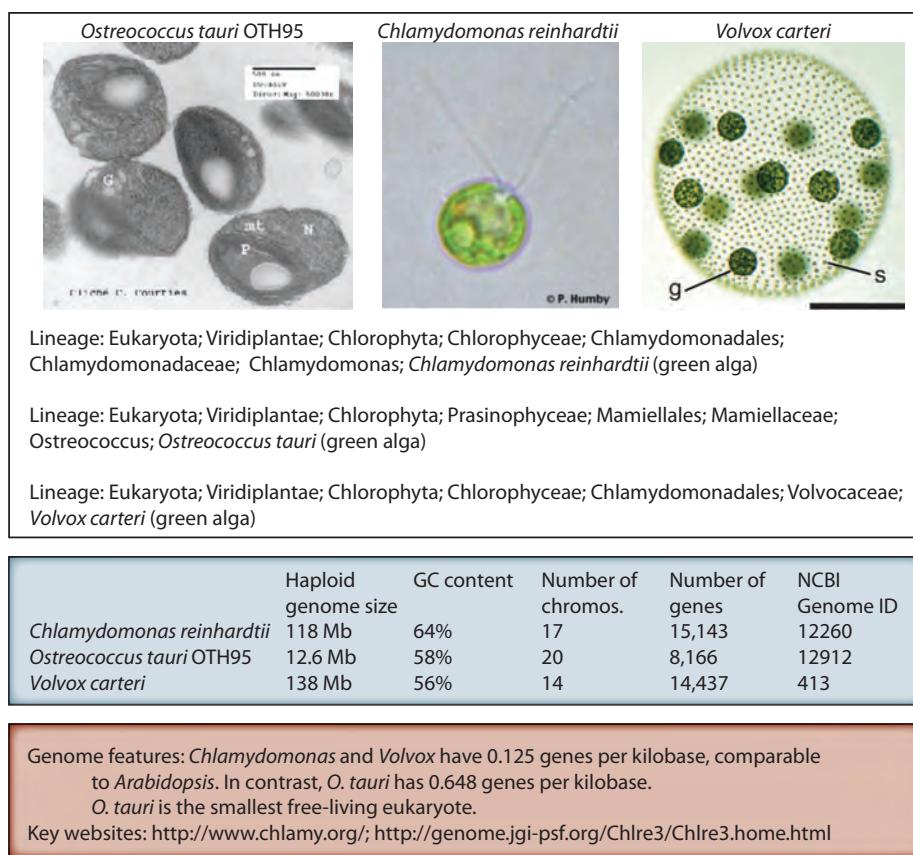
Rubisco is ribulose-1, 5-diphosphate carboxylase. It is an enzyme localized to chloroplasts that catalyzes the first step of carbon fixation in photosynthesizing plants. The enzyme irreversibly converts ribulose diphosphate and carbon dioxide ( $\text{CO}_2$ ) to two 3-phosphoglycerate molecules. The gene name for rubisco is *rbcl*; for a typical example, see the rice protein (RefSeq accession NP\_039391.1).

(Fig. 19.12). After their divergence, in another endosymbiotic event a cyanobacterium occupied plant cells to ultimately form the chloroplast. This occurred independently several times, but it has proven difficult to date these events. The first appearance of most animal phyla in the fossil record occurs in many samples dated at around 530 MYA, the “Cambrian explosion.”

We begin our bioinformatics and genomics approaches to plants by exploring their position among the eukaryotes (Fig. 19.1) and from a phylogenetic tree based on sequences of a key plant enzyme, rubisco (Fig. 19.13). The two main groups of Viridiplantae are *Chlorophyta* (green algae such as the genus *Chlamydomonas*) and *Streptophyta* (Ruhfel *et al.*, 2014). *Streptophyta* is further subdivided into groups such as mosses, liverworts, and the angiosperms (flowering plants), including the familiar monocots and eudicots. We begin with the green algae, then proceed to the flowering plants.

### Green Algae (*Chlorophyta*)

*Chlamydomonas reinhardtii* is a unicellular alga that lives in soil and water. Among the unicellular green algae, *Chlamydomonas* has served as a model organism for studying photosynthesis and chloroplast biogenesis (unlike flowering plants, it grows in the dark). The genome is 121 Mb with a very high GC content (64%) and contains about 15,000 protein-coding genes (Merchant *et al.*, 2007; Fig. 19.14). We can perform comparative genomic analyses of the *Chlamydomonas* genome to infer the properties of the ancestor of the green plants (Viridiplantae) and the opisthokonts (animals, fungi (Chapter 18), and Choanozoa). Many genes are shared by *Chlamydomonas* and animals but have been lost in angiosperms, such as those encoding the flagellum (or cilium) and the basal body (or centriole). For example, the *Chlamydomonas* genome encodes 486 membrane transporters including many shared in common with animals (e.g., voltage-gated ion channels involved in flagellar function). We explore further examples in computer laboratory exercise (19.6) at the end of this chapter. There are several possible explanations for the proteins that occur in *Chlamydomonas* and plants but not animals: (1) they may have been present in the common plant–animal ancestor and lost or diverged in the animal lineage; (2) they may have been horizontally transferred between



**FIGURE 19.14** One major division of the plants (Viridiplantae) is the green algae including *Chlamydomonas* (see Fig. 19.1). Photographs are from the NCBI Genomes website (of *Ostreococcus tauri* courtesy of O.O. Banyuls-CNRS Courties (<http://www.cs.us.es/~fran/students/julian/organisms/organisms.html>); of *Chlamydomonas reinhardtii* by Dr Durnford, University of New Brunswick). *Volvox carteri* photograph by Prochnik et al. (2010). s: somatic cells (of which there are ~2000); g: gonidia (there are ~16 of these large germline cells).

plants and *Chlamydomonas*; or (3) they may have arisen in the plant lineage before the divergence of *Chlamydomonas*. Such proteins include many involved in chloroplast function (Merchant *et al.*, 2007).

The multicellular green alga *Volvox carteri* has a genome of comparable size and complexity (Prochnik *et al.*, 2010). Comparison of these genomes may give clues to the origins of multicellularity. *Volvox* has two cell types, as shown in Figure 19.14: ~2000 small, biflagellate somatic cells and ~16 gonidia that are large reproductive cells.

Another unicellular green alga, *Ostreococcus tauri*, is thought to be the smallest free-living eukaryote (Fig. 19.14). *O. tauri* presents a simple, naked, nonflagellated cell with a nucleous, mitochondrion, and chloroplast. It is distributed throughout the oceans and was first identified in 1994 as a common component of marine phytoplankton. Derelle *et al.* (2006) sequenced its 12.6 Mb genome which is distributed on 20 chromosomes. There are 8166 protein-coding genes with a density of 1.3 kilobases per gene, greater than any other eukaryote sequenced to date. The genome therefore has an extraordinary degree of compaction with very short intergenic regions, many gene fusion events, and a reduction in the size of gene families. Another remarkable, unexplained feature of the genome is that two of the chromosomes (a large portion of 2 and all of 19) differ from all others in GC content (52–54% rather than 59% on the other chromosomes), and these two loci also contain most of the transposable elements in the genome (321 of 417). Chromosome 2 also

employs a different frequency of codon utilization and has much smaller introns (40–65 base pairs in contrast to an average of 187 base pairs elsewhere). The origin of these various differences is unknown, but these data suggest horizontal transfer from another organism.

### *Arabidopsis thaliana* Genome

The Angiosperm Phylogeny website is at <http://www.mobot.org/MOBOT/Research/APweb/welcome.html> (WebLink 19.28). It includes dozens of phylogenetic trees with access to text, photographs of plants, and extensive references.

In contrast to angiosperms, gymnosperms develop their seeds in cones. The eudicots (such as *Arabidopsis*) diverged from the monocots (such as *Oryza sativa*) about 200 MYA. Among the eudicots, the rosids and the asterids diverged about 100–150 MYA (Allen, 2002). The rosids include *Arabidopsis*, *Glycine max* (soybean), and *Medicago trunculata*. The asterids include *Lycopersicon esculentum* (tomato).

Angiosperms are flowering plants in which the seeds are enclosed in an ovary that ripens into a fruit. Monocots are characterized by an embryo with a single cotyledon (seed leaf); examples are rice, wheat, and oats. Eudicots (also called dicotyledons) have an embryo with two seed leaves; examples are tomato and potato. Eudicots include the majority of flowers and trees (but not conifers).

*Arabidopsis thaliana* is a thale cress and eudicot that is prominent as having the first plant genome to be sequenced (Fig. 19.15). *Arabidopsis* has been adopted by the plant research community as a model organism to study because it is small (about 12 inches tall),

<i>Populus trichocarpa</i> (black cottonwood)	<i>Vitis vinifera</i> (wine grape)	<i>Arabidopsis thaliana</i> (mouse-ear cress)
		
<i>Medicago truncatula</i> (barrel medic) <i>Oryza sativa</i> (rice) <i>Physcomitrella patens</i> (moss)		
Selected lineages: Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids; eurosids II; Brassicales; Brassicaceae; Arabidopsis; <i>Arabidopsis thaliana</i> Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; Vitales; Vitaceae; <i>Vitis</i> ; <i>Vitis vinifera</i>		

	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Arabidopsis thaliana</i>	125 Mb	34.9%	5	~25,498	13190
<i>Glycine max</i>	974 Mb	35%	20	10,337	5
<i>Lycopersicon esculentum</i>	782 Mb	34.9%	12	27,398	7
<i>Medicago truncatula</i>	314 Mb	35.9%	8	~19,000	10791
<i>Oryza sativa</i>	389 Mb	43.3%	12	37,544	13139, 13174
<i>Physcomitrella patens</i>	480 Mb	34%	27	35,938	13064
<i>Populus trichocarpa</i>	485 Mb	37.4%	19	45,555	10772
<i>Vitis vinifera</i>	487 Mb	35%	19	30,434	18357, 18785
<i>Zea mays</i>	2,067 Mb	46.8%	10	38,999	12

Key dates: Emergence of flowering plants 200 million years ago (MYA). *Arabidopsis* and the moss *P. patens* diverged ~450 MYA; *Arabidopsis* and *Populus* diverged ~120 MYA.  
 Disease relevance: Worldwide, up to 30% of crop yield is lost to pathogens. Plant genome sequencing projects can reveal disease resistance mechanisms.  
 Genome features: While the *Arabidopsis* genome is ~93% euchromatin, *Populus* is ~70% euchromatin. *Populus* has far more genes than *Vitis vinifera* although the two genomes have a similar size.  
 Key website: <http://www.medicago.org> (*Medicago*).

**FIGURE 19.15** Overview of selected plant genomes. Photographs are from the NCBI Genome website (*P. trichocarpa* by J.S. Peterson, USDA-NRCS PLANTS Database; *V. vinifera* by Kurt Stueber, Max Planck Institute for Plant Breeding Research, Cologne; *A. thaliana* by Luca Comai, University of Washington, Seattle, WA).

has a short generation time (about 5 weeks), has many offspring, and is convenient for genetic manipulations. It is a member of the Brassicaceae (mustard) family of vegetables, which includes horseradish, broccoli, cauliflower, and turnips. It is one of about 250,000 species of flowering plants, a group that emerged 200 MYA (Walbot, 2000). Comparative genomics analyses allow the comparison of the *Arabidopsis* genome to the genomes of other flowering plants in order to learn more about plant genomics (Hall *et al.*, 2002).

The *Arabidopsis* genome is about 125 Mb; its genome size is therefore very small compared to agriculturally important plants such as wheat and barley (see “Giant and Tiny Plant Genomes” below), making it an attractive choice as the first plant genome to be sequenced. The *Arabidopsis* Genome Initiative (2000) reported the sequence of most (115 Mb) of the genome. There are five chromosomes, initially predicted to contain 25,498 protein-coding genes. The *Arabidopsis* genome has an average density of one gene per 4.5 kb.

The estimated number of predicted genes in *Arabidopsis* has increased slightly to ~27,400, following reannotation of the genome (Crowe *et al.*, 2003; TAIR database; see Chapter 14). *Arabidopsis* has considerably more genes than *Drosophila* (about 14,000 protein-coding genes) and *C. elegans* (about 20,500 coding genes). The larger number of plant genes can be accounted for by a far greater extent of tandem gene duplications and segmental duplications. There is a core of about 11,600 distinct proteins, while the remaining genes are paralogs (*Arabidopsis* Genome Initiative, 2000).

Many plants have undergone whole-genome duplication, a phenomenon we saw with *Paramecium* and which also occurs with other eukaryotes and many fungi (Chapter 18). For an overview of ploidy in plants, see Box 19.3.

There are two main approaches to identifying whole-genome duplications (Paterson *et al.*, 2010; Fig. 19.16). The first is a bottom-up approach in which DNA or protein sequences are searched within the genome to find evidence of duplications. We described this approach for *S. cerevisiae* in Chapter 18, and it was also adopted for studies of *Paramecium* and *Arabidopsis*. Within the genome there are 24 large, duplicated segments of 100 kb or more, spanning 58% of the genome (*Arabidopsis* Genome Initiative, 2000). A second approach is called top-down in which the genome of interest is compared to a reference (Fig. 19.16b). A comparison of tomato genomic DNA with *Arabidopsis* revealed conserved gene content and gene order with four different *Arabidopsis* chromosomes (Ku *et al.*, 2000). The presence of duplicated and triplicated genomic regions suggests that two (or more) large-scale genome duplication events occurred. One event was ancient, while another occurred about 112 MYA. Following whole-genome duplication, gene loss occurred frequently. This reduces the amount of gene colinearity observed today and hinders our ability to decipher the nature and timing of past polyploidization events (Simillion *et al.*, 2002). The pattern of gene loss following genome duplication is typical

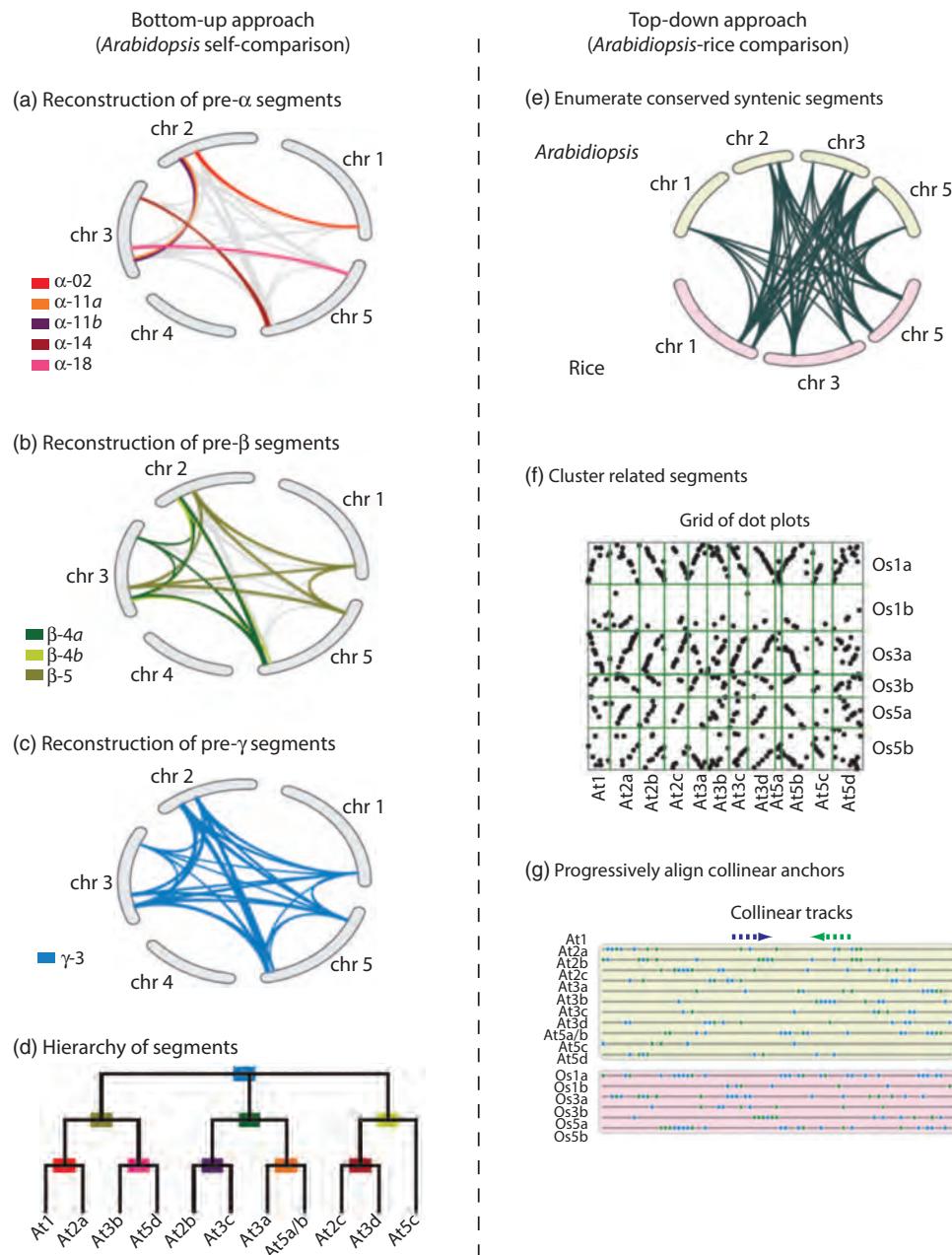
Online databases are available for model plant genome projects, such as MtDB for *Medicago trunculata* (Lamblin *et al.*, 2003; <http://www.medicago.org/>, WebLink 19.29) and MaizeGDB for maize (<http://www.maizegdb.org/>, WebLink 19.30). More comprehensive plant genomics databases include Unité de Recherche Génomique Info (URGI) (<http://urgi.versailles.inra.fr/>, WebLink 19.31) and Sputnik (Rudd *et al.*, 2003) in Turku (Åbo; <http://sputnik.btk.fi/>, WebLink 19.32). GrainGenes, a database for wheat, barley, rye, and oat, is available at <http://wheat.pw.usda.gov/GG2/index.shtml> (WebLink 19.33; Matthews *et al.*, 2003).

### BOX 19.3 PLOIDY IN PLANTS

Many plants are polyploid, that is, the nuclear genome is more than diploid. This includes autopolyploids such as *Saccharum* spp. (sugarcane) and *Medicago sativa* (alfalfa). Such species are often intolerant of inbreeding (see Paterson, 2006). Allopolyploids include wheat and cotton. In many naturally occurring allotetraploids (such as the tetraploid *Arabidopsis suecica*), the flowers are distinctly different from those of the diploid parents (*Cardaminopsis* and *Arabidopsis*). Polyploid plants are usually bigger and more vigorous than diploid plants. Examples of polyploid species include banana and apple (triploid), potato, cotton, tobacco, and peanut (all tetraploid), wheat and oat (hexaploid), and sugar cane and strawberry (octoploid).

Plant genome sequencing projects have allowed paralogs to be identified. Whole-genome duplication events have been inferred, including two or three events in both *Arabidopsis* and the poplar *Populus*, and one or two in the rice genome.

For a database of plant DNA C values, including data on polyploidy, see <http://data.kew.org/cvalues/> (Bennett and Leitch, 2011).



**FIGURE 19.16** Two strategies to detect ancient whole-genome duplications. In a bottom-up approach, (a) duplicated regions are identified to infer the most recent genome duplication event, then (b, c) successively more ancient duplications are identified. (d) A hierarchical interpretation of genome duplication events is then reconstructed. In a top-down approach, (e) conserved syntentic regions are identified between two genomes. (f) Those segments, derived from a common ancestral sequence, are then clustered. (g) Progressive alignment of the shared segments can be used to create multiple sequence alignments using software such as MCscan. Redrawn from Paterson *et al.* (2010). Reproduced with permission from Annual Reviews.

of fungi (Chapter 18) and fish (see “450 MYA: Vertebrate Genomes of Fish” below) but not *Paramecium*.

The advent of next-generation sequence technology inspired the 1001 Genomes Project which aims to sequence that many accessions. Hundreds of *Arabidopsis* genomes have already been sequenced, typically complemented by RNA-seq studies

## BOX 19.4 DATABASES FOR EUKARYOTIC GENOMES

The main *Arabidopsis* database, TAIR, uses a database template shared by other major sequencing projects (Table 19.1). We already explored EcoCyc in Chapter 14 and the yeast database SGD in Chapters 14 and 18. These databases offer both detailed and extremely broad views of the genomic landscape. The Genomics Unified Schema (GUS) is another commonly used platform. Many databases use a distributed annotation system (DAS) that allows a computer server to integrate genomic data from a variety of external computer systems. DAS, written by Lincoln Stein and Robin Dowell, is described at <http://www.biodas.org/>. It is employed at WormBase, FlyBase, Ensembl, and JCVI sites, among others.

(Cao *et al.*, 2011; Gan *et al.*, 2011). Crosses between a reference genome (Col-0) and 18 sequenced accessions were used to generate 700 strains called the Multiparent Advanced Generation Inter-Cross (MAGIC) collection. Analogous to the Collaborative Cross for mice, these many strains will have both detailed phenotypic characterization and whole-genome sequences, facilitating studies of the genetic basis of plant properties from seed type, disease susceptibility and environmental adaptation to growth properties. Corresponding data for 1001 proteomes is being collected (Joshi *et al.*, 2012).

The most comprehensive *Arabidopsis* genomics resource is TAIR, with a wide range of services (Lamesch *et al.*, 2012). This site includes a genome browser that provides access to genomic DNA sequence from the broadest chromosome-level view to descriptions of single-nucleotide polymorphisms (Fig. 14.5). The format of this site, GBrowse, is shared by a variety of genome projects (Box 19.4; Table 19.1). Other databases include SeedGenes, which describes essential genes of *Arabidopsis* that give a seed phenotype when disrupted by mutation (Tzafrir *et al.*, 2003).

### The Second Plant Genome: Rice

By some estimates, rice (*O. sativa*) is the staple food for half the human population. The rice genome was the second plant genome to be sequenced (Fig. 19.15). At approximately 389 Mb, this genome size is about one-eighth that of the human genome. It is however one of the smallest genomes among the grasses, and rice is studied as a model monocot species.

The genus *Oryza* includes 23 species, and efforts are underway to sequence all these genomes (Jacquemin *et al.*, 2013). The cultivated species are *O. glaberrima*

The 1001 Genomes website is <http://1001genomes.org/> (WebLink 19.34). The 1001 Proteomes portal is at <http://1001proteomes.masc-proteomics.org/> (WebLink 19.35). It is a nonsynonymous SNP browser.

The Arabidopsis Information Resource (TAIR) is online at <http://www.arabidopsis.org/> (WebLink 19.36). The MIPS plantsDB set of databases is at <http://mips.helmholtz-muenchen.de/plant/genomes.jsp> (WebLink 19.37). SeedGenes, describing essential genes in *Arabidopsis* development, is at <http://www.seedgenes.org> (WebLink 19.38).

Grasses include rice, wheat, maize, sorghum, barley, sugarcane, millet, oat, and rye. There are over 10,000 species of grasses (Bennetzen and Freeling, 1997). Cereals are seeds of flowering plants of the grass family (Gramineae, also called Poaceae) that are cultivated for the food value of their grains. Grasses are monocotyledonous plants that range from small, twisted, erect, or creeping annuals to perennials.

**TABLE 19.1 Variety of databases employing template from Generic Model Organism Project (GMOD, <http://www.gmod.org/>) under the terms of the GNU Free Documentation License 1.2.**

Database	Comment	URL
EcoCyc	Encyclopedia of <i>Escherichia coli</i> Genes and Metabolism	<a href="http://EcoCyc.org/">http://EcoCyc.org/</a>
FlyBase	<i>Drosophila</i> site	<a href="http://www.flybase.org/">http://www.flybase.org/</a>
Mouse Genome Informatics	Main mouse resource	<a href="http://www.informatics.jax.org/">http://www.informatics.jax.org/</a>
Rat Genome Database (RGD)	Rat resource	<a href="http://rgd.mcw.edu/">http://rgd.mcw.edu/</a>
SGD	See Chapter 18	<a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a>
TAIR	The <i>Arabidopsis</i> Information Resource	<a href="http://www.arabidopsis.org/">http://www.arabidopsis.org/</a>
Wormbase	<i>C. Elegans</i> Site	<a href="http://www.wormbase.org/">http://www.wormbase.org/</a>

(in Africa) and *O. sativa* with three cultivars (*japonica*, *indica*, and *javanica*). Four groups generated draft versions of the rice genome. A consortium led by the Beijing Genomics Institute reported a draft sequence of the rice genome (*O. sativa* L. ssp. *Indica*; Yu *et al.*, 2002). Another consortium reported a draft genome sequence of *O. sativa* L. ssp. *japonica* (Goff *et al.*, 2002), and Monsanto generated another genome sequence.

The Rice Annotation Project Database (RAP-DB) is at <http://rapdb.dna.affrc.go.jp/> (WebLink 19.39).

A finished quality sequence was reported separately by Yu *et al.* (2005) and by the International Rice Genome Sequencing Project (2005) for a single inbred cultivar, *O. sativa* L. ssp. *japonica* cv. Nipponbare. Yu *et al.* (2005) reported that, relative to their 2002 initial publication, they achieved a lower error rate and a 1000-fold improvement in long-range contiguity. The N50 sequence (the length above which half the total length of the sequence dataset is found) improved to 8.3 Mb, about a 1000-fold improvement, as the coverage increased from 4.2 $\times$  to 6.3 $\times$ . Annotation is a continuing process. The current Rice Annotation Project Database relies on next-generation sequence data, RNA-seq, and comparisons to sequence data from 150 monocot species (Sakai *et al.*, 2013).

The rice genome (subspecies *indica*) displays an unusual feature of a gradient in GC content. The mean GC content is 43.3%, higher than in *Arabidopsis* (34.9%) or human (41.1%) (Yu *et al.*, 2002). A plot of the number of 500 base pair sequences (y axis) versus the percent GC content (x axis) revealed a tail of many GC-rich sequences. These GC-rich regions occurred selectively in rice exons (rather than introns), and at least one exon of extremely high GC content was found in almost every rice gene (Yu *et al.*, 2002). The GC content of the 5' end of each gene was typically 25% more GC rich than the 3' end. These unique features of the rice genome present another major challenge for the use of *ab initio* gene-finding software.

From where did cultivated rice originate? Huang *et al.* (2012) sequenced 446 geographically diverse accessions of *Oryza rufipogon* (a wild rice species) and 1083 cultivated *indica* and *japonica* varieties, identifying a region of southern China where rice was first domesticated several thousand years ago. Huang *et al.* performed 1 $\times$  to 2 $\times$  coverage of these genomes (many of which were sequenced in separate studies), and identified single-nucleotide polymorphisms (SNPs) which were used for phylogenetic and population genetics studies.

### Third Plant: Poplar

The black cottonwood tree *Populus trichocarpa* was the third plant genome to be sequenced (Fig. 19.15). *Populus* was selected for sequencing because its haploid nuclear genome is relatively small (480 Mb), it grows quickly relative to other trees (~5 years), and is economically important as a source of wood and paper products.

Analysis of the genome indicates that *Populus* underwent a relatively recent whole-genome duplication about 65 million years ago, as well as experiencing tandem duplications and chromosomal rearrangements (Tuskan *et al.*, 2006). In contrast to *Arabidopsis*, *Populus* is predominantly dioecious (having male and female reproductive structures on separate plants), such that it must outcross and achieves high levels of heterozygosity. Tuskan *et al.* (2006) identified 1.2 million SNPs and, with insertion/deletion events, estimated 2.6 polymorphisms per kilobase. These were intended to enable further genetics and population biology studies. Highlighting the revolution brought by next-generation sequencing, Tuskan and colleagues then proceeded to sequence 16 *P. trichocarpa* genomes and genotype 120 trees from 10 subpopulations (Slavov *et al.*, 2012). They reported extensive linkage disequilibrium (to a greater extent than expected from previous smaller-scale studies), suggesting that genome-wide association studies may be feasible in undomesticated trees.

## Fourth Plant: Grapevine

The gravevines are highly heterozygous, with as much as 13% sequence divergence between alleles. The French–Italian Public Consortium for Grapevine Genome Characterization (Jaillon *et al.*, 2007) bred a grapevine variety derived from Pinot Noir to a high level of homozygosity and then determined its genome sequence (Fig. 19.15). Using an inbred variety was necessary to facilitate the assembly process. There are ~30,000 protein-coding genes predicted, which is fewer than in *Populus* even though the two organisms have similar-sized genomes. Genes are evenly distributed across the genome in *Arabidopsis* and rice, while in *V. vinifera* as in *Populus* there are gene-rich and gene-poor regions with transposable elements (such as SINEs) occupying complementary positions.

One notable feature of the *V. vinifera* genome is that encodes more than twice as many proteins related to terpene synthesis as other sequenced plant genomes. There are tens of thousands of terpenes in nature, typically containing two to four isoprene units, and many of these are highly odorous.

Analysis of the haploid grapevine genome showed that most gene regions have two different paralogous regions, therefore forming homologous triplets and suggesting that the present genome derives from three ancestral genomes (Jaillon *et al.*, 2007). There may have been three successive whole-genome duplications or a single hexaploidization event. To address this question they compared the *Vitis* gene order to poplar (its closest relative), *Arabidopsis* (a more distantly related dicotyledon), and rice (as a monocotyledon, its most distant relative). Grapevine aligned with two poplar segments, consistent with a recent whole-genome duplication in poplar (described above). Also, the grapevine homologous triplets aligned with different pairs of poplar segments, suggesting that a hexaploidy of ancient origin was already present in the common ancestor of grapevine and poplar.

## Giant and Tiny Plant Genomes

Plant genomes can be enormous in terms of size. An extreme example is *Paris japonica*, a small white flower with a 150 gigabase genome (50 times larger than humans; Pellicer *et al.*, 2010). Angiosperms in particular can also exhibit very high ploidy. While many are diploid or triploid, far higher levels can occur (Bennett and Leitch, 2011; Box 19.3).

In terms of genome size, the largest sequenced genome is that of the loblolly pine, *Picea glauca* (Neale *et al.*, 2014; Wegrzyn *et al.*, 2014). This conifer's genome is 23,564 Mb (~23.6 Gb), requiring novel approaches to genome assembly (Birol *et al.*, 2013). The hexaploid genome of bread wheat (*Triticum aestivum*) is also large at 17 Gb. It was sequenced and assembled by Brenchley *et al.* (2012).

At another extreme, a carnivorous family of angiosperms features small genomes as low as 63 Mb (Greilhuber *et al.*, 2006). Ibarra-Laclette *et al.* (2013) reported the 82 Mb genome of one such plant, *Utricularia gibba*. It has typical gene content ( $n = 28,500$ ), but reduced intergenic regions.

## Hundreds More Land Plant Genomes

As hundreds of plant genomes are sequenced, each offers a fascinating avenue to understanding crop production, genome evolution, disease susceptibility, and many other aspects of plant biology. You can browse these at NCBI Genome. Examples of recently sequenced genomes include: legumes *Glycine max* (soybean; Schmutz *et al.*, 2010) and *medicago* (Young *et al.*, 2011; for a review of legume genomes see Young and Bharti, 2012); the woodland strawberry (Shulaev *et al.*, 2011); the autotetraploid tuber crop potato *Solanum tuberosum* L. (Potato Genome Sequencing Consortium *et al.*, 2011); maize (Schnable *et al.*, 2009); and the tomato (Tomato Genome Consortium, 2012).

## Moss

The Moss Genome website is <http://www.cosmoss.org> (WebLink 19.40). A Joint Genomes Initiative website on *P. patens* is at [http://genome.jgi-psf.org/Phypa1\\_1/Phypa1\\_1/home.html](http://genome.jgi-psf.org/Phypa1_1/Phypa1_1/home.html) (WebLink 19.41).

The bryophytes, encompassing mosses, hornworts, and liverworts diverged from the embryophytes (land plants) about 450 MYA (near the time of divergence of the fish and human lineages). Rensing *et al.* (2008) sequenced the genome of the bryophyte moss *Physcomitrella patens*. Through comparisons to the genomes of water-dwelling plants, they propose that the movement of plants from aquatic to land environments involved the following components: (1) loss of genes that are associated with aquatic environments, such as those involved in flagellar function; (2) loss of dynein-mediated transport (as stated above, *Chlamydomonas* and animals share dyneins); (3) gain of genes involved in signaling capabilities such as auxin, many of which are absent in *Chlamydomonas* and *O. tauri* genomes; (4) capability of adapting to conditions of drought, radiation, and temperature extremes; (5) gain of transport capabilities; and (6) gain in gene family complexity, reflected in the large numbers of genes in the moss and other plant genomes.

## SLIME AND FRUITING BODIES AT THE FEET OF METAZOANS

As we examine the upper part of the tree of the eukaryotes in **Figure 19.1**, we see three great clades: the Mycetozoa, the Metazoa (animals), and the Fungi (Chapter 18). The metazoans are familiar to us as animals, including worms, insects, fish, and mammals. The Mycetozoa form a sister clade. The slime mold *Dictyostelium discoideum* is a social amoeba that is of great interest as a member of an outgroup of the metazoa.

### Social Slime Mold *Dictyostelium discoideum*

A principal website for information on *Dictyostelium* is <http://www.dictybase.org/> (WebLink 19.42). See also <http://amoebadb.org/amoeba/> (WebLink 19.43; Aurrecoechea *et al.*, 2011). The social, multicellular lifestyle of this eukaryote is reminiscent of the similar behavior of the proteobacterium *Myxococcus xanthus* (Chapter 17).

Chromosome 2 is characterized by an inverted 1.51 Mb duplication that is present in only some wildtype isolates.

Biologists have studied *Dictyostelium* because of its remarkable life cycle. In normal conditions it is a single-celled organism that occupies a niche in soil. Upon conditions of starvation, it emits pulses of cyclic AMP (cAMP), promoting the aggregation of large numbers of amoebae. This results in the formation of an organism having the properties of other multicellular eukaryotes: it differentiates into several cell types, responds to heat and light, and undergoes a developmental profile.

The *Dictyostelium* genome is 34 Mb, localized on six chromosomes (**Fig. 19.17**), and was sequenced by Eichinger *et al.* (2005). In addition to six chromosomes (and the standard mitochondrial genome of 55 kb), there is one 88 kb palindromic extrachromosomal element that occurs in ~100 copies per nucleus and contains ribosomal RNA genes.

Lineage: Eukaryota; Mycetozoa; Dictyosteliida; *Dictyostelium*; *Dictyostelium discoideum* AX4 (social amoeba; slime mold)



<i>Dictyostelium discoideum</i>	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
	34 Mb	22.4%	6	~12,500	201

Disease relevance: *Dictyostelium* has many hundreds of orthologs of human disease genes, and can reveal principles of the evolution of these genes.  
Genome features: The GC content is extraordinarily low and impacts many features of the genome.  
Key website: <http://www.dictybase.org>

**FIGURE 19.17** The slime mold *D. discoideum* is closely related to the metazoans, as shown in **Figure 19.1**. This summary includes a photograph from the NHGRI (<http://www.genome.gov/17516871>). Source: National Human Genome Research Institute and Dr Jonatha Gott.

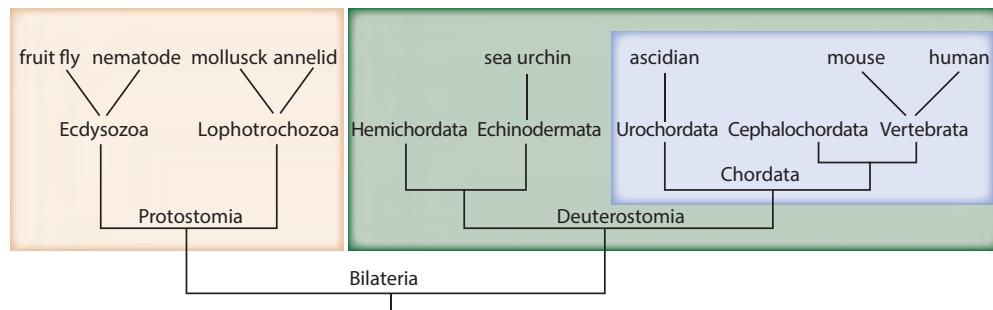
Because the genome consists of about 78% AT content – similar to *P. falciparum* – as well as many repetitive DNA sequences, large-insert bacterial clones are unstable, and a whole-chromosome shotgun strategy was adopted. The genome is compact: the gene density is high (there are ~12,500 genes, with one gene per 2.6 kb, spanning 62% of the genome), there are relatively few introns (1.2 per gene), and both introns and intergenic regions are short. The introns have an AT content of 87%, while in exons the AT content is 72%. This discrepant compositional bias may represent a mechanism by which introns are spliced out (Glockner *et al.*, 2002). Reflecting the AT richness of the genome, the codons NNT or NNA are used preferentially relative to the synonymous codons NNG or NNC. Amino acids encoded by codons having A or T in the first two positions and any nucleotide in the third position (asn, lys, ile, tyr, and phe) are far more common in *Dictyostelium* proteins than human proteins.

An unusual feature of the genome is that 11% comprises simple sequence repeats (Chapter 8), more than for any other sequenced genome. There is a bias toward repeat units of 3–6 base pairs. Noncoding simple sequence repeats and homopolymer tracts have 99.2% AT content.

## METAZOANS

### Introduction to Metazoans

The metazoans include the animals that are familiar to us, in particular the bilaterians, that is, bilaterally symmetric animals (Fig. 19.18). The bilaterian animals are further divided into two major groups. (1) The protostomes include the Ecdysozoa (arthropods and nematodes), as well as the Lophotrochozoa (annelids and mollusks). We survey the first protostome genomes that have been sequenced such as the insects *D. melanogaster* and *A. gambiae*, and the nematode *C. elegans*. (2) The deuterostomes form a superclade consisting of the phylum of echinoderms (such as the sea urchin *Strongylocentrotus purpuratus*), the phylum of the hemichordates (such as acorn worms), and the chordates (vertebrates as well as the invertebrate cephalochordates and urochordates). These three deuterostome phyla descended from a common ancestor about 550 MYA, the time of the Cambrian explosion. We discuss a basal member of the deuterostomes (the sea urchin *S. purpuratus*) and a basal member of the chordates (the urochordate sea squirt *Ciona intestinalis*), and we examine the vertebrate genomes such as the fish, mouse, and chimpanzee.



**FIGURE 19.18** Phylogenetic relationships of the bilaterians which have a bilateral body organization. The Protostomia include the arthropods or insects such as the fruit fly *Drosophila melanogaster* and the nematode worms such as *Caenorhabditis elegans*, as well as the mollusks and annelids. The Deuterostomia include the sister phyla Hemichordata and Echinodermata (including the sea urchin *Strongylocentrotus purpuratus*) as well as the Chordata. The chordates are further divided into three groups including the vertebrates. This figure was redrawn from the Sea Urchin Genome Sequencing Consortium *et al.* (2006). Used with permission.

The phylogeny of Figure 19.18 is consistent with those of Figures 19.1 and 19.12, although the trees differ in the placement of nematodes as an outgroup. For discussions of bilaterian phylogeny see Lartillot and Philippe (2008). For alternative classification systems, see Cavalier-Smith (1998) and Margulis and Schwartz (1998). Karl Leuckart (1822–1898) first divided the metazoan into six phyla. For a table describing the metazoan (animal) kingdom superphyla and phyla, see Web Document 19.1 at <http://www.bioinfbook.org/chapter19>. For a table describing the phylum bilateria, including the Coelomata (animals with a body cavity), Acoelomata (animals lacking a body cavity such as flatworms) and Pseudocoelomata (such as the roundworm *C. elegans*), see Web Document 19.2.

As we seek to understand the human genome and what makes us unique as a species from a genomic perspective, one approach has been to determine whether our complexity and advanced features can be accounted for by a relatively large collection of genes. It is now clear that this is not the case; our gene numbers are comparable to those of other species across the eukaryotic domain. Another notion has been that humans, and vertebrates in general, have a large collection of unique genes that are not present in invertebrates. This notion is correct to a limited extent, but it too is being challenged. As metazoan genomes become sequenced, we find many vertebrate genetic features shared with simpler animals (from insects to the invertebrate sea urchin to the sea squirt, a simple chordate).

Time Tree can be viewed at  
<http://www.timetree.org>  
 (WebLink 19.44).

The title of the next section begins “900 million years ago,” referring to the approximate date of a last common ancestor with the human lineage; in the remaining sections of this chapter we continue to track the relatedness of each group to humans. The numbers of years since humans last shared a common ancestor with each species are from the primary literature or the Time Tree project (Hedges *et al.*, 2006).

### 900 MYA: the Simple Animal *Caenorhabditis elegans*

The soma of an adult hermaphrodite worm consists of 959 cells, including 302 cells in the central nervous system. About 300 species of parasitic worms infect humans (Cox, 2002). While 20,000 nematode species have been described, it is thought that there may be one million species (Blaxter, 1998, 2003).

*Caenorhabditis elegans* is a free-living soil nematode. It has served as a model organism because it is small (about 1 mm in length), easy to propagate (its life cycle is three days), has an invariant cell lineage that is fully described, and is suitable for many genetic manipulations. Furthermore, it has a variety of complex physiological traits characteristic of higher metazoans such as vertebrates, including an advanced central nervous system. Many nematodes are parasitic, and an understanding of *C. elegans* biology may lead to treatments for a variety of human diseases.

Another advantage of studying *C. elegans* is that its genome size of ~100 Mb is relatively small (Fig. 19.19). This genome was the first of an animal and the first of a multicellular organism to be sequenced (*C. elegans* Sequencing Consortium, 1998). The

Genus, species	<i>Brugia malayi</i> <i>Caenorhabditis briggsae</i> <i>Caenorhabditis elegans</i>	<i>Brugia malayi</i>
Selected lineages:	Eukaryota; Metazoa; Nematoda; Chromadorea; Spirurida; Filarioidea; Onchocercidae; <i>Brugia</i> ; <i>Brugia malayi</i>	
Lineage:	Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Rhabditoidea; Rhabditidae; Pelerinae; Caenorhabditis; <i>Caenorhabditis briggsae</i>	
Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida; Rhabditoidea; Rhabditidae; Pelerinae; Caenorhabditis; <i>Caenorhabditis elegans</i>		

	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Brugia malayi</i>	90-95 Mb	30.5%	6	11,508	10729
<i>Caenorhabditis briggsae</i>	~104 Mb	37.4%	6	19,500	10731
<i>Caenorhabditis elegans</i>	100 Mb	35.4%	6	18,808	13758

Divergence dates: Nematodes diverged from arthropods (insects) 800-1000 million years ago (MYA).  
*C. elegans* diverged from *C. briggsae* ~80-110 MYA.

Disease relevance: *Brugia malayi* is the agent of lymphatic filariasis which infects 120 million people.  
 Key website: <http://www.wormbase.org>

**FIGURE 19.19** Overview of roundworm genomes. Image of the anterior end of a *Brugia malayi* microfilaria in a thick blood smear using Giemsa stain is from the CDC (<http://phil.cdc.gov/phil/home.asp>; content provider Mae Melvin).

genome sequencing was based on physical maps of the five autosomes and single X chromosome. The GC content is an unremarkable 36%. It was predicted that there are 19,099 protein-coding genes, with 27% of the genome consisting of exons (the current Ensembl tally is 20,400).

The *C. elegans* proteome contains a large number of predicted seven-transmembrane-domain (7TM) receptors of both the chemoreceptor family and rhodopsin family. This illustrates the principle that new protein functions can emerge following gene duplication (Sonnhammer and Durbin, 1997). It is also notable that many nematode proteins are absent from nonmetazoan species (plants and fungi).

The principal web resource for *C. elegans* is WormBase, a comprehensive database (Harris *et al.*, 2014). WormBase features a variety of data, including: genomic sequence data; the developmental lineage; the connectivity of the nervous system; mutant phenotypes, genetic markers, and genetic map data; gene expression data; and bibliographic resources.

After *C. elegans*, the genome of the related soil nematode *Caenorhabditis briggsae* was sequenced (Stein *et al.*, 2003; reviewed in Gupta and Sternberg, 2003). Remarkably, these organisms speciated about 100 MYA, but they are indistinguishable by eye. Each genome is about 100 Mb and encodes a comparable number of genes. The availability of *C. briggsae* sequence facilitated an improved annotation of the *C. elegans* genome and the discovery of about 1300 novel *C. elegans* genes. The genomes share extensive colinearity.

*Brugia malayi* was the first parasitic nematode to have its genome sequenced (Ghedin *et al.*, 2007). This parasite causes lymphatic filariasis, a chronic disease that is debilitating although associated with low mortality. The *B. malayi* genome contains fewer genes than *C. elegans* (~11,500 versus ~18,500), primarily because of lineage-specific expansions in *C. elegans*. There is a need for drugs to treat filariasis, and Ghedin *et al.* identified a number of gene products that are potential targets for therapeutic intervention. For example, *B. malayi* lacks most enzymes required for *de novo* purine biosynthesis, heme biosynthesis, and *de novo* riboflavin synthesis, probably obtaining these compounds from its host or its endosymbiont *Wolbachia*. Drugs that interfere with these synthetic pathways are potential targets.

The genomes of several dozen nematodes are currently being sequenced, and the UCSC Genome Browser currently includes assemblies for six nematodes (*C. elegans*, *C. briggsae*, *C. brenneri*, *C. japonica*, *C. remanei*, *Pristionchus pacificus*). Annotation tracks are available for conservation among these worms and to human proteins. The Million Mutation Project lists >2000 mutagenized strains with >180,000 nonsynonymous changes in ~20,000 genes.

## 900 MYA: *Drosophila melanogaster* (First Insect Genome)

The arthropods may be the most successful set of eukaryotes on the planet in terms of the number of species. They include the Chelicerates – such as the scorpions, spiders, and mites – and the Mandibulata, animals with modified appendages (mandibles) such as the insects (Table 19.2; Fig. 19.20). While insects first appear in the fossil record from about 350 MYA, their lineage is thought to have emerged as long as 900 MYA.

The fruit fly *D. melanogaster* has been an important model organism in biology for a century (Rubin and Lewis, 2000). The fly is ideal for studies of genetics because of its short life cycle (two weeks), varied phenotypes (from changes in eye color to changes in behavior, development, or morphology), and large polytene chromosomes that are easily observed under a microscope.

The *Drosophila* genome was sequenced based in large part upon the whole-genome shotgun sequencing strategy (Adams *et al.*, 2000). Prior to this effort, the whole-genome

The 2002 Nobel Prize in Physiology or Medicine was awarded to three researchers who pioneered the use of *C. elegans* as a model organism: Sydney Brenner, H. Robert Horvitz, and John E. Sulston. See [http://www.nobelprize.org/nobel\\_prizes/medicine/laureates/2002/](http://www.nobelprize.org/nobel_prizes/medicine/laureates/2002/) (WebLink 19.45).

WormBase is available at <http://www.wormbase.org> (WebLink 19.46).

Visit the Million Mutation Project at <http://genome.sfu.ca/mmp/> (WebLink 19.47).

Thomas Hunt Morgan was awarded a Nobel Prize in 1933 "for his discoveries concerning the role played by the chromosome in heredity." See [http://www.nobelprize.org/nobel\\_prizes/medicine/laureates/1933/](http://www.nobelprize.org/nobel_prizes/medicine/laureates/1933/) (WebLink 19.48). In 1995, Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus shared a Nobel Prize "for their discoveries concerning the genetic control of early embryonic development." These studies concerned *Drosophila* development ([http://www.nobelprize.org/nobel\\_prizes/medicine/laureates/1995/](http://www.nobelprize.org/nobel_prizes/medicine/laureates/1995/), WebLink 19.49). About 1 million arthropod species have been described, but there are an estimated 3–30 million species (Blaxter, 2003).

**TABLE 19.2 Arthropods (*Phylum arthropoda*) as classified at NCBI (<http://www.ncbi.nlm.nih.gov/Taxonomy/>). Arthropods are invertebrate protostomes (see Fig. 19.18). Pancrustacea is further divided into the superclasses Crustacea (crustaceans) and Hexapoda (insects). Insecta includes *D. melanogaster* and *A. gambiae*.**

Subphylum	Class
Chelicerata	Arachnida (mites, ticks, spiders)
	Merostomata (horseshoe crabs)
	Pycnogonida (sea spiders)
Mandibulata	Myriapoda (centipedes)
	Pancrustacea (crustaceans, insects)

Source: NCBI Taxonomy Browser, NCBI.

The *Drosophila* genome was sequenced through a collaborative effort that included Celera Genomics, the Berkeley *Drosophila* Genome Project (BDGP; <http://www.fruitfly.org>, WebLink 19.50), and the European *Drosophila* Genome Project (EDGP) (Adams *et al.*, 2000).

shotgun strategy had only been applied to far smaller genomes; this success therefore represented a significant breakthrough. The 180 Mb genome is organized into an X chromosome (numbered 1), two principal autosomes (numbered 2 and 3), a very small third autosome (numbered 4; about 1 Mb in length), and a Y chromosome. Approximately one-third of the genome contains heterochromatin (mostly simple sequence repeats as well as transposable elements and tandem arrays of rRNA genes). This heterochromatin is distributed around the centromeres and across the length of the Y chromosome. The transition zones at the boundary of heterochromatin and euchromatin contain many protein-coding genes that were previously unknown.

Two dozen *Drosophila*-related genomes were sequenced (Fig. 19.20). Following the sequencing of *D. melanogaster* and *D. pseudoobscura* (Richards *et al.*, 2005), a consortium of 250 researchers sequenced ten more genomes (*Drosophila* 12 Genomes Consortium, 2007). Seven genomes were sequenced to deep coverage (8.4–11.0×) and others to intermediate or low coverage to provide population variation data. These include several species that are closely related (e.g., *D. yakuba* and *D. erecta*, or *D. pseudoobscura* and *D. persimilis*) as well as some distantly related (e.g., *D. grimshawi* is a species restricted to Hawaii). Total genome size varies less than three-fold among the 12 species, and the gene content ranges from ~14,000 to ~17,000. Based on comparative annotation of protein coding genes, Stark *et al.* (2007) identified almost 1200 new protein-coding exons and resulted in a modification of 10% of the annotated protein-coding genes in *D. melanogaster*.

The availability of so many related genome sequences permits a deeper understanding of many areas of evolution including genomic rearrangements, the acquisition of transposable elements, and protein evolution. Most genes evolve under evolutionary constraint at most of their sites, so that the ratio  $\omega$  of nonsynonymous to synonymous mutations ( $d_N/d_S$ ) tends to be low. Of all *D. melanogaster* proteins, the majority (77%) are conserved across all 12 species. The number of noncoding RNA genes is also conserved, ranging from ~600 to ~900.

The sequencing and analysis of multiple *Drosophila* genomes (as for multiple fungal genomes; Chapter 18) represent important, pioneering effort in eukaryotic comparative genomics. Such approaches will result in improved catalogs of coding and noncoding genes, regulatory features, and functional regions of genomic DNA. A clearer understanding of evolutionary events, including when species diverged and how and when genomes have been sculpted by forces from chromosomal alterations to lateral transfer of transposable elements, can also be achieved through these techniques.

<i>Tribolium castaneum</i>	<i>Anopheles gambiae</i>	<i>Aedes aegypti</i>
Selected lineages: Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Nematocera; Culicoidea; Culicidae; Anophelinae; Anopheles; <i>Anopheles gambiae</i> str. PEST (African malaria mosquito)		
Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydriodea; Drosophilidae; Drosophila; <i>Drosophila melanogaster</i> (fruit fly)		
Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Coleoptera; Polyphaga; Cucujiformia; Tenebrionidae; Tribolium; <i>Tribolium castaneum</i> (red flour beetle)		

	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Anopheles gambiae</i>	278 Mb	44%	3	13,683	1438
<i>Apis mellifera</i> DH4	262 Mb	33%	16	15,314	10625
<i>Bombyx mori</i> (silkworm)	481 Mb	39%	28	18,510	12259, 13125
<i>Danaus plexippus</i> (butterfly)	273 Mb	32%	29-30	16,866	11702
<i>Daphnia pulex</i> (water flea)	197 Mb	41%	12	30,613	288
<i>Drosophila ananassae</i>	217 Mb	42%	4	22,551	12651
<i>Drosophila erecta</i>	135 Mb	42%	4	16,880	12661, 12662
<i>Drosophila grimshawi</i>	231 Mb	38%	4	16,901	12678, 12679
<i>Drosophila melanogaster</i>	200 Mb	42%	4	13,733	13812
<i>Drosophila mojavensis</i>	130 Mb	40%	4	17,738	12682, 12685
<i>Drosophila persimilis</i>	193 Mb	45%	4	23,029	12705, 12708
<i>Drosophila pseudoobscura</i>	193 Mb	45%	5	17,328	10626
<i>Drosophila sechellia</i>	171 Mb	42%	4	21,332	12711, 12712
<i>Drosophila simulans</i>	162 Mb	41%	4	17,049	12464
<i>Drosophila virilis</i>	364 Mb	40%	4	17,679	12688
<i>Drosophila willistoni</i>	222 Mb	37%	4	20,211	12664
<i>Drosophila yakuba</i>	190 Mb	42%	4	18,816	12366
<i>Plutella xylostella</i> (moth)	393 Mb	38%	31	18,071	11570
<i>Tribolium castaneum</i>	210 Mb	38%	10	10,132	12540

Selected divergence dates: The insect lineage diverged from the human lineage ~900 million years ago (MYA). Hymenoptera (such as the honeybee *A. mellifera*) diverged from Lepidopterans (such as the silkworm *B. mori*) and dipterans (such as fruitfly and mosquito) 300 MYA; silkworm and fruitfly lineages split 280-350 MYA.

Disease association: mosquitos are vectors for many diseases including dengue and yellow fever

Organism-specific web resources: <http://www.flybase.org>; <http://www.anobase.org>.

**FIGURE 19.20** Overview of insect genomes. Photo of a mosquito (*Aedes*) and scanning electron micrograph of *Anopheles gambiae* from the CDC image library (✉ <http://phil.cdc.gov/phil/details.asp>) by CDC/Paul I. Howell, MPH and Frank Hadley Collins. *Tribolium* photo from the NHGRI (✉ <http://www.genome.gov/17516871>).

### 900 MYA: *Anopheles gambiae* (Second Insect Genome)

The mosquito *A. gambiae* is most well known as the malaria vector that carries the protozoan parasite *P. falciparum* (as well as *P. vivax*, *P. malariae*, and *P. ovale*). Mosquitoes are responsible for a variety of human diseases, although most of these (except the West Nile vector) are generally restricted to the tropics (Table 19.3).

Holt *et al.* (2002) reported the genomic sequence of a strain of *A. gambiae* using the whole-genome sequencing strategy. The genome is 278 Mb arranged in an X chromosome (numbered 1) and two autosomes (numbered 2 and 3). A particular challenge in

A haplotype is a combination of alleles of closely linked loci that are found in a single chromosome and tend to be inherited together.

**TABLE 19.3 Human diseases borne by mosquitoes. West Nile virus disease data are for the year 2012 in the United States (Centers for Disease Control and Prevention, <http://www.cdc.gov>). Adapted from Budiansky (2002) and Holt *et al.* (2002).**

Disease	Mosquito species	Number of cases
Malaria	<i>Anopheles gambiae</i>	500 million
Dengue	<i>Aedes aegypti</i>	50 million per year
Lymphatic filariasis	<i>Culex quinquefasciatus, Anopheles gambiae</i>	120 million
Yellow fever	<i>Aedes aegypti</i>	200,000 per year
West Nile virus disease	<i>Culex tarsalis, Culex pipiens, other</i>	5600 per year

sequencing this genome is the high degree of genetic variation, as manifested in “single-nucleotide discrepancies.” There is therefore a mosaic genome structure caused by two haplotypes of approximately equal abundance. In contrast, the *D. melanogaster* and *M. musculus* genomes are relatively homozygous.

Draft genome sequences were subsequently generated for the yellow fever mosquito *Aedes aegypti* and the southern house mosquito *Culex quinquefasciatus* (a vector for West Nile Virus; Arensburger *et al.*, 2010; reviewed in Severson and Behura, 2012). Additional *Anopheles* species have also been sequenced (e.g., Marinotti *et al.*, 2013).

The *A. gambiae* genome is more than twice the size of that of *Drosophila*. This difference is largely accounted for by intergenic DNA, and *Drosophila* appears to have undergone a genome size reduction relative to *Anopheles* species (Holt *et al.*, 2002). *Anopheles gambiae* and *D. melanogaster* diverged about 250 MYA (Zdobnov *et al.*, 2002). Almost half the genes in these genomes are orthologs, with an average amino acid sequence identity of 56%. By comparison, the lineage leading to modern humans and pufferfish (see below) diverged 450 MYA, but proteins from those two species share even slightly higher sequence identity (61%). Insect proteins therefore diverge at a faster rate than vertebrate proteins. An outstanding problem is to understand the ability of *Anopheles* to feed on human blood selectively and to identify therapeutic targets. For this effort, it is important to identify arthropod-specific and *Anopheles*-specific genes (Zdobnov *et al.*, 2002).

Another interesting aspect of *A. gambiae sensu stricto* is that it is currently undergoing speciation, with two molecular forms (M and S) differentiating. These are indistinguishable based on morphology, and they co-inhabit regions of West and Central Africa. Lawniczak *et al.* (2010) sequenced both genomes and observed fixed differences spanning the entire genomes (including, but not limited to, pericentromeric regions called “speciation islands”).

## 900 MYA: Silkworm and Butterflies

The cocoon of the domesticated silkworm *Bombyx mori* is the source of silk fibers. Xia *et al.* (2004) determined the sequence of its genome. At 429 Mb, it is 3.6 times larger than that of fruit fly and 1.5 times larger than mosquito; much of this size can be attributed to the presence of more genes (18,510 relative to ~13,700 in *D. melanogaster*) and also larger genes. Transposable elements have also shaped the genome, comprising 21% of the genome. Of that fraction, half arrived just 5 million years ago as a single *gypsy-Ty3*-like retrotransposon insertion. Analysis of the *B. mori* genome has helped to elucidate the function of the silk gland (a modified salivary gland) and, although silkworms do not fly or have colorful wing patterns, there are homologs of genes implicated in wing development and pattern formation.

AnoBase is a major resource for anopheline species (Topalis *et al.*, 2005; <http://anobase.vectorbase.org>, WebLink 19.51). VectorBase includes resources for *A. aegypti* and *C. quinquefasciatus* (<https://www.vectorbase.org/>, WebLink 19.52). The Ensembl genome browser for the mosquito is available at [http://www.ensembl.org/Anopheles\\_gambiae/](http://www.ensembl.org/Anopheles_gambiae/) (WebLink 19.53).

We described the *Drosophila* Down syndrome cell adhesion molecule (DSCAM) in Chapter 10, a gene that potentially encodes up to 38,000 distinct proteins through alternative splicing (NP\_523649.5). The *A. gambiae* ortholog appears to share the same potential for massive alternative splicing (Zdobnov *et al.*, 2002). See GenBank protein accession XP\_309810.4.

A diamondback moth called *Plutella xylostella* is a destructive pest that causes US\$ 4–5 billion in damages to food crops each year. You *et al.* (2013) characterized its genome and compared it to 11 other insect genomes. *Bombyx mori* is relatively closely related (having shared a common ancestor 125 MYA). They suggest that the insect herbivores coevolved with their mono- and dicotyledonous plant hosts, beginning some 300 MYA.

Several butterfly genomes have been sequenced, including the 273 Mb genome of the migratory monarch butterfly *Danaus plexippus* (Zhan *et al.*, 2011). Each autumn, millions of these butterflies migrate thousands of miles south to central Mexico until the next spring when they reproduce, fly north, and deposit fertile eggs on milkweed plants. Genome analysis hints at adaptations to this extraordinary lifestyle, such as microRNAs expressed selectively during summer and circadian clock components that are essential for migration.

Butterflies transfer closely related traits between species. The *Heliconius* Genome Consortium (2012) explored this phenomenon in a series of butterfly species whose genus is rapidly radiating. They sequenced the *Heliconius melpomene* genome and noted pervasive exchange of protective color pattern genes among species.

### 900 MYA: Honeybee

The western honeybee *Apis mellifera* is of special interest because of its highly social behavior. Bee hives are organized around a queen and her workers who transition from roles in the hive (such as nurses and hive maintainers) to the outside (such as foragers and defenders). The queens typically live ten times longer than the workers and lay up to 2000 eggs per day. The workers have brains with only a million neurons, but display highly intricate behaviors. Somehow all these differentiated phenotypes are directed by a single underlying genome. The Honeybee Genome Sequencing Consortium (2006) sequenced the *A. mellifera* genome. There are 15 acrocentric chromosomes and a large metacentric chromosome 1; as for human chromosome 2 (Chapter 20; Fan *et al.*, 2002), this is thought to represent a fusion of two acrocentrics. Relative to other insect genomes it has a lower GC content, and there were fewer predicted protein-coding genes.

Elsik *et al.* (2014) offered a strongly revised assembly and annotation of the genome, reporting ~15,300 genes rather than the initial estimate of ~10,100. This offers a case study regarding the challenges of assembly and annotation (see Chapter 15) and the need for continuing efforts in these areas for many genomes.

### 900 MYA: A Swarm of Insect Genomes

NCBI Genome currently lists nearly 100 completed insect genomes, and many more projects are underway. The i5K Initiative was launched to sequence and analyze 5000 arthropods (i5K Consortium, 2013). The many notable projects include the following:

- Insect specimens from museum collections, as well as plant and fungal material, can be sequenced. Staats *et al.* (2013) discuss some of the opportunities and challenges of this approach.
- The genomes of a leaf-cutting ant and a red harvester ant species have been sequenced (Nygaard *et al.*, 2011; Smith *et al.*, 2011). Bonasio *et al.* (2010) compared the genomes of two socially divergent ant species, *Harpegnathos saltator* (showing limited dimorphism between queen and workers) and *Camponotus floridanus* (showing extreme dimorphism between queen and workers).
- Werren *et al.* (2010) characterized the genomes of parasitoid wasps (*Nasonia vitripennis*, *N. giraulti*, and *N. longicornis*).

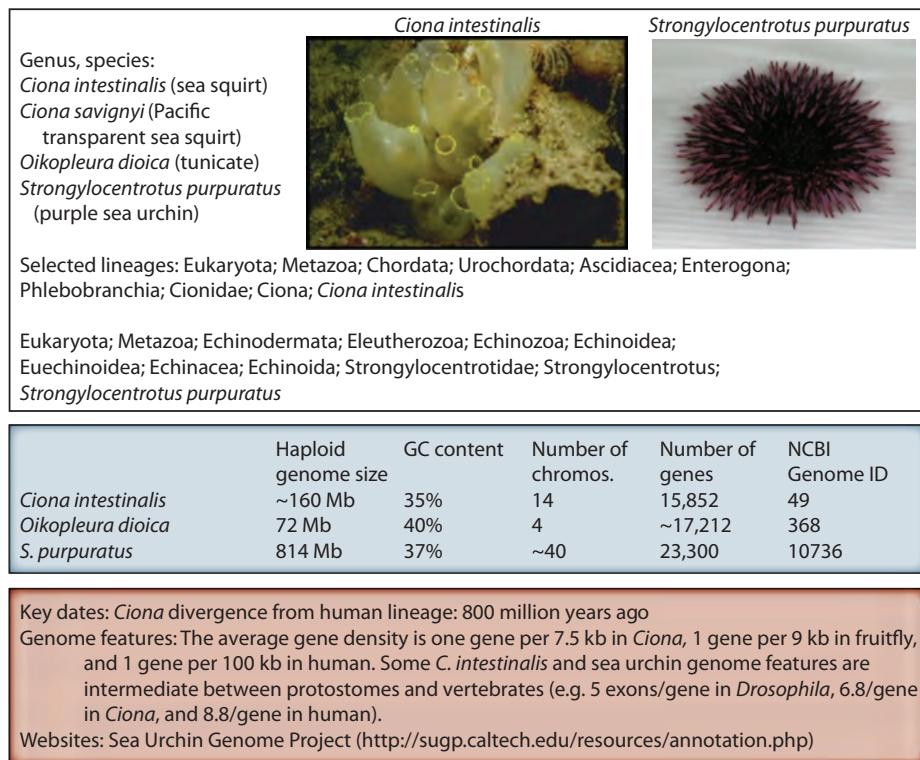
- J.B.S. Haldane is said to have commented that the Creator “has an inordinate fondness of beetles” because there are so many species. Characterized genomes include that of the mountain pine beetle *Dendroctonus ponderosae* (Keeling *et al.*, 2013) and the red flour beetle *Tribolium castaneum* (Tribolium Genome Sequencing Consortium *et al.*, 2008).
- *Daphnia pulex* is a common water flea, a crustacean found in shallow ponds (it is therefore a non-insect arthropod). Colbourne *et al.* (2011) sequenced its 197 Mb genome that harbors over 30,000 genes, with some estimates as high as 34,000. This large gene number could perhaps reflect its ability to reproduce sexually or instead to clone itself asexually.

## 840 MYA: A Sea Urchin on the Path to Chordates

For a brief and useful overview of how to interpret the relatedness of different species by inspection of a phylogenetic tree, see Baum *et al.* (2005).

As we survey the metazoan animals and move from the Protostomia (including the insects, nematodes, mollusks, and annelids) to the Deuterostomia (Fig. 19.18), we first come to the sister phyla of the hemichordates and echinoderms. The purple sea urchin *Strongylocentrotus purpuratus* is an echinoderm that has served as a model organism for studies of cell biology (including embryology and gene regulation) and evolution. The sea urchin serves as an outgroup for the chordates. This creature is a marine invertebrate, has a radial adult body plan (as shown in the photograph in Fig. 19.21), and has no apparent brain although there are neurons and brain functions. An individual can have a lifespan of over a century. It may be surprising to consider that it is more closely related to humans than nematodes or fruit flies with their well-defined brains and complex behaviors.

The assembled *S. purpuratus* genome is 814 megabases (Sea Urchin Genome Sequencing Consortium, 2006). Although linkage and cytogenetic maps are unavailable,



**FIGURE 19.21** Overview of simple (nonvertebrate) deuterostome genomes. Photograph of purple sea urchin from NCBI Genomes website (by Andy Cameron).

the number of chromosomes has been estimated to be ~40. There were several outstanding technical issues in sequencing the genome (reviewed in Sodergren *et al.*, 2006). One is that the sea urchin exhibits tremendous heterozygosity, with 4–5% nucleotide differences between single copy DNA of different individuals (this includes SNPs and insertions/deletions, and contrasts with ~0.5% heterozygosity in humans). The single male sea urchin that was sequenced displayed tremendous heterozygosity between its two haplotypes, making it challenging to distinguish sequencing errors from haplotype variants or from segmentally duplicated regions. One way this problem was overcome was to sequence bacterial artificial chromosome (BAC) clones of ~150,000 base pairs, in which each BAC corresponds to a single haplotype. A minimal tiling path of BAC clones spanned the genome and was sequenced at low (2 $\times$ ) coverage. This complemented a deep whole-genome shotgun assembly. This combined approach was introduced in the sequencing of the rat genome, and has become an increasingly common strategy for genome sequencing.

The Sea Urchin Genome Sequencing Consortium (2006) predicted about 23,300 genes for *S. purpuratus*. Some InterPro and Pfam domains (Chapter 12) are especially overrepresented in sea urchin relative to mouse, *Drosophila*, *C. elegans*, and sea squirt. Most dramatic are three families of receptor proteins that function in the innate immune response (toll-like receptors; NACHT and leucine-rich repeat-containing proteins; and scavenger receptor cysteine-rich domain proteins). Each of these genes is present in over 200 copies, while other animals from humans to fruit fly and nematode typically have about 0–20 copies. Another surprising finding is the presence of over 600 genes encoding G protein-coupled chemoreceptors as well as genes implicated in photoreception, expressed on the tube feet.

## 800 MYA: *Ciona intestinalis* and the Path to Vertebrates

The vertebrates include fish, amphibians, reptiles, birds, and mammals. All these creatures have in common a segmented spinal column. Where did the vertebrates originate? Vertebrates are members of the chordates, animals having a notochord (Fig. 19.18). The sea squirt *C. intestinalis* is a urochordate (also called tunicate), one of the subphyla of chordates but not a vertebrate. *Ciona* is a hermaphroditic invertebrate that offers us a window on the transition to vertebrates (Holland, 2002).

Dehal *et al.* (2002) produced a draft sequence of the *C. intestinalis* genome by the whole-genome shotgun strategy. At 160 Mb, it is about 12 times larger than typical fungal genomes and 20 times smaller than the human genome. There are 15,852 predicted genes organized on 14 chromosomes. Most of these predicted genes are supported by evidence from expressed sequence tags.

The availability of the *Ciona* genome sequence allows a comparison with protostomes and other deuterostomes and supports its position as related to an ancestral chordate (Dehal *et al.*, 2002). Almost 60% of *Ciona* genes have protostome orthologs; these presumably represent ancient bilaterian genes. Several hundred genes have invertebrate but not vertebrate homologs, such as the oxygen carrier hemocyanin. These comparative studies are augmented by the genome sequencing of the related urochordates *Ciona savignyi* and *Oikopleura dioica*. *O. dioica* has one of the smallest chordate genomes (about 72 Mb; Seo *et al.*, 2001; Denoeud *et al.*, 2010), and it is an attractive experimental organism because its lifespan is two to four days, it can be maintained in culture, and its females are fecund. *C. savignyi*, a sea squirt, exhibits considerable heterozygosity, with variable degrees of heterozygosity across the genome. Eric Lander and colleagues (Vinson *et al.*, 2005) introduced an algorithmic approach to assembling genome sequences from diploid genomes. This method assembles the two haplotypes separately, and therefore requires twice the sequencing depth of other whole-genome sequencing projects. The result is

The Sea Urchin Genome Database is available at <http://spbase.org> (WebLink 19.54); see also Cameron *et al.* (2009).

The phylum Cnidaria is an outgroup to the bilateria, having diverged about 600–750 MYA. Its members include sea anemones, hydras, corals, and jellyfishes. CnidBase organizes genomic and other information on diverse cnidarians (<http://cnidbase.bu.edu>, WebLink 19.55). See also Ryan and Finnerty (2003).

The Department of Energy Joint Genome Institute operates the *C. intestinalis* genome home page (<http://genome.jgi-psf.org/Cioin2/Cioin2.home.html>, WebLink 19.56). The GenBank accession number for the genome is AABS00000000.1, and you can find a *Ciona* BLAST server through the NCBI Genomes page of eukaryotic projects. The Ghost database, a *Ciona* EST project that includes a BLAST server and gene expression data, is available at <http://ghost.zool.kyoto-u.ac.jp/cgi-bin/gb2/gbrowse/kh/> (WebLink 19.57).

The Broad Institute offers a *Ciona savignyi* database at <http://www.broadinstitute.org/annotation/ciona/> (WebLink 19.58).

substantial improvement in sequence quality and contiguity. Such approaches will be increasingly useful as more outbred genomes are sequenced.

There are 2570 *Ciona intestinalis* genes (one-sixth) that have orthologs in vertebrates but none in protostomes; these genes arose in the deuterostome lineage before the last common ancestor diverged into vertebrates, cephalochordates, and urochordates (e.g., *Ciona*). There are 3399 *Ciona* genes (one-fifth) that have no identifiable homolog in vertebrates or invertebrates and may therefore be tunicate-specific genes that evolved after the divergence of the urochordate lineage.

A *Ciona* protein (NP\_001027621.1) has 46% identity to human choline acetyltransferase (NP\_065574.3) and 52% identity to a sea urchin ortholog (XP\_780154.3). A *Ciona* gene (accession AB071998.1) encodes a protein with 56% identity to a human vesicular acetylcholine transporter (NP\_003046.2). Many such genes also function in neurotransmission in invertebrates.

*Ciona* has genes involved in processes such as apoptosis (programmed cell death), thyroid function, neural function, and muscle action. This provides an opportunity for comparative analyses of fundamentally important genes within the chordate lineage. For example, nerves communicate with muscles by releasing the neurotransmitter acetylcholine from synaptic vesicles in presynaptic nerve terminals. This transmitter diffuses across the synapse (a gap between cells) to bind and activate postsynaptic receptors. *Ciona* has genes encoding proteins that function in neurotransmission, including a transferase enzyme that synthesizes acetylcholine, an acetylcholine transporter that pumps the neurotransmitter into vesicles, synaptic vesicle proteins, and neurotransmitter receptors. Similar genes are also present in sea urchin, such as the agrin protein that clusters acetylcholine receptors postsynaptically.

## 450 MYA: Vertebrate Genomes of Fish

The teleosts (or ray-finned fishes, *Actinopterygii*) are the largest group of vertebrates with ~24,000 known species (more than half the total number of vertebrate species). The ray-finned fishes diverged from the lobe-finned fishes (*Sarcopterygii*) about 450 million years ago. These relationships are depicted in the phylogenetic tree of **Figure 19.22a**. The teleosts are further shown in **Figure 19.22b**, including the first four sequenced fish genomes: those of the pufferfishes *Takifugu rubripes* and *Tetraodon nigroviridis*, the medaka *Oryzias latipes*, and the zebrafish *Danio rerio*. Dozens of fish genomes have now been sequenced, and selected lineages and genome features are presented in **Figure 19.23**.

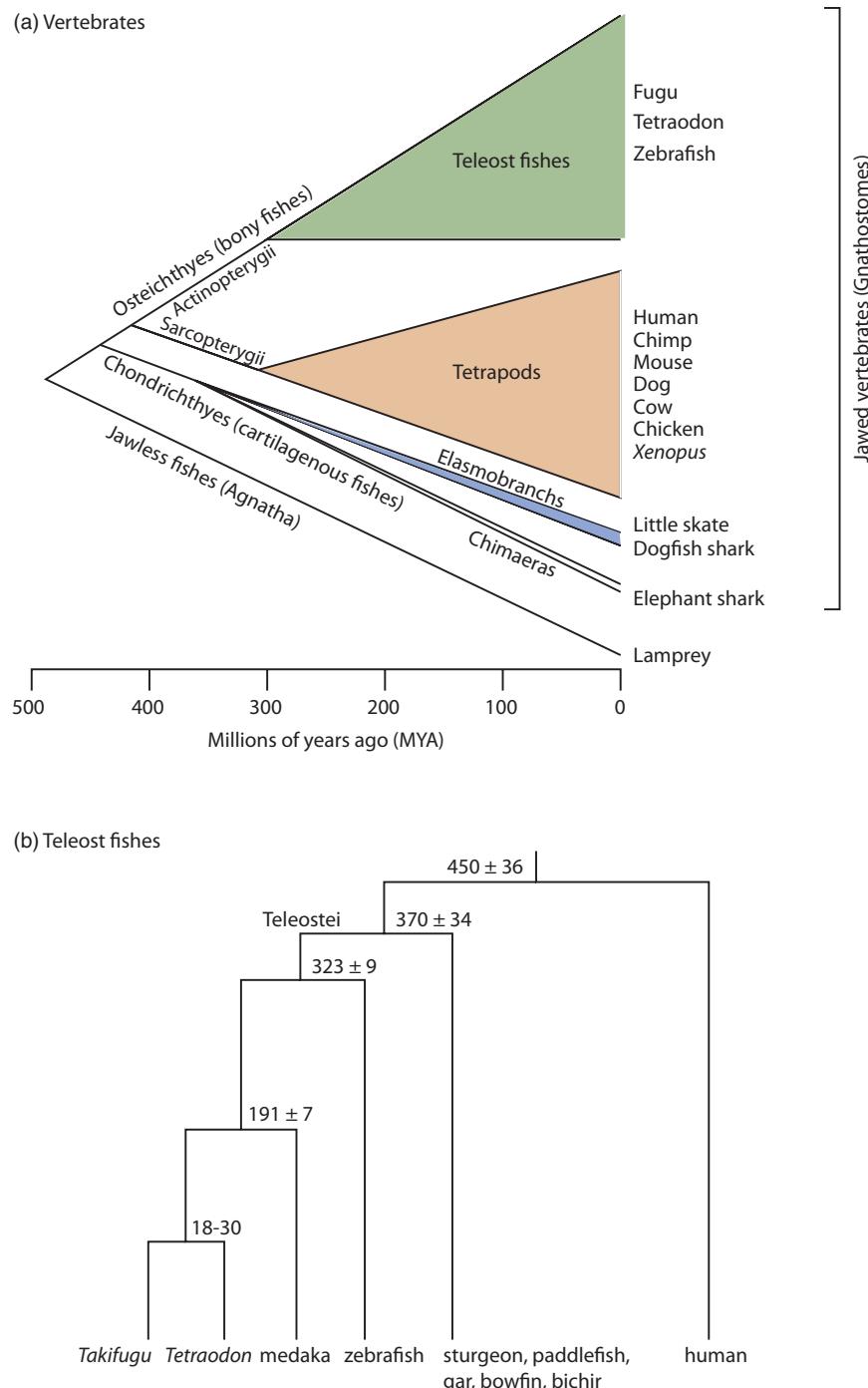
The second vertebrate genome sequencing project (after human) was that of the Japanese pufferfish *T. rubripes*, in part because it has a remarkably compact genome. This teleost fish has a genome size of 365 Mb, about one-ninth the size of the human genome (Aparicio *et al.*, 2002). However, *Takifugu* and humans have comparable numbers of predicted protein-coding genes.

There are several reasons that the *Takifugu* genome is relatively compact (Aparicio *et al.*, 2002):

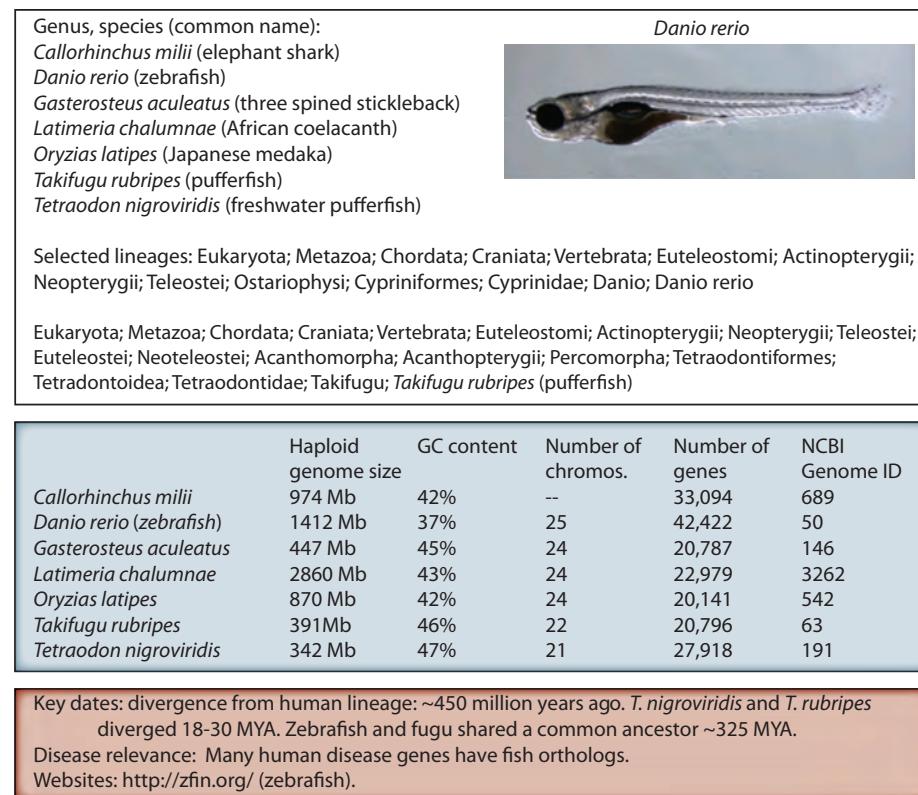
- Only 2.7% of the *Takifugu* genome consists of interspersed repeats, based on analyses with RepeatMasker. This contrasts with 45% interspersed repeats in the human genome (Chapter 20). Still, every known class of eukaryotic transposable elements is represented in *Takifugu*. The most common *Takifugu* repeat is the LINE-like element *Maui* (6400 copies), while in humans there are one million copies of the most common repeat, *Alu*.
- Introns are relatively short. Seventy-five percent of *Takifugu* introns are <425 base pairs in length, while in humans 75% of introns are <2609 base pairs. In *Takifugu*, about 500 introns have a length greater than 10 kb, while in humans more than 12,000 introns are greater than 10 kb.
- Gene loci occupy about 108 Mb of the total euchromatic DNA (320 Mb). This represents about one-third of the genome, a far higher fraction than in mouse or human.

After the *Takifugu* genome was completed, Jaillon *et al.* (2004) reported the sequence of another pufferfish, *Tetraodon nigroviridis*. This permitted comparative

*Fugu rubripes* is also called *Takifugu rubripes*. The International *Fugu* Genome Consortium was responsible for the sequencing of its genome. A *Takifugu* Browser is at [http://www.ensembl.org/Takifugu\\_rubripes/Info/Index](http://www.ensembl.org/Takifugu_rubripes/Info/Index) (WebLink 19.59). Produced by the Wellcome Trust Sanger Institute and the European Bioinformatics Institute, it is a major portal to this genome and others.



**FIGURE 19.22** (a) Phylogenetic tree of the vertebrates. The vertical axis corresponds to the abundance of extant species in each group, with representative names given. The Sarcopterygii (lobe finned-fishes) include coelacanths, lungfish, and tetrapods (amphibians, birds, reptiles, mammals); more detailed phylogenies of the tetrapods are presented in **Figures 19.24** and **19.26** below. The x axis shows the divergence times based on fossil records, which differ somewhat from estimates made by molecular sequence analyses. Redrawn from Venkatesh *et al.* (2007). Licensed under Creative Commons Attribution License 2.5. (b) Phylogenetic tree of the teleosts showing the relationships of the first four sequenced fish genomes. Adapted from Kasahara *et al.* (2007) with permission from Macmillan Publishers.



**FIGURE 19.23** Overview of fish genomes. The *D. rerio* image is from the NHGRI and Shawn Burgess (<http://www.genome.gov/17516871>).

analyses between *Takifugu* and human (at the time resulting in the prediction of ~900 novel human genes). A main focus of the genome analysis was on the evidence that telosts are descendants of an ancient whole-genome duplication. This was followed by massive gene loss, as described for separate whole-genome duplication events in fungi (Chapter 18). Jaillon *et al.* further inferred that the ancestral vertebrate genome had 12 chromosomes.

An emerging field studies the composition of ancestral karyotypes. Yuji Kohara and colleagues (Kasahara *et al.*, 2007) generated a draft sequence of the medaka (reviewed in Takeda and Shimada, 2010). Upon comparing the four available fish genomes with the human genome, they proposed a model of genome evolution in which a fish/human ancestor had 13 chromosomes. There are other models of the ancestral karyotype. However, there is a consensus that several whole-genome duplications occurred in the teleost lineage (e.g., Van de Peer, 2004; Christoffels *et al.*, 2004; Postlethwait, 2007). Once duplicate genes are identified within and between genomes (such as fish and human), the date of the duplication events can be estimated by using phylogenetic trees (e.g., neighbor-joining trees, assuming a constant molecular clock). About one-third of the duplicated genes in *Takifugu* seem to derive from a whole-genome duplication event that occurred ~320 MYA, as suggested by Ohno (1970). Approximately 1000 pairs of duplicated genes (paralogs) were identified in both *Tetraodon* and *Takifugu* and, based on  $K_s$  frequencies, 75% represent ancient duplications that occurred prior to the divergence of the *Takifugu* and *Tetraodon* lineages. Two other whole-genome duplication events occurred earlier (at the time of divergence of jawless and jawed vertebrates, ~500 MYA) and more recently in the salmonid lineage at ~50 MYA (reviewed in Postlethwait, 2007).

Other sequenced fish genomes include the following.

- Zebrafish remains a key organism for the study of vertebrate gene function. Howe *et al.* (2013) presented an updated reference genome sequence.
- The coelacanth was a lobe-finned fish that was known from the fossil record and thought to have become extinct 70 MYA; it was therefore a great surprise when a living specimen was found in 1938. Amemiya *et al.* (2013) reported the genome of the African coelacanth *Latimeria chalumnae*. This fish, along with lungfish, is the closest living fish relative to the tetrapods and therefore offers insight into the early evolution of land animals.
- The Pacific bluefin tuna fish (*Thunnus orientalis*) is a predator that relies on color vision to sense its prey. Nakamura *et al.* (2013) sequenced the genome and identified selective variants among its visual pigment (opsin) genes.
- The sex chromosomes in humans and other mammals are XX for females and XY for males. In fish and in birds, males are ZZ and females are ZW (i.e., females are heterogametic). It has been extraordinarily challenging to sequence the Y chromosome (see Chapter 20). Chen *et al.* (2014) selected a flatfish (the half-smooth tongue sole *Cynoglossus semilaevis*) for genome sequencing because of its evolutionarily young W chromosome, typical of all fish, that is less degenerate than that found in birds. They identified suppression of recombination, a driving force for sex chromosome evolution. The flatfish W chromosome has lost about two-thirds of its original protein-coding gene content, similar to the process of gene loss in the mammalian Y chromosome.
- The elephant shark *Callorhinichthys milii* genome was sequenced lightly (Venkatesh *et al.*, 2007), representing a cartilaginous fish that is an outgroup to the teleosts (Fig. 19.22a).

Visit Zfin, the principal zebrafish web resource, at <http://zfin.org/> (WebLink 19.60).

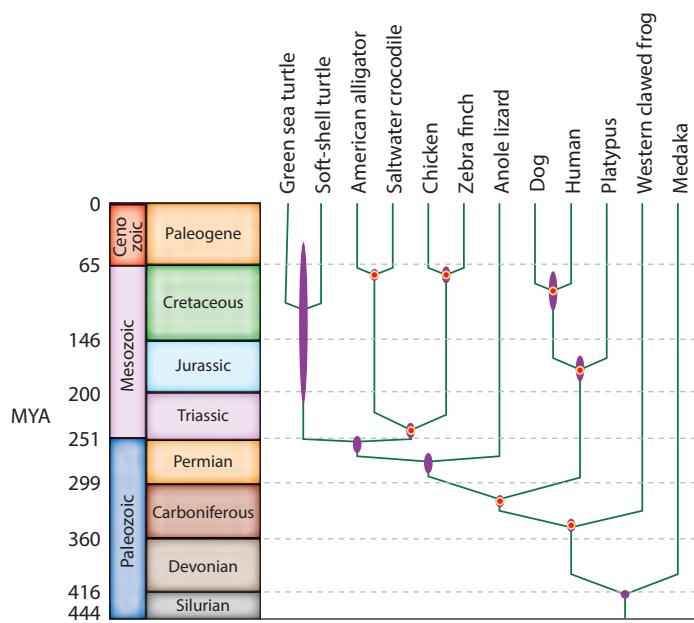
### 350 MYA: Frogs

The amphibians diverged from other vertebrates ~350 MYA (Fig. 19.24). *Xenopus laevis* has long been a model organism, particularly for embryological research. Its genome is tetraploid however, and so the diploid Western clawed frog *Xenopus tropicalis* was selected for sequencing. Its genome is estimated to be 3.1 Gb on 18 chromosomes. Hellsten *et al.* (2010) produced a draft sequence, annotating ~20,000 genes.

### 320 MYA: Reptiles (Birds, Snakes, Turtles, Crocodiles)

The amniotes are vertebrates that live on land including mammals, birds, and lizards. Some (such as cetaceans) have returned to the sea. This great group split into the two groups of modern mammals and reptiles about 320 MYA. The first sequencing of reptile genomes has provided fascinating insights into the birds, crocodiles and alligators, turtles, lizards, and snakes. A phylogenetic tree shows the relationships of reptiles to mammals (Fig. 19.24).

The first of these to be characterized was a bird. When the chicken genome was sequenced by the International Chicken Genome Sequencing Consortium (2004) it provided a unique perspective on the human genome because it is far closer to humans than fish, but farther than the rodents (diverged ~90 MYA). It therefore provided an excellent distance for identifying highly conserved functional elements (Chapter 8). The genome is 1200 megabases and is organized in 38 autosomes and a pair of sex chromosomes (ZW is the heterogametic female and ZZ is male; chromosome W is extremely small). The karyotype is therefore  $2n = 78$ . The autosomes include many minichromosomes, typically having a high GC content, a high gene content, and very high recombination rates (a median value of 6.4 cM per megabase; by comparison the human genome has a range of 1–2 cM/Mb and the mouse genome 0.5–1.0 cM/Mb).



**FIGURE 19.24** Phylogeny of the reptiles and other vertebrates. The phylogeny was constructed using first and second codon positions of 1113 single-copy protein-coding genes. Tree topology is supported by 100% bootstrap values (at most clades; not shown). Purple ellipses at the nodes correspond to 95% credibility intervals of the estimated posterior distributions of the divergence times. Red circles (with yellow outlines) indicate fossil calibration times. Redrawn and adapted from Wang *et al.* (2013), with permission from Macmillan Publishers.

The red jungle fowl, for which the genome was sequenced, is the precursor to the domesticated chicken.

The chicken genome is smaller than the human genome by a factor of three because it has relatively few repetitive elements. Interspersed repeats occur as transposable elements in decay. There is no evidence for active short interspersed line elements (SINEs) in the past 50 million years, in contrast to their active roles in the human genome. Expansions and reductions of protein-coding gene families occur; for example, an avian-specific family of keratins is used to create claws, scales, and feathers. One surprising expansion is a family of 218 genes that are predicted to encode olfactory receptors and are orthologous to two human genes (*OR5UI* and *OR5BF1*).

Additional bird genomes that have been characterized include the duck (*Anas platyrhynchos*; Huang *et al.*, 2013); domestic turkey (*Meleagris gallopavo*; Dalloul *et al.*, 2010); and zebra finch (*Taeniopygia guttata*; Warren *et al.*, 2010) (Fig. 19.25). The Assemblathon 2 (an assessment of assembly strategies; Chapter 9) featured the bird *Melopsittacus undulatus*, the fish *Maylandia zebra*, and the snake *Boa constrictor constrictor* (Bradnam *et al.*, 2013).

The turtle lineage diverged from the avian/crocodilian group ~250 MYA at the start of the Triassic period (Fig. 19.24). The western painted turtle genome displays a slow rate of evolution (*Chrysemys picta*; Shaffer *et al.*, 2013). It lost the ability to form teeth ~150–200 MYA (birds lost this ability ~80–100 MYA) and genes associated with tooth formation became pseudogenes. Genome analysis may help explain turtles' striking longevity as well as low temperature and low oxygen tolerance. Wang *et al.* (2013) sequenced two different turtle genomes, confirming that turtles are more closely related to birds and crocodilians than to lizards and snakes.

The crocodilians, sister group to birds, include 23 species in the three major groups Alligatoridae, Crocodylidae, and Gavialidae. Wan *et al.* (2013) sequenced the genome of the Chinese alligator (*Alligator sinensis*), annotating 22,200 genes. These alligators can

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus; <i>Gallus gallus</i> (red jungle fowl)	
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Testudines + Archosauria group; Testudines; Cryptodira; Trionychoidea; Trionychidae; <i>Pelodiscus sinensis</i> (Chinese soft-shelled turtle)	
Eukaryota; ... Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Testudines + Archosauria group; Archosauria; Crocodylia; Alligatoridae; Alligatorinae; <i>Alligator sinensis</i> (Chinese alligator)	

	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Alligator sinensis</i> (Chinese alligator)	2271 Mb	44.6%			
<i>Anas platyrhynchos</i> (mallard)	1105 Mb	41.2%	10	16,376	2793
<i>Anolis carolinensis</i> (green anole)	1799 Mb	40.8%	13	16,822	708
<i>Gallus gallus</i> (chicken)	1047 Mb	41.9%	39	21,211	111
<i>Meleagris gallopavo</i> (turkey)	1063 Mb	41.6%	32	13,282	112
<i>Pelodiscus sinensis</i> (turtle)	2202 Mb	44.5%	—	21,252	14578
<i>Taeniopygia guttata</i> (zebra finch)	1232 Mb	41.3%	32	15,287	367

Key dates: chicken divergence from human lineage: ~320 MYA. Duck and chicken lineages diverged ~90 MYA. Turtles last shared a common ancestor with bird/crocodile lineage ~250 MYA.  
 Disease relevance: chicken is an important non-mammalian vertebrate model organism for studies of embryonic development, virus infection (the first tumor virus, Rous sarcoma virus, and the first oncogene, src were identified in the chicken).  
 Genome features: the chicken genome is ~three-fold smaller than other mammalian genomes, and has a relatively small proportion of interspersed repeat content. About 70 Mb of the sequence is alignable with human. While mammals display XY-type sex determination, birds display ZW-type; non-avian reptiles exhibit XY, ZW, or temperature-dependent sex determination.  
 Key website: <http://aviangenomes.org/>

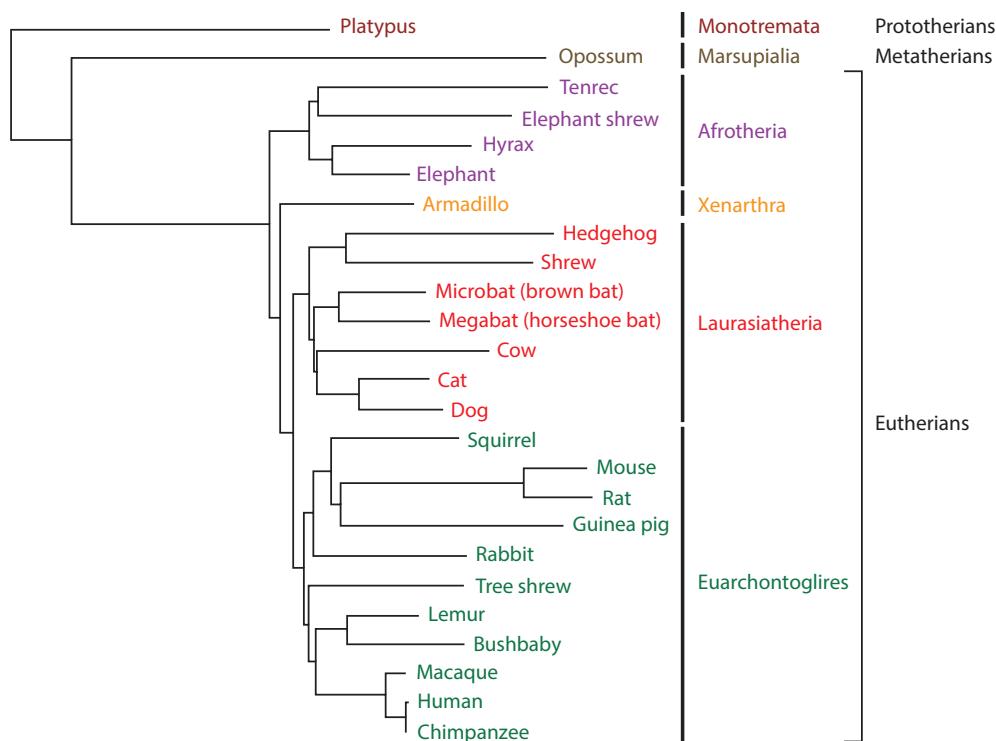
**FIGURE 19.25** Overview of reptilian genomes. Photograph from the NHGRI and Bill Payne ( <http://www.genome.gov/17516871>).

dive underwater for ~12 minutes at a time (sometimes reportedly for 1–2 hours), or for shorter more active dives during mating or killing prey. Wan *et al.* (2013) identified four crocodilian-specific hemoglobin genes (an alpha subunit *HBA1* and three beta subunits *HBB2*, *HBB4*, and *HBB5*). Several of these are mutated to a form that facilitates oxygen binding, likely helping these creatures hold their breath.

### 180 MYA: The Platypus and Opossum Genomes

There are three main groups of mammals: (1) eutherians (placental mammals such as humans); (2) metatherians (marsupials) such as the opossum, koala, and kangaroo; and (3) prototherians such as the platypus (Fig. 19.26; also shown in Fig. 19.24). Let's look at the draft genomes of the very unusual platypus and opossum (summarized in Fig. 19.27).

The platypus (*Ornithorhynchus anatinus*) has features that seem to place it in between mammals and reptiles: males have reptile-like venom and females lactate like mammals but lay eggs like reptiles. Males have five X and five Y chromosomes (sperm having 5X 5Y), and multiple sex chromosomes share limited homology with the bird Z chromosome; sex determination mechanisms as well as sex chromosome dosage compensation mechanisms are unknown. Warren *et al.* (2008) produced a draft sequence of the platypus genome. There were typical numbers of protein-coding genes and noncoding genes, as well as an expansion of small nucleolar RNAs (snoRNAs). Microsatellite content is comparable to that of reptiles while interspersed repeats are typical of mammalian



**FIGURE 19.26** Phylogenetic tree depicting mammalian genomes. The genomes of many of these organisms have been sequenced. Note that the branch lengths of the rat and mouse lineages are long relative to other members of the clade containing humans (the Euarchontoglires), reflecting a faster evolutionary rate. Data from <http://www.ncbi.nlm.nih.gov/genomes/> and Margulies *et al.* (2005).



	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Monodelphis domestica</i>	3600 Mb	37.7%	9	18-20,000	220
<i>Ornithorhynchus anatinus</i>	1996 Mb	45.7%	52	19,365	110

Selected divergence dates: The platypus lineage diverged from the human lineage ~180 MYA, while the marsupials diverged ~160 MYA.  
 Genome features: Opposum autosomes are extremely large (the smallest, at 257 Mb, is larger than human chromosome 1).  
 Disease association: *M. domestica* is a model for radiation-induced malignant melanoma. Newborn opossums are unique in their ability to recover from complete transections of the spinal cord.  
 Organism-specific web resources: <http://www.broad.mit.edu/mammals/opossum>

**FIGURE 19.27** Overview of the genome of the short-tailed opossum *Monodelphis domestica*, a marsupial. Photograph from the NHGRI (<http://www.genome.gov/17516871>).

genomes. The overall GC content (45.7%) is far higher than that observed in most mammalian genomes (~41%).

The genome sequence of the gray, short-tailed opossum *Monodelphis domestica* is the first from the metatherians (Mikkelsen *et al.*, 2007; **Fig. 19.27**). Its genome size is comparable to that of humans, organized into eight autosomes (257 megabases to 748 megabases). These autosomes are extremely large (the shortest one is longer than the longest in humans, chromosome 1). In contrast, the opossum X chromosome is extremely short (~76 megabases), smaller than that of any known eutherian.

The GC content of the *M. domestica* genome is 37.7%, lower than that of other amniote genomes (40.9–41.8%). Mikkelsen *et al.* note that the average recombination rate for the autosomes (~0.2–0.3 centimorgans per megabase) is lower than in other amniotes, consistent with a model in which the genome has undergone limited recombination.

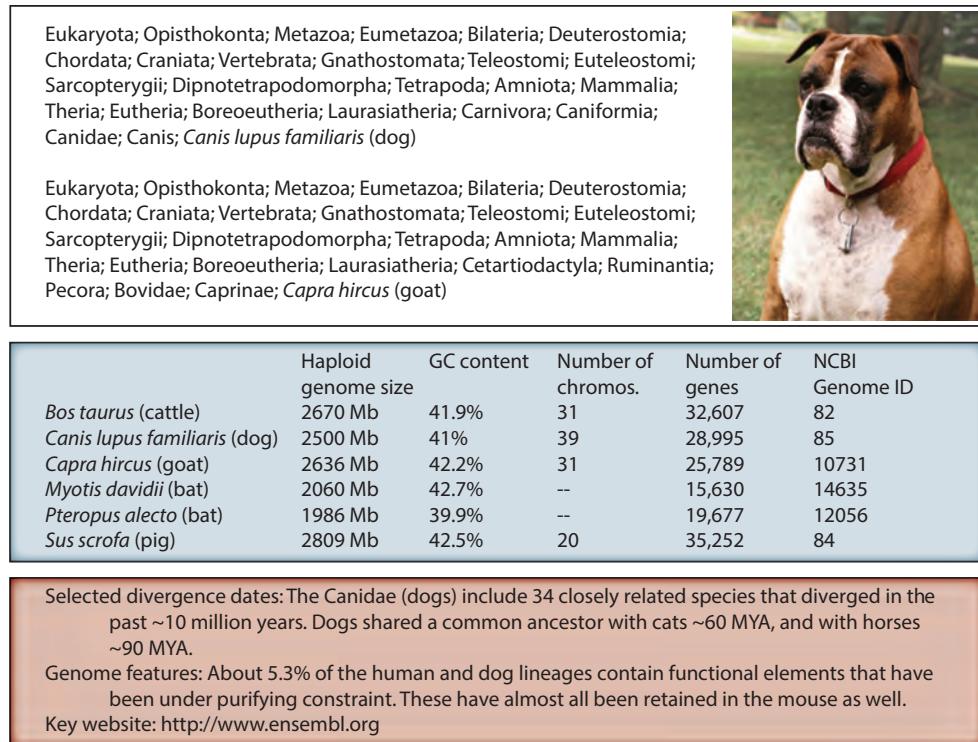
In eutherian mammals, females achieve dosage compensation of the X chromosome by the random inactivation of either the maternal or paternal X in female embryos. This is accomplished by an X inactivation center (XIC) that includes the *XIST* gene. Its RNA product coats and silences one X chromosome copy. In contrast, metatherian mammals such as the opossum inactive the paternal X. Mikkelsen *et al.* found no evidence for *XIST* in the opossum genome. While the human X chromosome has undergone remarkably little change since the eutherian radiation ~100 million years ago (**Fig. 20.4**), the opossum X chromosome has undergone large-scale rearrangements (affecting the XIC and X-linked pseudoautosomal region).

The predicted gene content of *M. domestica* (22,443) is comparable to that of humans, with relatively small numbers of organism-specific genes. Conserved noncoding elements (Chapter 8), rather than genes, comprise the majority of the well-conserved sequence elements.

## 100 MYA: Mammalian Radiation from Dog to Cow

A spectacular radiation of mammalian species occurred approximately 100–95 MYA. The tree of **Figure 19.26** shows the primates and rodents as part of a group called euar-chontoglires. The other eutherian organisms include dogs and cats, bats, armadillos, and elephants. We highlight several draft genome sequences. In each case the number of genes and repetitive elements are catalogued; accelerated evolution of particular genes is determined; expansion and contraction of gene families is noted; comparative analysis (including phylogeny) is performed. Almost all the draft genome sequence reports note that most genes have human counterparts; often the core eukaryotic genes approach of CEGMA (Chapter 15) is employed to confirm adequate annotation. We summarize some of the findings in **Figure 19.28**.

- Bats are notable as the only mammals capable of sustained flight and as reservoirs for highly pathogenic viruses. Zhang *et al.* (2013) sequenced the genomes of the fruit bat *Pteropus alecto* and the insectivorous bat *Myotis davidii*.
- Following a 1.5× coverage of the dog genome by Craig Venter and colleagues (Kirkness *et al.*, 2003), Lindblad-Toh *et al.* (2005) reported a high-quality draft genome sequence. There are ~400 modern dog breeds and many have a high prevalence of particular diseases due to breeding. A boxer was selected because that breed has relatively high homozygosity.
- There are >830 million goats in the world and >1000 goat breeds. Analysis of a black goat genome by Dong *et al.* (2013), complemented by RNA-seq from hair follicles, revealed a large family of keratin genes that may contribute to the production of cashmere.
- Pigs were domesticated beginning ~10,000 years ago. Analysis of a draft genome sequence (including further sequencing of ten other unrelated wild boars) pro-



**FIGURE 19.28** Overview of the dog genome. The photograph is of a boxer (Tasha) whose genome was sequenced. Photograph from the NHGRI website and Paul Samollow (<http://www.genome.gov/17516871>).

vided evidence for a split into Asian and European lineages in the mid-Pleistocene (1.6–0.8 MYA).

- We discussed challenges associated with the assembly and annotation of genomes in Chapter 15, and highlighted the taurine cattle (cow) genome as an example. The Bovine Genome Sequencing and Analysis Consortium *et al.* (2009) reported a draft genome sequence including evidence for five metabolic genes that are deleted or diverged relative to their human orthologs. Changes to *PLA2G4C*, *FAAH2*, *IDI2*, *GSTT2*, and *TYMP* could reflect adaptations of fatty acid metabolism.

For a discussion of the methodology involved in sequencing the genomes of many mammals to produce the tree presented in Figure 19.26, see Web Document 19.3.

## 90 MYA: The Mouse and Rat

The sequencing and analysis of the mouse genome represents a landmark in the history of biology. Following the human, the mouse was the second mammal to have its genome sequenced. Two groups independently sequenced the mouse genome: the Mouse Genome Sequencing Consortium (Mouse Genome Sequencing Consortium *et al.*, 2002) and Celera Genomics. Subsequently the Mouse Genome Sequencing Consortium produced a high-quality finished assembly (Church *et al.*, 2009). This defined 20,210 protein-coding genes; closed 175,000 gaps; added 139 Mb of novel sequence; and corrected numerous assembly errors.

The mouse is an excellent model for understanding human biology (Fig. 19.29):

- Forty percent of all mammalian species are rodents (Churakov *et al.*, 2010), highlighting their importance.

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus; *Mus musculus*

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Rattus; *Rattus norvegicus* (see photo)



	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Mus musculus</i>	2600 Mb	42%	20	23,049	13183
<i>Rattus norvegicus</i>	2750 Mb	~42%	21	20,973	10629

Selected divergence dates: Mouse and rat last shared a common ancestor 12-24 MYA. The rodent lineage diverged from the human lineage ~90 MYA.

Disease relevance. There are over 450 inbred mouse strains, and many of these serve as disease models. Knockouts and other manipulations of mouse genes allow studies of human diseases. Rats (like mice) are host to many pathogens, and are carriers for over 70 human diseases.

Web resources: Mouse Genome Informatics (<http://www.informatics.jax.org/>); Rat Genome Database (<http://rgd.mcw.edu/wg/>)

**FIGURE 19.29** Overview of rodent genomes. The photograph of a rat is from the NHGRI website (<http://www.genome.gov/17516871>).

- Remarkably, although these two organisms diverged about 90 MYA, most annotated genes in the mouse genome have an ortholog in the human genome. Church *et al.* detected 15,187 human and mouse genes with simple 1:1 orthologous relationships. These have mean nucleotide and amino acid identities of 85.3% and 88.2%, respectively.
- In addition to sharing thousands of orthologous protein-coding genes, the mouse and human genomes have large tracts of homologous nonprotein-coding DNA. These conserved sequences provide insight into regulatory regions of the genome or non-coding genes (Hardison *et al.*, 1997; Dermitzakis *et al.*, 2002).
- The mouse and human share many physiological features. Mice therefore make an important model for hundreds of human diseases, from infectious diseases to complex disorders.
- There are over 450 inbred strains of mice, and >1000 mouse strains having spontaneous mutations. Mutations can be introduced into the mouse through random mutagenesis approaches such as chemical mutagenesis or radiation treatment (Chapter 14). Mutations and other genetic modifications can also be introduced through directed approaches such as transgenic, knockout, and knock-in technologies.

Mouse Genome Sequencing Consortium *et al.* (2002) described 11 main conclusions of the mouse genome-sequencing project:

1. The total length of the euchromatic mouse genome is 2.5 Gb in size, about 14% smaller than the human genome (2.9 Gb). In contrast to other, more compact genomes we have discussed, the mouse genome (like the human genome) averages about one gene every 100,000 base pairs of genomic DNA. The GC content is comparable, with mean values of 42% (mouse) versus 41% (human). There are 15,500 CpG islands, about half the number observed in humans (see Chapter 20).
2. Over 90% of the mouse and human genomes can be aligned into conserved synteny regions. After the divergence of mouse and human about 90 MYA, chromosomal

The mouse genome is accessible through the three main genome browser sites: [http://www.ensembl.org/Mus\\_musculus/](http://www.ensembl.org/Mus_musculus/) (WebLink 19.61), <http://genome.ucsc.edu/> (WebLink 19.62), and <http://www.ncbi.nlm.nih.gov> (WebLink 19.63). The GenBank accession number of the 2002 project is CAAA01000000.1.

The December 2011 GRCm38 mouse build includes ~16,000 CpG islands and a 41.7% GC content. For human there are 30,477 CpG islands in the human GRCh38 build of December 2013.

The mouse sequencing consortium (Mouse Genome Sequencing Consortium *et al.*, 2002, p. 526) defined a synteny segment as “a maximal region in which a series of landmarks occur in the same order on a single chromosome in both species.” They identified 558,000 orthologous and highly conserved landmarks in the mouse assembly, comprising 7.5% of the mouse assembly.

DNA was shuffled in each species. However, large regions of DNA obviously correspond. As an example of how to visualize this, Ensembl offers a human/mouse conserved synteny viewer (e.g., Fig. 20.14).

3. About 40% of the human genome can be aligned to the mouse genome at the nucleotide level. This represents most of the orthologous sequence shared by these genomes. For 12,845 orthologous gene pairs, 70.1% of the corresponding amino acid residues were identical.
4. The neutral substitution rate in each genome can be estimated by comparing thousands of repetitive DNA elements to the inferred ancestral consensus sequence. The average substitution rate is 0.17 per site in humans and 0.34 per site in mouse. The mouse genome also shows a twofold higher rate of acquisition of small (<50 base pair) insertions and deletions.
5. The proportion of small (50–100 base pair) segments in the mammalian genome that is under purifying selection is about 5%. This is estimated by comparing the neutral rate to the extent of sequence conservation in the genome. Since this 5% value is greater than the proportion of protein-coding genes in the genome, genomic regions that do not code for genes must be selected for, such as regulatory elements. Regulatory regions such as those that control liver-specific and muscle-specific expression were conserved between mouse and human to a greater extent than regions of neutral DNA, although less than regions that are protein coding.
6. The mammalian genome is evolving in a nonuniform manner, with variation in the rates of sequence divergence across the genome. The neutral substitution rate varied across all chromosomes (and was lowest on the X chromosome), with a higher substitution rate associated with extremes of GC content.
7. The mouse and human genomes each contain about 30,000 protein-coding genes. (Note that these 2002 estimates have been revised with ongoing annotation and comparative genomics efforts, as summarized in Fig. 19.29.) About 80% of mouse genes have a single identifiable human ortholog. Less than 1% of human genes have no identifiable ortholog in the mouse, and vice versa. The sequencing effort revealed the existence of 9000 previously unknown mouse genes, as well as 1200 new human genes.
8. Dozens of local gene family expansions have occurred in the mouse genome, such as the olfactory receptor gene family. About 20% of this family are pseudogenes in mouse, suggesting a dynamic interplay between gene expansion and gene deletion. The lipocalins also underwent a mouse lineage-specific expansion. For example, the mouse X chromosome contains a cluster of genes related to odorant-binding protein that is absent in humans. Such expansions may account in part for the physiological differences between primates and rodents in terms of reproductive processes.
9. Particular proteins evolve at a rapid rate in mammals. For example, genes involved in the immune response appear to be under positive selection, which drives their evolution.
10. Similar types of repetitive DNA sequences are found in both human and mouse. (We discuss human repetitive sequences in Chapter 20.)
11. The public consortium described 80,000 single-nucleotide polymorphisms (SNPs). We introduced SNPs in Chapter 8 and discuss them further in Chapter 20. GRCm38 currently lists >8 million common SNPs.

A fundamental problem is to understand the genetic variation that underlies the phenotype differences of different mouse strains. Frazer *et al.* (2007) resequenced the genomes of 15 mouse subspecies or strains. These included four wild-derived strains (*M. m. musculus*, *M. m. castaneus*, *M. m. domesticus*, and *M. m. molossinus*). They also sequenced 11 wild-derived strains which were genetically more pure because they have been bred to homozygosity. Frazer *et al.* resequenced almost 1.5 billion bases (58%) of

these genomes and, by comparing them to the reference strain C57BL/6J, they identified 8.3 million SNPs. (The false positive rate of discovery was 2%, the accuracy of genotype calls was >99%, and the false negative rate was assessed as roughly half.) They generated a haplotype map across the mouse genome, defining ancestry breakpoints at which pairwise comparisons indicated a transition to (or from) high SNP densities. The genome-wide SNP map included over 40,000 segments with an average length of 58 kb and a range of 1 kb to 3 Mb. The significance of this project is that it describes the genetic basis of variation in these 15 strains, all of which have unique properties such as behaviors or disease susceptibilities. The C57BL/6J and C57BL/6N mouse strains exhibit significant phenotypic differences. Simon *et al.* (2013) reported 34 SNPs and two indels that could distinguish them.

The most comprehensive mouse resource is the Mouse Genome Informatics (MGI) database and its associated sites (see Chapter 14; Blake *et al.*, 2013).

Rats and mice last shared a common ancestor about 12–24 MYA. The Rat Genome Sequencing Project Consortium (2004) described a high-quality draft genome sequence of the Norway rat, allowing comparisons of the rat, mouse, and human genomes. All have comparable sizes (2.6–2.9 billion bases) and encode similar numbers of genes (see Fig. 19.29). Some properties differ: segmental duplications span over 5% of the human genome (Chapters 8 and 20) but just 3% of the rat genome and 1–2% of the mouse genome. About 40% of the euchromatic rat genome (or ~1 billion bases) aligns to orthologous regions of both mouse and human, containing most exons and known regulatory elements. A portion of this alignable sequence, spanning about 5% of each genome, is under selective constraint (negative selection) while the remainder evolves at the neutral rate. Another 30% of the rat genome aligns only with the mouse but not human, and largely comprises rodent-specific repeats.

The rodent lineage is evolving at a faster rate than the human lineage, as indicated by the longer rodent branch lengths in Figure 19.26. This includes a three-fold higher rate of nucleotide substitution in neutrally evolving DNA, based on analyses of repetitive elements shared since the last common ancestor of humans and rodents.

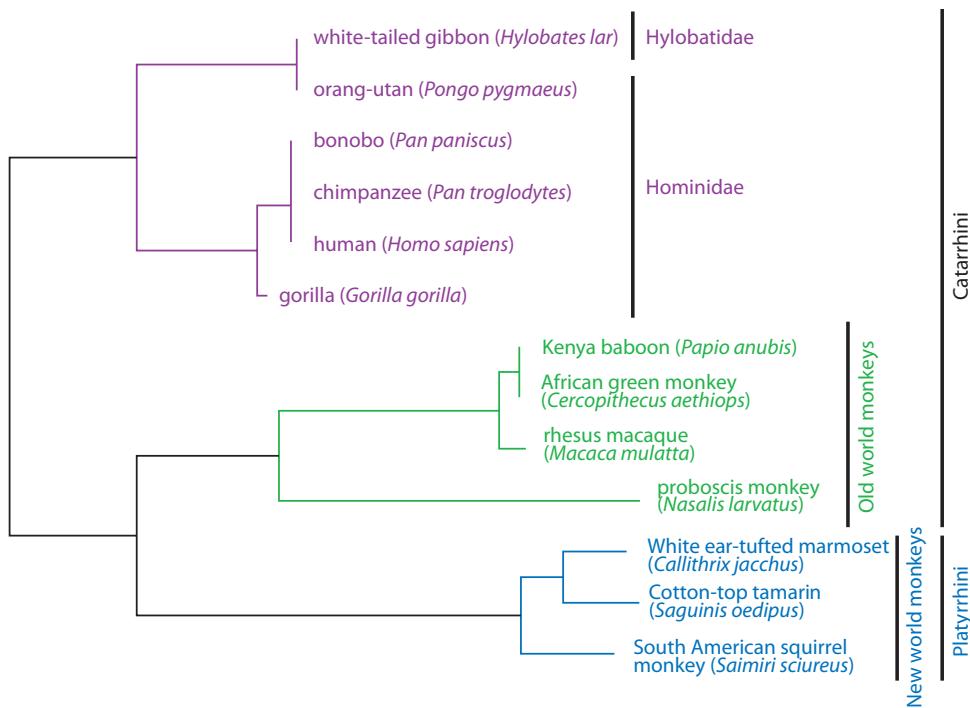
The rodents may be divided into three main groups: a mouse-related clade; the Ctenohystrica including the guinea pig (*Cavia porcellus*); and a squirrel-related clade (Churakov *et al.*, 2010). You can access tracks at the UCSC Genome and Table Browser site for seven rodents (rat, kangaroo rat, naked mole rat, guinea pig, squirrel, rabbit, and pika) in comparison to the mouse genome.

## 5–50 MYA: Primate Genomes

How did humans evolve from other primates? What features of the human genome account for our distinct traits, such as language and higher cognitive skills? A comparison of several primate genomes may elucidate the molecular basis of our unique traits; alternatively, depending on perspective, such a comparison may highlight how closely similar we are to the great apes at a genetic level.

For an overview of primates, we can begin with phylogenetic analyses. The tree of Figure 19.26 provides a glimpse of primates as a sister group to rodent-related species. We can focus on the primates by making a phylogenetic tree with lysozyme protein sequences (Fig. 19.30). The chimpanzee (*Pan troglodytes*) and the bonobo (pygmy chimpanzee, *Pan paniscus*) are the two species most closely related to humans. These three species diverged from a common ancestor  $5.4 \pm 1.1$  MYA, based on analyses of 36 nuclear genes (Stauffer *et al.*, 2001). Our next closest species is the gorilla, which diverged an estimated  $6.4 \pm 1.5$  MYA (or 8.8 MYA, according to timetree.org). Next in the branching order are the orang-utan *Pongo pygmaeus* ( $11.3 \pm 1.3$  MYA) and the gibbon ( $14.9 \pm 2.0$  MYA) (Stauffer *et al.*, 2001). The hominoids diverged from the Old World monkeys

MGI is available at <http://www.informatics.jax.org> (WebLink 19.64) and is operated by The Jackson Laboratory. MGI has multiple components, including the Mouse Genome Database (MGD), the Gene Expression Database (GDX), the Mouse Genome Sequencing (MGS) project, and the Mouse Tumor Biology (MTB) database.



**FIGURE 19.30** Phylogeny of the primates. A neighbor-joining tree representing primate phylogeny based on lysozyme protein sequences. These sequences were aligned using ClustalW and displayed as a neighbor-joining tree. The accession numbers are as follows: gibbon (P79180), orang-utan (P79180), bonobo (AAB41214), chimpanzee (AAB41209), human (P00695), gorilla (P79179), Kenya baboon (P00696), African green monkey (P00696), rhesus macaque (P30201), proboscis monkey (P79811), marmoset (P79158), tamarin (P79268), and South American squirrel monkey (P79294). The sequences are available in Web Document 19.5 at <http://www.bioinfbook.org/chapter19>.

(e.g., the macaque and baboon), having common ancestry as the Catarrhini. This divergence occurred 30–23 MYA, close to the age of the earliest extant hominoid fossils. New World monkeys (such as the tamarin) are even more distantly related.

Note that the method of making a phylogenetic tree using a single protein sequence may be considered simplistic compared to using a supermatrix of many multi-locus sequences or a coalescent approach that accounts for all alleles of a gene in a population to a single ancestral allele (Ting and Sterner, 2013). Nonetheless, the tree in Figure 19.30 is consistent with other reported phylogenies of the primates such as the excellent study by Perelman *et al.* (2011). An overview of the features of primate genomes shows that they have similar sizes, GC content, and some variability in chromosome number (Fig. 19.31).

Following humans, the next two genomes to be sequenced were the chimpanzee and the rhesus macaque (Fig. 19.30). The Chimpanzee Sequencing and Analysis Consortium (2005) described the genome sequence of Clint, a captive-born male. By comparing a human reference to an individual chimpanzee, the analysis focused on those relatively few differences that could be found. (In contrast, comparisons of the human genome to the fish or chicken focused on the relatively few similarities that could be detected, such as ultraconserved regions or coding sequences.) The assembly represents a consensus of two haplotypes from the diploid individual (with one allele from heterozygous sites arbitrarily selected for the assembled sequence); the situation is similar to that of the first sequence of a diploid human individual genome (Chapter 20).

Nucleotide divergence was found to occur at a mean rate of 1.23%, with 35 million SNPs catalogued (including ~1.7 million high-quality SNPs determined by sequencing

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Gorilla; *Gorilla gorilla*

*Callithrix jacchus* (white-tufted-ear marmoset)  
*Macaca fascicularis* (crab-eating macaque)  
*Nomascus leucogenys* (northern white-cheeked gibbon)  
*Pan paniscus* (bonobo; pygmy chimpanzee)

*Pongo pygmaeus* (orang-utan)



	Haploid genome size	GC content	Number of chromos.	Number of genes	NCBI Genome ID
<i>Callithrix jacchus</i> (marmoset)	2915 Mb	41.3%	24	30,292	442
<i>Chlorocebus sabaeus</i> (green monkey)	2790 Mb	40.9%	31	35,027	13136
<i>Gorilla gorilla</i> (western Gorilla)	3036 Mb	41.2%	24	31,334	2156
<i>Homo sapiens</i>	3209 Mb	41.3%	24	41,507	51 (GRCh38)
<i>Macaca fascicularis</i>	2947 Mb	41.3%	21	35,895	776
<i>Macaca mulatta</i> (Rhesus monkey)	3097 Mb	41.5%	21	30,556	215
<i>Nomascus leucogenys</i> (gibbon)	2962 Mb	41.4%	26	28,405	480
<i>Pan paniscus</i> (bonobo)	2869 Mb	41.2%	24	29,392	10729
<i>Pan troglodytes</i> (chimpanzee)	3323 Mb	41.9%	25	31,114	202
<i>Papio anubis</i> (olive baboon)	2948 Mb	41.1%	21	30,956	394
<i>Pongo abelii</i> (Sumatran orang-utan)	3441 Mb	41.6%	24	30,998	325
<i>Tarsius syrichta</i> (Philippine tarsier)	3454 Mb	41%	41	--	766

Selected divergence dates: The rhesus macaque and human lineages diverged ~25 MYA; chimpanzee and human lineages diverged ~6 MYA, also at the time of divergence from the bonobo.  
 Genome features: In aligned regions, DNA shares ~98% identity from chimp to human, and 93.5% identity from macaque to human. High confidence macaque-human orthologs share an average of 97.5% identity.  
 Disease relevance: Macaques are a widely used model for human disease because of their recent divergence (25 MYA rather than 90 MYA for rodents), similar anatomy, physiology, susceptibility to infectious agents related to human pathogens.  
 Web resources: see the Ensembl database at <http://www.ensembl.org>.

**FIGURE 19.31** Overview of primate genomes. The photograph of an orang-utan is from the NHGRI website and Yerkes National Primate Research Center (<http://www.genome.gov/17516871>).

portions of seven additional chimpanzees). Most of these changes reflect random genetic drift rather than being shaped by positive or negative selection pressures. The 1.23% nucleotide divergence rate includes both fixed divergence between humans and chimpanzees (~1.06%) and polymorphic sites within each species. Variation in the nucleotide substitution rates was especially prominent in subtelomeric regions. Of all the observed substitutions, those at CpG dinucleotide sites were most common. Considering the chromosomes separately, the human/chimpanzee divergence is greatest for the Y chromosome (1.9%, perhaps reflecting the greater mutation rate in male) and least for the well-conserved X chromosome (0.94% divergence).

While the number of substitutions is large (35 million), insertion/deletion (indel) events are notable for being fewer (~5 million events) but spanning more of the genomes (there are 40–45 megabases of species-specific euchromatic DNA, totaling ~90 megabases and corresponding to a ~3% difference between the human and chimpanzee genomes).

Humans have a haploid set of 23 chromosomes and, in contrast, chimpanzees have one more, reflecting the fusion of two chromosomes corresponding to chimpanzee 2a and 2b. Additionally there have been nine pericentric inversions (Chapter 8). Many other features have been characterized; among the repetitive elements, SINEs have been three-fold more active in humans, while several new retroviral elements (PtERV1, PtERV2) have invaded the chimpanzee genome selectively. Most of the protein-coding genes are highly conserved, with ~29% being identical. However, 585 out of 13,454 chimpanzee–human

Accessions for glycophorin C are NM\_002101.4, NP\_002092.1 (human) and XM\_001135559.3, XP\_001135559.1 (chimpanzee).

ortholog pairs have a  $K_N/K_S$  ratio greater than 1, suggestive of positive selection. These include glycophorin C, which mediates a *P. falciparum* invasion pathway in human erythrocytes, and granulysin which is involved in defense against pathogens such as *Mycobacterium tuberculosis* (Chapter 17).

A comparison of sequences in humans and chimpanzees does not reveal which genes or other elements evolved rapidly. A phylogenetic reconstruction is necessary in order to infer lineage-specific changes that occurred, leading to the present-day sequences that we can observe. This is one reason that the sequencing of the second nonhuman primate, the rhesus macaque *Macaca mulatta*, was so significant. The rhesus macaque is an Old World monkey (superfamily Cercopithecoidea, family Cercopithecidae) that diverged from the human/chimpanzee lineage ~23–30 million years ago. Its DNA has an average nucleotide identity of ~93% compared to human, in contrast to the ~99% identity between human and chimpanzee. The Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* (2007) sequenced the genome using whole-genome shotgun sequences. They predicted ~20,000 genes, of which high-confidence orthologs share 97.5% identity to human sequences at the DNA and protein levels. Using the macaque as an outgroup, it was possible to analyze many features of the human and chimpanzee genomes. For example, of the 9 pericentric inversions that occurred, 7 could be assigned to the chimpanzee lineage and 2 to the humans (on chromosomes 1 and 18).

The sequencing consortium detailed many features of the rhesus macaque genome, including the facts that 66.7 megabases (2.3%) consist of segmental duplications and there are many lineage-specific expansions and contractions of gene families. Eventually, as for other genome sequencing projects outlined in this chapter, this may permit the analysis of the cellular processes that ultimately underlie the unique biology of this primate.

Additional draft genome sequencing projects were of the orang-utan (Locke *et al.*, 2011), the gorilla (Scally *et al.*, 2012), two more macaques (Yan *et al.*, 2011), and the bonobo (also called pygmy chimpanzee; Prüfer *et al.*, 2012). The ancestor to bonobos and chimpanzees diverged from the human lineage ~6 MYA (and bonobos and chimpanzees diverged ~2 MYA). Prüfer *et al.* showed that >3% of the human genome is more closely related to either of those apes than they are to each other. Prado-Martinez *et al.* (2013) sequenced 79 great ape genomes (from human, *Gorilla*, *Pan*, and *Pongo*), reporting ~89 million SNPs and characterizing inbreeding as well as loss-of-function variants in these populations. All these studies underscore our relatedness to primates and stress the need to understand genetic diversity to support the preservation of endangered species.

## PERSPECTIVE

One of the broadest goals of biology is to understand the nature of each species of life: what are the mechanisms of development, metabolism, homeostasis, reproduction, and behavior? Sequencing of a genome does not answer these questions directly. Instead, we must first try to annotate the genome sequence in order to estimate its contents, and then we try to interpret the function of these parts in a variety of physiological processes.

The genomes of representative species from all major eukaryotic divisions are now available. This will have dramatic implications for all aspects of eukaryotic biology. For pathogenic organisms, it is hoped that the genome sequence will lead to an understanding of their cellular mechanisms of toxicity, their mechanisms of host immune system evasion, and their pharmacological response to drug treatments. For studies of evolution, we will further understand the forces that shape genome evolution: mutation and selection. The reconstruction of ancestral karyotypes is a newly emerging discipline.

As complete genomes are sequenced, we are becoming aware of the nature of noncoding and coding DNA. Major portions of the eukaryotic genomic landscape are

occupied by repetitive DNA, including transposable elements. The number of protein-coding genes varies from about 2000 in some fungi to tens of thousands in plants and mammals. Many of these protein-coding genes are paralogous within each species, such that the “core proteome” size is likely to be on the order of 10,000 genes for many eukaryotes. New proteins are invented in evolution through expansions of gene families or through the use of novel combinations of DNA-encoded protein domains.

Complete genome sequences and assemblies provide insight into the biology of each organism and also phylogenetic relationships between species, population studies, and the history of life on Earth.

## PITFALLS

An urgent need in genomics research is the continued development of algorithms to find protein-coding genes, noncoding RNAs, repetitive sequences, duplicated blocks of sequence within genomes, and conserved syntenic regions shared between genomes. We may then characterize gene function in different developmental stages, body regions, and physiological states. Through these approaches we may generate and test hypotheses about the function, evolution, and biological adaptations of eukaryotes. We may therefore extract meaning from the genomic data.

We are now in the earliest years of the field of genomics. Many new lessons are emerging:

- Draft versions of genome sequences are extremely useful resources, but gene annotation always improves dramatically as a sequence becomes finished.
- It is extraordinarily difficult to predict the presence of protein-coding genes in genomic DNA *ab initio*. It is important to use complementary experimental data on gene expression, such as expressed sequence tag information. Comparative genomics to align orthologous sequences has become the norm.
- We still know relatively little about the nature of noncoding RNA molecules, but comparative genomic studies have demonstrated their conservation across hundreds of millions of years of evolution (e.g., between opossum and human).
- Large portions of eukaryotic genomes consist of repetitive DNA elements.
- Comparative genomics is extraordinarily useful in defining the features of each eukaryotic genome.

Most publications describing genomes (both eukaryotic and bacterial and archaeal) define orthologs as descended by speciation from a single gene in a common ancestor. Typically, the predicted proteins from an organism are searched by BLAST against the complete proteome of other species using an *E* value cutoff such as  $10^{-4}$ . However, two orthologous proteins could have species-specific functions.

## ADVICE FOR STUDENTS

This chapter presents an overview of the dazzlingly varied world of eukaryotic genomes. In providing a broad survey we focused on which genomes have been sequenced and their basic properties such as the number of chromosomes, number of genes, features that make each genome unique, and principles that relate the genome architecture to the phenotype of the species. One useful approach to this chapter is to select a genome that interests you, then apply the five perspectives we offered at the start of the chapter. The NCBI Genome page for each organism provides links to the Sequence Read Archive; use the SRA Toolkit, BEDtools, GenomeWorkbench, MUMmer, RepeatMasker, and other methods we have discussed to further explore the genome.

## WEB RESOURCES

We have presented key resources for many eukaryotic organisms and their genome-sequencing websites. An excellent starting point is the Ensembl website <http://www.ensembl.org/> (WebLink 19.65), which currently includes gateways for the mouse, rat, zebrafish, fugu, mosquito, and other genomes. The Department of Energy Joint Genome Institute (DOE JGI) includes web resources for many of the organisms discussed in this chapter; see [http://genome.jgi-psf.org/euk\\_home.html](http://genome.jgi-psf.org/euk_home.html) (WebLink 19.66).



## Discussion Questions

**[19-1]** If there were no repetitive DNA of any kind, how would the genomes of various eukaryotes (human, mouse, a plant, a parasite) compare in terms of size, gene content, gene order, nucleotide composition, or other features?

**[19-2]** Web Document 19.4 at <http://www.bioinfbook.org/chapter19> consists of a text document with 256,157 bases of DNA from a eukaryotic genome in the FASTA format. How could you identify the species? Assume you cannot use BLAST to directly identify the species. The accession number is given so that you can later look up the species, but assume you cannot use that information at first. What features distinguish the genomic DNA sequence of a protozoan parasite from an insect, or a plant from a human, or one fish from another?

### PROBLEMS/COMPUTER LAB

**[19-1]** *Giardia* has only a few introns. Study the *Giardia* ferredoxin gene at GenBank (DNA accession XM\_001705479.1). To find the intron, try using BLAST to compare the protein (or the DNA encoding the protein) to the genomic DNA. Note that the project accession number for this organism (given in Fig. 19.3; AACB02000000) points to a set of whole-genome shotgun sequence reads (accessions AACB02000001–AACB02000306). To perform the BLAST search, go to BLASTN, use the query XM\_001705479.1, set the database to WGS, and include the Entrez Query AACB02000001:AACB02000306[PACC]. Set the database to reference genomic sequences (restricted to *Giardia lamblia*, taxid:5741).

**[19-2]** Circos software is used to plot genomes in a circular fashion (Krywinski *et al.*, 2009). Visit the Circos website (<http://circos.ca/?home>) (WebLink 19.67). Download and install the software (on a PC, Mac, or Linux), and follow the tutorial accompanying the software to create a genome plot.

**[19-3]** A universal minicircle binding protein (GenBank accession Q23698) has been purified from a trypanosome that infects insects, *Criithidia fasciculata*. A DELTA-BLAST search reveals that there are homologous proteins in plants, fungi, and metazoans (such as the worm *Caenorhabditis elegans*). How is this protein named

in various organisms? What is its presumed function? What is its domain called in the Conserved Domain Database?

**[19-4]** *Leishmania major* has repetitive DNA elements (e.g., accession AF421497). How can you determine how common this element is and where it is localized (e.g., to a particular chromosome or to a chromosomal region)?

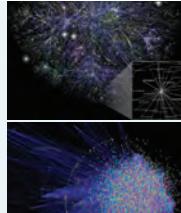
**[19-5]** The soybean pathogen *Phytophthora sojae* and the sudden oak death pathogen *Phytophthora ramorum* are oomycetes, cellular pathogens having different host ranges. Tyler *et al.* (2006) searched for genes encoding secreted proteins. Of the more than 1000 predicted secreted proteins in each organism, many show evidence of rapid diversification in terms of sequence conservation and the evolution of multigene families. These include secreted proteases that could relate to necrotrophic growth, that is, feeding on dead plants after infection of living plant tissue. Of particular note is the Avh (avirulence homolog) family of genes that has 350 members in each genome whose products suppress plant defense responses. Investigate this gene family. How related are its members within and between each species? For an example of an Avh protein from *P. sojae*, see AAR05402.1. Try this in a DELTA-BLAST query. While one iteration of DELTA-BLAST is usually recommended (see Chapter 5), in this case try 5–10 iterations, manually adding all appropriately named sequences that are both above and below threshold.

**[19-6]** The green algae (such as *Chlamydomonas* and *Ostreococcus*) are Viridiplantae that share some genes in common with the animals but not the angiosperms. Use TaxPlot at NCBI (from the home page, select Tools on the left sidebar). Set the query genome to *Ostreococcus lucimarinus*, then set the comparison genomes to *Homo sapiens* and *Arabidopsis thaliana* (as examples of an animal and a plant). Several proteins are dramatically absent from either human or *Arabidopsis*. What are they? What is their function?

**[19-7]** The *C. elegans* and *C. briggsae* genomes share extensive collinearity; study this using <http://www.wormbase.org>. Try to find a 100,000 base pair region including a globin gene on *C. briggsae* chromosome I within position 200,000–300,000.

**[19-8]** This exercise uses phylogenetic shadowing (Fig. 15.8) to evaluate genomic DNA regions under selection. Boffelli (2008) wrote a tutorial on comparing primate genomic DNA sequences using the VISTA server (<http://genome.lbl.gov/vista/index.shtml>, WebLink 19.68). At

the time of that tutorial (2008) there were fewer genomic sequences available. Follow the outline of the tutorial to align alpha globin sequences and use RankVISTA to determine the probability that 10 kilobase segments are evolving neutrally or are under selection.



## Self-Test Quiz

**[19-1]** The *Giardia lamblia* genome is unusual because:

- (a) it contains hardly any transposable elements or introns;
- (b) it is circular;
- (c) it contains extremely little nonrepetitive DNA; or
- (d) its AT content is nearly 80%.

**[19-2]** The genome of the trypanosome *T. brucei*:

- (a) has an intricate network of circular rings of genomic DNA;
- (b) almost completely lacks introns;
- (c) almost completely lacks pseudogenes; or
- (d) varies in size by up to 25% in different isolates.

**[19-3]** The genome of the malaria parasite *Plasmodium falciparum* is notable for having an AT content of 80.6%. Which amino acids are overrepresented in its encoded proteins?

- (a) F, L, I, Y, N, K;
- (b) F, L, I, Y, V, M;
- (c) A, P, C, G, T, R; or
- (d) N, S, Y, I, M, H.

**[19-4]** The *Paramecium tetraurelia* genome has the following properties except for:

- (a) it has about 800 macronuclear chromosomes;
- (b) it has two nuclei, each with distinct functions;
- (c) its genome encodes about twice as many proteins as the human genome; or
- (d) it has undergone whole-genome duplication with massive gene loss.

**[19-5]** Plant genomes from species such as *Arabidopsis* (125 Mb) and the black cottonwood tree *Populus trichocarpa* (480 Mb) were selected because they are relatively small. Nonetheless, each of these genomes is characterized by large amounts of repetitive DNA, and each whole-genome duplicated one or more times:

- (a) true; or
- (b) false.

**[19-6]** Which of these pairs of organisms diverged the longest time ago?

- (a) *Caenorhabditis elegans* and *Caenorhabditis briggsae*;
- (b) *Drosophila melanogaster* (fruit fly) and *Anopheles gambiae* (mosquito);
- (c) *Homo sapiens* and *Canis familiaris* (dog); or
- (d) *Arabidopsis thaliana* and *Oryza sativa* (rice).

**[19-7]** What do the *Takifugu rubripes* (pufferfish) and *Gallus gallus* (chicken) genomes have in common that distinguishes them from the human genome?

- (a) They have genome sizes 3–10-fold smaller than that of human, but a comparable number of genes.
- (b) They have a smaller total genome size but dozens more chromosomes.
- (c) They have smaller genome sizes and approximately half as many protein-coding genes.
- (d) They have a series of minichromosomes of variable size.

**[19-8]** How are the mouse and human genomes different?

- (a) The mouse genome has a lower GC content.
- (b) The mouse genome has more protein-coding genes.
- (c) The mouse genome has undergone specific expansions of genes encoding particular protein families such as olfactory receptors.
- (d) The mouse genome has fewer telomeric repeats per chromosome, on average.

**[19-9]** Many features distinguish the chimpanzee and human genomes, including all of the following except for:

- (a) chimpanzees have more chromosomes;
- (b) about 35 million nucleotide substitutions have been described;
- (c) there have been hundreds of pericentric inversions; or
- (d) over 500 chimpanzee–human ortholog pairs may be under positive selection.

## SUGGESTED READING

I recommend recent books by Eugene Koonin (*The Logic of Chance: The Nature and Origin of Biological Evolution*, 2012) and Michael Lynch (*The Origins of Genome Architecture*, 2007).

We presented a phylogenetic tree from Baldauf *et al.* (2000). For an evolutionary analysis of eukaryotic evolution, including a discussion of models of eukaryotic origins and the role of mitochondria, see Embley and Martin (2006). For a brief review of the significance of Apicomplexan genome projects, see Winzeler (2008). Paterson (2006) provides an excellent overview of plant genomics. Church *et al.* (2009) present the finished genome assembly of the mouse, showing the importance of continued genome assembly and annotation efforts.

## REFERENCES

- Abrahamsen, M.S., Templeton, T.J., Enomoto, S. *et al.* 2004. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* **304**, 441–445. PMID: 15044751.
- Adam, R.D. 2001. Biology of *Giardia lamblia*. *Clinical Microbiology Reviews* **14**, 447–475.
- Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195. PMID: 10731132.
- Akman, L., Yamashita, A., Watanabe, H. *et al.* 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genetics* **32**, 402–407. PMID: 12219091.
- Allen, K.D. 2002. Assaying gene content in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **99**, 9568–9572.
- Amemiya, C.T., Alföldi, J., Lee, A.P. *et al.* 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**(7445), 311–316. PMID: 23598338.
- Aparicio, S., Chapman, J., Stupka, E. *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310. PMID: 12142439.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.
- Archibald, J.M., Lane, C.E. 2009. Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction. *Journal of Heredity* **100**(5), 582–590. PMID: 19617523.
- Arensburger, P., Megy, K., Waterhouse, R.M. *et al.* 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* **330**(6000), 86–88. PMID: 20929810.
- Arkhipova, I., Meselson, M. 2000. Transposable elements in sexual and ancient asexual taxa. *Proceedings of the National Academy of Sciences, USA* **97**, 14473–14477.
- Arkhipova, I.R., Morrison, H.G. 2001. Three retrotransposon families in the genome of *Giardia lamblia*: Two telomeric, one dead. *Proceedings of the National Academy of Sciences, USA* **98**, 14497–14502.
- Armbrust, E.V., Berge, J.A., Bowler, C. *et al.* 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86. PMID: 15459382.
- Aurrecoechea, C., Brestelli, J., Brunk, B.P. *et al.* 2009a. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Research* **37**(Database issue), D526–530. PMID: 18824479.
- Aurrecoechea, C., Brestelli, J., Brunk, B.P. *et al.* 2009b. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Research* **37**(Database issue), D539–543. PMID: 18957442.
- Aurrecoechea, C., Barreto, A., Brestelli, J. *et al.* 2011. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Research* **39**(Database issue), D612–619. PMID: 20974635.
- Aury, J.M., Jaillon, O., Duret, L. *et al.* 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178. PMID: 17086204.

- Baldauf, S.L., Roger, A. J., Wenk-Siefert, I., Doolittle, W. F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972–977.
- Baum, D.A., Smith, S.D., Donovan, S.S. 2005. Evolution. The tree-thinking challenge. *Science* **310**, 979–980.
- Bennett, M.D., Leitch, I.J. 2011. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany* **107**(3), 467–590. PMID: 21257716.
- Bennetzen, J.L., Freeling, M. 1997. The unified grass genome: Synergy in synteny. *Genome Research* **7**, 301–306.
- Berriman, M., Ghedin, E., Hertz-Fowler, C. et al. 2005. The genome of the African trypanosome *Trypanosoma brucei*. *Science* **309**, 416–422. PMID: 16020726.
- Birol, I., Raymond, A., Jackman, S.D. et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**(12), 1492–1497. PMID: 23698863.
- Blake, J.A., Bult, C.J., Eppig, J.T. et al. 2014. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Research* **42**, D810–817. PMID: 24285300.
- Blaxter, M. 1998. *Caenorhabditis elegans* is a nematode. *Science* **282**, 2041–2046.
- Blaxter, M. 2003. Molecular systematics: Counting angels with DNA. *Nature* **421**, 122–124.
- Boffelli, D. 2008. Phylogenetic shadowing: sequence comparisons of multiple primate species. *Methods in Molecular Biology* **453**, 217–231. PMID: 18712305.
- Bonasio, R., Zhang, G., Ye, C. et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**(5995), 1068–1071. PMID: 20798317.
- Bovine Genome Sequencing and Analysis Consortium, Elsik, C.G., Tellam, R.L. et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**(5926), 522–528. PMID: 19390049.
- Bowler, C., Allen, A.E., Badger, J.H. et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**(7219), 239–244. PMID: 18923393.
- Bowler, C., Vardi, A., Allen, A.E. 2010. Oceanographic and biogeochemical insights from diatom genomes. *Annual Review of Marine Science* **2**, 333–365. PMID: 21141668.
- Bradnam, K.R., Fass, J.N., Alexandrov, A. et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**(1), 10. PMID: 23870653.
- Brayton, K.A., Lau, A.O., Herndon, D.R. et al. 2007. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathogens* **3**, 1401–1413. PMID: 17953480.
- Brenchley, R., Spannagl, M., Pfeifer, M. et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**(7426), 705–710. PMID: 23192148.
- Budiansky, S. 2002. Creatures of our own making. *Science* **298**, 80–86.
- C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**, 2012–2018.
- Cameron, R.A., Samanta, M., Yuan, A., He, D., Davidson, E. 2009. SpBase: the sea urchin genome database and web site. *Nucleic Acids Research* **37**(Database issue), D750–754. PMID: 19010966.
- Cao, J., Schneeberger, K., Ossowski, S. et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics* **43**(10), 956–963. PMID: 21874002.
- Carlton, J.M., Hirt, R.P., Silva, J.C. et al. 2001. Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Molecular and Biochemical Parasitology* **118**, 201–210. PMID: 17218520.
- Carlton, J.M., Muller, R., Yowell, C.A. et al. 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512–519. PMID: 11738710.
- Carlton, J.M., Hirt, R.P., Silva, J.C. et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* **315**, 207–212. PMID: 17218520.
- Carlton, J.M., Adams, J.H., Silva, J.C. et al. 2008a. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* **455**(7214), 757–763. PMID: 18843361.

- Carlton, J.M., Escalante, A.A., Neafsey, D., Volkman, S.K. 2008b. Comparative evolutionary genomics of human malaria parasites. *Trends in Parasitology* **24**(12), 545–550. PMID: 18938107.
- Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biological Reviews of the Cambridge Philosophical Society* **73**, 203–266.
- Cheeseman, I.H., Miller, B.A., Nair, S. *et al.* 2012. A major genome region underlying artemisinin resistance in malaria. *Science* **336**(6077), 79–82. PMID: 22491853.
- Chen, S., Zhang, G., Shao, C. *et al.* 2014. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* **46**, 253–260. PMID: 24487278.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87.
- Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., Venkatesh, B. 2004. *Fugu* genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Molecular Biology and Evolution* **21**, 1146–11451.
- Churakov, G., Sadasivuni, M.K., Rosenbloom, K.R. *et al.* 2010. Rodent evolution: back to the root. *Molecular Biology and Evolution* **27**(6), 1315–1326. PMID: 20100942.
- Church, D.M., Goodstadt, L., Hillier, L.W. *et al.* 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology* **7**(5), e1000112. PMID: 19468303.
- Cock, J.M., Sterck, L., Rouzé, P. *et al.* 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**(7298), 617–621. PMID: 20520714.
- Colbourne, J.K., Pfrender, M.E., Gilbert, D. *et al.* 2011. The ecoresponsive genome of *Daphnia pulex*. *Science* **331**(6017), 555–561. PMID: 21292972.
- Collins, K., Gorovsky, M.A. 2005. *Tetrahymena thermophila*. *Current Biology* **15**, R317–318.
- Conrad, M.D., Bradic, M., Warring, S.D., Gorman, A.W., Carlton, J.M. 2013. Getting trichy: tools and approaches to interrogating *Trichomonas vaginalis* in a post-genome world. *Trends in Parasitology* **29**(1), 17–25. PMID: 23219217.
- Cowman, A.F., Crabb, B. S. 2002. The *Plasmodium falciparum* genome: a blueprint for erythrocyte invasion. *Science* **298**, 126–128.
- Cox, F. E. 2002. History of human parasitology. *Clinical Microbiology Reviews* **15**, 595–612.
- Coyne, R.S., Lhuillier-Akakpo, M., Duhartcourt, S. 2012. RNA-guided DNA rearrangements in ciliates: is the best genome defence a good offence? *Biology of the Cell* **104**(6), 309–325. PMID: 22352444.
- Crowe, M.L., Serizet, C., Thareau, V. *et al.* 2003. CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Research* **31**, 156–158.
- Curtis, B.A., Tanifuji, G., Burki, F. *et al.* 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**(7427), 59–65. PMID: 23201678.
- Dalloul, R.A., Long, J.A., Zimin, A.V. *et al.* 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biology* **8**(9), pii: e1000475. PMID: 20838655.
- Dávila, A.M., Mended, P. N., Wagner, G. *et al.* 2008. ProtozoaDB: dynamic visualization and exploration of protozoan genomes. *Nucleic Acids Research* **36**(Database issue), D547–552. PMID: 17981844.
- Dehal, P., Satou, Y., Campbell, R.K. *et al.* 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298**, 2157–2167. PMID: 12481130.
- Denoeud, F., Henriet, S., Mungpakdee, S. *et al.* 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**(6009), 1381–1385. PMID: 21097902.
- Derelle, E., Ferraz, C., Rombauts, S. *et al.* 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences, USA* **103**, 11647–11652. PMID: 16868079.
- Dermitzakis, E.T., Reymond, A., Lyle, R. *et al.* 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582. PMID: 12466853.

- Doak, T.G., Cavalcanti, A.R., Stover, N.A., Dunn, D.M., Weiss, R., Herrick, G., Landweber, L.F. 2003. Sequencing the *Oxytricha trifallax* macronuclear genome: a pilot project. *Trends in Genetics* **19**, 603–607.
- Donelson, J.E. 1996. Genome research and evolution in trypanosomes. *Current Opinion in Genetics and Development* **6**, 699–703.
- Dong, Y., Xie, M., Jiang, Y. et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nature Biotechnology* **31**(2), 135–141. PMID: 23263233.
- Douglas, S.E., Penny, S. L. 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: Complete sequence and conserved synteny groups confirm its common ancestry with red algae. *Journal of Molecular Evolution* **48**, 236–244.
- Douglas, S., Zauner, S., Fraunholz, M. et al. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* **410**, 1091–1096. PMID: 11323671.
- Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218.
- Eichinger, L., Pachebat, J.A., Glöckner, G. et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43–57. PMID: 15875012.
- Eisen, J.A., Coyne, R.S., Wu, M. et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biology* **4**, e286. PMID: 16933976.
- Ekdahl, S., Sonnhammer, E.L. 2004. ChromoWheel: a new spin on eukaryotic chromosome visualization. *Bioinformatics* **20**, 576–577.
- El-Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E., Donelson, J. E. 2000. The African trypanosome genome. *International Journal of Parasitology* **30**, 329–345.
- El-Sayed, N.M., Myler, P.J., Blandin, G. et al. 2005a. Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404–409. PMID: 16020724.
- El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C. et al. 2005b. The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* **309**, 409–415. PMID: 16020725.
- Elsik, C.G., Worley, K.C., Bennett, A.K. et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15**(1), 86. PMID: 24479613.
- Embley, T.M., Hirt, R. P. 1998. Early branching eukaryotes? *Current Opinion in Genetics and Development* **8**, 624–629.
- Embley, T.M., Martin, W. 2006. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630.
- Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., Trask, B.J. 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Research* **12**, 1651–1662.
- Florens, L., Washburn, M.P., Raine, J.D. et al. 2002. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526. PMID: 12368866.
- Florent, I., Maréchal, E., Gascuel, O., Bréhélin, L. 2010. Bioinformatic strategies to provide functional clues to the unknown genes in *Plasmodium falciparum* genome. *Parasite* **17**(4), 273–283. PMID: 21275233.
- Fraser, C.M., Eisen, J. A., Salzberg, S. L. 2000. Microbial genome sequencing. *Nature* **406**, 799–803.
- Frazer, K.A. et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053.
- Gajria, B., Bahl, A., Brestelli, J. et al. 2008. ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Research* **36**(Database issue), D553–556.
- Gamo, F.J., Sanz, L.M., Vidal, J. et al. 2010. Thousands of chemical starting points for antimalarial lead identification. *Nature* **465**(7296), 305–310. PMID: 20485427.
- Gan, X., Stegle, O., Behr, J. et al. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**(7365), 419–423. PMID: 21874022.
- Gardner, M.J., Hall, N., Fung, E. et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511.

- Gardner, M.J., Bishop, R., Shah, T. *et al.* 2005. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* **309**, 134–137. PMID: 15994558.
- Ghedin, E., Wang, S., Spiro, D. *et al.* 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760. PMID: 17885136.
- Gilson, P. R., McFadden, G. I. 2001. A grin without a cat. *Nature* **410**, 1040–1041.
- Gilson, P. R., McFadden, G. I. 2002. Jam packed genomes: a preliminary, comparative analysis of nucleomorphs. *Genetica* **115**, 13–28.
- Gilson, P.R., Su, V., Slamovits, C.H. *et al.* 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proceedings of the National Academy of Sciences, USA* **103**, 9566–9571.
- Glockner, G., Eichinger, L., Szafranski, K. *et al.* 2002. Sequence and analysis of chromosome 2 of *Dicystostelium discoideum*. *Nature* **418**, 79–85. PMID: 12097910.
- Goff, S.A., Ricke, D., Lan, T.H. *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**, 92–100. PMID: 11935018.
- Greilhuber, J., Borsch, T., Müller, K. *et al.* 2006. Smallest angiosperm genomes found in lentinulariaceae, with chromosomes of bacterial size. *Plant Biology (Stuttgart)* **8**(6), 770–777. PMID: 17203433.
- Gupta, B.P., Sternberg, P.W. 2003. The draft genome sequence of the nematode *Caenorhabditis briggsae*, a companion to *C. elegans*. *Genome Biology* **4**, 238.
- Hall, A. E., Fiebig, A., Preuss, D. 2002. Beyond the *Arabidopsis* genome: Opportunities for comparative genomics. *Plant Physiology* **129**, 1439–1447.
- Hall, N., Karras, M., Raine, J.D. *et al.* 2005. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82–86. PMID: 15637271.
- Hardison, R.C., Oeltjen, J., Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Research* **7**, 959–966.
- Harris, T.W., Baran, J., Bieri, T. *et al.* 2014. WormBase 2014: new views of curated biology. *Nucleic Acids Research* **42**: D789–793. PMID: 24194605.
- Hedges, S.B., Dudley, J., Kumar, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23), 2971–2972. PMID: 17021158.
- Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**(7405), 94–98. PMID: 22722851.
- Hellsten, U., Harland, R.M., Gilchrist, M.J. *et al.* 2010. The genome of the Western clawed frog *Xenopus tropicalis*. *Science* **328**(5978), 633–636. PMID: 20431018.
- Hirt, R.P., de Miguel, N., Nakjang, S. *et al.* 2011. *Trichomonas vaginalis* pathobiology new insights from the genome sequence. *Advances in Parasitology* **77**, 87–140. PMID: 22137583.
- Hoffman, S.L., Subramanian, G. M., Collins, F. H., Venter, J. C. 2002. *Plasmodium*, human and *Anopheles* genomics and malaria. *Nature* **415**, 702–709.
- Holland, P.W. 2002. *Ciona*. *Current Biology* **12**, R609.
- Holt, R.A., Subramanian, G.M., Halpern, A. *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149. PMID: 12364791.
- Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949.
- Howe, K., Clark, M.D., Torroja, C.F. *et al.* 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**(7446), 498–503. PMID: 23594743.
- Huang, X., Kurata, N., Wei, X. *et al.* 2012. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**(7421), 497–501. PMID: 23034647.
- Huang, Y., Li, Y., Burt, D.W. *et al.* 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genetics* **45**(7), 776–783. PMID: 23749191.
- i5K Consortium. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity* **104**(5), 595–600. PMID: 23940263.

- Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G. *et al.* 2013. Architecture and evolution of a minute plant genome. *Nature* **498**(7452), 94–98. PMID: 23665961.
- International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716.
- International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436**, 793–800.
- Ivens, A.C., Peacock, C.S., Worthey, E.A. *et al.* 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309**, 436–442. PMID: 16020728.
- Jacquemin, J., Bhatia, D., Singh, K., Wing, R.A. 2013. The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Current Opinion in Plant Biology* **16**(2), 147–156. PMID: 23518283.
- Jaillon, O., Aury, J.M., Brunet, F. *et al.* 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946–957. PMID: 15496914.
- Jaillon, O. *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467.
- Jeffares, D.C., Pain, A., Berry, A. *et al.* 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nature Genetics* **39**, 120–125. PMID: 17159978.
- Jiang, R.H., de Brujin, I., Haas, B.J. *et al.* 2013. Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen *Saprolegnia parasitica*. *PLoS Genetics* **9**(6), e1003272. PMID: 23785293.
- Johnson, P.J. 2002. Spliceosomal introns in a deep-branching eukaryote: The splice of life. *Proceedings of the National Academy of Sciences USA* **99**, 3359–3361.
- Jomaa, H., Wiesner, J., Sanderbrand, S. *et al.* 1999. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285**, 1573–1576. PMID: 10477522.
- Joshi, H.J., Christiansen, K.M., Fitz, J. *et al.* 2012. 1001 Proteomes: a functional proteomics portal for the analysis of *Arabidopsis thaliana* accessions. *Bioinformatics* **28**(10), 1303–1306. PMID: 22451271.
- Kasahara, M., Naruse, K., Sasaki, S. *et al.* 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719. PMID: 17554307.
- Kaye, P., Scott, P. 2011. Leishmaniasis: complexity at the host-pathogen interface. *Nature Reviews Microbiology* **9**(8), 604–615. PMID: 21747391.
- Keeling, C.I., Yuen, M.M., Liao, N.Y. *et al.* 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biology* **14**(3), R27. PMID: 23537049.
- Keeling, P.J. 2007. Genomics. Deep questions in the tree of life. *Science* **317**, 1875–1876.
- Kiene, R.P. 2008. Marine biology: Genes in the glass house. *Nature* **456**(7219), 179–181. PMID: 19005540.
- Kirkness, E.F., Bafna, V., Halpern, A.L. *et al.* 2003. The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903.
- Koonin, E.V. 2012. *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT Press, New Jersey.
- Krzywinski, M., Schein, J., Birol, I. *et al.* 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**(9), 1639–1645. PMID: 19541911.
- Ku, H. M., Vision, T., Liu, J., Tanksley, S. D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proceedings of the National Academy of Sciences USA* **97**, 9121–9126.
- Lamblin, A. F., Crow, J.A., Johnson, J.E. *et al.* 2003. MtDB: A database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Research* **31**, 196–201. PMID: 12519981.
- Lamesch, P., Berardini, T.Z., Li, D. *et al.* 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**(Database issue), D1202–D1210. PMID: 22140109.

- Lartillot, N., Philippe, H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philosophical Transactions of the Royal Society of London, B: Biological Sciences*, doi: 10.1098/rstb.2007.2236.
- Laszsonczi, E., Ishihama, Y., Andersen, J.S. et al. 2002. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542. PMID: 12368870.
- Lawniczak, M.K., Emrich, S.J., Holloway, A.K. et al. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* **330**(6003), 512–514. PMID: 20966253.
- Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S. et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819. PMID: 16341006.
- Lloyd, D., Harris, J. C. 2002. *Giardia*: Highly evolved parasite or early branching eukaryote? *Trends in Microbiology* **10**, 122–127.
- Locke, D.P., Hillier, L.W., Warren, W.C. et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* **469**(7331), 529–533. PMID: 21270892.
- Logan-Klumpler, F.J., De Silva, N., Boehme, U. et al. 2012. GeneDB: an annotation database for pathogens. *Nucleic Acids Research* **40**(Database issue), D98–108. PMID: 22116062.
- Long, C. A., Hoffman, S. L. 2002. Malaria: from infants to genomics to vaccines. *Science* **297**, 345–347.
- Lynch, M. 2007. *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.
- Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proceedings of the National Academy of Sciences, USA* **102**, 4795–4800.
- Margulies, L., Schwartz, K. V. 1998. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth*. W. H. Freeman, New York.
- Marinotti, O., Cerqueira, G.C., de Almeida, L.G. et al. 2013. The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Research* **41**(15), 7387–7400. PMID: 23761445.
- Matthews, D. E., Carollo, V. L., Lazo, G. R., Anderson, O. D. 2003. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Research* **31**, 183–186.
- Merchant, S.S., Prochnik, S.E., Vallon, O. et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250. PMID: 17932292.
- Meyerowitz, E. M. 2002. Plants compared to animals: The broadest comparative study of development. *Science* **295**, 1482–1485.
- Mikkelsen, T.S., Wakefield, M.J., Aken, B. et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177. PMID: 17495919.
- Moore, C.E., Archibald, J.M. 2009. Nucleomorph genomes. *Annual Review of Genetics* **43**, 251–264. PMID: 19686079.
- Moore, C.E., Curtis, B., Mills, T., Tanifugi, G., Archibald, J.M. 2012. Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. *Genome Biology and Evolution* **4**(11), 1162–1175. PMID: 23042551.
- Morris, J.C., Drew, M.E., Klingbeil, M.M. et al. 2001. Replication of kinetoplast DNA: An update for the new millennium. *International Journal of Parasitology* **31**, 453–458. PMID: 11334929.
- Morrison, H.G., McArthur, A.G., Gillin, F.D. et al. 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**, 1921–1926. PMID: 17901334.
- Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562. PMID: 12466850.
- Murray, C.J., Rosenfeld, L.C., Lim, S.S. et al. 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* **379**(9814), 413–431 (2012). PMID: 22305225
- Myler, P.J., Audleman, L., deVos, T. et al. 1999. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proceedings of the National Academy of Sciences, USA* **96**, 2902–2906. PMID: 10077609.

- Myler, P.J., Sisk, E., McDonagh, P.D. *et al.* 2000. Genomic organization and gene function in *Leishmania*. *Biochemistry Society Transactions* **28**, 527–531. PMID: 11044368.
- Nakamura, Y., Mori, K., Saitoh, K. *et al.* 2013. Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proceedings of the National Academy of Sciences, USA* **110**(27), 11061–11066. PMID: 23781100.
- Neafsey, D.E., Galinsky, K., Jiang, R.H. *et al.* 2012. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nature Genetics* **44**(9), 1046–1050. PMID: 22863733.
- Neale, D.B., Wegrzyn, J.L., Stevens, K.A. *et al.* 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**(3), R59. PMID: 24647006.
- Niklas, K.J., Newman, S.A. 2013. The origins of multicellular organisms. *Evolution and Development* **15**(1), 41–52. PMID: 23331916.
- Nirujogi, R.S., Pawar, H., Renuse, S. *et al.* 2014. Moving from unsequenced to sequenced genome: Reanalysis of the proteome of *Leishmania donovani*. *Journal of Proteomics* **97**, 48–61. PMID: 23665000.
- Nixon, J.E., Wang, A., Morrison, H.G. *et al.* 2002. A spliceosomal intron in *Giardia lamblia*. *Proceedings of the National Academy of Sciences, USA* **99**, 3701–3705. PMID: 11854456.
- Nygaard, S., Zhang, G., Schiøtt, M. *et al.* 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Research* **21**(8), 1339–1348. PMID: 21719571.
- Ohno, S. 1970. *Evolution by Gene Duplication*. SpringerVerlag, Berlin.
- Pain, A., Renauld, H., Berriman, M. *et al.* 2005. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* **309**, 131–133. PMID: 15994557.
- Pain, A., Böhme, U., Berry, A.E. *et al.* 2008. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455**(7214), 799–803. PMID: 18843368.
- Pardue, M.L., DeBaryshe, P. G., Lowenhaupt, K. 2001. Another protozoan contributes to understanding telomeres and transposable elements. *Proceedings of the National Academy of Sciences, USA* **98**, 14195–14197.
- Paterson, A.H. 2006. Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nature Reviews Genetics* **7**, 174–184.
- Paterson, A.H., Freeling, M., Tang, H., Wang, X. 2010. Insights from the comparison of plant genome sequences. *Annual Review of Plant Biology* **61**, 349–372. PMID: 20441528.
- Patron, N.J., Rogers, M.B., Keeling, P.J. 2006. Comparative rates of evolution in endosymbiotic nuclear genomes. *BMC Evolutionary Biology* **6**, 46.
- Peacock, C.S., Seeger, K., Harris, D. *et al.* 2007. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nature Genetics* **39**, 839–847. PMID: 17572675.
- Pellicer, J., Fay, M.F., Leitch, I.J. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164**, 10–15.
- Perelman, P., Johnson, W.E., Roos, C. *et al.* 2011. A molecular phylogeny of living primates. *PLoS Genetics* **7**(3), e1001342. PMID: 21436896.
- Philippe, H., Laurent, J. 1998. How good are deep phylogenetic trees? *Current Opinion in Genetics and Development* **8**, 616–623.
- Postlethwait, J.H. 2007. The zebrafish genome in context: ohnologs gone missing. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **308**, 563–577.
- Potato Genome Sequencing Consortium, Xu, X., Pan, S. *et al.* 2011. Genome sequence and analysis of the tuber crop potato. *Nature* **475**(7355), 189–195. PMID: 21743474.
- Prado-Martinez, J., Sudmant, P.H., Kidd, J.M. *et al.* 2013. Great ape genetic diversity and population history. *Nature* **499**(7459), 471–475. PMID: 23823723.
- Prochnik, S.E., Umen, J., Nedelcu, A.M. *et al.* 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**(5988), 223–226. PMID: 20616280.

- Prüfer, K., Munch, K., Hellmann, I. *et al.* 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**(7404), 527–531. PMID: 22722832.
- Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521. PMID: 15057822.
- Rensing, S.A., Lang, D., Zimmer, A.D. *et al.* 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**, 64–69. PMID: 18079367.
- Reyes, A., Pesole, G., Saccone, C. 2000. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* **259**, 177–187.
- Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234.
- Richards, S., Liu, Y., Bettencourt, B.R. *et al.* 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Research* **15**, 1–18. PMID: 15632085.
- Roos, D.S. 2005. Themes and variations in apicomplexan parasite biology. *Science* **309**, 72–73.
- Rubin, G.M., Lewis, E. B. 2000. A brief history of *Drosophila*'s contributions to genome research. *Science* **287**, 2216–2218.
- Rudd, S., Mewes, H.W., Mayer, K. F. 2003. Sputnik: A database platform for comparative plant genomics. *Nucleic Acids Research* **31**, 128–132.
- Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., Burleigh, J.G. 2014. From algae to angiosperms—inferring the phylogeny of green plants (*Viridiplantae*) from 360 plastid genomes. *BMC Evolutionary Biology* **14**(1), 23. PMID: 24533922.
- Ryan, J.F., Finnerty, J. R. 2003. CnidBase: The Cnidarian Evolutionary Genomics Database. *Nucleic Acids Research* **31**, 159–163.
- Sakai, H., Lee, S.S., Tanaka, T. *et al.* 2013. Rice Annotation Project Database (RAP–DB): an integrative and interactive database for rice genomics. *Plant Cell Physiology* **54**(2), e6. PMID: 23299411.
- Scally, A., Dutheil, J.Y., Hillier, L.W. *et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**(7388), 169–175. PMID: 22398555.
- Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**(7278), 178–183. PMID: 20075913.
- Schnable, P.S., Ware, D., Fulton, R.S. *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956), 1112–1115. PMID: 19965430.
- Sea Urchin Genome Sequencing Consortium *et al.* 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941–952.
- Seo, H.C., Kube, M., Edvardsen, R.B. *et al.* 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506. PMID: 11752568.
- Severson, D.W., Behura, S.K. 2012. Mosquito genomics: progress and challenges. *Annual Review of Entomology* **57**, 143–166. PMID: 21942845.
- Shaffer, H.B., Minx, P., Warren, D.E. *et al.* 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology* **14**(3), R28. PMID: 23537068.
- Shapiro, T. A., Englund, P. T. 1995. The structure and replication of kinetoplast DNA. *Annual Review of Microbiology* **49**, 117–143.
- Shulaev, V., Sargent, D.J., Crowhurst, R.N. *et al.* 2011. The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics* **43**(2), 109–116. PMID: 21186353.
- Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA* **99**, 13627–13632.
- Simon, M.M., Greenaway, S., White, J.K. *et al.* 2013. A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biology* **14**(7), R82. PMID: 23902802.
- Simpson, A. G., MacQuarrie, E. K., Roger, A. J. 2002. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419**, 270.

- Slavov, G.T., DiFazio, S.P., Martin, J. *et al.* 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytology* **196**(3), 713–725. PMID: 22861491.
- Smith, C.R., Smith, C.D., Robertson, H.M. *et al.* 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proceedings of the National Academy of Sciences, USA* **108**(14), 5667–5672. PMID: 21282651.
- Sodergren, E., Shen, Y., Song, X. *et al.* 2006. Shedding genomic light on Aristotle's lantern. *Developmental Biology* **300**, 2–8.
- Sonnhammer, E.L., Durbin, R. 1997. Analysis of protein domain families in *Caenorhabditis elegans*. *Genomics* **46**, 200–216.
- Staats, M., Erkens, R.H., van de Vossenberg, B. *et al.* 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One* **8**(7), e69189. PMID: 23922691.
- Stark, A., Lin, M.F., Kheradpour, P. *et al.* 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232. PMID: 17994088.
- Stauffer, R.L., Walker, A., Ryder, O. A., Lyons-Weiler, M., Hedges, S. B. 2001. Human and ape molecular clocks and constraints on paleontological hypotheses. *Journal of Heredity* **92**, 469–474.
- Stein, L.D., Bao, Z., Blasiar, D. *et al.* 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biology* **1**, E45. PMID: 14624247.
- Stothard, P., Wishart, D.S. 2005. Circular genome visualization and exploration using CGView. *Bioinformatics* **21**, 537–539.
- Swart, E.C., Bracht, J.R., Magrini, V. *et al.* 2013. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biology* **11**(1), e1001473. PMID: 23382650.
- Takeda, H., Shimada, A. 2010. The art of medaka genetics and genomics: what makes them so unique? *Annual Review of Genetics* **44**, 217–241. PMID: 20731603.
- Tanifuji, G., Onodera, N.T., Wheeler, T.J. *et al.* 2011. Complete nucleomorph genome sequence of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. *Genome Biology and Evolution* **3**, 44–54. PMID: 21147880.
- Tanifuji, G., Onodera, N.T., Moore, C.E., Archibald, J.M. 2014. Reduced nuclear genomes maintain high gene transcription levels. *Molecular Biology and Evolution* **31**, 625–635. PMID: 24336878.
- Taylor, J.E., Rudenko, G. 2006. Switching trypanosome coats: what's in the wardrobe? *Trends in Genetics* **22**, 614–620.
- Tekle, Y.I., Parfrey, L.W., Katz, L.A. 2009. Molecular data are transforming hypotheses on the origin and diversification of eukaryotes. *Bioscience* **59**(6), 471–481. PMID: 20842214.
- Ting, N., Sterner, K.N. 2013. Primate molecular phylogenetics in a genomic era. *Molecular Phylogenetics and Evolution* **66**(2), 565–568. PMID: 22960143.
- Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**(7400), 635–641. PMID: 22660326.
- Topalis, P., Koutsos, A., Dialynas, E. *et al.* 2005. AnoBase: a genetic and biological database of anophelines. *Insect Molecular Biology* **14**, 591–597.
- Tribolium* Genome Sequencing Consortium, Richards, S., Gibbs, R.A. *et al.* 2008. The genome of the model beetle and pest *Tribolium castaneum*. *Nature* **452**(7190), 949–955. PMID: 18362917.
- Tuskan G.A., Difazio, S., Jansson, S. *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604. PMID: 16973872.
- Tyler, B.M., Tripathy, S., Zhang, X. *et al.* 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261–1266. PMID: 16946064.
- Tzafrir, I., Dickerman, A., Brazhnik, O. *et al.* 2003. The *Arabidopsis* SeedGenes Project. *Nucleic Acids Research* **31**, 90–93. PMID: 12519955.
- Upcroft, J.A., Krauer, K.G., Upcroft, P. 2010. Chromosome sequence maps of the *Giardia lamblia* assemblage A isolate WB. *Trends in Parasitology* **26**(10), 484–491. PMID: 20739222.

- Van de Peer, Y. 2004. *Tetraodon* genome confirms *Takifugu* findings: most fish are ancient polyploids. *Genome Biology* **5**, 250.
- Van de Peer, Y., Baldauf, S. L., Doolittle, W. F., Meyer, A. 2000. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *Journal of Molecular Evolution* **51**, 565–576.
- Venkatesh, B., Kirkness, E.F., Loh, Y.H. *et al.* 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinus mili*) genome. *PLoS Biology* **5**, e101. PMID: 17407382.
- Vinson, J.P., Jaffe, D.B., O'Neill, K. *et al.* 2005. Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Research* **15**, 1127–1135. PMID: 16077012.
- Walbot, V. 2000. *Arabidopsis thaliana* genome. A green chapter in the book of life. *Nature* **408**, 794–795.
- Walzer, K.A., Adomako-Ankomah, Y., Dam, R.A. *et al.* 2013. *Hammonia hammondi*, an avirulent relative of *Toxoplasma gondii*, has functional orthologs of known *T. gondii* virulence genes. *Proceedings of the National Academy of Sciences, USA* **110**(18), 7446–7451. PMID: 23589877.
- Wan, Q.H., Pan, S.K., Hu, L. *et al.* 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell Research* **23**(9), 1091–1105. PMID: 23917531.
- Wang, D. Y., Kumar, S., Hedges, S. B. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proceedings of the Royal Society of London B: Biological Sciences* **266**, 163–171.
- Wang, Z., Pascual-Anaya, J., Zadissa, A. *et al.* 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nature Genetics* **45**(6), 701–706. PMID: 23624526.
- Warren, W.C., Hillier, L.W., Marshall Graves, J.A. *et al.* 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**(7192), 175–183. PMID: 18464734.
- Warren, W.C., Clayton, D.F., Ellegren, H. *et al.* 2010. The genome of a songbird. *Nature* **464**(7289), 757–762. PMID: 20360741.
- Wegrzyn, J.L., Liechty, J.D., Stevens, K.A. *et al.* 2014. Unique Features of the Loblolly Pine (*Pinus taeda* L.) Megagenome Revealed Through Sequence Annotation. *Genetics* **196**(3), 891–909. PMID: 24653211.
- Werren, J.H., Richards, S., Desjardins, C.A. *et al.* 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* **327**(5963), 343–348. PMID: 20075255.
- Williams, B. A., Hirt, R. P., Lucocq, J. M., Embley, T. M. 2002. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. *Nature* **418**, 865–869.
- Winzeler, E.A. 2008. Malaria research in the post-genomic era. *Nature* **455**(7214), 751–756. PMID: 18843360.
- Wolfe, K. H., Morden, C. W., Palmer, J. D. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proceedings of the National Academy of Sciences USA* **89**, 10648–10652.
- Xia, Q., Zhou, Z., Lu, C. *et al.* 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* **306**, 1937–1940. PMID: 15591204.
- Xiong, J., Lu, Y., Feng, J. *et al.* 2013. Tetrahymena functional genomics database (TetraFGD): an integrated resource for *Tetrahymena* functional genomics. *Database (Oxford)* **2013**, bat008. PMID: 23482072.
- Xu, P., Widmer, G., Wang, Y. *et al.* 2004. The genome of *Cryptosporidium hominis*. *Nature* **431**, 1107–1112. PMID: 15510150.
- Yan, G., Zhang, G., Fang, X. *et al.* 2011. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology* **29**(11), 1019–1023. PMID: 22002653.
- Yang, N., Farrell, A., Niedelman, W. *et al.* 2013. Genetic basis for phenotypic differences between different *Toxoplasma gondii* type I strains. *BMC Genomics* **14**, 467. PMID: 23837824.
- You, M., Yue, Z., He, W. *et al.* 2013. A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics* **45**(2), 220–225. PMID: 23313953.

- Young, N.D., Bharti, A.K. 2012. Genome-enabled insights into legume biology. *Annual Review of Plant Biology* **63**, 283–305. PMID: 22404476.
- Young, N.D., Debelle, F., Oldroyd, G.E. et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**(7378), 520–524. PMID: 22089132.
- Yu, J., Hu, S., Wang, J. et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92. PMID: 11935017.
- Yu, J., Wang, J., Lin, W. et al. 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biology* **3**, e38. PMID: 15685292.
- Yuan, J., Cheng, K.C., Johnson, R.L. et al. 2011. Chemical genomic profiling for antimalarial therapies, response signatures, and molecular targets. *Science* **333**(6043), 724–729. PMID: 21817045.
- Zdobnov, E.M., von Mering, C., Letunic, I. et al. 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159. PMID: 12364792.
- Zhan, S., Merlin, C., Boore, J.L., Reppert, S.M. 2011. The monarch butterfly genome yields insights into long-distance migration. *Cell* **147**(5), 1171–1185. PMID: 22118469.
- Zhang, G., Cowled, C., Shi, Z. et al. 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* **339**(6118), 456–460. PMID: 23258410.
- Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I., Schneider, G. 2001. Deciphering apicoplast targeting signals: feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280**, 19–26.



Our current efforts to understand the human genome include a focus on the genetic similarities and differences among various geographic (ethnic) groups. The International HapMap Project initially generated detailed genotype information on 270 individuals from four groups with diverse geographic ancestry: Yoruba from Ibadan, Nigeria; Utah residents of Northern and Western European ancestry; Han Chinese in Beijing; and Japanese in Tokyo. In past centuries there have been many attempts to understand the bases of phenotypic differences among

humans. Baron Georges Cuvier (1769–1832) attempted a systematic classification of animals, describing four great divisions of the animal kingdom: vertebrate animals; molluscous animals; articulate animals; and radiate animals (also called Zoophytes). His work included a classification of humans based on anatomical differences. These images are from *The Animal Kingdom, Arranged According to Its Organization* by Cuvier (1849, plates I–IV) and depict varieties of human races.

Source: Cuvier (1849).

# Human Genome

# CHAPTER 20

*This project would greatly increase our understanding of human biology and allow rapid progress to occur in the diagnosis and ultimate control of many human diseases. As visualized, it would also lead to the development of a wide range of new DNA technologies and produce the maps and sequences of the genomes of a number of experimentally accessible organisms, providing central information that will be important for increasing our understanding of all biology.*

—National Research Council (1988, p. 11).

## LEARNING OBJECTIVES

After studying this chapter you should be able to:

- describe the main features of the human genome;
- provide an overview of all the human chromosomes, giving a general description of their size, number of genes, and key features; and
- explain the purpose and main conclusions of several key human genome efforts including the HapMap Project and 1000 Genomes Projects.

## INTRODUCTION

The human genome is the complete set of DNA in *Homo sapiens*. This DNA encodes the proteins and other products that define our cells and ultimately define who we are as biological entities. Through the genomic DNA, protein-coding genes are expressed that form the architecture of the trillions of cells that comprise each of our bodies. It is variations in the genome that in large part account for the differences between people, from physical features to personality to disease states.

The initial sequencing of the human genome in 2003 was a triumph of science. It followed 50 years exactly after the publication of the double-stranded helical structure of DNA by Crick and Watson (1953). The genome sequence was achieved through an international collaboration involving hundreds of scientists. (In the case of the publicly funded version, this was the International Human Genome Sequencing Consortium (IHGSC), described in “Human Genome Project” below.) This project could not have been possible without fundamental advances in the emerging fields of bioinformatics and genomics.

In this chapter we first summarize some of the major findings of the human genome project. Second, we review resources for the study of the human genome at three sites: the National Center for Biotechnology Information (NCBI); the Ensembl project; and the genome center at the University of California, Santa Cruz.

In 2001 the sequencing and analysis of a draft version of the human genome was reported by the IHGSC (2001) and Celera Genomics (Venter *et al.*, 2001). In the next part of this chapter, we follow the outline of the public consortium's 62-page article to describe the human genome from a bioinformatics perspective. We also describe subsequent findings on finishing the euchromatic sequence (IHGSC, 2004) and characterizing each of the 22 autosomes and two sex chromosomes (as well as the mitochondrial genome). Finally, we describe variation in the human genome, including the HapMap Project, the 1000 Genomes Project, and the analysis of individual genomes.

## MAIN CONCLUSIONS OF HUMAN GENOME PROJECT

These findings are summarized from several sources, including IHGSC (2001), Venter *et al.* (2001), and the Wellcome Trust Sanger Institute ([WebLink 20.1](http://www.sanger.ac.uk/about/history/hgp/)).

An Ensembl estimate of the number of genes (as well as many other human genome statistics) is at [http://www.ensembl.org/Homo\\_sapiens/Info/Annotation](http://www.ensembl.org/Homo_sapiens/Info/Annotation) (WebLink 20.2). Release 79 lists 20,300 human protein-coding genes.

We introduced various types of repetitive elements in Chapter 8, and further define them in "Repeat Content of Human Genome" below.

As an introduction to the Human Genome Project, we begin with a summary of its main findings. These are from the IHGSC (2001) paper, supplemented with more recent observations:

1. There were reported to be about 30,000–40,000 predicted protein-coding genes in the human genome. However, the initial sequencing and annotation were incomplete, and in subsequent years a variety of new tools were developed (Chapters 8 and 9) as well as comparative approaches as more vertebrate genomes were sequenced. A revised estimate suggests that there are ~20,300 protein-coding genes (IHGSC, 2004; Ensembl.org). This estimate is surprising because we have about the same number of genes as much simpler organisms such as *Arabidopsis thaliana* (~27,000 protein-coding genes according to TAIR) and pufferfish (~18,500 protein-coding genes according to Ensembl), and marginally more genes than are found in many nematode and insect genomes.
2. The human proteome is far more complex than the set of proteins encoded by invertebrate genomes. Vertebrates have a more complex mixture of protein domain architectures. Additionally, the human genome displays greater complexity in its processing of mRNA transcripts by alternative splicing.
3. Hundreds of human genes were acquired from bacteria by lateral gene transfer, according to the initial report (IHGSC, 2001; Ponting, 2001). Subsequently Salzberg *et al.* (2001) suggested a revised estimate of 40 genes that underwent horizontal transfer. These genes are homologous to bacterial sequences, but appear to lack orthologous genes in other vertebrate and invertebrate species. In recent years the emphasis has changed from laterally acquired genes (discussed in Chapter 17) to the vast number of bacterial, archaeal, and viral genes from organisms living inside the human body, called the human microbiome. We described this project in Chapter 17.
4. More than 98% of the human genome does not consist of exons that code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs; 20%), short interspersed elements (SINEs; 13%), long-terminal-repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Half the human genome is therefore derived from transposable elements. However, there has been a decline in the activity of these elements in the hominid lineage. In recent years the ENcyclopdia of DNA Element (ENCODE) project has created a deep, rich catalog of functional elements in the human genome (Chapter 8; ENCODE Project Consortium *et al.*, 2012). This project has defined coding and noncoding gene structures, catalogued pervasive transcriptional activity, and identified a wide range of biochemical signatures such as chromatin modifications.
5. Segmental duplication is a frequent occurrence in the human genome, particularly in pericentromeric and subtelomeric regions. This phenomenon is more common than in yeast, fruit fly, or worm genomes. There are three principal ways in which gene duplications arise in the human genome (Green and Chakravarti, 2001). First,

tandem duplications (created from sequence repeats in a localized region) occur rarely. Second, processed mRNAs are duplicated by retrotransposition. This produces intronless paralogs that are present at one or many sites. Third, and most commonly, segmental duplications occur in which large sections of a chromosome transfer to a new site. We introduced these concepts in Chapter 8.

6. There are several hundred thousand *Alu* repeats in the human genome. These have been thought to represent elements that replicate promiscuously. However, their distribution is nonrandom: they are retained in GC-rich regions and may therefore confer some benefit on the human genome.
7. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males. Recent whole-genome sequencing of members of families has established the mutation rate at  $\sim 1.2 \times 10^{-8}$  per base pair per generation (reviewed in Scally and Durbin, 2012).
8. More than 1.4 million single-nucleotide polymorphisms (SNPs) were identified. SNPs are single-nucleotide variations that occur once every 100–300 base pairs. The International HapMap Consortium *et al.* (2007) reported a haplotype map of 3.1 million SNPs, and today the genotype and copy number of one million SNPs are routinely measured on a single sample using a microarray. This is already having a profound impact on studies of variation in the human genome.

The NCBI database of SNPs currently lists ~113 million RefSNP clusters having rs identifiers, of which over 88 million have been validated (dbSNP build 142, March 2015; <http://www.ncbi.nlm.nih.gov/SNP/>, WebLink 20.3).

## GATEWAYS TO ACCESS THE HUMAN GENOME

There are many ways to access information about the human genome, including three principal browsers at NCBI, Ensembl, and UCSC.

### NCBI

The NCBI offers several main ways to access data on the human genome. From the Genome page select “human genome resources,” which provides links to each chromosome and a variety of web resources. Alternatively, select the Map Viewer (**Fig. 20.1**). This page allows searches by clicking on a chromosome or by entering a text query. The human Map Viewer integrates human sequence and data from cytogenetic maps, genetic linkage maps, radiation hybrid maps, and YAC chromosomes.

By visiting the NCBI Gene page for a gene such as hemoglobin beta (*HBB*), you can use another sequence viewer (**Fig. 20.2**). We previously showed this viewer within the Genome Workbench (**Fig. 9.18**). This viewer allows the addition of tracks such as RNA-seq data supporting gene models. Many other NCBI features are available through the NCBI Gene page, including links to UniGene entries and human-mouse-rat homology maps.

### Ensembl

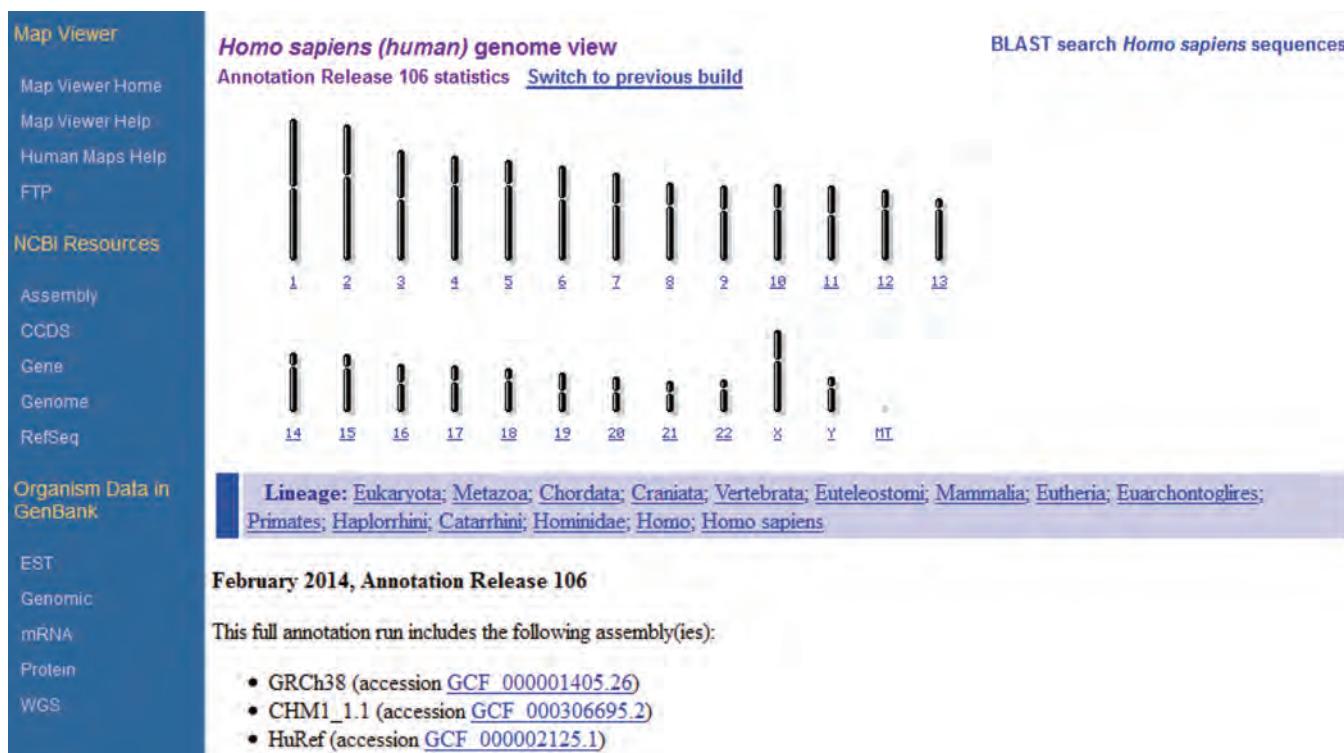
Ensembl is a comprehensive resource for information about the human genome as well as other genomes (Flicek *et al.*, 2014). This resource effectively interconnects a wide range of genomics tools with a focus on annotation of known and newly predicted genes. In addition to making annotation information on genes easily accessible, Ensembl provides access to the underlying data that support models of gene prediction. This is described in the following. The current statistics for the contents of the Ensembl human build are shown in **Table 20.1**.

From the main page of Ensembl, you can type a text query (such as *HBB* for human beta globin), perform a BLAST search, or browse by chromosome (see **Fig. 8.2**). There are several main entry points to access the Ensembl database. Note that the top bar of the

A Human Genome Resources page is available at <http://www.ncbi.nlm.nih.gov/genome/guide/human/> (WebLink 20.4). The Map Viewer (for human and other organisms) is accessed via <http://www.ncbi.nlm.nih.gov/projects/mapview/> (WebLink 20.5).

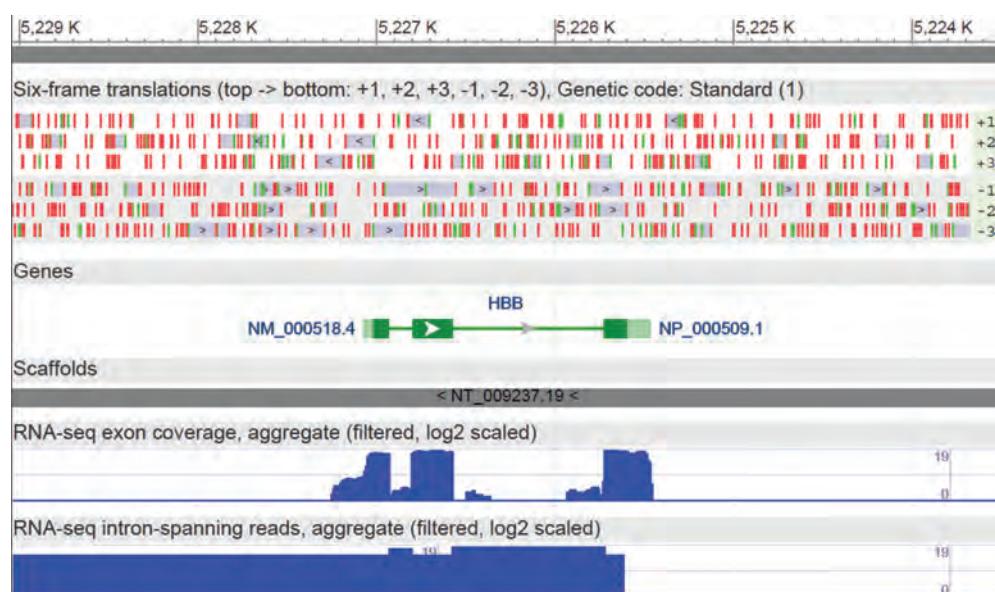
Ensembl, a joint project between the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and the Sanger Institute is available at <http://www.ensembl.org> (WebLink 20.6). The human database is at [http://www.ensembl.org/Homo\\_sapiens/](http://www.ensembl.org/Homo_sapiens/) (WebLink 20.7). We described Ensembl projects for mouse, rat, zebrafish, fugu, mosquito, and other organisms in Chapter 19.

We saw an example of the Ensembl BLAST server in Figures 5.1 and 5.2.



**FIGURE 20.1** The Human Map Viewer is accessible from NCBI. This resource displays cytogenetic, genetic, physical, and radiation hybrid maps of human genome sequence. A search box (not shown) allows you to enter a query such as “hbb” for a graphical view of beta globin on chromosome 11.

Source: Human Map Viewer, NCBI.



**FIGURE 20.2** The sequence viewer from NCBI Gene shows the region of chromosome 11 containing the beta globin gene. A variety of tracks can be added; shown here are six-frame translations, scaffolds, and RNA-seq data.

Source: NCBI Gene, NCBI.

**TABLE 20.1 Human genome statistics from Ensembl. Note that base pairs refers to sum of lengths of DNA table. Golden Path length refers to sum of nonredundant top-level sequence regions.**

Coding genes	20,364
Small noncoding genes	9,673
Long noncoding genes	14,817
Pseudogenes	14,415
Gene transcripts	196,345
Genscan gene predictions	50,117
Short variants (SNPs, indels, somatic mutations)	65,897,584
Structural variants	4,168,103
Base pairs	3,381,944,086
Golden Path length	3,096,649,726

*Source:* Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

result page includes three tabs for the human genome, for the location, and for the gene; each offers different viewing and analysis options.

1. We introduced a chromosome view (**Fig. 8.2a**), providing summary information across an entire chromosome.
2. We also introduced a region overview (accessed via a Location tab; **Fig. 8.2b**). This uses a Javascript-based, scrollable, zoomable browser called Genoverse.
3. Viewing the “region in detail” provides a detailed view of a chromosomal region (e.g., a gene, as shown for *HBB* in **Fig. 20.3a**). The left sidebar includes an option to “configure this page,” allowing you to select among hundreds of tracks to display (**Fig. 20.3b**).
4. The Genetic Variation link (on the left sidebar) includes a table listing types of variation across each gene, including information on SNPs and SIFT and PolyPhen-2 scores (described in Chapter 9).
5. A synteny view accessed from the location tab shows the corresponding region of chromosomes from other organisms where a gene such as *HBB* is localized (**Fig. 20.4**). This figure shows the correspondence of four mouse chromosomal regions to human chromosome 11. For contrast, the figure also shows the most-conserved chromosome (the X chromosome) and the least-conserved chromosome (the Y chromosome).
6. Cyto view displays genes, BAC end clones, repetitive elements, and the tiling path across genomic DNA regions.

The UCSC Genome Bioinformatics site is accessible at <http://genome.ucsc.edu/> (WebLink 20.8). It was developed by David Haussler’s group (Karolchik *et al.*, 2014).

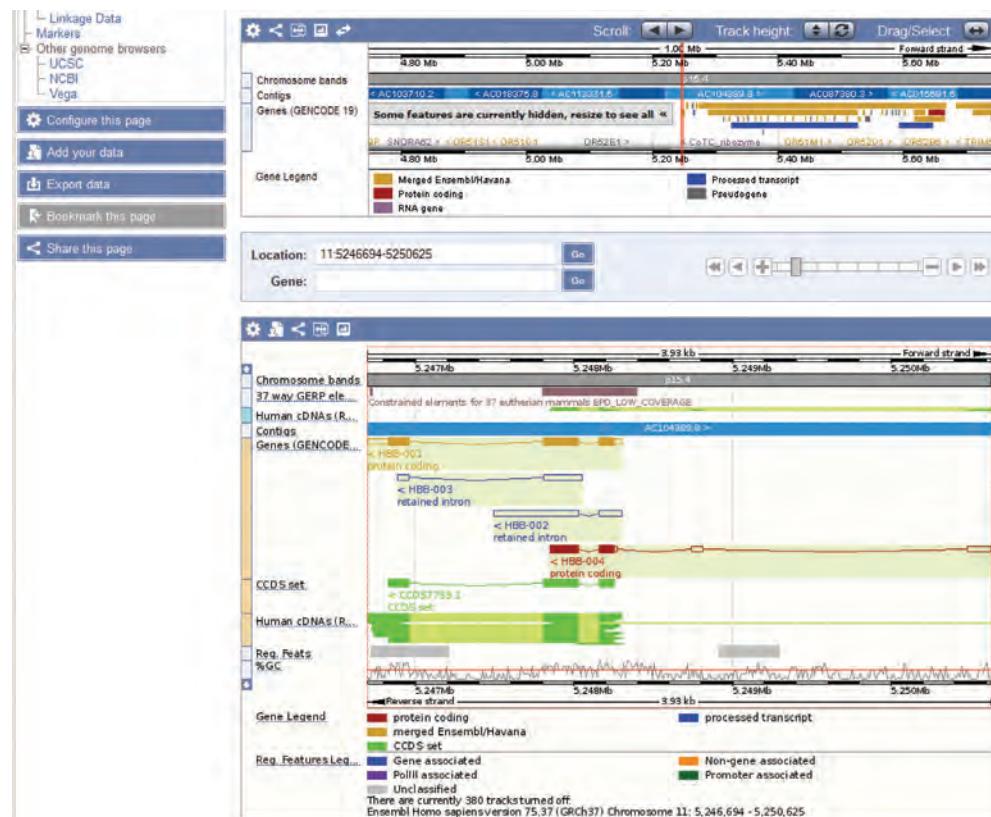
An NHGRI introduction to the human genome project is available at <http://www.genome.gov/10001772> (WebLink 20.9). A document describing a 2003 NHGRI vision for the future of genomics research can be viewed at <http://www.genome.gov/11007524> (WebLink 20.10, Collins *et al.*, 2003). A 2011 vision by Eric Green *et al.* is available at <http://www.genome.gov/pages/about/planning/2011nhgristrategicplan.pdf> (WebLink 20.11). We discussed that article at the start of Chapter 15.

## University of California at Santa Cruz Human Genome Browser

The “Golden Path” is the human genome sequence annotated at UCSC. Along with the Ensembl and NCBI sites, the human genome browser at UCSC is one of the three main web-based sources of information for both the human genome and other vertebrate genomes. It has become a basic resource in the fields of bioinformatics and genomics, and we have relied on it throughout this book.

## NHGRI

The National Human Genome Research Institute (NHGRI) has a leading role in genome sequencing, coordinating pilot-scale and large-scale sequencing efforts, technology development, and policy development.

(a) Ensembl location view of *HBB*

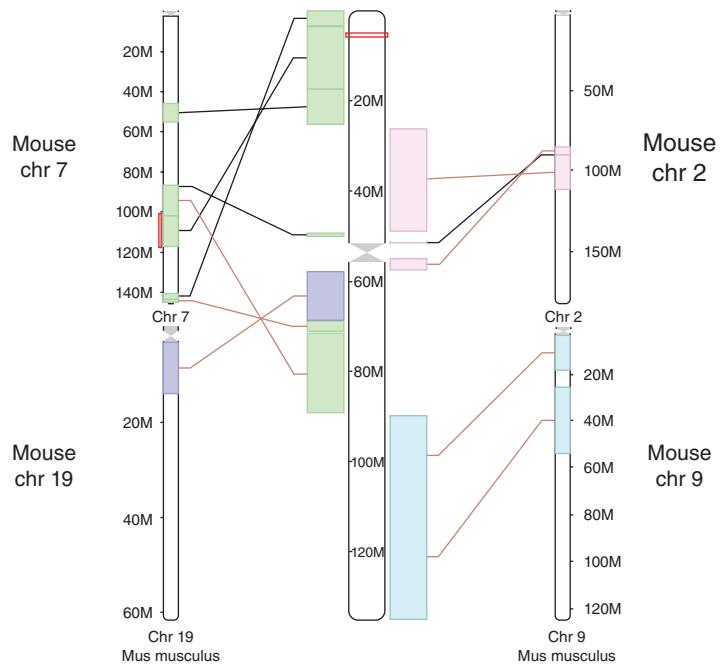
(b) Ensembl configure options

Active tracks	
Favourite tracks	
Track order	
Search results	
Sequence and assembly	(4/26)
Sequence	(2/4)
Markers	(0/1)
GRC alignments	(2/3)
Ditag features	(0/2)
Simple features	(0/4)
Clones & misc. regions	(0/12)
Genes and transcripts	(2/85)
Genes	(2/6)
Prediction transcripts	(0/2)
RNASeq models	(0/77)
mRNA and protein alignments	(1/14)
mRNA alignments	(1/3)
EST alignments	(0/1)
Protein alignments	(0/4)
Protein features	(0/6)
ncRNA	(0/1)
Variation	(0/91)
dbSNP	(0/2)
1000 Genomes & HapMap	(0/13)
Phenotype and curated variants	(0/17)
Individual genomes	(0/14)
Arrays and other	(0/13)
Failed variants	(0/1)
Sequence variants	(0/2)
Structural variants	(0/26)
Recombination & Accessibility	(0/3)
Somatic mutations	(0/5)
Somatic variants	(0/3)
Somatic structural variants	(0/2)
Regulation	(1/117)
Regulatory features	(1/20)
Open chromatin & TFBS	(0/14)
Histones & polymerases	(0/13)
DNA Methylation	(0/65)
Other regulatory regions	(0/5)
Comparative genomics	(1/73)
Multiple alignments	(0/4)
Conservation regions	(1/5)
BLASTz/LASTz alignments	(0/47)
Translated blat alignments	(0/17)

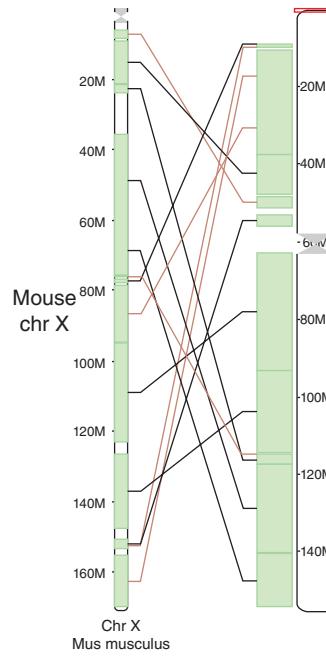
**FIGURE 20.3** The Ensembl human genome browser offers a wealth of resources. A direct way to begin searching the site is to enter a search term such as HBB (top) for beta globin. (a) A “Gene” tab provides a link to the “region in detail.” (b) The left sidebar includes a link to configure the page. Hundreds of tracks can be selected and displayed on the browser.

Source: Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

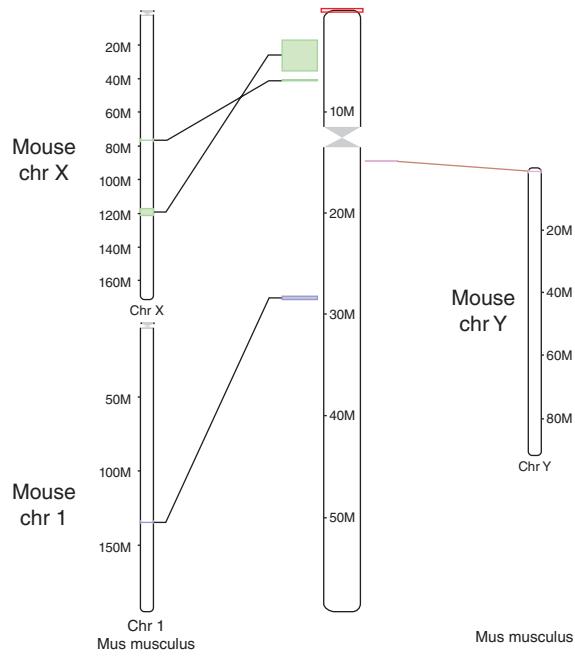
(a) Conserved synteny between human chromosome 11 and mouse chromosomes



(b) Human chromosome X compared to mouse



(c) Human chromosome Y compared to mouse



**FIGURE 20.4** Ensembl location tab links to conserved synteny maps including those for human/mouse. (a) Human chromosome 11 (including the *HBB* gene, red box) is shown in the center as an ideogram. It corresponds to mouse chromosomes 7, 2, 19, and 9. Although the lineages leading to modern humans and mice diverged about 90 million years ago, it is still straightforward to identify regions of conserved synteny. (b) The human X chromosome (ideogram at right) is extremely closely conserved with the mouse X chromosome. (c) The Y chromosomes of human (ideogram at center) and mouse are extraordinarily poorly conserved.

Source: Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

## Wellcome Trust Sanger Institute

The website for human genetics at the Wellcome Trust Sanger Institute (WTSI) is <http://www.sanger.ac.uk/research/areas/human-genetics/> (WebLink 20.12). The Human Genome Project gateway is at <http://www.sanger.ac.uk/about/history/hgp/> (WebLink 20.13).

The euchromatin is the primary gene-containing part of the genome, although there are also genes in heterochromatin.

The National Human Genome Research Institute describes the finishing process at <http://www.genome.gov/10000923> (WebLink 20.14).

The National Academy Press (<http://www.nap.edu>, WebLink 20.15) offers this 1988 book free online at [http://www.nap.edu/catalog.php?record\\_id=1097](http://www.nap.edu/catalog.php?record_id=1097) (WebLink 20.16).

You can read about ELSI at <http://www.genome.gov/10001618> (WebLink 20.17) or [http://web.ornl.gov/sci/techresources/Human\\_Genome/elsi/index.shtml](http://web.ornl.gov/sci/techresources/Human_Genome/elsi/index.shtml) (WebLink 20.18).

## HUMAN GENOME PROJECT

The two articles on the human genome project that appeared in February 2001 provided an initial glimpse of the genome (IHGSC, 2001; Venter *et al.*, 2001). In the next portion of this chapter, we will follow the outline of the public consortium paper (IHGSC, 2001). We do not summarize all the major findings, but focus on selected topics. The sequence reported in 2001 represented 90% completion of the human genome.

Finishing the human genome is a process that involves producing finished maps (with continuous, accurate alignments of large-insert clones spanning euchromatic loci) and producing finished clones (completely, accurately sequenced). Additional publications have described the sequence of all 25 human chromosomes in more detail (22 autosomes, the two sex chromosomes, and the mitochondrial genome); we summarize the findings in the following. The IHGSC (2004) reported finishing the euchromatic sequence of the human genome. Even at that stage 341 gaps remained, spanning about 1% of the euchromatic genome. Furthermore, heterochromatic regions which are far harder to sequence contain many genes and other elements of interest. Although the human genome was sequenced, finishing and annotating this sequence are ongoing processes.

## Background of Human Genome Project

The Human Genome Project was first proposed by the US National Research Council (1988). This report proposed the creation of genetic, physical, and sequence maps of the human genome. At the same time, parallel efforts were supported for model organisms (bacteria, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*).

The major goals of the Human Genome Project are listed in **Table 20.2**. One component of the human genome project is the Ethical, Legal and Social Issues (ELSI) initiative. A portion of the annual budget (3–5%) has been devoted to ELSI, making it the world's largest bioethics project.

Examples of issues addressed by ELSI include:

- Who owns genetic information?
- Who should have access to genetic information?
- How does genomic information affect members of minority communities?
- What societal issues are raised by new reproductive technologies?
- How should genetic tests be regulated for reliability and validity?
- To what extent do genes determine behavior?
- Are there health risks associated with genetically modified foods?

All these issues are becoming increasingly important, particularly as we begin to obtain the nearly complete genomic DNA sequence of hundreds of thousands of individuals (described in “Variation: Sequencing Individual Genomes” below). In Chapter 21 we ask other questions raised by whole-genome and whole-exome sequencing. How should “incidental” findings be handled, such as finding a mutation that predisposes a patient to cancer when the sequencing was performed to identify genetic variants that underlie a completely different condition?

**TABLE 20.2 Eight goals of Human Genome Project (1998–2003). Adapted from**  
**✉ [http://www.ornl.gov/sci/techresources/Human\\_Genome/hg5yp/goal.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/hg5yp/goal.shtml).**

1. Human DNA sequence	<ul style="list-style-type: none"> <li>• Finish the complete human genome sequence by the end of 2003.</li> <li>• Achieve coverage of at least 90% of the genome in a working draft based on mapped clones by the end of 2001.</li> <li>• Make the sequence totally and freely accessible.</li> </ul>
2. Sequencing technology	<ul style="list-style-type: none"> <li>• Continue to increase the throughput and reduce the cost of current sequencing technology.</li> <li>• Support research on novel technologies that can lead to significant improvements in sequencing technology.</li> <li>• Develop effective methods for the development and introduction of new sequencing technologies.</li> </ul>
3. Human genome sequence variation	<ul style="list-style-type: none"> <li>• Develop technologies for rapid, large-scale identification and/or scoring of single-nucleotide polymorphisms and other DNA sequence variants.</li> <li>• Identify common variants in the coding regions of the majority of identified genes during this five-year period.</li> <li>• Create a SNP map of at least 100,000 markers.</li> <li>• Create public resources of DNA samples and cell lines.</li> </ul>
4. Functional genomics technology	<ul style="list-style-type: none"> <li>• Generate sets of full-length cDNA clones and sequences that represent human genes and model organisms.</li> <li>• Support research on methods for studying functions of nonprotein-coding sequences.</li> <li>• Develop technology for comprehensive analysis of gene expression.</li> <li>• Improve methods for genome-wide mutagenesis.</li> <li>• Develop technology for large-scale protein analyses.</li> </ul>
5. Comparative genomics	<ul style="list-style-type: none"> <li>• Complete the sequence of the roundworm <i>C. elegans</i> genome and the fruit fly <i>Drosophila</i> genome</li> <li>• Develop an integrated physical and genetic map for the mouse, generate additional mouse cDNA resources, and complete the sequence of the mouse genome by 2008.</li> </ul>
6. Ethical, legal, and social issues	<ul style="list-style-type: none"> <li>• Examine issues surrounding completion of the human DNA sequence and the study of genetic variation.</li> <li>• Examine issues raised by the integration of genetic technologies and information on health care and public health activities.</li> <li>• Examine issues raised by the integration of knowledge about genomics and gene–environment interactions in nonclinical settings.</li> <li>• Explore how new genetic knowledge may interact with a variety of philosophical, theological, and ethical perspectives.</li> <li>• Explore how racial, ethnic, and socioeconomic factors affect the use, understanding, and interpretation of genetic information, the use of genetic services, and the development of policy.</li> </ul>
7. Bioinformatics and computational biology	<ul style="list-style-type: none"> <li>• Improve content and utility of databases.</li> <li>• Develop better tools for data generation, capture, and annotation.</li> <li>• Develop and improve tools and databases for comprehensive functional studies.</li> <li>• Develop and improve tools for representing and analyzing sequence similarity and variation.</li> <li>• Create mechanisms to support effective approaches for producing robust, exportable software that can be widely shared.</li> </ul>
8. Training and manpower	<ul style="list-style-type: none"> <li>• Nurture the training of scientists skilled in genomics research.</li> <li>• Encourage the establishment of academic career paths for genomic scientists.</li> <li>• Increase the number of scholars who are knowledgeable in both genomic and genetic sciences and in ethics, law, or the social sciences.</li> </ul>

## Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence

The public consortium approach to sequencing the human genome was to employ the hierarchical shotgun sequencing strategy. The rationale for taking this approach was as follows:

- Shotgun sequencing can be applied to DNA molecules of many sizes, including plasmids (typically several kilobases), cosmid clones (40 kb), yeast, and BACs (up to 1–2 Mb).
- The human genome has large amounts of repetitive DNA (about 50% of the genome; see “Repeat Content of Human Genome” below). Whole-genome shotgun sequencing, the main approach taken by Celera Genomics, was not adopted by the public consortium because of the difficulties associated with assembling repetitive DNA fragments. In the public consortium approach, large-insert clones (typically 100–200 kb) from defined chromosomes were sequenced.
- The reduction of the sequencing project to specific chromosomes allowed the international team to reduce and distribute the sequencing project to a set of sequencing centers. These centers are listed in Web Document 20.1.

The 2001 draft version of the human genome was based on the sequence and assembly of over 29,000 BAC clones with a total length of 4.26 billion base pairs (Gb). There were 23 Gb of raw shotgun sequence data.

Early in the evolution of the Human Genome Project, it was thought that breakthroughs in DNA sequencing technology would be necessary to allow the completion of such a large-scale project. This did not occur. Instead, the basic principles of dideoxynucleotide sequencing by the method of Sanger (see Chapter 9) were improved upon. Some innovations to Sanger sequencing (see Green, 2001) included capillary electrophoresis-based sequencing machines for the automated detection of DNA molecules, improved thermostable polymerases, and fluorescent dye-labeled dideoxynucleotide terminators.

### Human Genome Assemblies

The human genome is organized into chromosomes that range in size from about 50 to 250 megabases. Most sequencing technologies produce reads that are well under 1000 base pairs in length and, in some cases, closer to 100 base pairs. Assembly is the process of building fragments of a genome to represent the genomic sequence. Sequencing reads are overlapped, a multiple sequence alignment is generated, and the consensus sequence is called a contig. (These contain no gaps, although they may contain N (unknown) base calls due to sequence ambiguity.) With each successive genome build, the fraction of all contigs that are small (e.g., less than 5 Mb) continues to decline (Table 20.3). Scaffolds are defined as contigs that have been ordered and oriented; they contain gaps whose number and length are estimated. We discussed assembly strategies in Chapter 9.

The public consortium draft genome sequence was generated by selecting, sequencing, and assembling BAC clones. Most libraries contained BAC clones or P1-derived artificial clones (PACs). These libraries were prepared from DNA obtained from anonymous donors. Selected clones were subjected to shotgun sequencing. In conjunction with sequencing of BAC and other large-insert clones, the sequence data were assembled into an integrated draft sequence. An example of the procedure is shown in Figure 15.11.

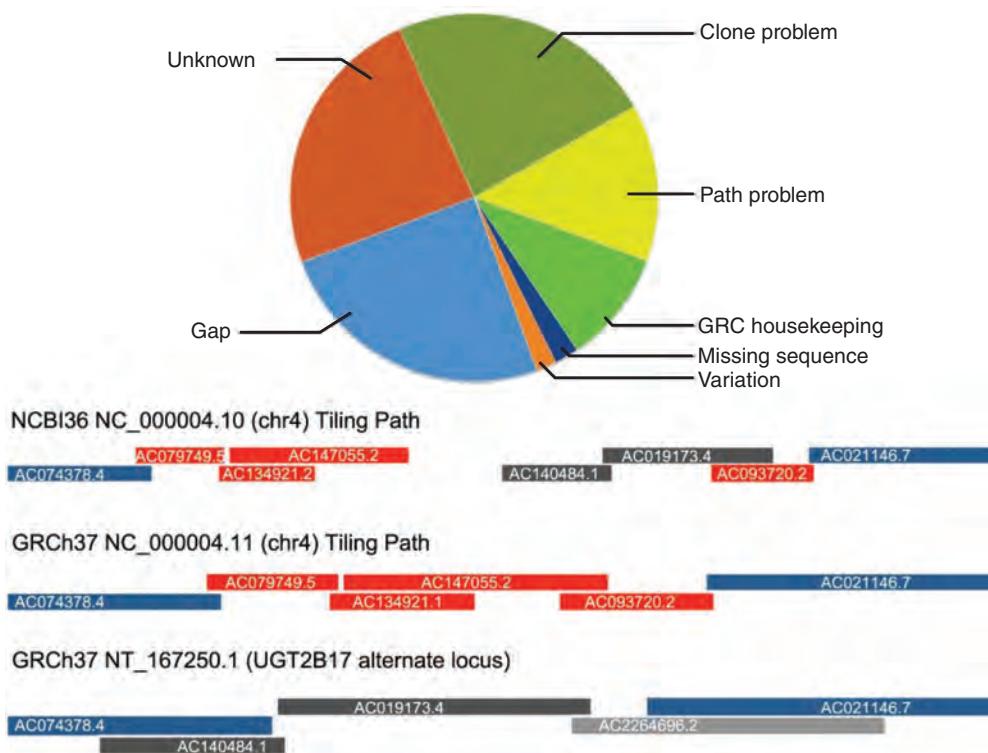
**TABLE 20.3 Contigs categorized by size. See Build 37 statistics (linked from [http://www.ncbi.nlm.nih.gov/genome/guide/human/release\\_notes.html](http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html))**

Range (kb)	Number	Length (base pairs)	Percent of total
< 300	37	5,818,180	0.2
300–1000	39	23,660,700	0.82
1000–5000	32	79,778,700	2.78
> 5000	82	2,757,100,000	96.18

The whole-genome shotgun assembly approach that was championed by Celera was proven successful by the sequencing of the *Drosophila melanogaster* genome in 2000 as well as by the initial sequence of the human genome (Venter *et al.*, 2001). Since then it has been widely adopted for thousands of bacterial, archaeal, and eukaryotic genome sequencing projects. A caveat noted by Evan Eichler and colleagues is that whole-genome shotgun sequencing and assembly performs poorly at correctly assembling repetitive DNA elements such as the segmental duplications that occupy over 5% of the human genome (She *et al.*, 2004). They compared a whole-genome shotgun sequence assembly to the assembly based on ordered clones and found that 38.2 megabases of pericentromeric DNA (about 80% of the size of a small autosome) was either not assembled, not assigned, or misassigned. Additionally, She *et al.* suggested that 40% of the duplicated sequence might be misassembled. Correctly resolving these structures will require a targeted approach to supplement whole-genome shotgun sequencing and assembly.

The Genome Reference Consortium (GRC) is responsible for coordinating new assemblies for human, mouse, and zebrafish genomes. Every few years a new genome assembly is released. The current assembly is Genome Reference Consortium Human Build 38 (abbreviated GRCh38). GRCh38 addresses some of the following issues (Fig. 20.5), many of which

GRCh38 was released in December 2013. Visit the GRC site at <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/#> (WebLink 20.22). GRC involves a collaboration between EMBL-EBI, the Wellcome Trust Sanger Institute, the Genome Institute at Washington University, and NCBI.



**FIGURE 20.5** Issues addressed by the Genome Reference Consortium (GRC) in the releases leading up to GRCh38 (December 2013). Top: categories of issues. Clone problem: a single clone has a single-nucleotide difference or misassembly. Path problem: the tiling path is incorrect and must be updated. GRC housekeeping: the tiling path must be regularized. Missing sequence: sequences need to be placed on the assembly. Variation: an alternate allele may need to be represented. Gap: a gap must be filled. Bottom: examples for a path problem. The NCBI representation was a mixed haplotype; the tiling path is shown. For GRCh37 the tiling paths include an alternate locus. Blue clones are anchors (included in all three paths). Red clones in GRCh37 correspond to an insertion path; dark-gray cones are in a deletion path (at bottom). A light-gray clone, not used in NCBI36, forms part of the GRCh37 alternate locus.

Source: Church *et al.* (2011).

**TABLE 20.4 Global statistics of human build GRCh38. The 25 chromosomes include 1-22, X, Y, and the mitochondrial genome.**

Number of regions with alternate loci or patches	207
Total sequence length	3,209,286,105
Total assembly gap length	159,970,007
Gaps between scaffolds	349
Number of scaffolds	735
Scaffold N50	67,794,873
Number of contigs	1,385
Contig N50	56,413,054
Total number of chromosomes	25

Source: [http://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.26/](http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/)

historically stemmed from the decentralized nature of the Human Genome Project and from inadequate models for the complexity of the human genome.

You can view GRC map contigs, a GRC incident database, patches, and haplotypes at the UCSC Genome Browser using the tracks in the “Mapping and Sequencing” section at <http://genome.ucsc.edu> (WebLink 20.23). For example, view a region of 6.5 Mb on chromosome 6 (chr6:27,500,001–34,000,000 of GRCh37). This provides access to the DNA sequence of alternate loci in the MHC region.

- The haploid assembly represents a mixture of haplotypes. Humans are diploid, and these sequences need to be represented. In a diploid assembly a chromosome assembly is available for both sets of chromosomes from an individual.
- Assemblies have errors. These are corrected in each new assembly release, and GRC also releases occasional “patches” updating information in scaffolds. Examples of errors that have been corrected in GRCh38 include missing sequences (e.g., the *TAS2R45* gene was absent from the reference assembly), and mismatches between transcript sequence and genomic sequences.
- Some loci exhibit allelic complexity. A single path cannot represent alternative haplotypes, and GRC creates alternate locus definitions. For example, explore the Major Histocompatibility Complex (MHC) on chromosome 6.
- Some regions, such as centromeres, have complex structures that are now assigned improved representation.

**Table 20.4** lists some global statistics of GRCh38. The total sequence length is about 3.2 billion base pairs, and there are still 160 Mb in gaps (typically involving highly repetitive regions that are challenging to sequence across).

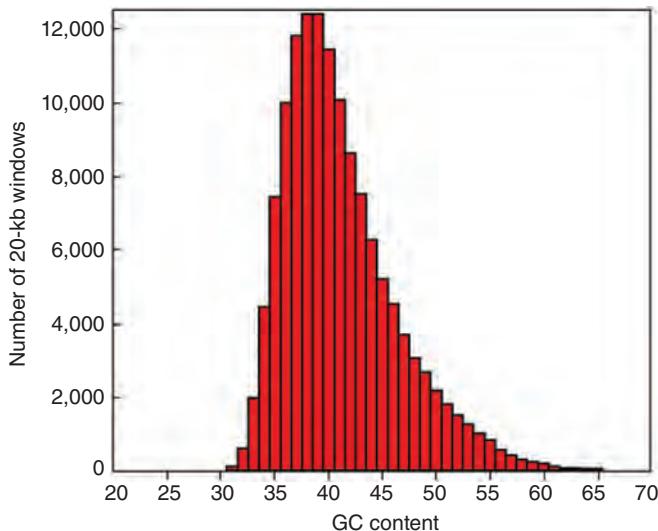
A key aspect of the sequence is the extent to which the sequenced fragments are contiguous. The average length of a clone or a contig is not a consistently useful measure of the extent to which a genome has been sequenced and assembled. Instead the N50 length describes the largest length  $L$  such that 50% of all nucleotides are contained in contigs or scaffolds of at least size  $L$ . For the draft version of the human genome sequence, half of all nucleotides were present in a fingerprint clone contig of at least 8.4 Mb. The N50 length rose to 38.5 Mb with the GRCh36 genome assembly, while currently the contig and scaffold N50 values are >56 Mb and ~68 Mb, respectively (**Table 20.4**).

## Broad Genomic Landscape

We discuss the 25 human chromosomes in more detail below (see “25 Human Chromosomes”), based on projects focused on finishing the sequence of each one. The autosomes are numbered approximately in order of size. The largest, chromosome 1, is 249 Mb in length; the smallest, chromosome 21, is 48 Mb.

Having a nearly complete view of the nucleotide sequence of the human genome, we can explore its broad features. These include:

- the distribution of GC content;
- CpG islands and recombination rates;



**FIGURE 20.6** Histogram of percent GC content versus the number of 20 kb windows in the draft human genome sequence. Note that the distribution is skewed to the right, with a mean GC content of 41%.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

- the repeat content; and
- the gene content.

We examine each of these four features of the genome in the following sections. Using the resources of UCSC, Ensembl, and NCBI, we can explore the genomic landscape from the level of single nucleotides to entire chromosomes.

#### Long-Range Variation in GC Content

The average GC content of the human genome is 41%. However, there are regions that are relatively GC rich and GC poor. A histogram of the overall GC content (in 20 kb windows) shows a broad profile with skewing to the right (Fig. 20.6). Fifty-eight percent of the GC content bins are below the average while 42% are above the average, including a long tail of highly GC-rich regions.

Giorgio Bernardi and colleagues have proposed that mammalian genomes are organized into a mosaic of large DNA segments (e.g., >300 kb) called isochores. These isochores are fairly homogeneous compositionally and can be divided into GC-poor families (L1 and L2) or GC-rich families (H1, H2, and H3). The IHGSC (2001) report did not identify clearly defined isochores, and Haring and Kypr (2001) did not detect isochores in human chromosomes 21 and 22. Subsequent analyses by Bernardi and colleagues (Bernardi, 2001; Costantini and Bernardi, 2008; Arhondakis *et al.*, 2011) do support the mosaic organization of the human genome by GC content. The discrepancies depend in part on the size of the window of genomic DNA that is analyzed.

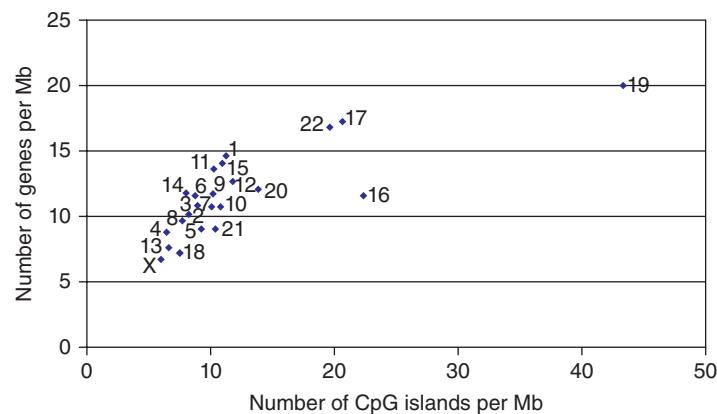
#### CpG Islands

The dinucleotide CpG is greatly underrepresented in genomic DNA, occurring at about one-fifth its expected frequency (we introduced this topic in Chapter 8). Many CpG dinucleotides are methylated on the cytosine and are subsequently deaminated to thymine bases. However, the genome contains many “CpG islands” which are typically associated with the promoter and exonic regions of housekeeping genes (Gardiner-Garden and Frommer, 1987). CpG islands have roles in processes such as gene silencing, genomic imprinting, and X-chromosome inactivation (Tycko and Morison, 2002; Jones, 2012; Smith and Meissner, 2013).

You can view GC content across any chromosome in the NCBI, Ensembl, or UCSC genome browsers. For example, in the Ensembl browser click “configure” to add a GC content layer. In the UCSC Table Browser you can output the GC content in a BED file, summarized by chromosome.

The L (light) and H (heavy) designations for isochores refer to the sedimentation behavior of genomic DNA in cesium chloride gradients. Genomic DNA fragments migrate to different positions based on their percent GC content.

Gene silencing refers to transcriptional repression. We briefly described MeCP2, a protein that binds to methylated CpG islands, in Chapter 12 (Fig. 12.9, 12.10). MeCP2 further recruits proteins such as a histone deacetylase that alters chromatin structure and represses transcription. Mutations in *MECP2*, the X-linked gene encoding MeCP2, cause Rett syndrome (Amir *et al.*, 1999). This disease causes distinctive neurological symptoms in girls, including loss of purposeful hand movements, seizures, and autistic-like behavior (Chapter 21). X-chromosome inactivation is a dosage compensation mechanism in which cells in a female body selectively silence the expression of genes from either the maternally or paternally derived X chromosome (Avner and Heard, 2001).



**FIGURE 20.7** The number of CpG islands per megabase is plotted versus the number of genes per megabase as a function of chromosome. Note that chromosome 19, the most gene-rich chromosome, has the greatest number of CpG islands per megabase.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

The UCSC Table Browser lists 28,691 CpG islands in the human genome. To see this, visit the Table Browser at <http://genome.ucsc.edu> (WebLink 20.24). Set the clade to vertebrate, the genome to human, the assembly to GRCh37 (or another assembly), the group to Regulation, the track to CpG islands, and click summary statistics. CpG islands are defined as having GC content  $\geq 50\%$ , length  $> 200$  base pairs, and ratio  $> 0.6$  of observed number of CG dinucleotides to the expected number in that segment. Genomic imprinting is the differential expression of genes from maternal and paternal alleles. Tycko and Morison (2002) offer a database of imprinted genes (<http://igc.otago.ac.nz/home.html>, WebLink 20.25).

The NCBI, Ensembl, and UCSC genome browsers allow you to view both physical maps and genetic maps.

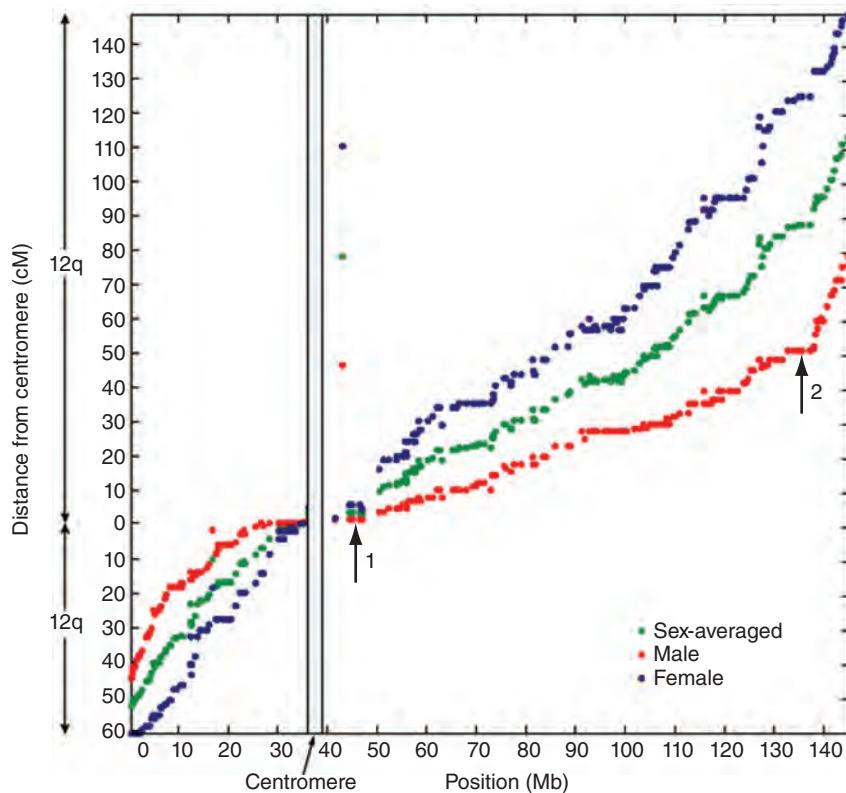
You can display predicted CpG islands in genomic DNA at the NCBI, Ensembl, and UCSC genome browser websites. According to the IHGSC (2001), there are 50,267 predicted CpG islands in the human genome. After blocking repetitive DNA sequences with RepeatMasker, there were 28,890 CpG islands, matching the number currently listed in the UCSC Table Browser for GRCh37. (This lower number reflects the high GC content of *Alu* repeats.) There are 5–15 CpG islands per megabase of DNA on most chromosomes, although chromosome 19 (the most gene-dense chromosome) contains 43 CpG islands per megabase (Fig. 20.7).

#### Comparison of Genetic and Physical Distance

It is possible to compare the genetic maps and physical maps of the chromosomes to estimate the rate of recombination per nucleotide (Yu *et al.*, 2001). Genetic maps, also known as linkage maps, are chromosome maps based on meiotic recombination. During meiosis the two copies of each chromosome present in each cell are reduced to one. The homologous parental chromosomes recombine (exchange DNA) during this process. Genetic maps describe the distances between DNA sequences (genes) based on their frequency of recombination. Genetic maps therefore describe DNA sequences in units of centimorgans (cM), which describe relative distance. One centimorgan corresponds to 1% recombination.

In contrast to genetic maps, physical maps describe the physical position of nucleotide sequences along each chromosome. With the completion of the human genome sequence, it became possible to compare genetic and physical maps.

Figure 20.8 shows a plot of genetic distance (y axis; in centimorgans) versus physical distance for human chromosome 12 (x axis; in megabases) (IHGSC, 2001). There are two main conclusions. First, the recombination rate tends to be suppressed near the centromeres (note the flat slope in Fig. 20.8, arrow 1), while the recombination rate is far higher near the telomeres. This effect is especially pronounced in males. Second, long chromosome arms tend to have an average recombination rate of 1 cM/Mb, while the shortest arms have a much higher average recombination rate ( $> 2$  cM/Mb). The range of the recombination rate throughout the genome varies from 0 to 9 cM/Mb (Yu *et al.*, 2001). These researchers identified 19 recombination “deserts” (up to 5 Mb in length with sex-average recombination rates  $< 0.3$  cM/Mb) and 12 recombination “jungles” (up to 6 Mb in length with sex-average recombination rates  $> 3.0$  cM/Mb). In computer laboratory exercise (20.4) at the end of this chapter, we identify regions of high (or low) recombination on the UCSC Genome Browser.



**FIGURE 20.8** Comparison of physical distance (in megabases, *x* axis) with genetic distance (in centimorgans, *y* axis) for human chromosome 12. Note that the recombination rate tends to be lower near the centromere (arrow 1) and higher near the telomeres (distal portion of each chromosome). The recombination is especially high in the male meiotic map (arrow 2).

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

## Repeat Content of Human Genome

Repetitive DNA occupies over 50% of the human genome. The origin of these repeats and their function present fascinating questions. What are the different kinds of repeats which occur? From where did they originate and when? Is there a logic to their promiscuous growth in our genomes or do they multiply without purpose? One of the outcomes of the Human Genome Project is that we are beginning to understand the extent and nature of the repeat content of our genome.

There are five main classes of repetitive DNA in humans (Jurka, 1998; IHGSC, 2001), as discussed in Chapter 8:

1. interspersed repeats (transposon-derived repeats);
2. processed pseudogenes: inactive, partially retroposed copies of protein-coding genes;
3. simple sequence repeats: microsatellites and minisatellites, including short sequences such as  $(A)_n$ ,  $(CA)_n$ , or  $(CGG)_n$ ;
4. segmental duplications, consisting of blocks of 10–300 kb that are copied from one genomic region to another; and
5. blocks of tandemly repeated sequences such as are found at centromeres, telomeres, and ribosomal gene clusters.

We briefly explore each of these types of repeats in the following sections.

Classes of interspersed repeat in the human genome						
				Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA		6-8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA		100-300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)		6-11 kb	450,000	8%
	Non-autonomous	(gag)		1.5-3 kb		
DNA transposon fossils	Autonomous	transposase		2-3 kb	300,000	3%
	Non-autonomous			80-3,000 bp		

**FIGURE 20.9** There are four types of transposable elements in the human genome: LINEs, SINEs, LTR transposons, and DNA transposons.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

### Transposon-Derived Repeats

Incredibly, 45% of the human genome or more consists of repeats derived from transposons. These are often called interspersed repeats. Many transposon-derived repeats replicated in the human genome in the distant past (hundreds of millions of years ago); because of sequence divergence, it is possible that the 45% value is an underestimate. Transposon-derived repeats can be classified as one of four categories (Jurka, 1998; Ostertag and Kazazian, 2001):

- LINEs occupy 21% of the human genome;
- SINEs occupy 13% of the human genome;
- LTR transposons account for 8% of the human genome; and
- DNA transposons comprise about 3% of the human genome.

The structure of these repeats is shown in **Figure 20.9**, as well as their abundance in the human genome. LINEs, SINEs, and LTR transposons are all retrotransposons that encode a reverse transcriptase activity. They integrate into the genome through an RNA intermediate. In contrast, DNA transposons have inverted terminal repeats and encode a bacterial transposon-like transposase activity.

Retrotransposons can further be classified into those that are autonomous (encoding activities necessary for their mobility) and those that are nonautonomous (depending on exogenous activities such as DNA repair enzymes). The most common nonautonomous retrotransposons are *Alu* elements.

Interspersed repeats occupy a far greater proportion of the human genome than in other eukaryotic genomes (**Table 20.5**). The total number of interspersed repeats is estimated to be 3 million. These repeats offer an important opportunity to study molecular evolution. Each repeat element, even if functionally inactive, represents a “fossil record” that can be used to study genome changes within and between species. Transposons accumulate mutations randomly and independently. It is possible to perform a multiple sequence alignment of transposons and to calculate the percent sequence divergence. Transposon evolution is assumed to behave like a molecular clock, which can be calibrated based on the known age of divergence of species such as humans and Old World monkeys (23 million years ago or MYA). Based on such phylogenetic analyses, several conclusions can be made (IHGSC, 2001; **Fig. 20.10**):

- Most interspersed repeats in the human genome are ancient, predating the mammalian eutherian radiation 100 MYA. These elements are removed from the genome only slowly.

The number of interspersed repeats was estimated using RepeatMasker to search RepBase (see Chapter 8).

*Alu* elements are so named because the restriction enzyme *Alu*I digests them in the middle of the sequence. In mice, these are called B1 elements.

**TABLE 20.5** Interspersed repeats in four eukaryotic genomes. “Bases” refers to percentage of bases in the genome, “families” to approximate number of families in the genome. Adapted from IHGSC (2001) with permission from Macmillan Publishers.

	Human		<i>Drosophila</i>		<i>C. elegans</i>		<i>A. thaliana</i>	
	Bases (%)	Families	Bases (%)	Families	Bases (%)	Families	Bases (%)	Families
LINE/SINE	33.4	6	0.7	20	0.4	10	0.5	10
LTR	8.1	100	1.5	50	0	4	4.8	70
DNA	2.8	60	0.7	20	5.3	80	5.1	80
Total	44.4	170	3.1	90	6.5	90	10.5	160

- SINEs and LINEs have long lineages, some dating back to 150 MYA.
- There is no evidence for DNA transposon activity in the human genome in the past 50 million years; they are therefore extinct fossils.

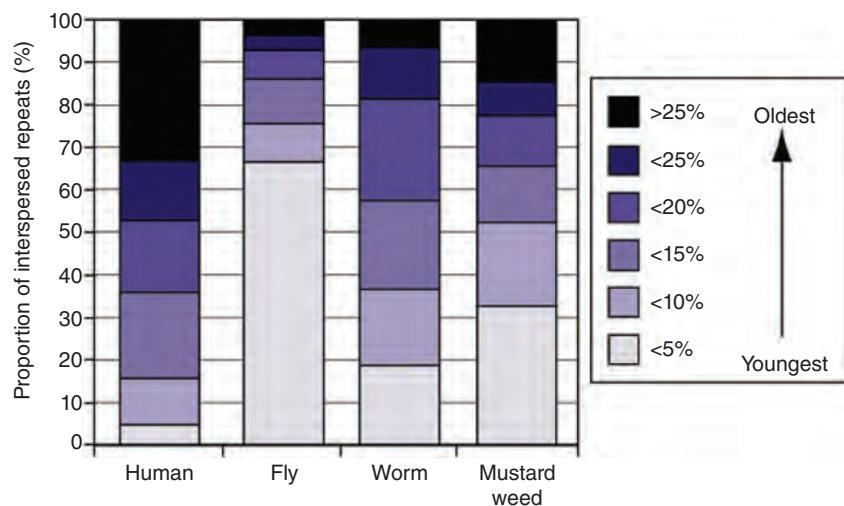
#### Simple Sequence Repeats

Simple sequence repeats are repetitive DNA elements that consist of a perfect (or slightly imperfect) tandem repeats of  $k$ -mers. When the repeat unit is short ( $k$  is about 1–12 bases), the simple sequence repeat is called a microsatellite. When the repeat unit is longer (from about 12–500 bases), it is called a minisatellite (Toth *et al.*, 2000).

Micro- and minisatellites comprise about 3% of the human genome (IHGSC, 2001). The most common repeat lengths are shown in Table 20.6. The most common repeat units are the dinucleotides AC, AT, and AG. We saw examples of these with the RepeatMasker program (Fig. 8.8).

#### Segmental Duplications

About 5.7% of the human genome consists of segmental duplications. These occur when the genome contains duplicated blocks of 1–200 kb of sequence (the typical size is 10–50 kb; Bailey *et al.*, 2001). Many of these duplication events are recent, because both introns and coding regions are highly conserved. (For ancient duplication



**FIGURE 20.10** Comparison of the age of interspersed repeats in four eukaryotic genomes. Humans have a small proportion of recent interspersed repeats.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

**TABLE 20.6 Simple sequence repeats (microsatellites) in human genome. SSR: simple sequence repeat.**

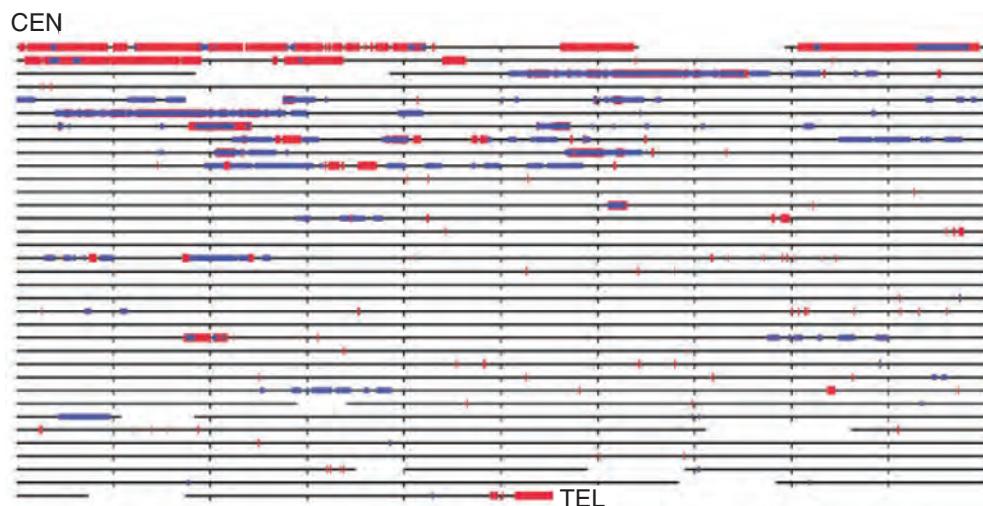
Length of repeat	Average bases per megabase	Average number of SSR elements per megabase
1	1660	33.7
2	5046	43.1
3	1013	11.8
4	3383	32.5
5	2686	17.6
6	1376	15.2
7	906	8.4
8	1139	11.1
9	900	8.6
10	1576	8.6
11	770	8.7

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

events, less conservation is expected between duplicated intronic regions.) Segmental duplications may be interchromosomal or intrachromosomal. The centromeres contain large amounts of interchromosomal duplicated segments, with almost 90% of a 1.5 Mb region containing these repeats (Fig. 20.11). Smaller regions of these repeats also occur near the telomeres.

### Gene Content of Human Genome

It is of great interest to characterize the gene content of the human genome because of the critical role of genes in human biology. However, the genes are the hardest



**FIGURE 20.11** The centromeres consist of large amounts of interchromosomal duplicated segments. The size and location of intrachromosomal (black) and interchromosomal (red) segmental duplications are indicated. Each horizontal line represents 1 Mb of chromosome 22q; the tick marks indicate 100 kb intervals. The centromere is at top left, and the telomere is at the lower right. Adapted from IHGSC (2001). Reproduced with permission from Macmillan Publishers.

features of genomic DNA to identify (see Chapter 8). This is a challenging task for many reasons:

- The average exon is only 50 codons (150 nucleotides). Such small elements are hard to unambiguously identify as exons.
- Exons are interrupted by introns, some many kilobases in length. In the extreme case, the human dystrophin gene extends over 2.4 Mb, the size of an entire genome of a typical bacterial genome. The use of complementary DNAs and RNA-seq therefore continues to provide an essential approach to gene identification.
- There are many pseudogenes that may be difficult to distinguish from functional protein-coding genes.
- The nature of noncoding genes is poorly understood (see Chapter 10 and the following section).

We described cDNA projects in Chapter 10.

### Noncoding RNAs

There are many classes of human genes that do not encode proteins. Noncoding RNAs can be difficult to identify in genomic DNA because they lack open reading frames, they may be small, and they are not polyadenylated (they are therefore not enriched by oligo(dT) capture methods used to purify mRNA). They are difficult to detect by gene-finding algorithms, and they are not present in cDNA libraries. These noncoding RNAs include the following:

- transfer RNAs, required as adapters to translate mRNA into the amino acid sequence of proteins;
- ribosomal RNAs, required for mRNA translation;
- small nucleolar RNAs (snoRNAs), required for RNA processing in the nucleolus; and
- small nuclear RNAs (snRNAs), required for spliceosome function.

Hundreds of noncoding RNAs were identified in the draft version of the human genome (**Table 20.7**). The tRNA genes were most predominant, with 497 such genes and an additional 324 tRNA-derived pseudogenes. The tRNA genes associated with the human genetic code can now be described. This version of the genetic code includes the frequency of codon utilization for each amino acid and the number of tRNA genes that are associated with each codon. The total number of tRNA genes is comparable to that observed in other eukaryotes (**Table 10.2**).

### Protein-Coding Genes

Protein-coding genes are characterized by exons, introns, and regulatory elements. These basic features are summarized in **Table 20.8**. The average coding sequence for human genes is 1340 bp (IHGSC, 2001), comparable to the size of an average coding sequence in nematode (1311 bp) and *Drosophila* (1497 bp). Most internal exons are about 50–200 bp in length in all three species (**Fig. 20.12a**), although worm and fly have a greater proportion of longer exons (note the flatter tail in **Fig. 20.12a**). However, the size of human introns is far more variable (**Fig. 20.12b, c**). This results in a more variable overall gene size in humans than in worm and fly.

Protein-coding genes are associated with a high GC content (**Fig. 20.13**). While the overall GC content of the human genome is about 41%, the GC content of known genes (having RefSeq identifiers) is higher (**Fig. 20.13a**). Gene density increases 10-fold as GC content rises from 30 to 50% (**Fig. 20.13b**).

Currently Ensembl lists 20,300 protein-coding genes (**Table 20.1**). The 10 most common InterPro hits include immunoglobulin domains and protein kinases (**Table 20.9**).

### Comparative Proteome Analysis

The importance of comparative analyses has emerged as one of the fundamental tenets of genomics. The IHGSC (2001) analyzed functional groups of these proteins based on InterPro and Gene Ontology (GO) Consortium classifications. Humans have relatively

The longest coding sequence is titin (104,301 bp; NM\_001256850.1). The gene for titin, on chromosome 2q24.3, has 178 exons and encodes a muscle protein of 34,350 amino acids (about 3.8 million Da). By contrast, a typical protein encoded by an mRNA of 1340 bp is about 50,000 Da.

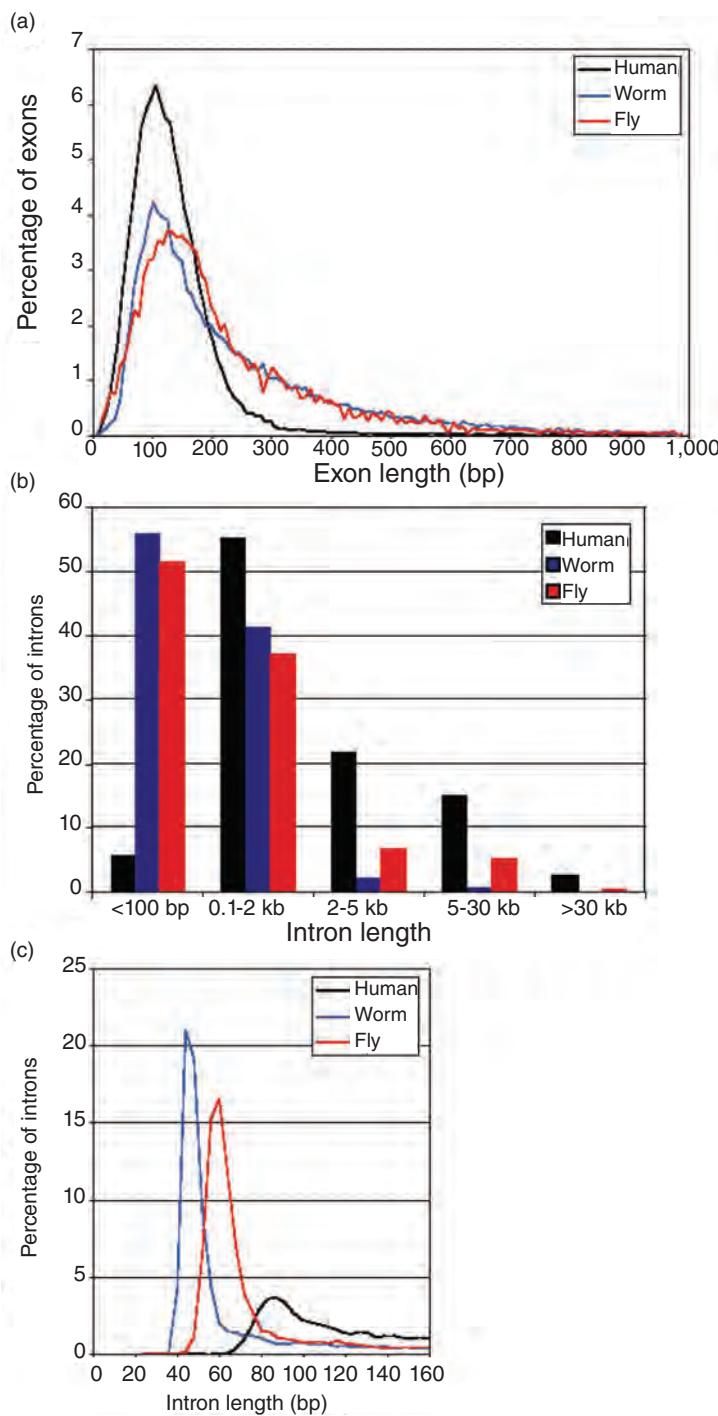
Chromosome 19, the most GC-rich chromosome, also houses the greatest density of genes (26.8 per megabase). The average density of gene predictions across the genome is 11.1 per megabase. The Y chromosome is least dense, having 6.4 predicted genes per megabase.

**TABLE 20.7 Noncoding genes in human genome. Adapted from IHGSC (2001). Reproduced with permission from Macmillan Publishers.**

RNA gene	Number of noncoding genes	Number of related genes	Function
tRNA	497	324	Protein synthesis
SSU (18S) RNA	0	40	Protein synthesis
5.8S rRNA	1	11	Protein synthesis
LSU (28S) rRNA	0	181	Protein synthesis
5S RNA	4	520	Protein synthesis
U1	16	134	Spliceosome component
U2	6	94	Spliceosome component
U4	4	87	Spliceosome component
U4atac	1	20	Minor (U11/U12) spliceosome component
U5	1	31	Spliceosome component
U6	44	1135	Spliceosome component
U6atac	4	32	Minor (U11/U12) spliceosome component
U7	1	3	Histone mRNA 3' processing
U11	0	6	Minor (U11/U12) spliceosome component
U12	1	0	Minor (U11/U12) spliceosome component
SRP (7SL) RNA	3	773	Component of signal recognition particle
RNase P	1	2	tRNA 5' end processing
RNase MRP	1	6	rRNA processing
Telomerase RNA	1	4	Template for addition of telomeres
hY1	1	353	Component of Ro RNP, function unknown
hY3	25	414	Component of Ro RNP, function unknown
hY4	3	115	Component of Ro RNP, function unknown
hY5 (4.5S RNA)	1	9	Component of Ro RNP, function unknown
Vault RNAs	3	1	Component of 13 Mda vault RNP
7SK	1	330	Unknown
H19	1	2	Unknown
Xist	1	0	Initiation of X chromosome inactivation
Known C/D snoRNAs	69	558	Pre-rRNA processing or site-specific ribose methylation of rRNA
Known H/ACA snoRNAs	15	87	Pre-rRNA processing or site-specific pseudouridylation of rRNA

**TABLE 20.8 Characteristics of human genes. aa: amino acids; bp: base pairs; kb: kilo base pairs. Adapted from IHGSC (2001). Reproduced with permission from Macmillan Publishers.**

Feature	Size (median)	Size (mean)
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1023 bp	3365 bp
3' untranslated region	400 bp	770 bp
5' untranslated region	240 bp	300 bp
Coding sequence	1100 bp	1340 bp
Coding sequence	367 aa	447 aa
Genomic extent	14 kb	27 kb



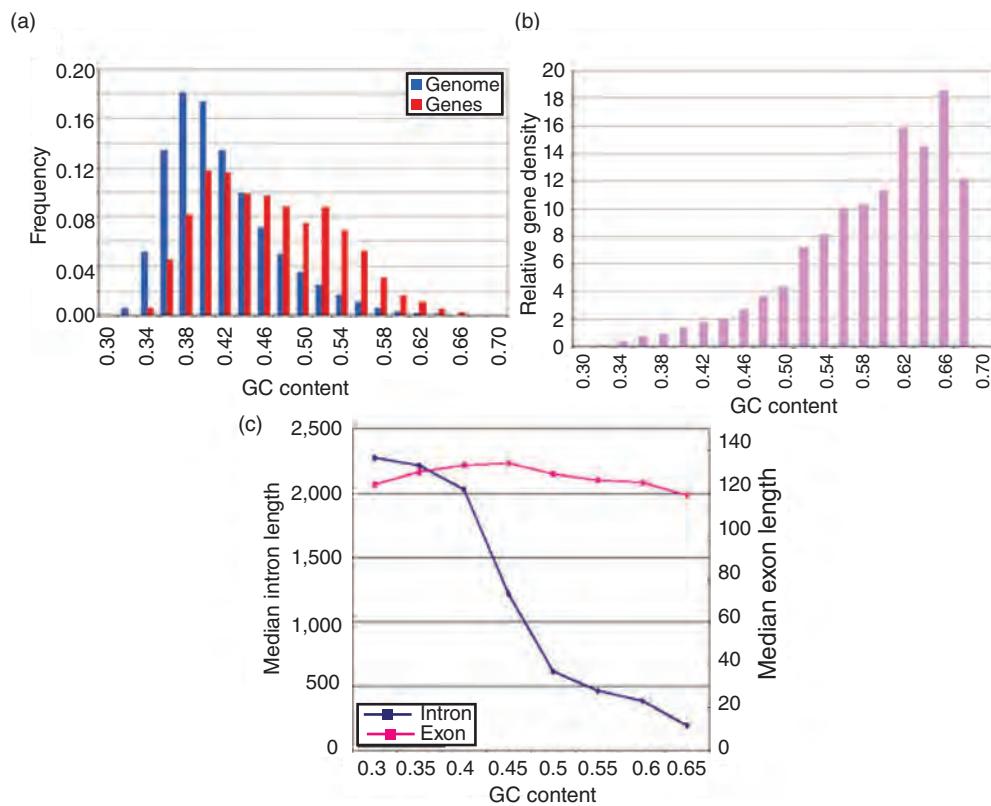
**FIGURE 20.12** Size distribution of (a) exons, (b) introns, and (c) short introns (enlarged from (b)) in human, worm, and fly.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

more genes that encode proteins predicted to function in cytoskeleton, transcription/translation, and defense and immunity.

The human proteome was further studied by BLASTP searching every predicted protein against the nonredundant database. Overall, 74% of the proteins were significantly related to other known proteins. As more sequences are accumulated in databases over time, the matches between human proteins and other eukaryotes (and bacteria and archaea) continue to increase.

We discussed the GO Consortium and InterPro in Chapter 12.



**FIGURE 20.13** (a) Distribution of GC content in genes and in the genome shows that protein-coding genes are associated with a higher GC content. (b) The gene density (the ratio of the values in (a)) is plotted as a function of the GC content. As GC content rises, the relative gene density increases dramatically. (c) Mean exon length is unaffected by GC content, but introns are far shorter as GC content rises.

Source: IHGSC (2001). Reproduced with permission from Macmillan Publishers.

### Complexity of Human Proteome

The number of protein-coding genes in humans is comparable to the number of genes in other metazoans and plants and only three-fold greater than the number in unicellular fungi. Nonetheless, the human proteome may be far more complex for several reasons (IHGSC, 2001):

**TABLE 20.9** Ten most common InterPro hits for *Homo sapiens*.

InterPro	InterPro name	Number of genes
IPR007110	Immunoglobulin-like domain	7199
IPR027417	P-loop containing nucleoside triphosphate hydrolase	3901
IPR011009	Protein kinase-like domain	2543
IPR015880	Zinc finger, C2H2-like	2500
IPR007087	Zinc finger, C2H2	2414
IPR000719	Protein kinase domain	2283
IPR003599	Immunoglobulin subtype	1645
IPR017452	GPCR, rhodopsin-like, 7TM	1631
IPR000276	G protein-coupled receptor, rhodopsin-like	1567
IPR001909	Krueppel-associated box	1519

Source: Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

**TABLE 20.10 Human paralogous genes with largest cluster sizes involved in recent gene duplications ( $K_s \leq 0.3$ ).**  
Modified from IHGSC (2004) with permission from Macmillan Publishers.

Cluster size	Minimum size in ancestral genome	Genes involved in recent duplications	Chromosome	Gene family
64	50	23	11	Olfactory receptor
59	54	10	11	Olfactory receptor
34	25	13	1	Olfactory receptor
30	8	26	2	Immunoglobulin K chain V
23	5	19	19	KRAB zinc-finger protein
23	19	6	11	Olfactory receptor
21	9	15	14	Immunoglobulin heavy chain
20	11	12	22	Immunoglobulin $\lambda$ chain V-region
18	9	13	19	Leukocyte and NK cell immunoglobulin-like receptors
18	14	6	19	Gonadotropin-inducible transcription repressor-2-like
16	4	13	9	Interferon $\alpha$
16	10	7	19	FDZF2-like KRAB zinc-finger protein
14	8	7	12	Taste receptor, type 2
13	3	11	1	PRAME/MAPE family (cancer/germ line antigen)
13	9	8	17	Olfactory receptor
11	2	11	16	Immunoglobulin heavy chain
10	1	10	19	Pregnancy-specific $\beta$ -1-glycoprotein

- There are relatively more domains and protein families in humans than in other organisms.
- The human genome encodes relatively more paralogs, potentially yielding more functional diversity.
- There are relatively more multidomain proteins having multiple functions.
- Domain architectures tend to be more complex in the human proteome.
- Alternative RNA splicing may be more extensive in humans.

There may be a synergistic effect among these factors, leading to a substantially greater complexity of the human proteome that could account for the phenotypic complexity of vertebrates, including humans.

In its reannotation of the human genome, the IHGSC (2004) identified the largest clusters of human paralogous genes that involve recent gene duplications (Table 20.10); these genes are neighboring (indicating local gene duplication). The selected sites displayed near neutrality (estimated substitution rate per synonymous site  $K_s < 0.30$ , such that each homolog differs from a common ancestral gene by an average  $K_s < 0.15$ ). These represent genes that were recently born in the human lineage (after the divergence from rodents), and many have functions in olfaction, immune function, and the reproductive system.

## 25 HUMAN CHROMOSOMES

Each human chromosome was finished (or nearly finished) by a dedicated research team. For each chromosome, there is a publication in the journal *Nature* (or *Science*). There are seven traditional cytogenetic groups A–F which categorize the chromosomes (other than the mitochondrial genome) according to morphological properties (Table 20.11). We briefly summarize key aspects of each chromosome, following this organization (Tables 20.12–20.18).

A list of accession numbers for the human chromosomes is given at [ftp://ftp.ncbi.nlm.nih.gov/genomes/H\\_sapiens/Assembled\\_chromosomes/chr\\_NC\\_gi](ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/Assembled_chromosomes/chr_NC_gi) (WebLink 20.26).

**TABLE 20.11 Human chromosome groups.**

Group	Chromosomes	Description
A	1–3	Largest chromosomes; 1, 3 are metacentric; 2 is submetacentric
B	4, 5	Large chromosomes; submetacentric
C	6–12, X	Medium-size chromosomes; submetacentric
D	13–15	Medium-size chromosomes; acrocentric with satellites
E	16–18	Small; 16 is metacentric; 17, 18 are submetacentric
F	19, 20	Small, metacentric chromosomes
G	21, 22, Y	Smallest chromosomes; acrocentric; satellites on 21 and 22

**TABLE 20.12 Group A chromosomes. Length is from NCBI build 37; gap sizes are from GRCh37; and chromosome length is from NCBI. Adapted from Hillier *et al.* (2005), Gregory *et al.* (2006), Muzny *et al.* (2006), NCBI build 37 (February 2014), GRCh37, [http://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.26/#/st](http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/#/st).**

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
1	249	3141	991	24.0	NC_00001.10
2	243	1346	1239	5.0	NC_00002.11
3	198	1463	122	3.2	NC_00003.11

**TABLE 20.13 Group B chromosomes. Adapted from Schmutz *et al.* (2004); Hillier *et al.* (2005). Length is from NCBI build 37; gap sizes are from GRCh37.**

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
4	191	796	778	3.5	NC_00004.11
5	181	923	577	3.2	NC_00005.9

**TABLE 20.14 Group C chromosomes. Adapted from Hillier *et al.* (2003), Mungall *et al.* (2003), Deloukas *et al.* (2004), Humphray *et al.* (2004), Ross *et al.* (2005), Nusbaum *et al.* (2006), Taylor *et al.* (2006). Length is from NCBI build 37; gap sizes are from GRCh37.**

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
6	171	1557	633	3.7	NC_00006.11
7	159	1150	941	3.8	NC_00007.13
8	146	793	301	3.5	NC_00008.10
9	141	1149	426	21.1	NC_00009.11
10	136	816	430	4.2	NC_00010.10
11	135	1524	765	3.9	NC_00011.9
12	134	1342	93	3.4	NC_00012.11
X	155	1098	700	4.1	NC_00023.10

**TABLE 20.15 Group D chromosomes. Adapted from Heilig *et al.* (2003), Dunham *et al.* (2004), Zody *et al.* (2006b). Length is from NCBI build 37; gap sizes are from GRCh37.**

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
13	115	633	296	19.6	NC_00013.10
14	107	1050	393	19.1	NC_00014.8
15	103	695	250	20.8	NC_00015.9

**TABLE 20.16 Group E chromosomes.** Adapted from Martin *et al.* (2004), Nusbaum *et al.* (2005), Zody *et al.* (2006a). Length is from NCBI build 37; gap sizes are from GRCh37.

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
16	90	796	778	11.5	NC_000016.9
17	81	1266	274	3.4	NC_000017.10
18	78	337	171	3.4	NC_000018.9

**TABLE 20.17 Group F chromosomes.** Adapted from Deloukas *et al.* (2001), Grimwood *et al.* (2004). Length is from NCBI build 37; gap sizes are from GRCh37.

Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
19	59	1461	321	3.3	NC_000019.9
20	63	727	168	3.5	NC_000020.10

**TABLE 20.18 Group G chromosomes.** Adapted from Dunham *et al.* (1999), Hattori *et al.* (2000), Skaletsky *et al.* (2003). Length is from NCBI build 37; gap sizes are from GRCh37.

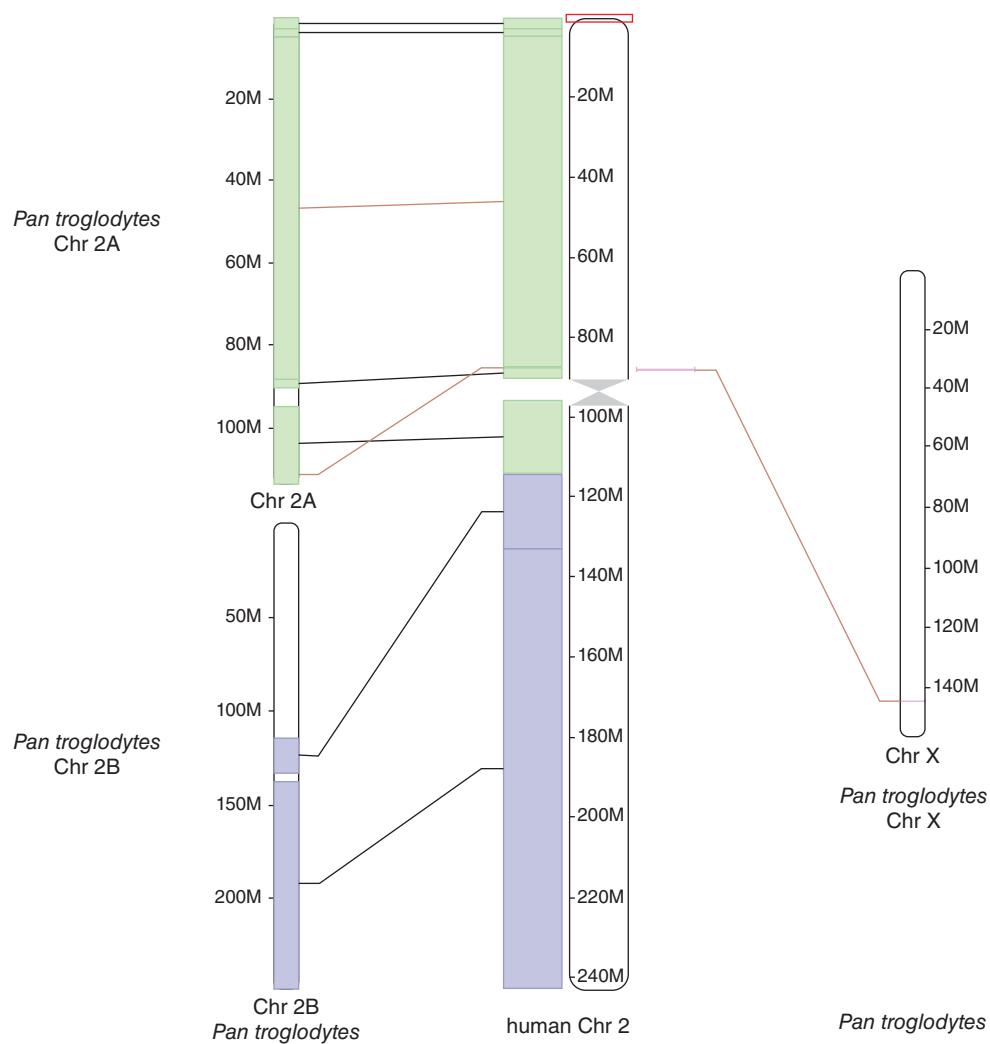
Chromosome	Length (Mb)	# Genes	# Pseudogenes	Gap size (Mb)	Accession
21	48	796	778	13.0	NC_000021.8
22	51	545	134	16.4	NC_000022.10
Y	59	78	n/a	33.7	NC_000024.9

The exact number of genes is not yet known (almost all have been annotated). The EGASP competition, described in Chapter 8, highlights the computational challenges in correctly identifying genes with good sensitivity and specificity. The values of gap lengths in **Tables 20.12–20.18** typically decrease over time. In almost every case they represent regions that are refractory to cloning and sequencing because of the highly repetitive nature of the underlying DNA sequence, even when up to 100-fold coverage of the chromosome is obtained. Overall, the finished euchromatic portion of the human genome included 250 gaps spanning 25 Mb, while the heterochromatic portion had far fewer gaps (just 33) spanning a vast size (200 Mb; IHGSC, 2004). In 2015 the total gap size (for GRCh38.p2) was 160 Mb.

### Group A (Chromosomes 1–3)

Chromosome 1, the largest chromosome, has 3141 genes and 991 pseudogenes (Gregory *et al.*, 2006; **Table 20.12**). Its gene density (14.2 genes per megabase) is nearly twice the genome-wide average (7.8 genes per megabase). Typical for essentially all the chromosome finishing projects, sequence integrity and completeness were assessed three ways: (1) by determining whether all RefSeq genes assigned to the chromosome were accounted for; (2) by comparing the order of hundreds of chromosome markers to the DeCode genetic map to search for discrepancies; and (3) by aligning over 32,000 pairs of fosmid end sequences to unique positions in the sequence. This resulted in the identification of several misassemblies caused by low-copy repeats. In some cases, naturally occurring polymorphisms confound the analysis; for example, 50% of individuals lack the *GSTM1* gene.

Chromosome 2, the second largest chromosome, is remarkable because it corresponds to two intermediate-sized ancestral, acrocentric chromosomes that fused end-to-end. In



**FIGURE 20.14** Conserved synteny between human chromosome 2 and two smaller chimpanzee chromosomes provides evidence that two ancestral human acrocentric chromosomes fused. This image is from the Ensembl synteny viewer ([http://www.ensembl.org/Homo\\_sapiens/Location/Genome, WebLink 20.53](http://www.ensembl.org/Homo_sapiens/Location/Genome,WebLink 20.53)).

Source: Ensembl Release 75; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

other primates these chromosomes remain separate, as in the case of chimpanzee chromosomes 2A and 2B (Fig. 20.14). In its finished sequence, the fusion site is in 2q13-2q14.1 (Hillier *et al.*, 2005). One of the two centromeres (at 2q21) became inactivated, and contains  $\alpha$ -satellite remnants.

Although chromosome 3 is large, it contains the lowest rate of segmental duplications in the genome (1.7% compared to a genome-wide average of 5.3% of nucleotides segmentally duplicated; Muzny *et al.*, 2006). Chromosomes 3 and 21 derive from a larger ancestral chromosome that split. It also includes a large pericentric inversion (also present in chimpanzee and gorilla, but not orang-utan or Old World monkeys).

### Group B (Chromosomes 4, 5)

Chromosome 4 has an unusually low GC content of 38.2%, compared to the genome-wide average of 41% (Hillier *et al.*, 2005; Table 20.13). Over 19% of the chromosome has a GC content of <35%. However, portions of the chromosome have a GC content >70%.

You can view these using the UCSC Genome Browser's GC content annotation track, or the Table Browser.

Chromosome 5 has both a very low gene density and a very high rate of intrachromosomal duplications (Schmutz *et al.*, 2004). It includes 923 gene loci, and 577 pseudogenes. There are many gene-poor loci that are highly conserved and are therefore thought to be functionally constrained.

### Group C (Chromosomes 6–12, X)

The largest transfer RNA gene cluster is localized to chromosome 6p, with 157 tRNA genes out of 616 across the entire genome (Mungall *et al.*, 2003). Chromosome 6 (**Table 20.14**) also contains HLA-B, the most polymorphic gene in the human genome. We explore this polymorphism further in computer laboratory exercise (20.6) at the end of this chapter.

Chromosome 7 was sequenced by the public consortium (Hillier *et al.*, 2003) and by Scherer *et al.* (2003) using a mixture of Celera whole-genome scaffolds and International Human Genome Sequencing Consortium data. The centromere is polymorphic with a range of 1.5–3.8 Mb at one locus (marker D7Z1) and 100–500 kb at another site (D7Z2). There is an unusually large amount of segmentally duplicated sequence (8.2%). As an example of the consequence of this, Williams–Beuren syndrome results from the hemizygous deletion of 1.5 million base pairs on chromosome 7q11.23, a region containing about 17 genes. There are flanking repeats that mediate unequal meiotic recombination (**Fig. 8.19**) or, in some cases, hemizygous inversions (Osborne *et al.*, 2001).

Other group C chromosomes are 8 (Nusbaum *et al.*, 2006), 9 (Humphray *et al.*, 2004), 10 (Deloukas *et al.*, 2004), 11 (Taylor *et al.*, 2006), 12 (Scherer *et al.*, 2006), and X (Ross *et al.*, 2005). Chromosome 9 contains the largest autosomal block of heterochromatin. Chromosome 11 is notable for having the beta globin gene cluster as well as the insulin gene.

The X chromosome joins group C chromosome because of its comparable size. It is unique in many ways. Mammals are classified into three groups, in all of which males have X and Y chromosomes: the eutherians (placental mammals); the metatheria (marsupials); and the prototheria (egg-laying mammals). Females undergo X chromosome inactivation (XCI) in which one copy is silenced early in development. In contrast to the autosomes, the male X chromosome does not recombine during meiosis, except for short pseudoautosomal regions at the tips (PAR1 on Xp and PAR2 on Xq) that recombine with corresponding portions of the Y chromosome. Since males have only a single copy of the X chromosome (it is therefore hemizygous), recessive phenotypes are exposed and many X-linked diseases have been described from hemophilia to X-linked intellectual disability syndromes. The X and Y chromosomes derive from an ancient autosomal chromosome pair that began transforming into sex chromosomes over 300 million years ago, and sequencing of the X (and Y) chromosomes has revealed traces of evolutionary conservation between the two (Ross *et al.*, 2005 and see “Group G (Chromosomes 21, 22, Y)” below).

### Group D (Chromosomes 13–15)

The five human acrocentric chromosomes are 13, 14, and 15 (**Table 20.15**) as well as 21 and 22. For each, the p arm is almost entirely heterochromatic. These regions have a highly repetitive structure, and all five include arrays of ribosomal DNA genes as shown in **Figure 10.7**. Sequencing and accurately assembling these regions is so challenging that they were not targeted by the Human Genome Project and are still not part of the standard human genome assemblies.

### Group E (Chromosomes 16–18)

Of this group of chromosomes, 16 and 17 are notable for above-average levels of segmental duplication (**Table 20.16**). Chromosome 18 has the lowest gene density of any autosome (4.4 genes per megabase) and encodes only 337 genes (about one-quarter of the number of the similar-sized chromosome 17). One region of chromosome 18 has only 3 genes across 4.5 Mb. The sparse number of genes may partly explain why some individuals with trisomy 18 (Edwards syndrome) survive to birth, while all other autosomal trisomies (except trisomy 13 and trisomy 21) are embryonic lethal.

### Group F (Chromosomes 19, 20)

Chromosome 19 has the highest gene density with 26 protein-coding genes per megabase (**Table 20.17**). It also has an unusually high density of repeats (55% of the chromosome, in contrast to a genome-wide average of about 45%). Almost 26% of the chromosome is composed of *Alu* repeats, consistent with the high gene density.

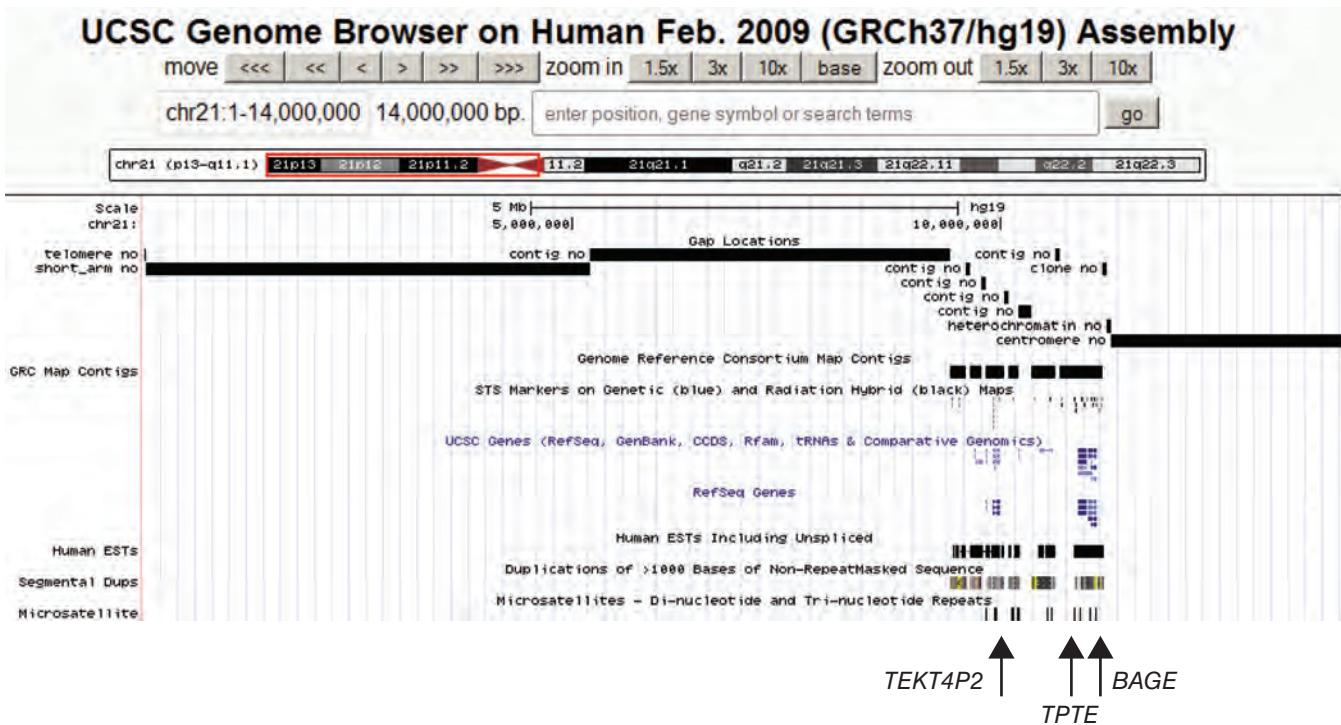
### Group G (Chromosomes 21, 22, Y)

Group G chromosomes are the smallest (**Table 20.18**). While the short arms of the five acrocentric chromosomes are nearly entirely heterochromatic, an exception is 21p11.2 which includes a very small euchromatic region. A view of 14 megabases extending across the p arm of chromosome 21 and its centromere highlights how little annotated information is currently available (**Fig. 20.15**). Only two protein-coding genes are annotated there: *TPTE* (transmembrane phosphatase with tensin homology) and *BAGE* (B melanoma antigen) as well as *TEKT4P2* (a pseudogene). The other acrocentric arms have no protein-coding genes annotated.

The Y chromosome was the most technically difficult to sequence because of its extraordinarily repetitive nature (Skaletsky *et al.*, 2003). It has short pseudoautosomal regions at the ends that recombine with the X chromosome. A large central region, spanning 95% of its length, is termed the male-specific region (MSY). There are 23 Mb of eukaryotic DNA including 8 Mb on Yp and 14.5 Mb on Yq. There are three notable heterochromatic regions: (1) a centromeric region of about 1 Mb; (2) a block of ~40 Mb on the long arm; and (3) an island of 400 kb comprising over 3000 tandem repeats of 125 base pairs. Of 156 transcription units, about half encode proteins. Skaletsky *et al.* defined three classes of euchromatic sequences:

1. X-transposed sequences total 3.4 Mb and share 99% identity to Xq21 DNA sequences. Just 3–4 million years ago, after the human–chimpanzee divergence, there was a massive transposition of X chromosome sequences to the Y chromosome, followed by an inversion that dispersed these sequences on the Y.
2. X-degenerate sequences share 60–90% identity to 27 different X chromosome genes, and represent relics of the ancient autosomes from which X and Y evolved.
3. Ampliconic sequences span over 10 Mb and consist of blocks of sequences sharing as much as 99.9% nucleotide identity over spans of tens or hundreds of kilobases. The amplicons are the most gene-dense regions of the Y chromosome, and have a low content of interspersed repeats. The ampliconic regions contain eight giant palindromes, collectively spanning 5.7 Mb, each with two long arms interrupted by a unique, central spacer.

The extraordinary conservation of the palindromic arms is due to gene conversion, the nonreciprocal transfer of sequences from one DNA duplex to another (Rozen *et al.*, 2003; Skaletsky *et al.*, 2003).



**FIGURE 20.15** View of the p arm of the acrocentric chromosome 21. A region of 14 million base pairs is shown, extending across the centromere. It is notable that essentially no features are annotated, other than those in a small euchromatic region containing two protein-coding genes (*TPTE* and *BAGE*) as well as a pseudogene (*TEKT4P2*) and two microRNAs. The p arm is filled with ribosomal DNA genes but these are difficult to sequence, and highly similar across the acrocentric chromosomes and between adjacent clusters (discussed in Chapter 10). Note the lack of data across most of 21p for various tracks such as contigs, sequence tag site (STS) markers, UCSC and RefSeq genes, expressed sequence tags (ESTs), duplications, and microsatellites. In contrast, gap locations are assigned across most of the chromosome arm.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

## Mitochondrial Genome

In addition to 22 autosomes and two sex chromosomes, humans have a mitochondrial genome. Mitochondrial genomes have a number of fascinating properties that also make them useful for phylogenetic studies (reviewed in Pakendorf and Stoneking, 2005). They are present in high copy number, typically with hundreds or even thousands of genomes per cell. They are maternally inherited; all (or almost all) sperm-derived mitochondria are targeted for destruction in the fertilized oocyte. One consequence is that molecular phylogenetic studies of mitochondria follow the history of the maternal lineage, and have therefore been traced to a “mitochondrial Eve” or proposed earliest human female ancestor. Another consequence of maternal inheritance is that mitochondrial DNA does not undergo recombination. The mutation rate is higher than in nuclear DNA, providing a useful signal for molecular phylogenetic studies. Excluding the D-loop (which has not evolved at a constant rate across human lineages), Ingman *et al.* (2000) estimated the mitochondrial mutation rate to be  $1.70 \times 10^{-8}$  substitutions per site per year (although this rate is higher in hypervariable regions).

While the mitochondrial genome does not undergo recombination, it is polymorphic. There are 18 known mitochondrial haplogroups or lineages. For the HapMap project (described in “Human Genome Variation” below), subjects of various geographic origins were assigned to 15 of these known groups (Table 20.19).

The reference genome, called the Revised Cambridge Reference Sequence, is 16,569 base pairs in a circular genome obtained from a Yoruba individual (from Ibadan, Nigeria).

**TABLE 20.19 mtDNA haplogroups. YRI: Yoruba in Ibadan, Nigeria; CEU: Utah residents with ancestry from northern and western Europe; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan.**

MtDNA haplogroup	DNA sample (number of chromosomes)			
	YRI (60)	CEU (60)	CHB (45)	JPT (44)
L1	0.22	–	–	–
L2	0.35	–	–	–
L3	0.43	–	–	–
A	–	–	0.13	0.04
B	–	–	0.33	0.30
C	–	–	0.09	0.07
D	–	–	0.22	0.34
M/E	–	–	0.22	0.25
H	–	0.45	–	–
V	–	0.07	–	–
J	–	0.08	–	–
T	–	0.12	–	–
K	–	0.03	–	–
U	–	0.23	–	–
W	–	0.02	–	–

*Source:* International HapMap Consortium (2005). Reproduced with permission from Macmillan Publishers.

The RefSeq human mitochondrial accession number is NC\_012920.1. The MitoMap website includes a reanalysis of the Cambridge reference sequence based on resequencing efforts. See <http://www.mitomap.org/MITOMAP/CambridgeReanalysis> (WebLink 20.27).

The GC content is 44.5%, higher than for the other human chromosomes. The genome includes 37 annotated genes, spanning 68% of the genome. These include 13 protein-coding genes (encoding proteins involved in oxidative phosphorylation) and 24 structural RNAs (two ribosomal RNAs and 22 transfer RNAs; see Chapter 10). A region of about 1100 base pairs called the control region has regulatory functions.

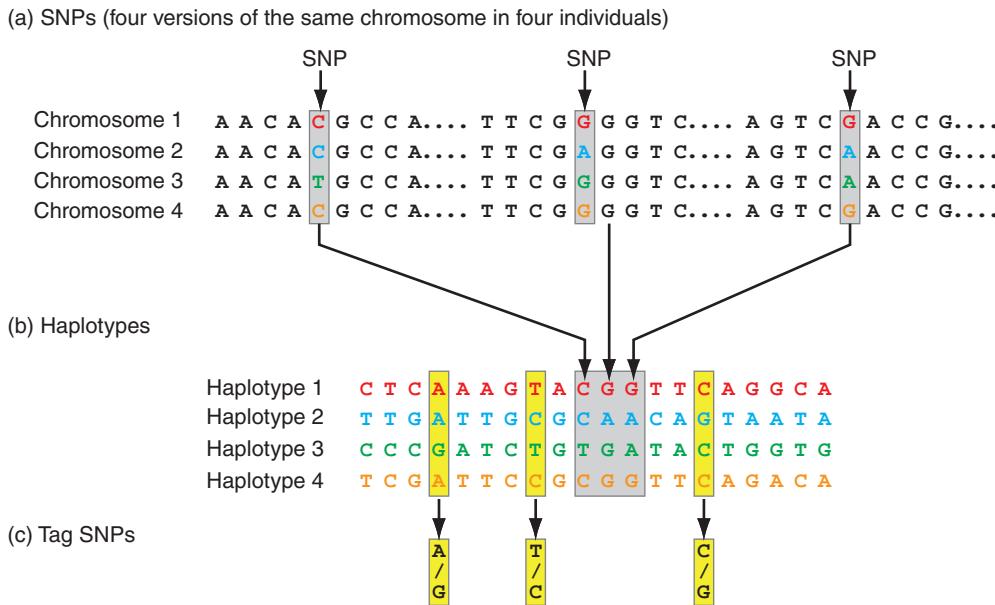
Behar *et al.* (2012) analyzed >18,000 human mitochondrial sequences and performed phylogenetic analyses. They proposed using a new Reconstructed Sapiens Reference Sequence that includes references to *Homo neanderthalensis* and catalogs changes relative to an ancestral reference sequence.

## HUMAN GENOME VARIATION

We conclude this chapter by considering several aspects of variation in the human genome: single-nucleotide polymorphisms (SNPs) and the International HapMap project; the 1000 Genomes Project; and the sequencing of individual human genomes. A goal of the Human Genome Project was to define a consensus human genome sequence with a focus on the >>99% nucleotide identity we all share. A goal of HapMap and the 1000 Genomes Projects has been complementary, seeking to define the <<1% of difference that characterizes each of our individual genomes, including both common and rare variants.

### SNPs, Haplotypes, and HapMap

SNPs represent a fundamental form of variation in the human population. The International HapMap Project began in 2002 and reported the genotypes of 1.3 million SNPs in four geographically diverse populations (International HapMap Consortium, 2003, 2005): (1) 30 trios (consisting of mother, father, and an adult child) from the Yoruba tribe in Ibadan, Nigeria, abbreviated YRI; (2) 30 trios of northern and western European



**FIGURE 20.16** Single-nucleotide polymorphisms (SNPs), haplotypes, and tag SNPs. (a) A SNP is a difference between chromosomes occurring at a particular site in genomic DNA. Four versions of the same chromosomal region are shown (from different individuals), and three SNPs are indicated (arrows). Each SNP has two alleles (assuming it is biallelic); for the first SNP the alleles are C and T. (b) A haplotype is composed of a particular combination of neighboring SNPs. The observed genotypes are shown for 20 SNPs, all of which are variable bases occurring in a region of 6000 bases of DNA. Each row corresponds to a different haplotype. (c) Three tag SNPs are indicated. By genotyping just these three SNPs (rather than genotyping all 20 SNPs or sequencing all 6000 bases of DNA), it is possible to uniquely identify the four haplotypes in the region. Using tag SNPs is possible because SNP alleles are co-inherited, leading to associations called linkage disequilibrium (LD). Redrawn from International HapMap Consortium, 2003. Reproduced with permission from Macmillan Publishers.

ancestry living in Utah and obtained from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (abbreviated CEU); (3) 45 unrelated Han Chinese individuals in Beijing, China (CHB); and (4) 45 unrelated Japanese individuals in Tokyo, Japan (abbreviated JPT). In some studies, data from the Chinese and Japanese populations are pooled to yield three groups of 90 (YRI, CEU, CHB+JPT). A second generation haplotype map increased the number of characterized SNPs to 3.1 million (International HapMap Consortium *et al.*, 2007). The third generation of HapMap extended genotyping to more individuals (1184 reference individuals) from 11 global populations, and also sequenced a series of ten 100 kilobase regions from almost 700 of these individuals (International HapMap 3 Consortium *et al.*, 2010).

The sequencing of whole genomes confirms that each person has about 3.5 million or more SNPs. Each SNP corresponds to a specific nucleotide position having two alleles (in the case of biallelic SNPs; some SNPs are triallelic or even tetraallelic). Figure 20.16a shows a DNA region having three biallelic SNPs (arrows) that occur in various combinations along individual chromosomes (rows). For each SNP we can define at least three properties.

- We can define the sequence (e.g., the first SNP is either a C allele or a T allele).
- We can calculate the major allele frequency as well as the minor allele frequency (abbreviated MAF) in a given population. Common SNPs have a MAF >5%.
- The copy number of SNPs can be determined, allowing an assessment of deletions (copy number <2) or amplifications (copy number >2).

HaploView software can be downloaded from the lab of Mark Daly at the Broad Institute (<http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haplovie/haplovie>, WebLink 20.28). It is a Java-based program that can be conveniently run on a Mac or PC.

The HapMap website is <http://www.hapmap.org> (WebLink 20.29). The site features a browser and several options for data downloads. HapMap samples are also available as DNA aliquots or cell lines from the Coriell Cell Repositories (<http://ccr.coriell.org/>, WebLink 20.30). Accession numbers beginning with NA refer to genomic DNA samples, while GM accession numbers refer to cell lines.

Integrative Genomics Viewer (version 2.3) is available from <http://www.broadinstitute.org/software/igv> (WebLink 20.31, described in Chapter 9). After registration, it is accessible as a Java application. HapMap VCF files are available from NCBI; visit [ftp://ftp.ncbi.nih.gov/snp/organisms/human\\_9606/VCF/](ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/) (WebLink 20.32) for a directory. We specifically used both [ftp://ftp.ncbi.nih.gov/snp/organisms/human\\_9606/VCF/clinvar\\_20140211.vcf.gz](ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/clinvar_20140211.vcf.gz) (WebLink 20.33) and its index file [ftp://ftp.ncbi.nih.gov/snp/organisms/human\\_9606/VCF/clinvar\\_20140211.vcf.gz.tbi](ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/clinvar_20140211.vcf.gz.tbi) (WebLink 20.34). Another folder at the above NCBI FTP site includes VCF files organized by chromosome; we downloaded chromosome 11 VCF files and index files from the three HapMap populations CHB, MKK, and CEU.

Another key aspect of SNPs is that we can define their relationships to neighboring SNPs. **Figure 20.16b** shows a set of 20 variable positions (SNPs) that occur in the middle of a stretch of 6000 base pairs of DNA (most of which are invariant between individuals, reflecting the extremely high nucleotide identity shared by all humans). A haplotype is a specific combination of alleles that occur in neighboring SNPs. The HapMap project was designed to create a map of haplotypes occurring in human populations. SNPs are tightly linked to each other and form blocks in which the behavior of one SNP can serve as a proxy for the genotypes of neighboring SNPs. Such related blocks have linkage disequilibrium (LD), which is the association of co-inherited alleles in the population. Commonly used measures of LD include  $D'$ ,  $r^2$ , and LOD.  $D'$  has a value of 1 in the absence of historical recombination, and is <1 when recombination or recurrent mutation occur.  $r^2$  is the squared correlation coefficient between two SNPs, having a value of 1 when they share an evolutionary haplotype and are not disrupted by recombination. LOD is a logarithm of an odds score.

A subset of SNPs, called “tag SNPs,” may uniquely identify a larger haplotype block (Fig. 20.16c). Such tag SNPs can discriminate between possible haplotypes. This can be useful in practice because it is cost-effective to genotype just a subset of SNPs in a region.

## Viewing and Analyzing SNPs and Haplotypes

HapMap data can be viewed, downloaded, and analyzed at the Ensembl, NCBI, and UCSC websites, as well as at the HapMap website. Many software tools are useful to analyze SNPs. We describe six approaches in the following sections.

### *HaploView*

HaploView software provides linkage disequilibrium statistics and displays haplotype data from primary genotype data (Barrett *et al.*, 2005). It can also be used to import HapMap data as shown in **Figure 20.17** for two HapMap populations at the globin locus on chromosome 11. In a triangle plot (Fig. 20.17b), LD measures for every pair of SNPs are plotted along lines at 45° to the horizontal track. Here, red colors correspond to higher LD.

### *HapMap Browser*

The HapMap Project includes a website from which all SNP data can be downloaded or viewed in a browser (Thorisson *et al.*, 2005). The recombination rate (in cM/Mb) can be plotted, and recombination hotspots can be identified. This browser also links to HaploView for optimal viewing and analysis.

### *Integrative Genomics Browser (IGV)*

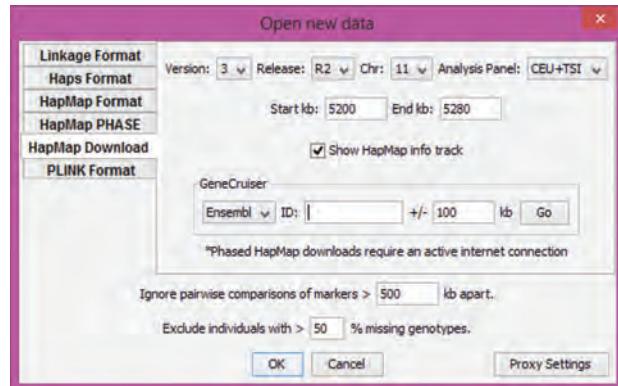
HapMap data can be downloaded from NCBI and visualized with Integrative Genomics Viewer (IGV) software. Let’s look at SNPs in the region of the *HBB* gene that are clinically relevant. First, download a VCF file and the corresponding indexed file from NCBI. Then run Integrative Genomics Viewer (IGV) software, load human hg19, and choose File > Load from file to upload the VCF. There are hundreds of SNPs in the vicinity of the *HBB* exons (Fig. 20.18a). Next, we’ll view a selection of SNPs from HapMap individuals. Since these participants are apparently normal, there is no overlap with clinically relevant SNPs, and instead the variants appear in intergenic regions (Fig. 20.18b). Some of these variants appear in just a subset of the geographic (ethnic) populations represented in HapMap.

### *NCBI dbSNP*

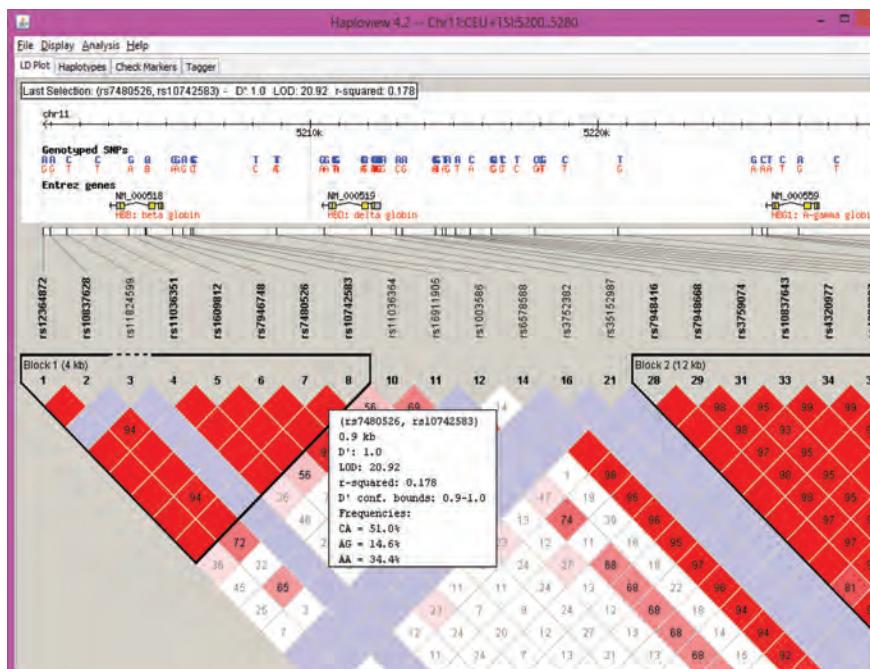
A search of the NCBI dbSNP resource with the query HBB leads to data in several formats.

- There is a list of individual SNPs (with identifiers such as rs334; Fig. 20.19a).
- GeneView displays SNPs overlapping a gene of interest, organized into functional groups (e.g., missense, synonymous, frameshift) and annotated by the amino acid and nucleotide coordinates of the genes (Fig. 20.19b).

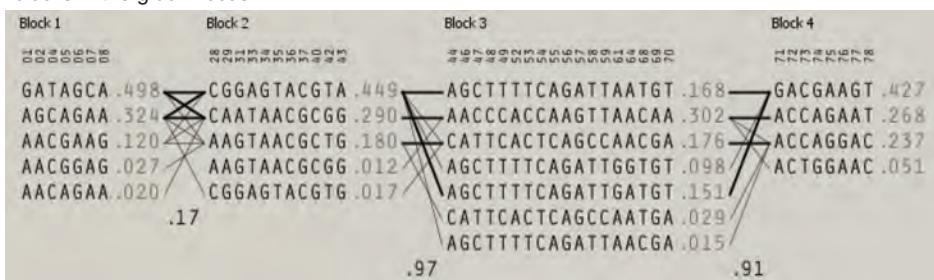
(a) HaploView data input



(b) Linkage disequilibrium (LD) plot

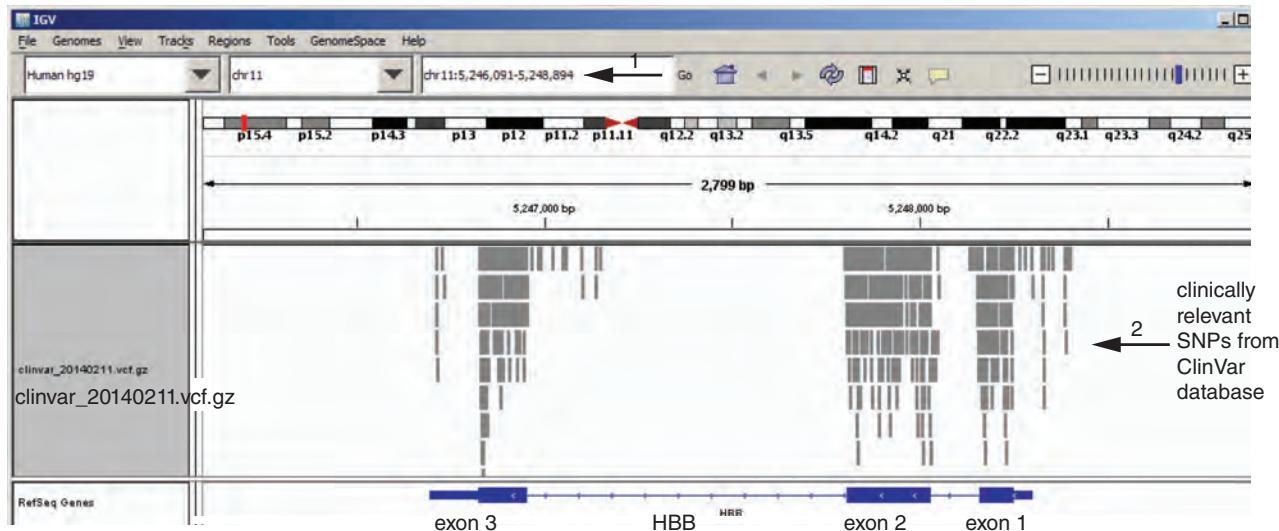
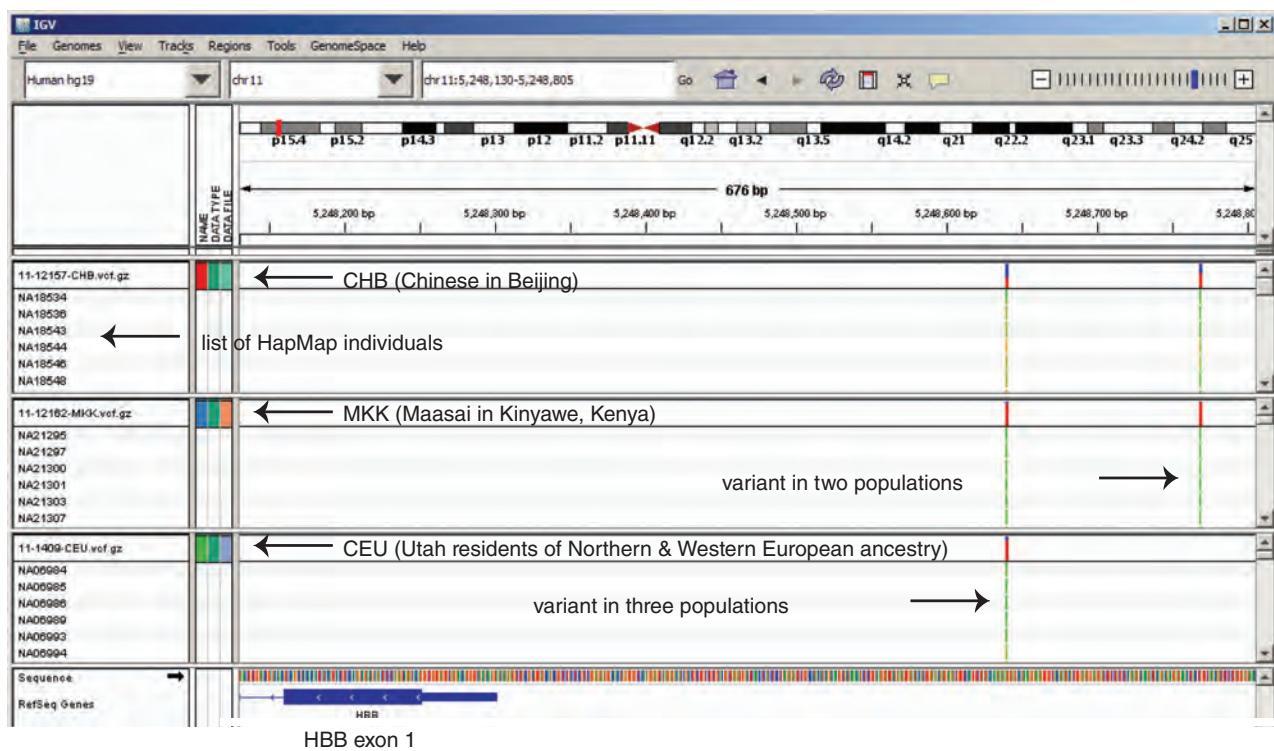


(c) LD blocks in the globin locus



**FIGURE 20.17** SNPs and linkage disequilibrium blocks are analyzed and visualized using HaploView software. (a) For this example data are imported from HapMap at the beta globin locus (chromosome 11, 5200–5800 kb) for the CEU+TSI populations. (b) A LD plot includes squares representing relatedness between SNPs, based on  $D'$  (values in boxes) or  $r^2$  values (not shown). LD statistics are available by right-clicking a box (one such box is open displaying data on two SNPs). This view is a portion of the selected 60 kilobase region. (c) Haplotype block definitions can be displayed. Population frequencies are shown to the right of each block. Lines show the most common crossings between blocks (thicker lines are more common crossings). The values below the lines (0.17, 0.97, 0.91) are the multilocus  $D'$  values which measure LD between two blocks. (Values closer to 0 have more historical recombination between blocks.)

Source: HaploView. Barrett *et al.* (2005).

(a) Visualizing clinically relevant single nucleotide polymorphisms (SNPs) at the *HBB* locus(b) Visualizing HapMap SNPs at the *HBB* locus from individuals of varying geographic origin

**FIGURE 20.18** Visualizing SNP data from Variant Call Format (VCF) files using Integrative Genomic Viewer (IGV). (a) A VCF file containing clinically relevant SNPs was downloaded from NCBI and uploaded to IGV. A search term such as HBB (or genomic coordinates of interest) can be entered (arrow 1) to view the beta globin locus. Around 2800 base pairs are shown here. SNPs from the ClinVar database are displayed (arrow 2). Most of these overlap the three exons of *HBB* and they represent deleterious variants. (b) Three VCF files including HapMap SNPs from chromosome 11 were downloaded from NCBI and uploaded to IGV. The populations were CHB (from China), MKK (from Africa), and CEU (of European ancestry). Data for many individuals can be displayed (data for six individuals per group are shown, with identifiers beginning with NA). A variant present in all three populations is indicated (arrow), as well as a SNP appearing in CHB and MKK but not CEU populations. Note that for these HapMap SNPs (derived from apparently normal individuals) there are no variants in *HBB* exons, and the intronic variants are presumably neutral rather than deleterious.

## (a) dbSNP listing of an individual SNP (query: HBB)

rs1135071 [Homo sapiens]

TATGGTCTATTTCCCACCCCTAG [c/G/t] CTGCTGGTGGCTACCCCTGGACCC

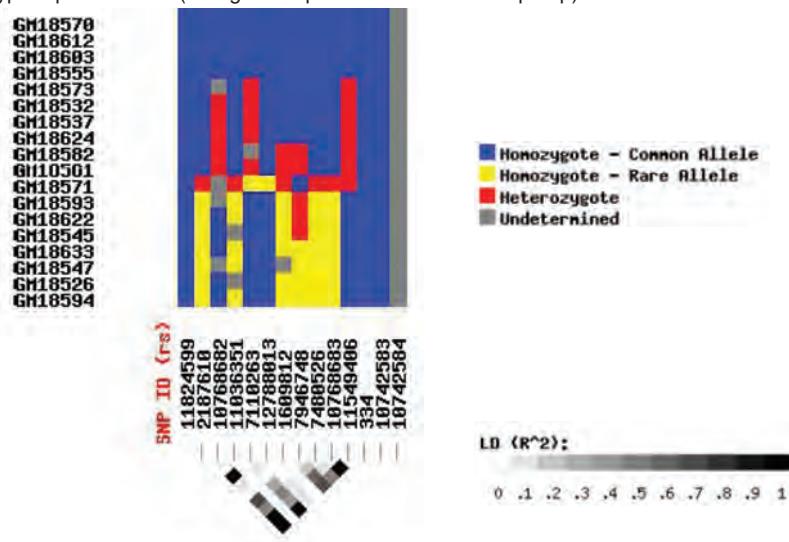
Chromosome: 11:5248029  
 Gene: **HBB** (GeneView)  
 Functional Consequence: missense  
 Allele Origin: G(germline)/T(germline)/C(germline)  
 Clinical significance: pathogenic  
 Validated: by 1000G, by cluster  
 Global MAF: A=0.0005/1  
 HGVS: NC\_000011.9:g.5248029C>A, NC\_000011.9:g.5248029C>G, NG\_000007.3:g.70817G>C, NG\_000007.3:g.70817G>T, NM\_000518.4:c.93G>C, NM\_000518.4:c.93G>T, NP\_000509.1:p.Arg31Ser, NT\_009237.18:g.5188029C>A, NT\_009237.18:g.5188029C>G

[PubMed](#) [Varview](#) [Protein3D](#)

## (b) NCBI GeneView report for HBB

Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	Validation	MAF	Allele origin	3D	Linkout	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
	5246883 439	rs369582912	N.D.				Yes			frame shift	CC	Leu [L]	2	131
										frame shift	AC	[DL]	2	130
										frame shift	AA	[EL]	2	130
	5247855 317	rs11549405	N.D.				Yes	↓		synonymous	C	Leu [L]	3	89
								↓		contig reference	G	Leu [L]	3	89
	5247878 294	rs11549406	0.005	H			Yes	↓		missense	G	Val [V]	1	82
								↓		contig reference	C	Leu [L]	1	82

## (c) dbSNP genotype report for HBB (linkage disequilibrium data from HapMap)



**FIGURE 20.19** SNP resources at NCBI include (a) lists of individual SNPs at dbSNP, one of which is shown here; (b) a dbSNP Gene view, listing SNPs and their functional consequences; and (c) a genotype report listing individuals and SNPs across HapMap populations and their linkage disequilibrium patterns.

Source: SNP resources, NCBI.

- A Variation Viewer reports SNPs and their clinical interpretation (e.g., pathogenic, probable-pathogenic, untested, unknown, other).
- A genotype report provides linkage disequilibrium analysis across HapMap populations (**Fig. 20.19c**).

SNP data are available at dbSNP at NCBI (<http://www.ncbi.nlm.nih.gov/snp/>, WebLink 20.35). The Variation Viewer entry for *HBB* is at <http://www.ncbi.nlm.nih.gov/sites/varvu?gene=3043> (WebLink 20.36). Information on SNP attributes is available from NCBI at [http://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs\\_attributes.html](http://www.ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html) (WebLink 20.37).

Note that in addition to the NCBI website, the Genome Workbench offers extensive resources to analyze SNP data.

The PLINK website is <http://pngu.mgh.harvard.edu/~purcell/plink/> (WebLink 20.38).

### PLINK

PLINK software, a command-line program, is a versatile, open-source set of tools for whole-genome association analysis (Purcell *et al.*, 2007). Input can include pedigree (.PED) and map files, such as files in those formats that can be downloaded from the HapMap website. The PED file includes identifiers for the family, individual, paternal and maternal identifiers, sex, and phenotype code. The MAP file consists of a set of rows with four columns describing the chromosome, rs# (SNP identifier), genetic distance (morgans), and base pair position. The types of analyses performed by PLINK include summary statistics, quality control steps, case/control and family-based association tests, permutation tests, linkage disequilibrium calculations, imputation of genotypes, and analysis of copy number variants.

When you download PLINK, test MAP and PED files are provided. We can look at their contents with `less`:

```
$ less test.map # this test set has just two SNPs
1.snp1 0 1
1.snp2 0 2
$ less test.ped # this PED file lists six individuals
# Three are affected, and three are unaffected.
1 1 0 0 1 1 A A G T
2 1 0 0 1 1 A C T G
3 1 0 0 1 1 C C G G
4 1 0 0 1 2 A C T T
5 1 0 0 1 2 C C G T
6 1 0 0 1 2 C C T T
```

We can measure the allele frequencies of the SNPs as follows:

```
$ ./plink --file test -freq
CHR SNP A1 A2 MAF NCHROBS
1 snp1 A C 0.3333 12
1 snp2 G T 0.4167 12
```

Next we can run an association test:

```
$ ./plink --file test -assoc
```

Six individuals are read from the PED file (three unaffected, three affected). The results are dumped to a file.

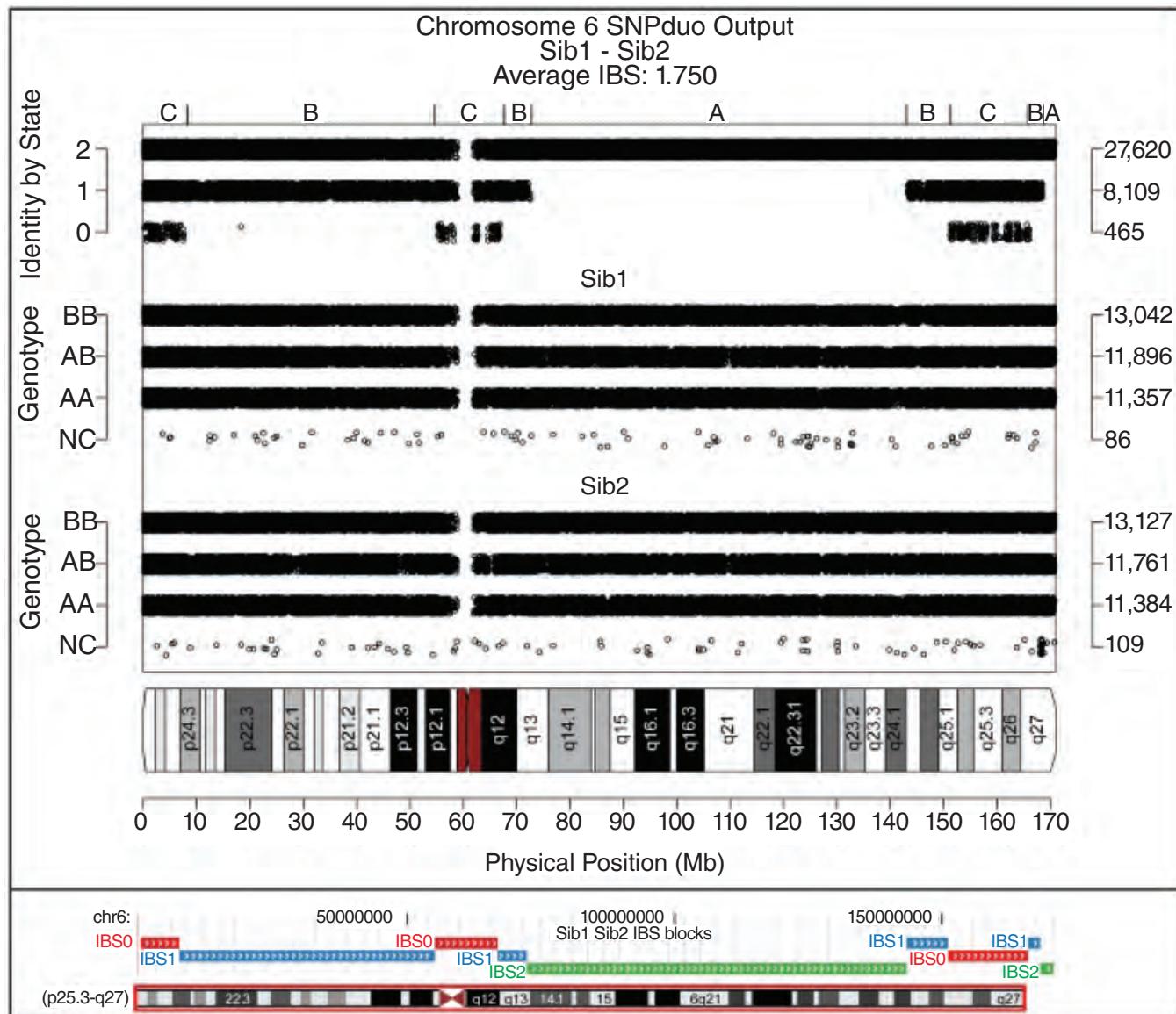
```
$ less plink.assoc
CHR SNP BP A1 F_A F_U A2 CHISQ P OR
1 snp1 1 A 0.1667 0.5 C 1.5 0.2207 0.2
1 snp2 2 G 0.1667 0.6667 T 3.086 0.07898 0.1
```

Here CHR is the chromosome, SNP is the SNP identifier, A1 is the minor allele name, F\_A is the frequency of this allele, F\_U is the frequency in controls, A2 is the major allele name, CHISQ is the basic allelic chi-square, P is a probability value, and OR is an estimated odds ratio for A1 (with A2 as the reference). In this way PLINK can perform a GWAS involving a million SNPs and thousands of individuals. The output can be integrated with HaploView, R packages, and a Java-based software package called gPLINK.

### SNPduo

While there are dozens (or perhaps hundreds) of excellent software tools, I will mention several developed by my lab to analyze SNP data. SNPduo performs pairwise comparisons

SNPduo, kcoeff, SNPtrio, pediSNP and other software are available at <http://pevsnnerlab.kennedykrieger.org> (WebLink 20.39). Eli Roberson maintains the latest version of SNPduo at a GitHub repository, <https://github.com/RobersonLab> (WebLink 20.40). trioPOD is available at <https://github.com/jdbaugh/tripod> (WebLink 20.41).



**FIGURE 20.20** SNPduo software visualizes genotype and identity-by-state data across chromosomes. Here chromosome 6 genotypes from two siblings are analyzed. For Sib1 and Sib2 there are homozygous (AA or BB) and heterozygous (AB) genotype calls, as well as some no calls (NC). The top panel indicates the identity-by-state which is either IBS2 (AA matches AA or BB/BB in the two siblings), IBS1 (AA/AB, BB/AB, AB/AA, or AB/BB), or IBS0 (AA matching BB or BB/AA). In positions of IBS2 (labeled A on top row) the two siblings inherited the same two alleles from their parents. In positions of IBS1 (labeled B) there is one shared allele, and at IBS0 positions (labeled C) there are no shared alleles. These inheritance patterns reflect meiotic recombination events. An ideogram of chromosome 6 is displayed. The output includes a BED file that can be uploaded to the UCSC (or Ensembl) genome browser as shown in the bottom panel. From Roberson and Pevsner (2009). Licensed under Creative Commons Attribution License 2.5.

between SNP datasets (Roberson and Pevsner, 2009; Fig. 20.20). SNPduo can be run on the command line (using PED and MAP files as input), or it can be run as a web-based application. kcoeff uses a windowed genome-wide approach to estimate identity-by-state and identity-by-descent (Stevens *et al.*, 2011). We have used it to analyze relatedness between all HapMap individuals, identifying many pedigrees that had been misspecified. triPOD identifies mosaic abnormalities in mother/father/child trio datasets with extremely high sensitivity and specificity (Baugher *et al.*, 2013).

You can learn more about the International HapMap Project at its NHGRI website, <http://www.genome.gov/page.cfm?pageID=10001688> (WebLink 20.42).

## Major Conclusions of HapMap Project

The three phases of the HapMap project contributed basic knowledge of human genetic variation. The observations and conclusions included the following (McVean *et al.*, 2005; International HapMap Consortium, 2005, 2007; International HapMap 3 Consortium, 2010):

1. Most variation is manifest in individuals of African descent. Asian and European populations emerged relatively recently in human history, and their genetic diversity largely represents a subset of African diversity.
2. Linkage disequilibrium displays a block-like structure. There are regions of high D' that are interrupted by regions of recombination. Any given SNP is therefore usually tightly associated with its neighboring SNPs. A typical block has a length of 30–50 kb (about 0.1 centiMorgans). Lower frequency variants tend to be younger than common variants, and they tend to have longer haplotype blocks.
3. LD blocks may span multiple recombination hotspots. HapMap phase 2 characterized regions of recent sharing in detail. Some recent sharing is due to autozygosity (recent inbreeding within a population).
4. Some regions are characterized by lack of recombination across extended haplotype structures. This is evident in centromeric regions where linkage disequilibrium is elevated (see **Fig. 20.8**).
5. SNPs are useful for genome-wide association studies (GWAS; see Chapter 21). A large set of SNPs (>1 million) can be genotyped at relatively low cost per individual. Comparison of two groups (affected and control populations) can reveal genomic regions harboring variation that segregates with a disease phenotype. This has led to the discovery of many disease risk factors and/or causative genes.
6. Natural selection can remove deleterious mutations and preserve (fix) advantageous variants. HapMap data reveal genes that have undergone recent adaptive evolution. Sabeti *et al.* (2007), including members of the HapMap Consortium, used three criteria to identify SNPs under strong positive selection: they were newly arisen (derived) alleles, based on comparisons to primate outgroups; they were highly differentiated between human populations, since recent positive selection is likely to reflect a local environmental adaptation; and they focused on nonsynonymous coding SNPs and SNPs in evolutionarily conserved sequences since those are most likely to have biological effects. Sabeti *et al.* described 300 candidate regions. In some cases they identified pairs of genes that have related functions and have undergone positive selection in the same populations (e.g., *LARGE* and *DMD* in the YRI population; both encode proteins implicated in Lassa fever virus binding and infection). Other examples include *HBB*, *LCT* encoding lactase (and promoting the ability to consume milk from other mammals after weaning), the human leukocyte antigen (HLA) region on chromosome 6, and an inversion on chromosome 17 (spanning 900 kb in individuals of European ancestry).
7. The prevalence of structural variation can be measured through SNP analysis. We showed an example of a hemizygous deletion in **Figure 8.23**, and we can see both available SNPs and a summary of structural variants (deletions, amplifications, inversions, and complex variants) at the UCSC browser (**Fig. 8.10, 21.15**) or the Ensembl or NCBI browsers. Phase 3 HapMap reported ~1600 genomic segments that varied in copy number with a minor allele frequency of at least 1%. The median size was 7.2 kb per copy number polymorphism and a cumulative 3.5 Mb of sequence per individual (about 0.1% of each genome). A total of 92% are deletions, 8% are duplications, and a third overlap RefSeq genes (International HapMap 3 Consortium, 2010).

According to the Database of Genomic Variants (<http://dgv.tcag.ca/dgv/app/home>, WebLink 20.43, MacDonald *et al.*, 2014), copy number variants span  $2.2 \times 10^9$  nucleotides or >71% of the human genome (March 2014).

## The 1000 Genomes Project

The goal of the 1000 Genomes Project is to create a comprehensive resource on human genetic variation. It is significant as the first publicly available whole-genome sequence dataset on the population scale. One specific aim was to identify most (>95%) of the genetic variants that have at least a 1% frequency in the populations being studied. In the pilot phase three approaches were taken (1000 Genomes Project Consortium *et al.*, 2010): (1) the genomes of two father/mother/daughter trios were sequenced to high coverage (average mapped coverage of 42× per individual); (2) whole-genome sequencing of 179 individuals (from four populations) was performed with average mapped coverage of 3.6× per individual; and (3) exon-targeted sequencing of 697 individuals was performed. Data for the project (including 4.9 terabases of sequence reported by the 1000 Genomes Project Consortium *et al.*, 2010) are available at the project website.

Each individual has two haplotypes at each autosomal locus (as discussed above). The three approaches taken by the 1000 Genomes Project offer different information about haplotypes. Trio sequencing allows haplotypes to be phased, so that the sequences of two haploid genomes in the child are inferred (Fig. 20.21a). Low-coverage whole-genome sequencing is far less expensive than whole-genome sequencing, and is especially able to identify common haplotypes (Fig. 20.21b). Whole-exome sequencing provides data on a limited portion of the genome (typically about 60 Mb rather than nearly 3000 Mb) without sufficient breadth of coverage to phase haplotypes (Fig. 20.21c).

The main conclusions of the 1000 Genome Project include the following (1000 Genomes Project Consortium *et al.*, 2010, 2012).

1. High rates of variation tended to occur at the HLA and subtelomeric regions. Lowest rates occurred in a 5 Mb, gene-dense region around 3p21.
2. 1000 Genomes Project data have been useful to impute SNPs for genome-wide association studies.

3. The number of variants has been described for different functional classes of variants.

In particular, variants at conserved sites were emphasized, having genomic evolutionary rate profiling (GERP) scores >2. The rationale is that these variants, summarized in Table 20.20, are more likely to have functional relevance. Each individual harbors ~2500 nonsynonymous variants at conserved sites (most of which are common variants having allele frequencies >5%), but a typical individual human genome harbors >10,000 nonsynonymous variants. Each person in the 1000 Genomes Project, and therefore each of us in general, has 20–40 variants at conserved sites that are identified as damaging; 10–20 loss of function variants; 2–5 damaging mutations; and 1–2 variants previously identified from cancer genome sequencing (1000 Genomes Project Consortium *et al.*, 2012). A separate analysis by the consortium suggested that each person harbors the following (Xue *et al.*, 2012):

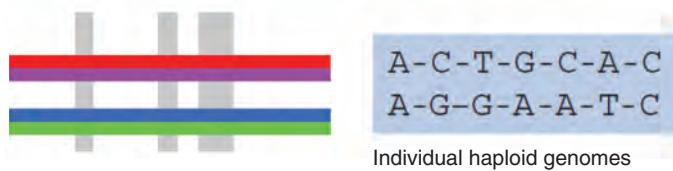
- 40–85 homozygous missense mutations that are predicted to be damaging;
- 40–110 variants classified as disease-causing by the Human Gene Mutation Database (HGMD; Chapter 21);
- 0–8 disease-causing mutations that are predicted to be highly damaging; and
- 0–1 of these mutations in the homozygous state.

To work with 1000 Genomes data you can visit the project website or its browser which is modeled closely on the Ensembl genome browser. Enter HBB and you can view variants. By clicking on an individual SNP you can view its alleles, population data (e.g., allele frequencies across a range of populations), and phenotype data (with clinical data). A link on the left sidebar allows you to “Get VCF data.” This presents the Data Slicer, a tool that allows you to upload BAM or VCF files (Chapter 9). By default, the selected

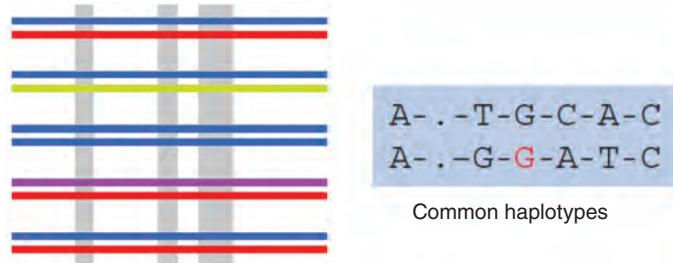
The 1000 Genomes Website is at  
🌐 <http://www.1000genomes.org/>  
(WebLink 20.44).

The 1000 Genomes Browser is online at🌐 <http://browser.1000genomes.org/>  
(WebLink 20.45). NCBI also offers a 1000 Genomes Browser at  
🌐 <http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>  
(WebLink 20.46).

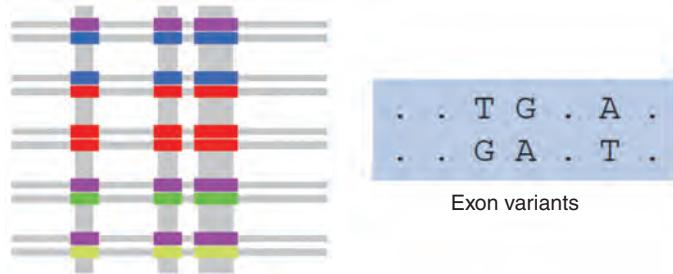
(a) Trio sequencing: haplotypes phased by transmission



(b) Low coverage whole-genome sequencing: statistical phasing of haplotypes



(c) Exon sequencing: haplotypes remain unphased



**FIGURE 20.21** Haplotype phasing in the 1000 Genomes Project. Each individual has two haplotypes at each autosomal locus; these are typically shared with others in the same population. Methods vary in their abilities to reconstruct these haplotypes. Colors (left side) indicate different haplotypes in individual genomes. Line widths indicate depth of coverage (not to scale). Shaded region (right) indicates examples of genotype data that could be observed or inferred for the same sample using three strategies. (a) Trio sequencing allows accurate discovery of variants and phasing of haplotypes across most of the genome. (b) Low-coverage sequencing identifies shared variants on common haplotypes (red, blue bars) but is less powered to detect rare haplotypes (e.g., light green) as well as associated variants (see dots indicating missing alleles). Some inaccurate genotypes occur (red allele is incorrectly assigned G; it should be A). (c) Exon sequencing includes less coverage of the genome. Common, rare, and low-frequency variation can be detected in targeted portions of the genome. Haplotypes cannot be phased. Redrawn from the 1000 Genomes Project Consortium *et al.* (2010). Reproduced with permission from Macmillan Publishers Ltd.

The URL for the VCF spanning the HBB region, for which one individual is selected, is [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis\\_results/integrated\\_call\\_sets/ALL.chr11.integrated\\_phase1\\_v3.20101123.snps\\_indels\\_svs.genotypes.vcf.gz](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ALL.chr11.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.vcf.gz). This file is available as Web Document 20.2.

region (HBB; in this case 11:5246694–5250625) can be analyzed. By following the Data Slicer steps you can select all 1000 Genomes data or filter by population or individual, and obtain a VCF file for download. You can analyze this VCF using VCFtools, IGV (as described above), Ensembl, NCBI, or UCSC browsers, or annotate its variants with VAAST or other software (Chapter 9).

An example of Data Slicer output for the HBB region is as follows, showing the VCF header lines and the first two rows of variants.

```
##fileformat=VCFv4.1
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
```

**TABLE 20.20 Variant load per individual at conserved sites from the 1000 Genomes Project. DAF: derived allele frequency across sample.**

Variant type	Number of derived variant sites per individual			Excess rare deleterious	Excess low-frequency deleterioius
	<0.5% DAF	0.5–5% DAF	>5% DAF		
All sites	30,000–150,000	120,000–680,000	3.6M–3.9M	ND	ND
Synonymous	29–120	82–420	1300–1400	ND	ND
Nonsynonymous	130–400	240–910	2300–2700	76–190	77–130
Stop-gain	3.9–10	5.3–19	24–28	3.4–7.5	3.8–11
Stop-loss	1.0–1.2	1.0–1.9	2.1–2.8	0.8–1.1	0.80–1.0
HGMD-DM	2.5–5.1	4.8–17	11–18	1.6–4.7	3.8–12
COSMIC	1.3–2.0	1.8–5.1	5.2–10	0.93–1.6	1.3–2.0
Indel frameshift	1.0–1.3	11–24	60–66	ND	3.2–11
Indel nonframeshift	2.1–2.3	9.5–24	67–71	ND	0–0.73

*Source:* The 1000 Genomes Project Consortium (2012; not all rows are displayed here.) Reproduced with permission from Macmillan Publishers Ltd.

```

##INFO=<ID=AVGPOST,Number=1,Type=Float,Description="Average posterior probability from MaCH/Thunder">
##INFO=<ID=RSQ,Number=1,Type=Float,Description="Genotype imputation quality from MaCH/Thunder">
##INFO=<ID=ERATE,Number=1,Type=Float,Description="Per-marker Mutation rate from MaCH/Thunder">
##INFO=<ID=THETA,Number=1,Type=Float,Description="Per-marker Transition rate from MaCH/Thunder">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Count">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total Allele Count">
##ALT=<ID=DEL,Description="Deletion">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/README">
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN">
##INFO=<ID=AMR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AMR based on AC/AN">
##INFO=<ID=ASN_AF,Number=1,Type=Float,Description="Allele Frequency for samples from ASN based on AC/AN">
##INFO=<ID=AFR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AFR based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from EUR based on AC/AN">
##INFO=<ID=VT,Number=1,Type=String,Description="indicates what type of variant the line represents">
```

```

##INFO=<ID=SNPSOURCE,Number=.,Type=String,Description="indicates if a snp
was called when analysing the low coverage or exome alignment data">
##reference=GRCh37
##source_20140302.1=/nfs/public/rw/ensembl/vcftools/bin/vcf-subset -c
NA18912 /net/isilonP/public/rw/ensembl/1000genomes/release-14/tmp/
slicer/11.5246694-5250625.ALL.chr11.integrated_phase1_v3.20101123.snps_
indels_svs.genotypes.vcf.gz
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA18912
11 5246794 rs200399660 C T 100 PASS
AA=c;AC=0;AF=0.0005;AN=2;ASN_AF=0.0017;AVGPOST=1.0000;
ERATE=0.0003; LDAF=0.0005;RSQ=1.0000;SNPSOURCE=EXOME; THETA=0.0003;
VT=SNP GT:DS:GL 0|0:0.000:0.00,-5.00,-5.00
11 5246840 rs36020563 G A 100 PASS
AA=g;AC=0; AF=0.0005;AFR_AF=0.0020;AN=2;AVGPOST=1.0000;
ERATE=0.0003;LDAF=0.0005;
RSQ=1.0000;SNPSOURCE=LOWCOV,EXOME;THETA=0.0006;VT=SNP GT:DS:GL
0|0:0.000:0.00,-5.00,-5.00

```

This example shows that it is easy to obtain and analyze 1000 Genomes data. The header provides a guide to the specific information offered for each variant. It is critical to examine whether the variant may be associated with artifacts (e.g., strand bias or low read depth). The variant information includes allele frequencies in different populations and the type of experiment (low coverage versus exome-targeted sequencing). In this example the source shows that we have chosen just one individual (NA18912) for analysis, although a single VCF can include variants from large numbers of individuals.

### Variation: Sequencing Individual Genomes

While sequencing the human genome was a massive project, one that has been compared in magnitude to landing a human on the moon, resequencing an individual human genome is easier. The National Human Genome Research Institute (NHGRI) of the National Institutes of Health has launched programs to reduce the cost of sequencing an individual genome from the early value (tens of millions of dollars in the late 2000s) to the current cost (nearing US\$ 1000) and eventually to under US\$ 1000 each.

The significance of individual genome sequencing is that it has the potential to facilitate the start of an era of individualized medicine in which DNA changes that are associated with a disease condition are identified. As discussed in Chapter 21, most diseases involve an interplay between genetic and environmental factors. Even for diseases that are seemingly caused by environmental factors, from traumatic brain injury to malnutrition to infectious disease, an individual's genetic constitution is likely to have a large effect on the disease process. Another significant aspect of individual genome sequencing is that it will help to elucidate the genetic diversity and history of the species.

In 2007 the first two individual human genome sequences were announced: those of J. Craig Venter (Levy *et al.*, 2007) and James Watson, Nobel laureate and co-discoverer of the structure of DNA (Wheeler *et al.*, 2008). The Venter genome was reported as the diploid sequence of an individual. In contrast, the Celera human genome sequence (Venter *et al.*, 2001) was based on a consensus of DNA sequences from five individuals, and the public consortium sequence (IHGSC, 2001) was also based on genomes from multiple individuals. These were composite efforts that represented sequence data that were essentially averaged to yield information on 23 pairs of chromosomes. They did not assess the variation that occurs in an individual having each autosome derived from maternal and paternal alleles. The surprising finding of Levy *et al.* was that there were four million variants between the parental chromosomes, about five-fold more than had been anticipated. It was not until the years 2004–2006 that the great diversity of copy number variants as well as smaller indels and SNPs became more fully appreciated.

The NHGRI genome technology program website is <http://www.genome.gov/10000368> (WebLink 20.47). The decline in sequencing costs is depicted in Figure 9.3.

In computer laboratory exercise (20.1) below, we perform BLAST searches against the Venter genome.

(e.g., Iafrate *et al.*, 2004; Sebat *et al.*, 2004; Redon *et al.*, 2006; Pinto *et al.*, 2007; Scherer *et al.*, 2007).

The strategy employed by Levy *et al.* (2007) to sequence, assemble, and analyze the genome included seven steps: (1) obtaining informed consent to collect the DNA sample; (2) genome sequencing; (3) genome assembly; (4) comparative mapping of the individual genome to an NCBI reference genome; (5) DNA variation detection and filtering; (6) haplotype assembly; and (7) data annotation and interpretation.

The assembly of Venter's genome was based on 32 million sequence reads generating ~20 billion base pairs of DNA sequence with a 7.5-fold depth of coverage. Sanger dideoxynucleotide sequencing technology was used because each read is longer than currently available 454 technology (used to sequence Watson's genome) or Illumina technology (see Chapter 9). The assembly included 2,782,357,138 bases of DNA. Comparison to the NCBI reference genome revealed 4.1 million variants. These included 3.2 million SNPs (slightly more than one per 1000 base pairs), over 50,000 block substitutions (of length 2–206 base pairs), almost 300,000 insertions/deletions (indels) of 1–571 base pairs, ~560,000 homozygous indels (ranging up to ~80,000 base pairs), 90 inversions, and many copy number variants. The majority of variants relative to the reference human genome were SNPs. Insertions and deletions accounted for a smaller proportion of the variable events (22%) but, because they tend to involve larger genomic regions, they accounted for 74% of the variant nucleotides relative to the reference NCBI genome. In an effort to identify the full spectrum of variation in this genome, Pang *et al.* (2010) reannotated the Venter genome using data from additional sequencing and microarray platforms. They reported thousands of new variants, in particular structural variants that are difficult to detect by sequencing alone.

In recent years thousands of individual genomes have been sequenced. Early examples include the genome of an Asian individual, YH (Wang *et al.*, 2008); the genome of James D. Watson (Wheeler *et al.*, 2008); Korean individuals SJK (Ahn *et al.*, 2009) and AK1 (Kim *et al.*, 2009); the genome of a Yoruba male, sample NA18507 (Bentley *et al.*, 2008); and the genomes of four indigenous Namibian hunter-gatherers (KB1, NB1, TK1 and MD8) and Archbishop Desmond Tutu (ABT) who is a Bantu (Schuster *et al.*, 2010). The focus of these early studies was on validating the technologies that were applied to the massive task of sequencing a 6 Gb diploid human genome at adequate depth of coverage; identifying SNPs, including those present in dbSNP; identifying structural variation; performing comparative genomics to learn which variants were shared; and predicting which variants were disease-associated.

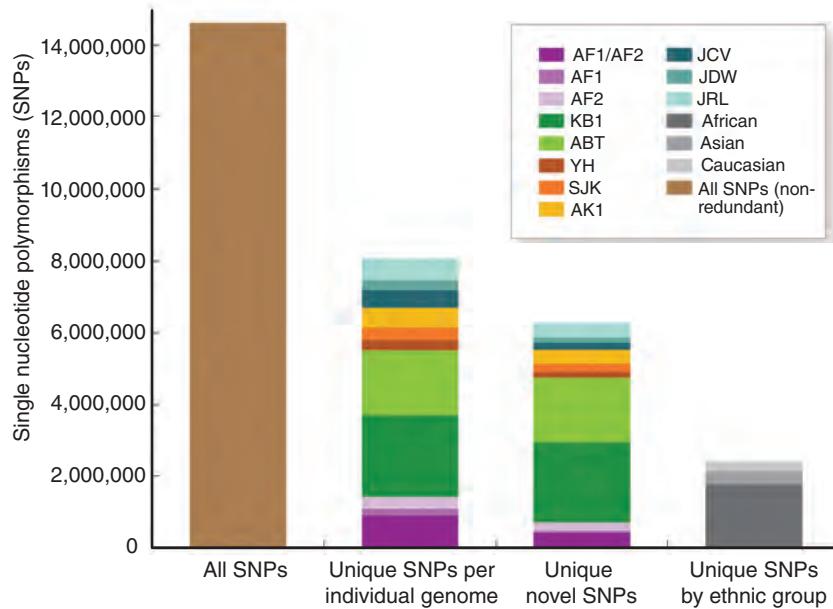
Gonzaga-Jauregui *et al.* (2012) analyzed variants in ten whole genomes (Fig. 20.22). Cumulatively, these 10 genomes included 14.6 million nonredundant SNPs. As expected, the greatest number of unique SNPs were contributed by individuals of African descent.

It is likely that the number of sequenced genomes (or exomes) is rapidly approaching 100,000 if not more (although the majority of these data are in the process of being collected and analyzed). Examples of large-scale projects, other than the 1000 Genomes Project, are the UK10K project (with a goal of sequencing the genomes of 10,000 individuals in the UK); the Autism Genome 10K project; the Personal Genome Project (an effort to sequence 100,000 human genomes); and several cancer initiatives.

## PERSPECTIVE

The sequencing of the human genome represents one of the great accomplishments in the history of science. This effort is the culmination of decades of work in an international effort. Two major technological advances enabled the human genome to be sequenced: (1) the invention of automated DNA sequencing machines in the 1980s allowed nucleotide data to be collected on a large scale; and (2) the computational

The UK10K project is described at <http://www.sanger.ac.uk/about/press/2010/100624-uk10k.html> (WebLink 20.48). The Autism Genome 10K Project website is <http://autismgenome10k.org/> (WebLink 20.49), and the Personal Genome Project website is <http://www.personalgenomes.org/> (WebLink 20.50).



**FIGURE 20.22** Comparison of single-nucleotide polymorphisms (SNPs) in 10 personal genomes. All SNPs in each genome were compared with the 9 others. First bar: there were 14,608,404 nonredundant SNPs (first bar). Second bar: SNPs that were unique to each genome. Third bar: SNPs that were unique in an individual genome and novel. Fourth bar: SNPs shared by individuals of the same ethnic group. Abbreviations: AF1: NA18507; AF2: NA18507; KB1: Khoisan genome; ABT: Archbishop Desmond Tutu; YH: Chinese genome; SJK: Korean genome 1; AK1: Korean genome 2; JCV: J. Craig Venter; JDW: James D. Watson; JRL: James R. Lupski. Redrawn from Gonzaga-Jauregui *et al.* (2012). Reproduced with permission from Annual Reviews.

biology tools necessary to analyze those sequence data were created by biologists and computer scientists.

By some estimates, the total number of sequenced human genomes is already approaching one hundred thousand in the year 2015. When the human genome project concluded in around 2003, very few people anticipated the new revolution that would be introduced by next-generation sequencing. In the coming years, we can expect the pace of DNA sequence to continue to increase. It is already becoming possible to compare the complete genome sequence of many individuals in an effort to relate genotype to phenotype. The genomic sequence permits analyses of sequence variation such as SNPs and copy number variants; disease-causing mutations; evolutionary forces; and genomic properties such as recombination, replication, and the regulation of gene function. While in the past we relied heavily on mouse and other model organisms to model gene function, it is becoming possible to envision a “human knockout collection” in which the phenotypes of large numbers of individuals with a particular homozygous gene knockout, copy number variant, or other genetic profile are explored.

## PITFALLS

As each chromosome has been finished, there have been many technical problems to solve regarding sequencing depth, assembly (particularly in regions with highly repetitive DNA), and annotation. There are discrepancies between the results of gene-finding algorithms (as revealed by the ENCODE project, Chapter 8) and there are often discrepancies between different databases. Copy number variants can be particularly difficult to identify and assemble because they are often associated with repetitive DNA, and segmental duplications are difficult to resolve using whole-genome shotgun assembly.

There are a number of outstanding problems that have yet to be solved:

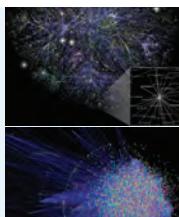
- How can we accurately determine the number of protein-coding genes?
- How can we determine the number of noncoding genes?
- How can we determine the function of genes and proteins?
- What is the evolutionary history of our species?
- What is the degree of heterogeneity between individuals at the nucleotide level?

As we take our initial look at the human genome, it is appropriate to see this moment as a beginning rather than an end. Having the sequence in hand, and having the opportunity to compare the human genome sequence to that of many other genomes, we are now in a position to pose a new generation of questions.

Regarding individual human genome sequences, perhaps the biggest pitfall is the misconception that there is a single, agreed-upon set of variants associated with that genome. The methods selected for alignment to a reference genome, variant calling, and annotation of variants will have a large impact on the final description of the genome.

## ADVICE FOR STUDENTS

Many individual human genome sequences are publicly available in the BAM format, including HapMap individuals. Following the computational methods outlined in Chapter 9, analyze this genome with different aligners, variant callers, and annotators. For one workflow, for example using SAMtools, perform alignment with different parameters to learn the effects.



## Discussion Questions

**[20-1]** If you had the resources and facilities to sequence the entire genome of 50 individuals, which would you select? Why? Describe how you would approach the data analysis.

**[20-2]** The *Saccharomyces cerevisiae* genome duplicated about 100 MYA, as indicated by BLAST searching (Chapter 18), and we discussed whole-genome duplication in fish, *Paramecium*, and plants (Chapter 19). Why is it not equally straightforward to identify large duplications of the human genome? Is it because they did not occur, because the evolutionary history of humans obscures such events, or because we lack the tools to detect such large-scale genomic changes?

## PROBLEMS/COMPUTER LAB

**[20-1]** Determine the sequence of beta globin in Craig Venter's genome. First, identify the accession number for beta globin (NM\_000518). Next, identify the accession number for the genome; Levy *et al.* (2007) list it as ABBA00000000. By viewing that record, note that ABBA00000000 itself does not directly refer to DNA sequences, but it lists the accessions ABBA01000001–ABBA01255300 that do

contain whole-genome shotgun sequence data. Perform a BLASTN search at NCBI, using the beta globin query NM\_000518 and setting the database to whole-genome shotgun reads (WGS). In the Entrez query box enter ABBA01000001:ABBA01255300[PACC] in order to limit the search to just Venter's genome sequences. (You can visit the Entrez help link to learn the appropriate formats for limits.) Extra problem: the *ABCC11* gene (ATP-binding cassette, subfamily C, member 11; NM\_032583) encodes a protein that Venter has in a variant form that predisposes someone to wet rather than dry earwax. Identify the variant nucleotides and/or amino acids.

**[20-2]** Go to NCBI Gene and select a human gene of interest, such as alpha 2 globin. Examine the features of this gene at the Ensembl, NCBI, and UCSC websites. Make a table of various properties (e.g., exon/intron structure, number of ESTs corresponding to the expressed gene, polymorphisms identified in the gene, neighboring genes). Are there discrepancies between the data reported in the three databases? It is also possible to compare the contents of these databases by searching and comparing within any one resource.

**[20-3]** How many protein-coding genes are on each human chromosome? Use EDirect (introduced in Chapter 2). This problem is adapted from <http://www.ncbi.nlm.nih.gov/books/NBK179288/>. Try the following code (in blue) and compare your answer to that below. Note the paucity of genes on chromosomes 18, 21, Y, and MT as well as the large number on chromosome 19.

```
for chr in {1..22} X Y MT
do
  esearch -db gene -query "Homo sapiens [ORGN]
AND $chr [CHR]" |
  efilter -query "alive [PROP] AND genotype
protein coding [PROP]" |
  efetch -format docsum |
  xtract -pattern DocumentSummary -NAME Name \
-block GenomicInfoType -match "ChrLoc:$chr" \
-tab "\n" -element ChrLoc,"&NAME" |
  grep '.' | sort | uniq | cut -f 1 |
  sort-uniq-count-rank
done
2063 1
1268 2
1081 3
766 4
871 5
1035 6
932 7
683 8
801 9
751 10
1292 11
1034 12
334 13
612 14
609 15
843 16
1195 17
275 18
1409 19
550 20
245 21
455 22
849 X
74 Y
13 MT
```

**[20-4]** The recombination rate is higher near the telomeres (see Fig. 20.8). Use the UCSC Table Browser to identify regions having very high recombination rates. (1) Go to <http://genome.ucsc.edu> (WebLink 20.51) and select Table Browser. Select the human genome, mapping and sequencing group, Recomb Rate track. Clicking on the summary statistics button shows that there are 2822 entries (one per megabase). (2) Select filter and set the decodeAvg (DeCode genetic map average value) to greater than 5. Try setting the filter using other genetic maps. (3) When you submit this, the summary statistics show that there are now just 12 entries (on chromosomes 4, 9, 10, 12, 14, 17, 19,

20, and X). You can also set the output to hyperlinks to the Genome Browser, showing that most of these regions are indeed subtelomeric. (4) Identify sites with the lowest recombination rate in the genome using a similar strategy. (5) Identify RefSeq genes that are close to the highest (or lowest) recombination rates. Use the intersection tool at the UCSC Table Browser site.

**[20-5]** Compare the extent of conserved synteny between human and the rhesus macaque (*Macaca mulatta*) on chromosomes 1 (the largest chromosome in humans), 21 (the smallest autosome), X, and Y. Which shows the most conservation? What specific genes are conserved between human and rhesus macaque on the Y chromosome? Why is the extent of conservation on that chromosome so low? One way to accomplish this exercise is to visit the Ensembl human genome browser ([http://www.ensembl.org/Homo\\_sapiens](http://www.ensembl.org/Homo_sapiens), WebLink 20.52), click on a chromosome (e.g., Y), then use the pull-down menu “View Chr Y Synteny” on the left sidebar.

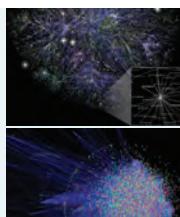
**[20-6]** *HLA-B* is the most polymorphic gene in the human genome. Explore its properties. (1) Set the UCSC Genome Browser (e.g., GRCh38 assembly) to coordinates chr6:31,353,872–31,357,212 and view the SNPs. You can see the spectacular amount of polymorphism. (2) Obtain a broader perspective by viewing the SNPs across a one million base pair region, chr6:31,000,001–32,000,000. (3) Use the Table Browser and its intersection feature to find the five most polymorphic genes across the entire genome.

**[20-7]** Human mitochondrial DNA (RefSeq identifier NC\_012920.1) has a bacterial origin. (1) Perform a BLASTN search of the nonredundant (nr) database, restricting the output to bacteria. To which group of bacteria is the human sequence most related? (You may view the Taxonomy Report for a convenient summary.) (2) To which genes is the human sequence most related? You may inspect your BLASTN results. (3) There is just one bacterial protein that is related to the proteins encoded by the human mitochondrial genome. What is it? You may inspect your BLASTN results or, to specifically search for proteins encoded by human mitochondrial DNA, use NC\_012920.1 as a query in a BLASTX search restricted to bacteria. (4) The UCSC Genome Browser includes a track in the Variation section called “NumtS Sequence.” These are nuclear mitochondrial sequences, that is, sequences that transferred from the bacterial endosymbiont into the human nuclear genome. How many entries are there? (Use the Table Browser to find out, selecting summary statistics.) Are they clustered at particular genomic loci? You can select Tools > Genome Graphs and display that (or any other) track (Fig. 20.23).



**FIGURE 20.23** The UCSC Genome Graphs tool allows you to plot any UCSC or custom-selected tracks on ideograms. The distribution of NumtS genes is shown, that is, nuclear genes of mitochondrial origin.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.



## Self-Test Quiz

- [20-1]** Approximately how large is the human genome?
- 3 Mb;
  - 300 Mb;
  - 3000 Mb; or
  - 30,000 Mb.
- [20-2]** Approximately what percentage of the human genome consists of repetitive elements of various kinds?
- 5%;
  - 25%;
  - 50%; or
  - 85%.
- [20-3]** What percentage of the human genome is devoted to the protein-coding regions?
- 1–5%;
  - 5–10%;
  - 10–20%; or
  - 20–40%.
- [20-4]** The human genome contains many transposon-derived repeats. These are described as:
- dead fossils;
  - young, active elements;
  - human-specific elements; or
  - inverted repeats.

**[20-5]** Approximately how much of the human genome do segmental duplications occupy?

- (a) <1%;
- (b) 5%;
- (c) 20–30%; or
- (d) 50%.

**[20-6]** In areas of high GC content of the human genome,

- (a) gene density tends to be low;
- (b) gene density tends to be high;
- (c) gene density is highly variable; or
- (d) genes tend to have fewer introns.

**[20-7]** In comparison to other metazoan genomes (such as nematodes, insects and mouse),

- (a) the human genome contains considerably more protein-coding genes;
- (b) the human genome has considerably more unique genes that lack identifiable orthologs;

- (c) the human genome has a higher GC content; or
- (d) the human genome has somewhat more multidomain proteins and alternative splicing.

**[20-8]** When the human genome project was completed by 2001–2004, how much of the genome remained impossible to sequence due to repetitive content and other technical challenges?

- (a) essentially none;
- (b) about 2 Mb;
- (c) about 25 Mb; or
- (d) about 225 Mb.

**[20-10]** Single-nucleotide polymorphisms (SNPs) are useful to characterize all aspects of the human genome except for:

- (a) disease association;
- (b) microduplications;
- (c) inverse selection; or
- (d) population migration.

## SUGGESTED READING

In this chapter, we discussed the public consortium description of the human genome (IHGSC, 2001) and the finishing of the euchromatic portion of the genome (IHGSC, 2004). The companion Celera article (Venter *et al.*, 2001) is also of great interest, as are the many accompanying articles in those issues of *Science* and *Nature*. We also discussed individual genomes. The first of these is the Levy *et al.* (2007) article on the genome of an individual, with an emphasis on variants of assorted sizes.

For each of the 22 autosomes and two sex chromosomes, there has been a paper published in *Nature* that describes the chromosome in detail. We provide links to these papers at <http://www.bioinfbook.org/chapter20>. These important papers describe the in-depth analyses of finished (or nearly finished) chromosomal sequences. They highlight the need for complete sequencing in order to perform more accurate annotation and comparative analyses.

## REFERENCES

- 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D. *et al.* 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073. PMID: 20981092.
- 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A. *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65. PMID: 23128226.
- Ahn, S.M., Kim, T.H., Lee, S. *et al.* 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**(9), 1622–1629. PMID: 19470904.
- Amir, R.E., Van den Veyver, I.B., Wan, M. *et al.* 1999. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nature Genetics* **23**, 185–188. PMID: 10508514.
- Arhondakis, S., Auletta, F., Bernardi, G. 2011. Isochores and the regulation of gene expression in the human genome. *Genome Biology and Evolution* **3**, 1080–1089. PMID: 21979159.

- Avner, P., Heard, E. 2001. X-chromosome inactivation: Counting, choice and initiation. *Nature Reviews Genetics* **2**, 59–67.
- Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., Eichler, E. E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research* **11**, 1005–1017.
- Barrett, J.C., Fry, B., Maller, J., Daly, M.J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2), 263–265. PMID: 15297300.
- Baugher, J.D., Baugher, B.D., Shirley, M.D., Pevsner, J. 2013. Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. *BMC Genomics* **14**, 367. PMID: 23724825.
- Behar, D.M., van Oven, M., Rosset, S. *et al.* 2012. A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics* **90**(4), 675–684. PMID: 2248206.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P. *et al.* 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218), 53–59. PMID: 18987734.
- Bernardi, G. 2001. Misunderstandings about isochores. Part 1. *Gene* **276**, 3–13. PMID: 11591466.
- Church, D.M., Schneider, V.A., Graves, T. *et al.* 2011. Modernizing reference genome assemblies. *PLoS Biology* **9**(7), e1001091. PMID: 21750661.
- Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. 2003. US National Human Genome Research Institute. A vision for the future of genomics research. *Nature* **422**, 835–847.
- Costantini, M., Bernardi, G. 2008. The short-sequence designs of isochores from the human genome. *Proceedings of the National Academy of Sciences, USA* **105**(37), 13971–13976. PMID: 18780784.
- Crick, F. H., Watson, J. D. 1953. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.
- Cuvier, G. 1849. *The Animal Kingdom, Arranged According to Its Organization*. William S. Orr & Co, London.
- Deloukas, P., Matthews, L.H., Ashurst, J. *et al.* 2001. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871. PMID: 11780052.
- Deloukas, P., Earthrow, M.E., Grafham, D.V. *et al.* 2004. The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429**, 375–382. PMID: 15164054.
- Dunham, I., Shimizu, N., Roe, B.A. *et al.* 1999. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495. PMID: 10591208.
- Dunham, A., Matthews, L.H., Burton, J. *et al.* 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428**, 522–528. PMID: 15057823.
- ENCODE Project Consortium, Bernstein, B.E., Birney, E. *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74. PMID: 22955616.
- Flicek, P., Amode, M.R., Barrell, D. *et al.* 2014. Ensembl 2014. *Nucleic Acids Research* **42**(1), D749–755. PMID: 24316576.
- Gardiner-Garden, M., Frommer, M. 1987. CpG islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261–282.
- Gonzaga-Jauregui, C., Lupski, J.R., Gibbs, R.A. 2012. Human genome sequencing in health and disease. *Annual Review of Medicine* **63**, 35–61. PMID: 22248320.
- Green, E.D. 2001. Strategies for the systematic sequencing of complex genomes. *Nature Reviews Genetics* **2**, 573–583.
- Green, E.D., Chakravarti, A. 2001. The human genome sequence expedition: Views from the “base camp.” *Genome Research* **11**, 645–651.
- Green, E.D., Guyer, M.S., National Human Genome Research Institute. 2001. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**(7333), 204–213. PMID: 21307933.
- Gregory, S.G., Barlow, K.F., McLay, K.E. *et al.* 2006. The DNA sequence and biological annotation of human chromosome 1. *Nature* **441**, 315–321. PMID: 16710414.
- Grimwood, J., Gordon, L.A., Olsen, A. *et al.* 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428**(6982), 529–535. PMID: 15057824.

- Haring, D., Kypr, J. 2001. Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Molecular Biology Reports* **28**, 9–17.
- Hattori, M., Fujiyama, A., Taylor, T.D. *et al.* 2000. The DNA sequence of human chromosome 21. *Nature* **405**, 311–319. PMID: 10830953.
- Heilig, R., Eckenberg, R., Petit, J.L. *et al.* 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607. PMID: 12508121.
- Hillier, L.W., Fulton, R.S., Fulton, L.A. *et al.* 2003. The DNA sequence of human chromosome 7. *Nature* **424**, 157–164. PMID: 12853948.
- Hillier, L.W., Graves, T.A., Fulton, R.S. *et al.* 2005. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434**, 724–731. PMID: 15815621.
- Humphray, S.J., Oliver, K., Hunt, A.R. *et al.* 2004. DNA sequence and analysis of human chromosome 9. *Nature* **429**, 369–375. PMID: 15164053.
- Iafrate, A.J., Feuk, L., Rivera, M.N. *et al.* 2004. Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.
- Ingman, M., Kaessmann, H., Pääbo, S., Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713.
- International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**(6968), 789–967. PubMed PMID: 14685227.
- International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**(7063), 1299–1320. PubMed PMID: 16255080.
- International HapMap Consortium, Frazer, K.A., Ballinger, D.G. *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164), 851–861. PMID: 17943122.
- International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A. *et al.* 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311), 52–58. PMID: 20811451.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
- Jones, P.A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**(7), 484–492. PMID: 22641018.
- Jurka, J. 1998. Repeats in genomic DNA: Mining and meaning. *Current Opinion in Structural Biology* **8**, 333–337.
- Karolchik, D., Barber, G.P., Casper, J. *et al.* 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research* **42**(1), D764–70. PMID: 24270787.
- Kim, J.I., Ju, Y.S., Park, H. *et al.* 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**(7258), 1011–1015. PMID: 19587683.
- Levy, S., Sutton, G., Ng, P.C. *et al.* 2007. The diploid genome sequence of an individual human. *PLoS Biology* **5**, e254. PMID: 17803354.
- MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., Scherer, S.W. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research* **42**(Database issue), D986–992. PMID: 24174537.
- Martin, J., Han, C., Gordon, L.A. *et al.* 2004. The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432**, 988–994. PMID: 15616553.
- McVean, G., Spencer, C.C., Chaix, R. 2005. Perspectives on human genetic variation from the HapMap Project. *PLoS Genetics* **1**, e54.
- Mungall, A.J., Palmer, S.A., Sims, S.K. *et al.* 2003. The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811. PMID: 14574404.
- Muzny, D.M., Scherer, S.E., Kaul, R. *et al.* 2006. The DNA sequence, annotation and analysis of human chromosome 3. *Nature* **440**, 1194–1198. PMID: 16641997.
- National Research Council. 1988. *Mapping and Sequencing the Human Genome*. National Academy Press, Washington, DC.

- Nusbaum, C., Mikkelsen, T.S., Zody, M.C. *et al.* 2006. DNA sequence and analysis of human chromosome 8. *Nature* **439**, 331–335. PMID: 16421571.
- Osborne, L.R., Li, M., Pober, B. *et al.* 2001. A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nature Genetics* **29**, 321–325.
- Ostertag, E. M., Kazazian, H. H., Jr. 2001. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Research* **11**, 2059–2065.
- Pakendorf, B., Stoneking, M. 2005. Mitochondrial DNA and human evolution. *Annual Review of Genomics and Human Genetics* **6**, 165–183.
- Pang, A.W., MacDonald, J.R., Pinto, D. *et al.* 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**(5), R52. PMID: 20482838.
- Pinto, D., Marshall, C., Feuk, L., Scherer, S.W. 2007. Copy-number variation in control population cohorts. *Human Molecular Genetics* **16**, R168–R173.
- Ponting, C. P. 2001. Plagiarized bacterial genes in the human book of life. *Trends in Genetics* **17**, 235–237.
- Purcell, S., Neale, B., Todd-Brown, K. *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3), 559–575. PMID: 17701901.
- Redon, R., Ishikawa, S., Fitch, K.R. *et al.* 2006. Global variation in copy number in the human genome. *Nature* **444**, 444–454. PMID: 17122850.
- Roberson, E.D., Pevsner, J. 2009. Visualization of shared genomic regions and meiotic recombination in high-density SNP data. *PLoS One* **4**(8), e6711. PMID: 19696932.
- Ross, M.T., Grafham, D.V., Coffey, A.J. *et al.* 2005. The DNA sequence of the human X chromosome. *Nature* **434**, 325–337. PMID: 15772651.
- Rozen, S., Skaletsky, H., Marszalek, J.D. *et al.* 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876. PMID: 12815433.
- Sabeti, P.C., Varilly, P., Fry, B. *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164), 913–918. PMID: 17943131.
- Salzberg, S. L., White, O., Peterson, J., Eisen, J. A. 2001. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292**, 1903–1906.
- Scally, A., Durbin, R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**(10), 745–753. PMID: 22965354.
- Scherer, S. E., Muzny, D.M., Buhay, C.J. *et al.* 2006. The finished DNA sequence of human chromosome 12. *Nature* **440**, 346–351. PMID: 16541075.
- Scherer, S.W., Cheung, J., MacDonald, J.R. *et al.* 2003. Human chromosome 7: DNA sequence and biology. *Science* **300**(5620), 767–772. PMID: 12690205.
- Scherer, S.W., Lee, C., Birney, E. *et al.* 2007. Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**, S7–S15.
- Schmutz, J., Martin, J., Terry, A. *et al.* 2004. The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274. PMID: 15372022.
- Schuster, S.C., Miller, W., Ratan, A. *et al.* 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**(7283), 943–947. PMID: 20164927.
- Sebat, J., Lakshmi, B., Troge, J. *et al.* 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528.
- She, X., Horvath, J.E., Jiang, Z. *et al.* 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430**, 857–864.
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J. *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837. PMID: 12815422.
- Smith, Z.D., Meissner, A. 2013. DNA methylation: roles in mammalian development. *Nature Reviews Genetics* **14**(3), 204–220. PMID: 23400093.
- Stevens, E.L., Heckenberg, G., Roberson, E.D. *et al.* 2011. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genetics* **7**(9), e1002287. PMID: 21966277.

- Taylor, T.D., Noguchi, H., Totoki, Y. *et al.* 2006. Human chromosome 11 DNA sequence and analysis including novel gene identification. *Nature* **440**, 497–500. PMID: 16554811.
- Thorisson, G.A., Smith, A.V., Krishnan, L., Stein, L.D. 2005. The International HapMap Project web site. *Genome Research* **15**, 1592–1593.
- Toth, G., Gaspari, Z., Jurka, J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research* **10**, 967–981.
- Tycko, B., Morison, I. M. 2002. Physiological functions of imprinted genes. *Journal of Cellular Physiology* **192**, 245–258.
- Venter, J.C., Adams, M.D., Myers, E.W. *et al.* 2001. The sequence of the human genome. *Science* **291**, 1304–1351. PMID: 11181995.
- Wang, J., Wang, W., Li, R. *et al.* 2008. The diploid genome sequence of an Asian individual. *Nature* **456**(7218), 60–65. PMID: 18987735.
- Wheeler, D.A., Srinivasan, M., Egholm, M. *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189), 872–876. PMID: 18421352.
- Xue, Y., Chen, Y., Ayub, Q. *et al.* 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics* **91**(6), 1022–1032. PMID: 23217326.
- Yu, A., Zhao, C., Fan, Y. *et al.* 2001. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953. PMID: 11237020.
- Zody, M.C., Garber, M., Adams, D.J. *et al.* 2006a. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature* **440**, 1045–1049. PMID: 16625196.
- Zody, M.C., Garber, M., Sharpe, T. *et al.* 2006b. Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**, 671–675. PMID: 16572171.



## CHAPITRE IX.

*Des modifications chimiques qu'éprouvent les matières albumineuses des solides et des fluides organiques chez l'homme malade.*

Nous venons de terminer l'étude des matières albumineuses, telles qu'on les trouve dans l'organisation saine. Nous avons remarqué qu'elles y existent en grande abondance, tant dans les tissus que dans les humeurs. Elles y ont donc une grande importance. Aussi aucun dérangement fonctionnel ne doit pas probablement s'effectuer, sans les lésier, soit dans leurs états isomériques, soit dans leur quantité habituelle, soit dans leurs combinaisons, soit même dans leur composition élémentaire. En effet, les désordres qui viennent si souvent tourmenter l'économie, se manifestent fréquemment dans ces matières; c'est ce que l'observation expérimentale nous apprend depuis quelques années. Nous avons déjà pu signaler chez l'homme malade, tantôt que l'albumine solide et non dissoute y passe vicieusement, en certains cas, à l'état de dissolution, ou bien qu'elle s'y convertit en albumin, mais que cependant, jamais ou que rarement du moins, l'albumin n'y devient albumine; tantôt que l'albumine combinée, soit l'albumine à laquelle on donne l'épithète de soluble, soit celle qu'on nomme caseine, y éprouve quelquefois, vicieusement encore, un changement qui la rend neutre ou acide, ou qui finit par lui ôter sa solubilité, et la laisse à l'état solide, désormais indissoutte; tantôt même, que cette albumine ou cet albumin, libre ou combiné, y augmente ou y diminue irrégulièrement, et par fois y disparaît dans quelques fluides, dans certains organes; tantôt, enfin, que les sels et l'acide ou l'acide unis à l'albumine soluble, y varient de proportion au-delà du terme normal. De telles connaissances déjà acquises, prouvent que nous parviendrons à découvrir, un jour, toutes les modifications chimiques que les maladies apportent dans les matières albumineuses. C'est par de semblables connaissances,

As soon as proteins were discovered, investigators studied their role in disease. Prosper-Sylvain Denis (1799–1863) wrote *Études Chimiques, Physiologiques, et Médicales, Faites de 1835 à 1840, sur les Matières Albumineuses* (Chemical, Physiological and Medical Studies, done from 1835 to 1840, on the Albuminous Materials) in 1842. His chapter 9 (p. 141) is entitled "On the chemical modifications that prove the albuminous materials of solids and fluids in the sick person." He wrote (see arrow 1): "In effect, the disorders that so often torment the economy, frequently manifest themselves in these materials." This passage includes a reference to caseine ("caséine") and concludes (arrow 2) "Such knowledge already acquired proves that we will come to discover, one day, all the chemical modifications that illnesses carry in the albuminous materials." The lower panel (from p. 144) shows a table comparing the water, proteins, alkali, and salts from healthy and diseased serum.

Source: Denis (1842).

SERUM SAIN réuni à sa fibrine.	SERUM DU SANG COUENNEUX réuni également à sa fibrine.
Eau... 900	Eau..... 900 .....
Matière albumineuse 79	{ Albumine..... 77 ..... 67 Matière albumineuse. 80. Fibrine..... 2 ..... 13 }
Alcali.. 1	Soude..... 1 ..... 2 Alcali... 2.
Sels... 6	{ Sulfates et phosphates à bases alcalines 2 ..... 2 Sels.... 4. Chlorure de sodium 4 ..... 2 }

# Human Disease

# CHAPTER 21

*Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life. While there are molecular diseases, there are no diseased molecules. At the level of the molecules we find only variations in structure and physicochemical properties. Likewise, at that level we rarely detect any criterion by virtue of which to place a given molecule “higher” or “lower” on the evolutionary scale. Human hemoglobin, although different to some extent from that of the horse, appears in no way more highly organized. Molecular disease and evolution are realities belonging to superior levels of biological integration. There they are found to be closely linked, with no sharp borderline between them. The mechanism of molecular disease represents one element of the mechanism of evolution. Even subjectively the two phenomena of disease and evolution may at times lead to identical experiences. The appearance of the concept of good and evil, interpreted by man as his painful expulsion from Paradise, was probably a molecular disease that turned out to be evolution. Subjectively, to evolve must most often have amounted to suffering from a disease. And these diseases were of course molecular.*

—Emile Zuckerkandl and Linus Pauling (1962, pp. 189–190)

## LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- describe major categories of human disease;
- explain different approaches to identifying disease-associated genes;
- compare and contrast the main disease databases; and
- describe how studies of model organisms elucidate disease-related variation.

## HUMAN GENETIC DISEASE: A CONSEQUENCE OF DNA VARIATION

Variation in DNA sequence is a defining feature of life on Earth. For each species, genetic variation is responsible for the adaptive changes that underlie evolution. Evolution is a process by which species adapt to their environment. When changes in DNA improve the fitness of a species, its population reproduces more successfully. When changes are relatively maladaptive, the species may become extinct. At the level of the individual within a species, some mutations improve fitness, most mutations have no effect on fitness, and some are maladaptive (relative to some norm). Disease may be defined as maladaptive

Mutation is the alteration of DNA sequence. The cause may be errors in DNA replication or repair, the effects of chemical mutagens, or radiation. While there may be negative connotations associated with the concept of mutations, mutation and fixation are the essential driving forces behind evolution.

changes that afflict individuals within a population. Disease is also defined as an abnormal condition in which physiological function is impaired. Our focus is on the molecular basis of physiological defects at the levels of DNA, RNA, and protein.

From a medical perspective, disease is “a pathological condition of the body that presents a group of clinical signs, symptoms, and laboratory findings peculiar to it and setting the condition apart as an abnormal entity differing from other normal or pathological condition” (Thomas, 1997, p. 552). Disorder is a “pathological condition of the mind or body” (Thomas, 1997, p. 559). A syndrome is “a group of symptoms and signs of disordered function related to one another by means of some anatomical, physiological, or biochemical peculiarity. This definition does not include a precise cause of an illness but does provide a framework of reference for investigating it” (Thomas, 1997, p. 1185). Costa *et al.* (1985) follow a World Health Organization definition of disease as a cause through a process (pathogenesis) resulting in manifestations.

There is a tremendous diversity to the nature of human diseases for several reasons:

- Mutations affect all parts of the human genome. There are limitless opportunities for maladaptive mutations to occur, and there are many mechanisms by which mutations can cause disease (summarized in **Table 21.1**). These include disruptions of gene function by point mutations that change the identity of amino acid residues; by deletions or insertions of DNA, ranging in size from one nucleotide to an entire chromosome that is over 100 million base pairs (Mb); or inversions of the orientation of a DNA fragment. In many cases, different kinds of mutations affecting the same gene cause distinct phenotypes.
- Protein-coding genes function by producing a protein as a gene product. A disease-causing mutation in a gene results in the failure to produce the gene product with normal function. This has profound consequences on the ability of the cells in which the gene product is normally expressed to function.
- The interaction of an individual with his or her environment has profound effects on disease phenotype. Genetically identical twins may have entirely different phenotypes. Such differences are attributable to environmental influences or to epigenetic effects. The concordance rate between monozygotic twins for a given clinical phenotype is an indication of the relative extent to which genetic and environmental effects influence disease. Even for highly genetic disorders, such as autism (see “Complex Disorders” below) and schizophrenia, the concordance rate is never 100%.

### A Bioinformatics Perspective on Human Disease

In Chapter 1, we defined bioinformatics as a discipline that uses computer databases and computer algorithms to analyze proteins, genes, and genomes. Our approach to human disease is reductionist, in that we seek to describe genes and gene products that cause disease. However, an appreciation of the molecular basis of disease may be integrated with a holistic approach to uncover the logic of disease in the entire human population (Childs and Valle, 2000). As we explore bioinformatics approaches to human disease, we are constantly faced with the complexity of all biological systems. Even when we uncover the gene that when mutated causes a disease, our challenge is to attempt to connect the genotype to the phenotype. We can only accomplish this by synthesizing information about the biological context in which each gene functions and in which each gene product contributes to cellular function (Childs and Valle, 2000; Dipple *et al.*, 2001).

The field of bioinformatics offers approaches to human disease that may help us to understand basic questions about the influence of genes and the environment on all aspects of the disease process. Some examples of ways in which this field can have an

**TABLE 21.1 Mechanisms of genetic mutation. AG/GT indicates mutations in the canonical first two and last two base pairs of an intron. Outside AG/GT indicates mutations in less canonical sequences. Adapted from Beaudet *et al.* (2001, p. 9) with permission from McGraw Hill.**

Mechanism	Usual effect	Example
<i>Large mutation</i>		
Deletion	Null	Duchenne dystrophy
Insertion	Null	Hemophilia A/LINE
Duplication	Null, gene disrupted	Duchenne dystrophy
Duplication	Dosage, gene intact	Charcot–Marie–Tooth
Inversion	Null	Hemophilia A
Expanding triplet	Null	Fragile X
Expanding triplet	Gain of function	Huntington
<i>Point mutation</i>		
Silent	None	Cystic fibrosis
Missense or in-frame deletion	Null, hypomorphic, altered function, benign	Globin
Nonsense	Null	Cystic fibrosis
Frame shift	Null	Cystic fibrosis
Splicing (AG/GT)	Null	Globin
Splicing (outside AG/GT)	Hypomorphic	Globin
Regulatory (TATA, other)	Hypomorphic	Globin
Regulatory (poly A site)	Hypomorphic	Globin

impact on our knowledge of disease are highlighted throughout the chapter, and include the following.

- To the extent that the genetic basis of disease is a function of variation in DNA sequences, DNA databases offer us the basic material necessary to compare DNA sequences. These databases include major, general repositories of DNA sequence such as GenBank/EMBL/DDBJ and SRA (Chapter 2), general resources such as Online Mendelian Inheritance in Man (OMIM), and locus-specific databases that provide data on sequence variations at individual loci.
- Geneticists who search for disease-causing genes through linkage studies, association studies, or other tests (described in “Approaches to Identifying Disease-Associated Genes and Loci” below) depend on physical and genetic maps in their efforts to identify mutant genes.
- When a protein-coding gene is mutated, there may be a consequence on the three-dimensional structure of the protein product. Bioinformatics tools described in Chapter 13 allow us to predict the structure of protein variants and, from such analyses, we may infer changes in function.
- Once a mutant gene is identified, we want to understand the consequence of that mutation on cellular function. We have described a variety of approaches to understanding protein function in Chapters 12–14. In our discussion of *Saccharomyces cerevisiae*, we discussed additional high-throughput approaches to understanding eukaryotic protein function (Chapter 14). Gene expression studies (Chapters 10 and 11) have been employed to study the transcriptional response to disease states.
- We may obtain great insight into the role of a particular human gene by identifying orthologs in simpler organisms. We discuss orthologs of human disease genes found in a variety of model systems.

This chapter is organized in six main sections. (1) We first provide an overview of human disease, including approaches to disease classification. We consider the subject of human disease at several levels (outlined in Fig. 21.1). (2) We describe categories of disease (monogenic, complex, and genomic disorders as well as environmental disease, somatic disease, and cancer). (3) We introduce disease databases including Online Mendelian Inheritance in Man (OMIM, a principal disease database), HGMD, and ClinVar. There are also several thousand locus-specific mutation databases, and we discuss these. (4) We describe approaches to identifying disease-associated genes such as linkage, genome-wide association studies, and human genome sequencing. (5) Human diseases have been studied in a variety of model organisms, and we introduce these projects. (6) Finally, we consider the functional classification of disease genes.

The OMIM entry for alkaptonuria is #20355; the # sign is defined in “OMIM: Central Bioinformatics Resource for Human Disease” below. The RefSeq accession of HGD is NP\_000178.2. The gene is localized to chromosome 3q21-q23. You can read Garrod’s 1902 paper on alkaptonuria online as Web Document 21.1 at <http://www.bioinfbook.org/chapter21>.

A trait is a characteristic or property of an individual that is the outcome of the action of a gene or genes.

### Garrod’s View of Disease

Sir Archibald Garrod (1857–1936) made important contributions to our understanding of the nature of human disease. In a 1902 paper, Garrod described his studies of alkaptonuria, a rare inherited disorder. In alkaptonuria, the enzyme homogentisate 1,2-dioxygenase (HGD) is defective or missing. As a result, the amino acids phenylalanine and tyrosine cannot be metabolized properly, and a metabolite (homogentisic acid) accumulates. This metabolite oxides in urine and turns dark. Garrod considered this phenotype from the perspective of evolution, noting the influence of natural selection on chemical processes. Variations in metabolic processes between individuals might include those changes that cause disease.

Garrod had the insight that, for each of the rare disorders he studied, the disease phenotype reflects the chemical individuality of the individual. He further realized that this trait was inherited; he proposed that alkaptonuria is transmitted by recessive Mendelian inheritance.

level	Bioinformatics resources
Molecular level	DNA general resources: OMIM locus-specific mutation databases
	RNA databases of gene expression
	protein UniProt; databases of mutant proteins
Systems level	organelles databases of peroxisomal, mitochondrial, lysosomal disease
	organs/systems disease databases focused on blood, neuromuscular, retinal, cardiovascular, gastrointestinal, other
Organismal level	clinical phenotype databases with information on data on age of onset; frequency; severity; malformations; tissue involvement; other features
	animal model human disease orthologs in various deuterostomes (mouse, sea urchin), protostomes (fly, worm), plants, other species
	organizations and foundations general organizations (NORD) disease-specific organizations

**FIGURE 21.1** Bioinformatics resources for the study of human disease are organized at a variety of levels.

At the time, it was thought that most diseases were caused by external forces such as bacterial infection. In studying this and related recessive disorders (such as cystinuria and albinism) he instead proposed that the manifestation of the disease is caused by an inherited enzyme deficiency or biochemical error (Scriver and Childs, 1989). He described this point of view in his first book, *Inborn errors of metabolism* (1909). Garrod wrote in 1923 (cited in Scriver and Childs, 1989, p. 7):

If it may be granted that the individual members of a species vary from the normal of the species in chemical structure and chemical behavior, it is obvious that such variations or mutations are capable of being perpetuated by natural selection; and not a few biologists of the present day assign to chemical structure and function a most important share in the evolution of species ... Very few individuals exhibit such striking deviations from normal metabolism as porphyrinurias and cystinurias show, but I suspect strongly that minimal deviations which escape notice are almost universal. How else can be explained the part played by heredity in disease? There are some diseases which are handed down from generation to generation ... which tend to develop in later childhood and early adult life ... It is difficult to escape the conclusion that although these maladies are not congenital, their underlying causes are inborn peculiarities.

Garrod thus presented a new view of how inborn factors cause disease. He worked at a time before Beadle and Tatum offered the hypothesis that one gene encodes one protein, and Garrod never used the word “gene.” We now understand that the “inborn peculiarities” he described are mutated genes. A main conclusion of his work is that chemical individuality, achieved through genetic differences, is a major determinant of human health and disease. Although the phrase “chemical individuality” is not used often today, the concept is of tremendous interest in the field of pharmacogenomics. Not everyone who is exposed to an infectious agent becomes sick, and it is imperative to understand why. Not everyone who takes a drug responds in a similar way.

Garrod further developed these ideas in a second book, *Inborn Factors in Disease* (1931). Here he addressed the question of why certain individuals are susceptible to diseases: whether the disease is clearly inherited or derives from another cause such as an environmental agent. He argued that chemical individuality predisposes us to disease. Every disease process is affected by both internal and external forces: our genetic complement and the environmental factors we face. In some cases, such as inborn errors of metabolism, genetic factors have a more prominent role. In other cases, such as multifactorial disease, mutations in many genes are responsible for the disease. In infectious disease, genes also have an important role in defining the individual’s susceptibility and bodily response to the infectious agent. We next proceed to discuss these various kinds of disease.

## Classification of Disease

We describe several general categories of disease below such as single-gene disorders, complex disorders, chromosomal disorders and environmental disease. From the perspective of bioinformatics, we are interested in understanding the mechanism of disease in relation to genomic DNA, genes, and their gene products. We are further interested in the consequences of mutations on cell function and on the comparative genomics of disease-causing genes throughout evolution. This perspective is complementary to and yet quite distinct from that of the clinician or epidemiologist.

For any study of disease a classification system is useful, and many approaches are available. One is to describe mortality statistics. These data (based on death certificates in the United States from the year 2010) include rankings of the cause of death (**Table 21.2**). This information is helpful in identifying the most common diseases, and projections of

**TABLE 21.2** Leading causes of death in United States in 2010. Cause of death is based on the International Classification of Diseases, Tenth Revision, 1992.

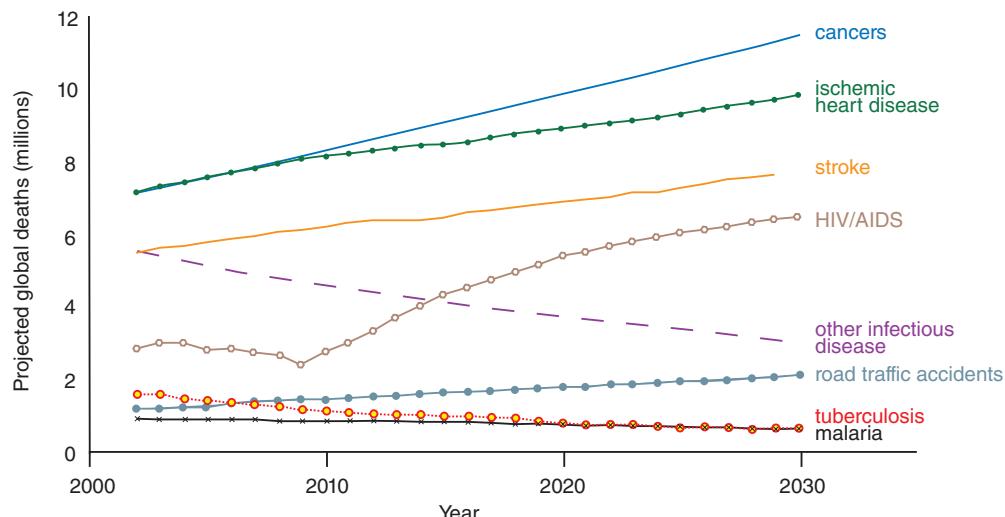
Rank	Cause of death	Number	Percent of all deaths
–	All causes	2,468,435	100.0
1	Diseases of heart	597,689	24.2
2	Malignant neoplasms	574,743	23.3
3	Chronic lower respiratory diseases	138,080	5.6
4	Cerebrovascular diseases	129,476	5.2
5	Accidents (unintentional injuries)	120,859	4.9
6	Alzheimer's disease	83,494	3.4
7	Diabetes mellitus	69,071	2.8
8	Nephritis, nephrotic syndrome, and nephrosis	50,476	2.0
9	Influenza and pneumonia	50,097	2.0
10	Intentional self-harm (suicide)	38,364	1.6

Source: National Vital Statistics Reports, 62(6) ([http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62\\_06.pdf](http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_06.pdf))

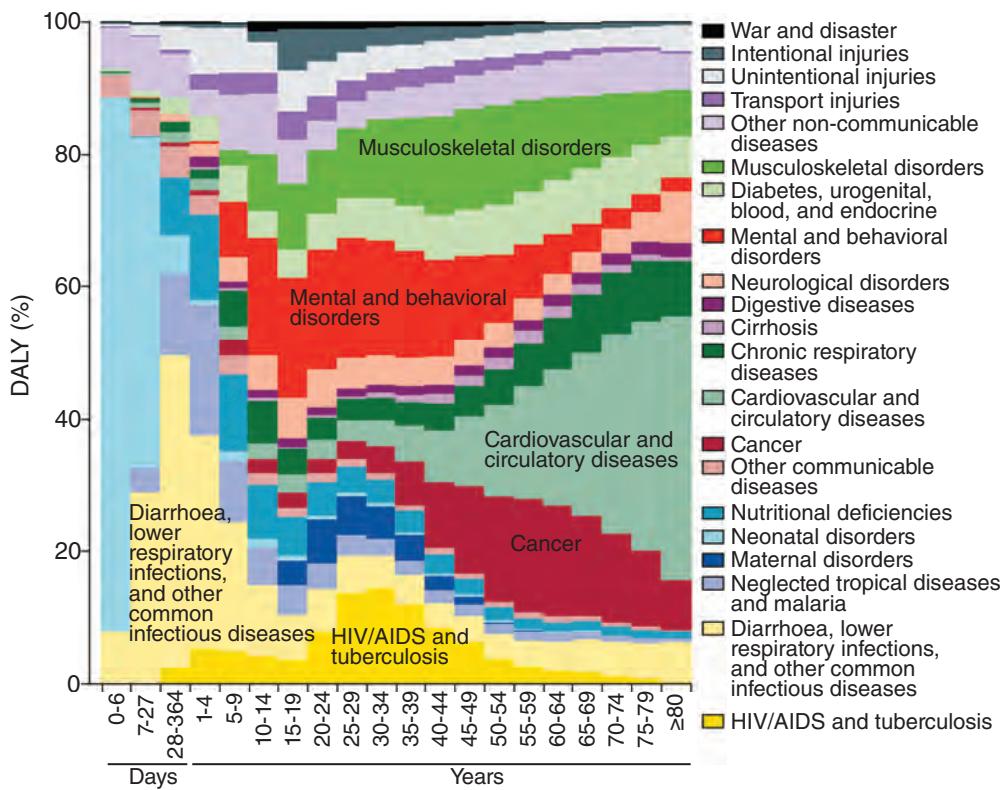
The data in Table 21.2 are available from the website of the National Center for Health Statistics (<http://www.cdc.gov/nchs/nvss.htm>, WebLink 21.1).

the most common causes of death in the future have been made (Fig. 21.2). According to the World Health Organization, the four leading causes of death globally in 2030 are projected to be ischemic heart disease, stroke, HIV/AIDS, and chronic obstructive pulmonary disease (Mathers and Loncar, 2006). Tobacco is projected to kill 50% more people than HIV/AIDS in 2015 and will be responsible for 10% of all deaths.

Another approach to describing the scope of human disease is to measure the global burden of disease in terms of the percentage of affected individuals or in terms of disability-adjusted life years (DALYs; Murray *et al.*, 2012). DALYs are a summary metric of population health, consisting of years of life lost due to premature mortality and years



**FIGURE 21.2** Projected global deaths for selected causes of death, 2002–2030. Redrawn from the World Health Organization (World Health Statistics 2007, <http://www.who.int/whosis/whostat2007.pdf>). Reproduced with permission from World Health Organization.



**FIGURE 21.3** Percentage of global disability-adjusted life years (DALY) for various causes in 2010. Data are for females; results for males (not shown) are similar. Redrawn from Murray *et al.* (2012). Reproduced with permission from Elsevier.

lived with disability. The causes of DALYs change across the lifespan (Fig. 21.3), with major differences across geographic locations. The causes of DALYs change over time; between 1990 and 2010 the rank of some disorders increased dramatically (e.g., HIV/AIDS, major depressive disorder, diabetes, low back pain) while others decreased (e.g., measles, meningitis, protein-energy malnutrition, tuberculosis).

A far more extensive listing of morbidity data is provided by the International Statistical Classification of Diseases and Related Health Problems (abbreviated ICD). This resource, published by the World Health Organization (WHO), is used to classify diseases (Table 21.3). It provides a standard for coding patients at most hospitals.

Mortality statistics list the most common diseases. We are interested in the full spectrum of disease, including rare diseases. These are defined as diseases affecting fewer than 200,000 people. In the United States, an estimated 25 million individuals (almost 10% of the population) suffer from one or more of 7000 rare diseases.

### NIH Disease Classification: MeSH Terms

The National Library of Medicine (NLM) has developed Medical Subjects Heading (MeSH) terms as a unified language for biomedical literature database searches. The current MeSH term system includes 23 disease categories (Fig. 21.4). PubMed at NCBI also uses this classification system for indexing articles.

MeSH terms are a controlled vocabulary thesaurus used to index MEDLINE (and PubMed, which is based on MEDLINE). A search for the term "Sturge-Weber syndrome" results in a description of that syndrome and a list of MeSH subheadings. Selecting one such as "genetics" allows you to use a PubMed Search Builder to create a PubMed query, "Sturge-Weber Syndrome/genetics" [Mesh]. This PubMed search is directed by the

A summary of the Global Burden of Disease findings is available at <http://www.thelancet.com/themed/global-burden-of-disease> (WebLink 21.2). DALYs are calculated by adding the years of life lost through all deaths in a year plus the years of life expected to be lived with a disability for all cases beginning in that year. The DALYs metric was introduced in the 1990 Global Burden of Disease study (Murray and Lopez, 1996).

The WHO ICD website is at <http://www.who.int/classifications/icd/en/> (WebLink 21.3). This resource was begun in 1893 as the International List of Causes of Death.

The Office of Rare Diseases at the National Institutes of Health (NIH) has a website that serves as a portal to information on rare diseases (<http://rarediseases.info.nih.gov/>, WebLink 21.4).

**TABLE 21.3 ICD Classification System (ICD-10 Version 2015).**

I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional, and metabolic diseases
V	Mental and behavioral disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth, and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations, and chromosomal abnormalities
XVIII	Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified
XIX	Injury, poisoning, and certain other consequences of external causes
XX	External causes of morbidity and mortality
XXI	Factors influencing health status and contact with health services
XXII	Codes for special purposes

Source: <http://apps.who.int/classifications/icd10/browse/2015/en>.

MeSH terms you select. The MeSH site also shows the hierarchical tree structure of the MeSH terms relevant to Sturge-Weber syndrome.

We have used EDirect to query NCBI Entrez databases on the command line (Chapter 2); we can do that now to explore MeSH. For information about MeSH, such as the number of records it contains and the fields you can search, try the following:

```
$ einfo -db mesh
```

You can begin constructing a query with a general search term such as disease:

```
$ esearch -db mesh -query "disease"
```

We next use as an example a paper from my lab by Shirley *et al.* (2013) reporting a mutation that causes a rare disease (Sturge-Weber syndrome) as well as a commonly occurring port-wine stain birthmark. The PubMed identifier (given in the reference list at the end of the chapter) is 23656586. MeSH database limiters include [MESH] (for all MeSH terms), [MAJR] (for MeSH major topics), and [SUBH] for MeSH subheadings. What are the MeSH headings and subheadings of the Shirley *et al.* paper? The main steps include `efetch` to download the PubMed record in the XML format, `xtract` to convert the XML into a table of values, the `-block` statement to explore each MeSH heading in the XML file for that PubMed entry, and the use of the UNIX stream editor called `sed` to format the output and add an asterisk to each major heading. You can perform this query

You can access the MeSH system at NLM (<http://www.nlm.nih.gov/mesh/MBrowser.html>, WebLink 21.5) or at NCBI (from PubMed, select MeSH terms, then enter a query such as “disease”).

1. + Anatomy [A]
2. + Organisms [B]
3. + Diseases [C]
4. + Chemicals and Drugs [D]
5. + Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. + Psychiatry and Psychology [F]
7. + Phenomena and Processes [G]
8. + Disciplines and Occupations [H]
9. + Anthropology, Education, Sociology and Social Phenomena [I]
10. + Technology, Industry, Agriculture [J]
11. + Humanities [K]
12. + Information Science [L]
13. + Named Groups [M]
14. + Health Care [N]
15. + Publication Characteristics [V]
16. + Geographicals [Z]

**FIGURE 21.4** The Medical Subject Heading (MeSH) term system at the National Library of Medicine includes 16 major categories (2015 version, upper panel). The disease category further includes the 26 headings shown in the lower panel.

Source: Medical Subject Heading (MeSH), NLM <http://www.nlm.nih.gov/mesh/>.

- Diseases [C]
  - [Bacterial Infections and Mycoses \[C01\]](#) +
  - [Virus Diseases \[C02\]](#) +
  - [Parasitic Diseases \[C03\]](#) +
  - [Neoplasms \[C04\]](#) +
  - [Musculoskeletal Diseases \[C05\]](#) +
  - [Digestive System Diseases \[C06\]](#) +
  - [Stomatognathic Diseases \[C07\]](#) +
  - [Respiratory Tract Diseases \[C08\]](#) +
  - [Otorhinolaryngologic Diseases \[C09\]](#) +
  - [Nervous System Diseases \[C10\]](#) +
  - [Eye Diseases \[C11\]](#) +
  - [Male Urogenital Diseases \[C12\]](#) +
  - [Female Urogenital Diseases and Pregnancy Complications \[C13\]](#) +
  - [Cardiovascular Diseases \[C14\]](#) +
  - [Hemic and Lymphatic Diseases \[C15\]](#) +
  - [Congenital, Hereditary, and Neonatal Diseases and Abnormalities \[C16\]](#) +
  - [Skin and Connective Tissue Diseases \[C17\]](#) +
  - [Nutritional and Metabolic Diseases \[C18\]](#) +
  - [Endocrine System Diseases \[C19\]](#) +
  - [Immune System Diseases \[C20\]](#) +
  - [Disorders of Environmental Origin \[C21\]](#) +
  - [Animal Diseases \[C22\]](#) +
  - [Pathological Conditions, Signs and Symptoms \[C23\]](#) +
  - [Occupational Diseases \[C24\]](#) +
  - [Chemically-Induced Disorders \[C25\]](#) +
  - [Wounds and Injuries \[C26\]](#) +

on a Mac operating system using its terminal (or in Cygwin on a PC); using a Linux operating system type `$ man` to learn more about a utility such as `sed` (i.e., type `man sed`).

```
$ efetch -db pubmed -id 23656586 -format xml | xtract -pattern
PubMedArticle -tab " " -element MedlineCitation/PMID -block
MeshHeading -pxf "\n|" -sep "|" -tab " " -element DescriptorName@
MajorTopicYN,DescriptorName -subset QualifierName -pxf "/" -sep " | "
-tab " " -element "@MajorTopicYN,QualifierName" | sed -e 's/|N//g' -e
's/|Y|/*/g'
23656586 # the start of the output lists the PubMed ID
Brain /pathology
Female
GTP-Binding Protein alpha Subunits /*genetics
Humans
Infant, Newborn
Magnetic Resonance Imaging
Male
*Mutation
Port-Wine Stain /*genetics
Sequence Analysis, DNA
Sturge-Weber Syndrome /*genetics
```

The script is available in text format at Web Document 21.2. This example is adapted from the online documentation for EDirect at <http://www.ncbi.nlm.nih.gov/books/NBK179288/> (WebLink 21.6). You can also use EDirect to count the number of MeSH disease entries with the command:

```
$ esearch -db pubmed
-query "disease [MESH]"
```

The result matches a web-based PubMed search for `disease[MESH]`.

Major MeSH categories therefore include genetics and mutation.

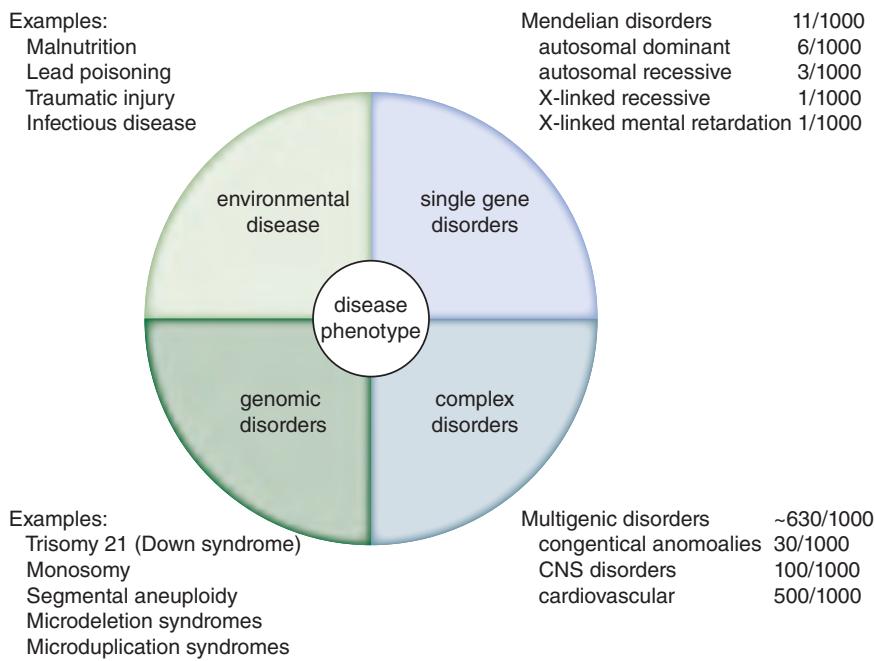
## CATEGORIES OF DISEASE

Pathology is the study of the nature and cause of disease. Pathophysiology is the study of how disease alters normal physiological processes.

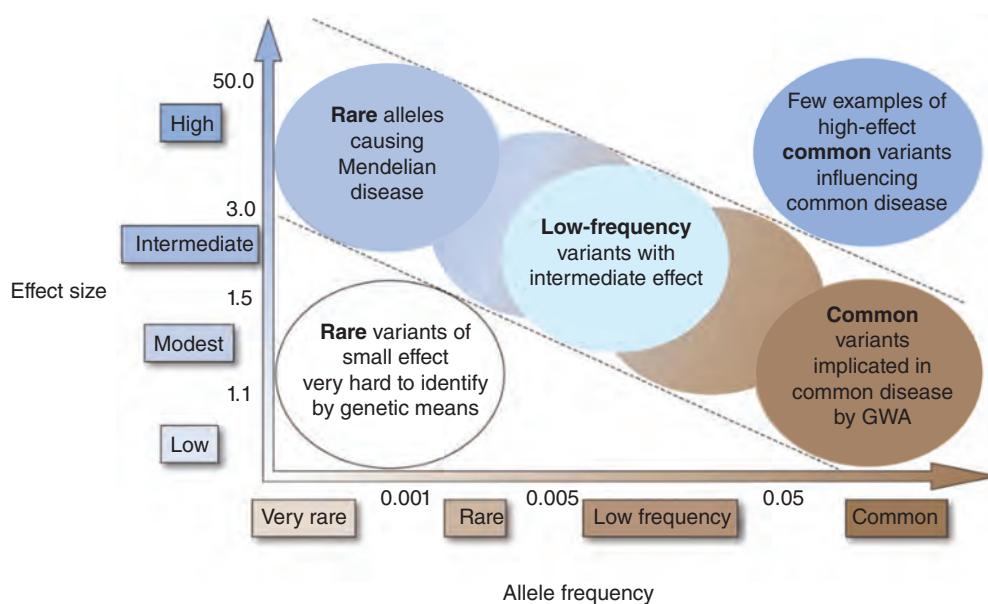
What kinds of diseases afflict humans? We can describe four main categories: single-gene (monogenic) disease, complex disease, genomic disease, and environmental disease (Fig. 21.5). Additionally there are somatic diseases (such as cancer), and mitochondrial disease. All these categories are interconnected in many ways, as we discuss next. Consistent with Garrod's perspective, the pathophysiology of any disease may be considered multigenic. Two individuals who are exposed to the same disease-causing stimulus – whether it is a virus or lead paint or a mutated gene – may have entirely different reactions. One person may become ill, while the other is unaffected. There is a large genetic component to the responses to any disease-causing condition.

### Allele Frequencies and Effect Sizes

As we begin thinking about disease we can consider two properties of disease alleles, illustrated in a plot from Manolio *et al.* (2009) based on an earlier version by McCarthy *et al.* (2008; Fig. 21.6). The first property is allele frequencies, shown on the *x* axis. These range from common (often defined as  $\geq 5\%$  minor allele frequency (MAF)) to low frequency ( $< 5\%$  MAF), rare ( $< 0.5\%$  MAF), or very rare ( $< 0.1\%$  MAF). The HapMap Project and the 1000 Genomes Project have helped to catalog millions of variants and reported each of their allele frequencies. A variety of SNP-based and sequencing-based approaches (described in “Approaches to Identifying Disease-Associated Genes and Loci” below) have shown which of these variants are likely to be pathogenic or neutral.



**FIGURE 21.5** Human disease can be categorized based on the cause. These include single-gene or monogenic disorders (caused primarily by mutations in a single gene; examples include phenylketonuria and sickle cell anemia); complex disorders (having mutations in two or more genes, such as cancer or schizophrenia); genomic disorders (such as Down syndrome, involving chromosomal abnormalities); and environmental disease (including infectious disease). The values for the incidence of these disorders are only approximate estimates. The four quadrants of the circle are not intended to reflect incidence. Overall, complex disorders are far more common than single-gene disorders. However, it is far easier to discover the genetic defect that underlies single-gene disorders. For all categories of disease, the pathophysiology (i.e., the disease-altered physiological processes) depends on the influence of many genetic and environmental factors.



**FIGURE 21.6** Risk allele frequencies (*x* axis) and strength of genetic effect (odds ratio; *y* axis) determine the feasibility of identifying disease-associated genetic variants. Most emphasis has been on the region within the dotted lines. Redrawn from Manolio *et al.* (2009). Reproduced with permission from Macmillan Publishers.

A second property of disease-associated variants is the effect size (*y* axis). This may be quantitated as an odds ratio (OR; Szumilas, 2010). An OR is a measure of association between an exposure (in our case a genetic variant) and an outcome (expression of a disease). An OR of 1 implies that the presence of a variant does not affect the odds of a disease outcome;  $OR > 1$  implies an association with higher odds of a disease occurrence.

A major goal of human genetics and genomics is to identify variants that cause disease (or confer risk for disease). We can focus on the region within the dashed lines of Figure 21.6. Rare alleles having large effects sizes tend to cause Mendelian diseases that are primarily monogenic (see figure, upper left). Low-frequency alleles tend to have effects that are less strong (see center of figure). Some common alleles have a low effect size yet still contribute to common disease (lower right of figure). Such common alleles have been captured by genome-wide association studies (GWAS; see “Genome-Wide Association Studies” below). There are very few examples of common variants that have large effects in contributing to common diseases (upper right). Rare variants having small effects (lower left) can be extremely difficult to identify.

We next describe several categories of disease. Considerations of the allele frequencies and effect sizes further impacts the choice of experimental approach used to study the causes of disease. For example, rare and very rare variants may be studied more effectively with whole-genome and/or whole-exome sequencing approaches rather than with SNP arrays which target common alleles. Both allele frequencies and effect sizes impact the sample size required to achieve statistical power for studies of disease-associated variation.

## Monogenic Disorders

Our perspectives on the molecular nature of disease have evolved in recent decades. Previously, geneticists recognized a dichotomy between simple traits and complex traits. More recently, all traits have come to be appreciated as part of a continuum. Simple traits are transmitted following the rules of Mendel. Online Mendelian Inheritance in Man (OMIM) database currently lists over 5000 phenotypes for which the molecular

**TABLE 21.4 Examples of monogenic disorders. Adapted from Beaudet *et al.* (2001) with permission from McGraw Hill.**

Mechanism	Disorder	Frequency
Autosomal dominant	BRCA1 and BRCA2 breast cancer	1 in 1000 (1 in 100 for Ashkenazim)
	Huntington chorea	1 in 2500
	Neurofibromatosis I	1 in 3000
	Tuberous sclerosis	1 in 15,000
Autosomal recessive	Albinism	1 in 10,000
	Sickle cell anemia	1 in 655 (US African-Americans)
	Cystic fibrosis	1 in 2500 (Europeans)
	Phenylketonuria	1 in 12,000
X linked	Hemophilia A	1 in 10,000 (males)
	Glucose 6-phosphate dehydrogenase deficiency	Variable; up to 1 in 10 males
	Fragile X syndrome	1 in 1250 males
	Color blindness	1 in 12 males
	Rett syndrome	1 in 20,000 females
	Adrenoleukodystrophy	1 in 17,000

We examined the structure of normal beta globin (HBB) as well as the most common mutated form (HBS) in Chapter 13. The E7V substitution (valine in place of glutamate as the seventh amino acid) adds a hydrophobic patch to the protein, promoting the aggregation of globin molecules and the formation of sickle-shaped red blood cells. Sickle cell anemia is unusually common for a single-gene disorder. This is presumably because of the protection it confers to heterozygotes exposed to malaria (Box 21.1).

Read the Pauling *et al.* (1949) article online at <http://profiles.nlm.nih.gov/MM/B/B/R/L/> (WebLink 21.7). The National Library of Medicine (NLM) offers online access to all the publications of several prominent biologists through its Profiles in Science site (<http://profiles.nlm.nih.gov/>, WebLink 21.8). The scientists include Linus Pauling and other Nobel Prize laureates such as Barbara McClintock, Julius Axelrod, and Oswald Avery.

basis is known. (Here a phenotype refers to single-gene Mendelian disorders, traits, some susceptibilities to complex disease, and some somatic conditions.) While each Mendelian disease tends to be rare in the population, cumulatively these conditions affect at least 1% of liveborn infants (Costa *et al.*, 1985). Over 90% of these disorders manifest by the age of puberty. Single-gene disorders are estimated to affect 25–30 million individuals in the United States (Cutting, 2014).

Several monogenic disorders are listed in Table 21.4. As an example of a single-gene disorder, consider sickle cell anemia (Box 21.1). In 1949 Linus Pauling and colleagues described the abnormal electrophoretic behavior of sickle cell hemoglobin (Pauling *et al.*, 1949). It was subsequently shown that a single amino acid substitution accounts for the abnormal behavior of the sickle cell and is the basis of sickle cell anemia. This is a single-gene disorder that is inherited in an autosomal recessive fashion. Single-gene disorders tend to be rare in the general population. Note that sickle cell disease is the outcome of having a particular mutant hemoglobin protein. While there are common features of sickle cell disease, such as sickling of the red blood cells, there is not a single disease phenotype. The pleiotropic phenotype is caused by the influence of other genes.

Rett syndrome is another example of a single-gene disorder (Katz *et al.*, 2012; Box 21.2). This disease affects girls almost exclusively. While they are apparently born healthy, Rett syndrome girls acquire a constellation of symptoms beginning at 6–18 months of age. They lose the ability to make purposeful hand movements, and they typically exhibit hand-wringing behavior. Whatever language skills they have acquired are lost, and they may display autistic-like behaviors. Rett syndrome is caused by mutations in the gene encoding MeCP2, a transcriptional repressor that binds methylated CpG islands (Amir *et al.*, 1999). It is not yet known why mutations affecting a transcriptional repressor that functions throughout the body cause a primarily neurological disorder.

While Rett syndrome is a disease caused by a mutation in a single gene, it exemplifies the extraordinary complexity of human disease and even monogenic disorders:

- The disease occurs primarily in females. It was thought that this could be explained by the location of the *MECP2* gene on the X chromosome: a mutation in this gene might be lethal for males in utero (having only a single X chromosome), while females might have the disease phenotype because they have one normal and one mutant

### BOX 21.1 SICKLE CELL ANEMIA AND THALASSEMIAS

Our cells depend on oxygen to live, and blood transports oxygen throughout the body. However, oxygen is a hydrophobic molecule that requires the carrier protein hemoglobin to transport it through blood. (The homologous protein myoglobin transports oxygen in muscle cells.) Adult hemoglobin is composed of two  $\alpha$  chains and two  $\beta$  chains. Other  $\alpha$ - and  $\beta$ -type chains are used at different developmental stages, such as  $\alpha_2/\gamma_2$  in fetal hemoglobin and  $\alpha_2/\epsilon_2$  in embryonic hemoglobin. Mutation in the  $\beta$  chain (NM\_000518 and NP\_000509) on chromosome 11p15.5 causes sickle cell anemia (OMIM 603903). Red blood cells in patients can assume a curved, “sickled” appearance that reflects hemoglobin aggregation in the presence of low oxygen levels.

Sickle cell anemia is the most common inherited blood disorder in the United States, affecting 1 in 500 African-Americans. It is inherited as an autosomal recessive disease. Heterozygotes (individuals with one normal copy of hemoglobin beta and one mutant copy; the HBS mutation) are somewhat protected against the malaria parasite, *Plasmodium falciparum*. This may be because normal red blood cells infected by the parasite are destroyed. There is therefore a selective evolutionary pressure to preserve the HBS mutation in the population that is at risk for malaria.

Red blood cells closely regulate the proportions of  $\alpha$  and  $\beta$  globin that are produced, as well as the heme moiety that is inserted into the globin tetramer to form hemoglobin. The absence of the  $\beta$  chain causes beta-zero-thalassemia, while the production of reduced amounts of  $\beta$  globin causes beta-plus-thalassemia. Reduced levels of  $\alpha$  globin cause alpha thalassemias. Thalassemia can cause severe anemia, in which hemoglobin levels are low.

Web resources for sickle cell disease include an NIH fact sheet (<http://www.nhlbi.nih.gov/health/health-topics/topics/sca/>), Genes and Disease at NCBI (<http://www.ncbi.nlm.nih.gov/books/NBK22183/>), and the Sickle Cell Disease Association of America (<http://www.sicklecelldisease.org/>).

copy of the gene. Instead, the more likely explanation is that most mutations occur in fathers. The father is healthy, but a new germline mutation arises and is passed to a daughter. Thus all sons (XY) receive a normal Y chromosome from the father, while a daughter may receive a mutant copy of the X chromosome from the father.

- After the discovery that mutations in *MECP2* cause Rett syndrome, it was discovered that some males with intellectual disability also have mutations in this gene (Hammer *et al.*, 2002; Zeev *et al.*, 2002). However, the phenotype of mutations in the male is distinctly different than in females, often involving severe neonatal encephalopathy. In males, having a single X chromosome means that the mutant gene is expected to adversely affect virtually every cell in the body. In contrast, females undergo random X-chromosome inactivation. Having two copies of the X chromosome, every cell expresses only one chromosome (either the maternal or paternal chromosome, randomly selected early in development). Females are therefore a mosaic in terms of

### BOX 21.2 RETT SYNDROME

Rett syndrome (RTT; OMIM #312750) is a developmental neurological syndrome that occurs almost exclusively in females (Hagberg *et al.*, 1983). Affected females are apparently normal through pre- and perinatal development, following which there is a developmental arrest. This is accompanied by decelerated head and brain growth, loss of speech and social skills, severe intellectual disability, truncal ataxia, and characteristic hand-wringing motions. Prominent neuropathological features include reductions in cortical thickness in multiple cerebral cortical regions, reduced neuronal soma size, and dramatically decreased dendritic arborization (Bauman *et al.*, 1995).

Mutations in the methyl-CpG-binding protein 2 (*MECP2*) gene located in Xq28 have been found in most cases of RTT (Amir *et al.*, 1999, 2000). MeCP2 binds to methylated CpG dinucleotides throughout the genome and is involved in methylation-dependent repression of gene expression via the recruitment of the corepressor mSin3A and the chromatin-remodeling histone deacetylases HDAC1 and HDAC2. The expression of MeCP2 mRNA in many tissues and its interaction with regulatory DNA elements in multiple chromosomes suggest that MeCP2 is a global repressor of gene expression (Nan *et al.*, 1997). DNA methylation-dependent repression of gene expression has been associated with genetic imprinting, X-chromosome inactivation, carcinogenesis, and tissue-specific gene expression.

Why is it that some tissues are spared the effects of *MECP2* mutations? There could be tissue-specific redundancy of gene function, or other compensation mechanisms. This provides an example of how the tools of bioinformatics are relevant to studying many different aspects of human disease.

X-chromosome allelic expression, and a Rett syndrome female has on average 50% normal cells throughout her body.

- MECP2 duplication syndrome is 100% penetrant in males and causes symptoms including infantile hypotonia, severe to profound intellectual disability, autism or autistic features, and poor speech development (Ramocki *et al.*, 2010).
- While Rett syndrome is caused by mutations in a gene encoding a transcriptional repressor, it is almost certain that the consequence of this mutation involves subsequent effects on the expression of many other genes. Like any other monogenic disorder, many other genes are involved and may influence the phenotype of the disease.
- Two females having the identical mutation in *MECP2* may have entirely different phenotypes (in terms of severity of the disease). There are two main explanations for this observation, which is also seen for many other single-gene disorders. (1) There may be modifier genes that influence the disease process (Dipple and McCabe, 2000). Modifier genes have been identified for patients with sickle cell anemia, adrenoleukodystrophy, cystic fibrosis, and Hirschsprung disease. Most (if not all) apparently monogenic disorders are complex. (2) A variety of epigenetic influences may drastically affect the clinical phenotype. For example, the methylation status of genomic DNA could determine the molecular consequences of mutations in *MECP2*. X chromosome inactivation is sometimes skewed, such that the phenotype is more severe (if the X chromosome copy with mutant *MECP2* is preferentially expressed) or less severe (if the healthy X chromosome is selectively expressed).
- While the disorder is neurodevelopmental, introduction of a mutation that deletes the protein beginning in adulthood recapitulates the germline knockout phenotype (McGraw *et al.*, 2011). The developmental effects of single-gene disorders are often complex.

## Complex Disorders

Complex disorders such as Alzheimer's disease and cardiovascular disease are caused by defects in multiple genes. These disorders are also called multifactorial, reflecting that they are expressed as a function of both genetic and environmental factors. In comparison to monogenic disorders, complex disorders tend to be highly prevalent (Todd, 2001). These traits do not segregate in a simple, discrete, Mendelian manner. Examples are asthma, autism (Box 21.3), depression, diabetes, high blood pressure, obesity, and osteoporosis. In the United States, chronic diseases such as heart disease, senile dementia, cancer, and diabetes are leading causes of death and disability. These all have some degree of genetic basis.

Complex disorders are characterized by the following features:

- Multiple genes are thought to be involved. It is the combination of mutations in multiple genes that defines the disease. In single-gene disorders, even if there are modifying loci, one gene has a dramatic influence on the disease phenotype.
- Complex diseases involve the combined effect of multiple genes, but they are also caused by both environmental factors and behaviors that elevate the risk of disease.
- Complex diseases are non-Mendelian: they show familial aggregation but not segregation. For example, autism is a highly heritable condition (if one identical twin is affected, there is a very high probability that the other is also affected).
- Susceptibility alleles have a high population frequency, that is, complex diseases are generally more frequent than single-gene disorders. Sickle cell anemia, a single-gene disorder, is unusually frequent in the African-American population, but the heterozygous condition confers a selective advantage (see Box 21.1).
- Susceptibility alleles have low penetrance. Penetrance is the frequency with which a dominant or homozygous recessive gene produces its characteristic phenotype in a population. At the extremes, it is an all-or-none phenomenon: a genotype is either expressed or it is not. In complex disorders, partial penetrance is common.

A quantitative trait locus (QTL) is an allele that contributes to a multifactorial disease.

Penetrance is the frequency of manifestation of a hereditary condition in individuals. Having the genotype for a disease does not imply that the phenotype will occur, especially if multiple genes have modifying effects on the presentation of the phenotype.

### BOX 21.3 AUTISM: COMPLEX DISORDER OF UNKNOWN ETIOLOGY

Autism (OMIM #209850) is a lifelong neurological disorder with onset before three years of age (Kanner, 1943; reviewed in Rapin, 1997). It is characterized by a triad of deficits: (1) an individual's failure to have normal reciprocal social interaction; (2) impaired language or communication skills; and (3) restricted, stereotyped patterns of interests and activities. Autistic children's play is abnormal beginning in infancy, and there is a notable lack of imaginative play. Approximately 30% of autistic children appear to develop normally, but then undergo a period of regression in language skills between 18 and 24 months of age. In addition, cognitive function may be impaired. Seventy-five percent of autistic individuals have intellectual disability. Approximately 10% of autistic individuals have savant-like superior abilities in areas such as mathematical calculation, rote memory, or musical performance. Autism is accompanied by seizures; by adulthood about one-third of autistic individuals will have had at least two unprovoked seizures (Olsson *et al.*, 1988; Volkmar and Nelson, 1990; Rossi *et al.*, 1995).

In the 1990s, the prevalence of autism was estimated to be between 0.2 and 2 per 1000 individuals (Smalley *et al.*, 1988; Rapin and Katzman, 1998; Fombonne, 1999; Gillberg and Wing, 1999). More recently, the prevalence has been estimated to be about 1:68. However, the definition of autism has broadened considerably in recent years, with a large number of patients formerly defined as having intellectual disability now diagnosed as having autism or autism spectrum disorder. About three to four times more males are affected than females (Fombonne, 1999).

The cause of autism is unknown, but there is strong evidence that the disorder is genetic (Smalley *et al.*, 1988; Szatmari *et al.*, 1998; Turner *et al.*, 2000). The concordance between monozygotic twins is approximately 60%, and >90% if coaffected twins are defined as having classically defined autism or more generalized impairments in social skills, language, and cognition (Bailey *et al.*, 1995). Autism has a far stronger genetic basis than most other common neuropsychiatric disorders such as schizophrenia or depression. Linkage, GWAS, and exome sequencing studies suggest that there is extreme locus heterogeneity: there are very few penetrant variants having a large effect on the phenotype.

## Genomic Disorders

Large-scale chromosomal abnormalities are extremely common causes of disease in humans. Lupski (1998) defined genomic disorders as those changes in the structure of the genome that cause disease. Some genomic disorders involve large-scale changes such as aneuploidies in which a chromosomal copy is gained (trisomy) or lost (monosomy). More rarely, two copies are gained (tetrasomy) or lost (nullsomy). Trisomies 13 (Patau syndrome), 18 (Edwards syndrome), and 21 (Down syndrome) are the only autosomal trisomies that are compatible with life (Table 21.5). Of these, trisomies 13 and 18 are typically fatal in the first years of life. A variety of X chromosome aneuploidies are compatible with life.

Many developmental abnormalities involve a portion of a chromosome. Some involve cytogenetically detectable changes and span millions of base pairs. If they are too small to be cytogenetically visible (e.g., smaller than about 3 Mb) they are usually referred to as cryptic changes. Examples of microdeletion syndromes include Cri-du-chat syndrome, Angelman syndrome, Prader Willi syndrome, Smith-Magenis syndrome, and various forms of intellectual disability that result from the gain (microduplication) or loss (microdeletion) of chromosomal regions. Table 21.6 lists

An aneuploidy is the condition of having an abnormal number of chromosomes. Segmental aneuploidy affects a portion of a chromosome.

**TABLE 21.5 Frequency of chromosomal aneuploidies among liveborn infants.**

Abnormalities	Disorder	Frequency
Autosomal	Trisomy 13 (Patau syndrome)	1 in 15,000
	Trisomy 18 (Edwards syndrome)	1 in 5000
	Trisomy 21 (Down syndrome)	1 in 600
Sex chromosome	Klinefelter syndrome (47,XXY)	1 in 700 males
	XYY syndrome (47, XYY)	1 in 800 males
	Triple X syndrome (47, XXX)	1 in 1000 females
	Turner syndrome (45, X or 45X/46XX or 45X/46, XY or isochromosome Xq)	1 in 1500 females

Source: Beaudet *et al.* (2001) with permission from McGraw Hill.

**TABLE 21.6 Examples of Mendelian genomic disorders.** AD: autosomal dominant; AR: autosomal recessive; del: deletion; inv/dup: inversion/duplication; OMIM: Online Mendelian Inheritance in Man; Orientation I: inverted; XL: X chromosome linked; G: gene;  $\Psi$ : pseudogene; S: segment of genome.

Disorders	OMIM	Inheritance pattern	Chrom. location	Gene(s)	Rearrangement			Recombination substrates		
					Type	Size (kb)	Repeat size (kb)	Identity %	Orientation	Type
Bartter syndrome type III	601678	AD	1p36	CLCNKA/B	del	11		91	D	$G/\Psi$
Gaucher disease	230800	AR	1q21	GBA	del	16			D	$G/\Psi$
Spinal muscular atrophy	253300	AR	5q13.2	SMN	inv/dup	500			I	
$\beta$ -thalassemia	141900	AR	11p15.5	$\beta$ -globin	del	4 (7?)			D	G
$\alpha$ -thalassemia	141800		16p13.3	$\alpha$ -globin	del	3.7 or 4.2			D	S
Polyzystic kidney disease 1	601313	AD	16p13.3	PKD1			50	95		
Charcot–Marie–Tooth (CMT1A)	118220	AD	17p12	MPZP22	dup	1400		98.7	D	S
Neurofibromatosis type 1	162200	AD	17q11.2	NF1	del	1500			D	G
Huntington syndrome (mucopolysaccharidosis type II)	309900	XL	Xq28	IDS	inv/del	20	3	>88		$G/\Psi$
Hemophilia A	306700	XL	Xq28	FB	inv	300-500	9.5	99.9	I	

Source: Stankiewicz and Lupski (2002). Reproduced with permission from Elsevier.

**TABLE 21.7 Common structural polymorphisms and disease. VNTR: variable number tandem repeats.**

Gene	Type	Locus	Size (kb)	Phenotype	Copy number variation
UGT2B17	Deletion	4q13	150	Variable testosterone levels, risk of prostate cancer	0–2
DEFB4	VNTR	8p23.1	20	Colonic Crohn's disease	2–10
FCGR3	Deletion	1q23.3	>5	Glomerulonephritis, systemic lupus erythematosus	0–14
OPN1LW/ OPN1MW	VNTR	Xq28	13–15	Red/green color blindness	0–4/0–7
LPA	VNTR	6q25.3	5.5	Altered coronary heart disease risk	2–38
CCL3L1/ CCL4L1	VNTR	17q12	Not known	Reduced HIV infection; reduced AIDS susceptibility	0–14
RHD	Deletion	1p36.11	60	Rhesus blood group sensitivity	0–2
CYP2A6	Deletion	19q13.2	7	Altered nicotine metabolism	2–3

Source: Human Genome Structural Variation Working Group (2007). Reproduced with permission from Macmillan Publishers.

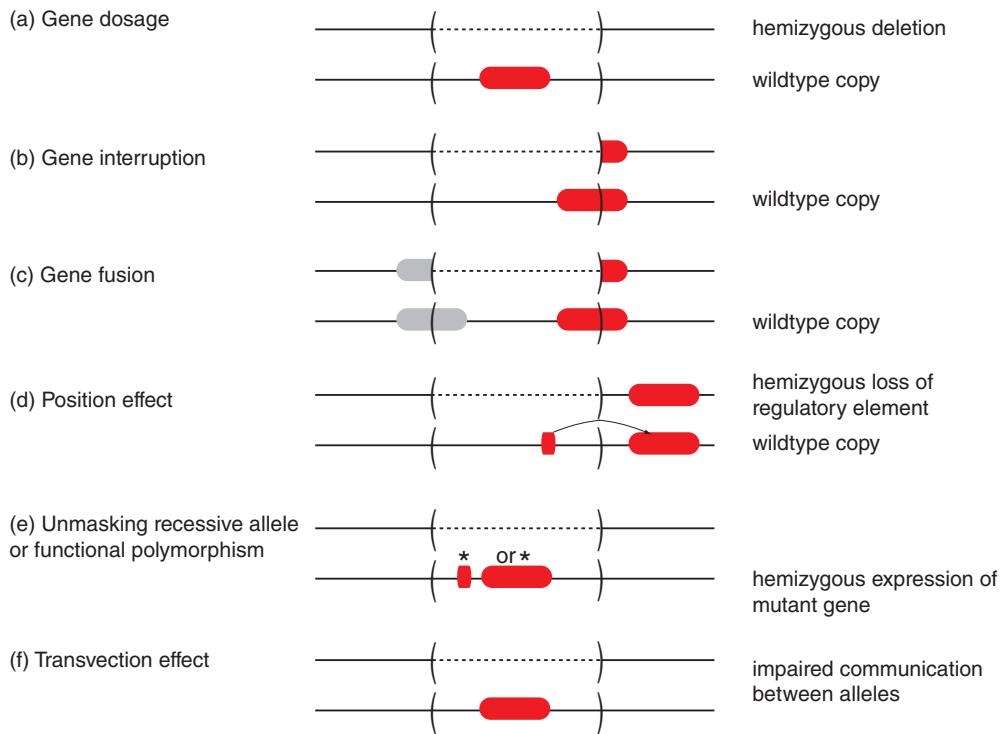
examples of genomic disorders that are inherited in a Mendelian fashion and involve only one or several genes (Stankiewicz and Lupski, 2002). Table 21.7 provides a similar list of common structural variations that are associated with disease, reported by the Human Genome Structural Variation Working Group *et al.* (2007). That group reported an initiative to characterize structural variation in phenotypically normal individuals using fosmid libraries.

We considered several mechanisms by which non-allelic homologous recombination causes deletions or duplications of chromosomal segments (see Fig. 8.19). Figure 21.7 shows six possible consequences of such events, such as loss of normal gene function, the fusion of two genes, or the exposure of a recessive allele.

Consistent with our view of allele frequencies in Figure 21.6, chromosomal alterations may be considered to occur along a spectrum from having little or no adverse effects to causing disease (Fig. 21.8). Copy number variants (described in Chapters 8 and 20) may have no phenotypic consequences and may be thought of as chromosomal alterations (in contrast to chromosomal abnormalities). Some copy number variants may increase disease susceptibility, perhaps contributing to common complex (multigenic) disorders. Some common and relatively benign traits such as color blindness can be attributed to copy number variants. At the extreme end of the spectrum, chromosomal changes may cause or contribute to a variety of genomic disorders including aneuploidies, microdeletion syndromes, and microduplication syndromes. Genomic disorders are also notably common in cancers, with occurrence of amplifications and deletions of loci. We discuss cancer in more detail in “Cancer: A Somatic Mosaic Disease” below.

Chromosomal disorders are an extremely common feature of normal human development. Humans have a very low fecundity even relative to other mammals, with perhaps 50–80% of all human conceptions resulting in miscarriage. This low fecundity is

The database of chromosomal imbalance and phenotype in humans using Ensembl resources (DECIPHER) is a major database resource for genomic disease. It is available at <http://decipher.sanger.ac.uk/> (WebLink 21.9).

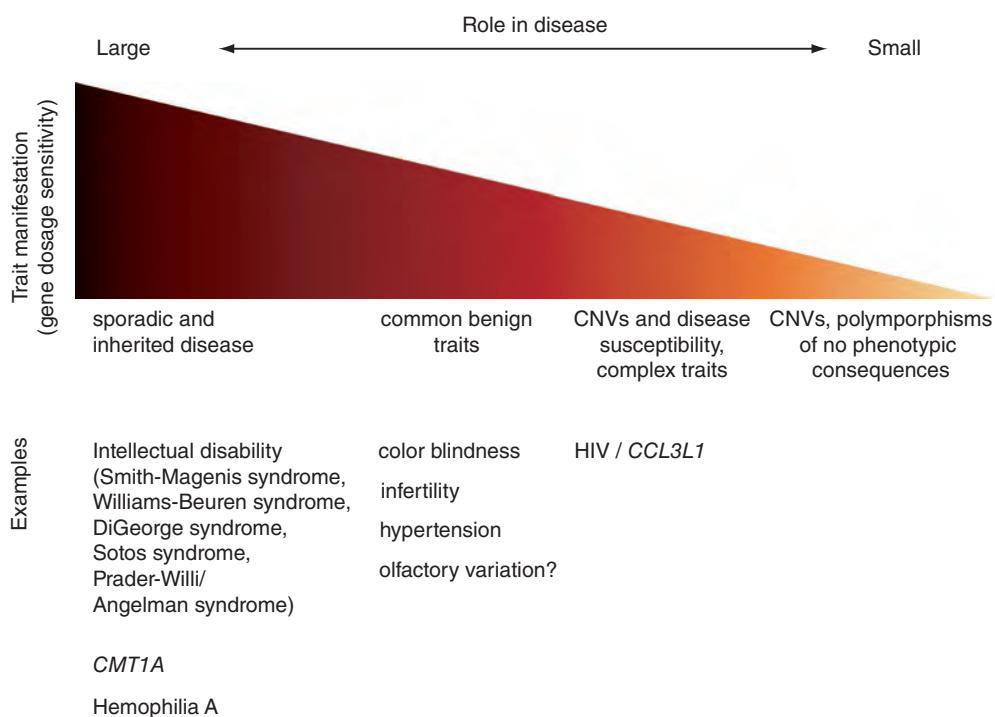


**FIGURE 21.7** Models for the molecular mechanisms of genomic disorders. For each, a hemizygous deletion is depicted (i.e., loss of one of the normal two copies of an allele) in brackets, and the two chromosomal homologs are indicated by horizontal lines. Note that duplications also potentially cause disease; homozygous deletions (resulting in zero copies of a gene) typically have more severe consequences than hemizygous deletions. (a) Gene dosage effect in which one (of two) copies is deleted. Genes vary in their dosage sensitivity. (b) Gene interruption. A rearrangement breakpoint interrupts a gene. (c) Gene fusion in which two genes (and/or regulatory elements such as enhancers or promoters) are fused following a deletion. (d) Position effect: the expression or function of a gene near a breakpoint is disrupted by loss of a regulatory element. (e) Unmasking a recessive allele. The deletion results in hemizygous expression of a recessive mutation (asterisk) in a gene or a regulatory sequence. (f) Transvection, in which a deletion impairs communication between two alleles. Genes are indicated as red (or gray) filled ovals, while regulatory sequences are smaller ovals. Adapted from Lupski and Stankiewicz (2005), with permission from J. R. Lupski.

primarily due to the common occurrence of chromosomal abnormalities (Vouillaire *et al.*, 2000; Wells and Delhanty, 2000):

- A woman who has already had one child (and is therefore of established fertility) has only a 25% chance of achieving a viable pregnancy in any given menstrual cycle.
- 52% of all women that conceive have an early miscarriage.
- Following *in vitro* fertilization, pregnancies that are confirmed positive in the first two weeks result in miscarriage 30% of the time.
- Over 60% of spontaneous abortions that occur at 12 weeks gestation or earlier are aneuploid, suggesting that early pregnancy failures are likely due to lethal chromosome abnormalities.

A review of 36 published studies showed that of 815 human preimplantation embryos, only 177 (22%) were diploid (van Echten-Arends *et al.*, 2011). A total of 73% were mosaic, meaning that not all cells contain the same chromosomal constitution. The majority of these were diploid-aneuploid mosaic embryos, having one or more diploid cells as well as other cells that were haploid or polyploid for a particular chromosome. Mitotic errors could account for the high rate of chromosomal mosaicism.



**FIGURE 21.8** Spectrum of effects of copy number variants. At one extreme, copy number variants cause genomic diseases such as microdeletion and microduplication syndromes. At the other extreme, copy number variants have no known phenotypic effects and occur in the apparently normal population. For example, many of the 270 HapMap individuals (who are defined as normal although everyone is susceptible to some diseases) have hemizygous and homozygous deletions as well as extended tracts of homozygosity. Adapted from Lupsik and Stankiewicz (2005), with permission from J. R. Lupsik.

## Environmentally Caused Disease

Environmental diseases are extremely common. We may consider two types.

1. *Infectious diseases* are caused by a pathogen (such as a virus, bacterium, protozoan, fungus, or nematode). From birth to old age, infectious disease is a leading cause of death worldwide. We described common, vaccine-preventable diseases caused by viruses in Chapter 16 (Table 16.2) and by bacteria in Chapter 17 (Table 17.7), and we discussed fungal pathogens and a variety of protozoan pathogens in Chapters 18 and 19, respectively.
2. Many diseases or other conditions are not caused by an infectious agent. These include malnutrition (whether maternal, fetal, or in an independent individual), poisoning by toxicants such as lead or mercury, or injury.

Genome-wide association studies (GWAS; see below) have been used to compare the genotypes of large numbers of individuals who are susceptible to an infectious disease relative to controls (reviewed in Chapman and Hill, 2012). Markers with strong evidence of association have been identified for diseases such as HIV-1/AIDS, Hepatitis B and C, dengue, malaria, tuberculosis, and leprosy. In many cases there is only a small increase in risk, and the biological relevance of the variants (often in intergenic loci) is uncertain. In some cases the variant confers resistance by impairing the function of a receptor for a pathogen, as in the case of the *CCR5* for HIV-2, *FUT2* for norovirus, and *DARC* for *Plasmodium vivax* (a malaria pathogen).

About 8% of all children in the United States have blood levels that are defined as “alarming,” according to the Centers for Disease Control and Prevention.

For more information, see

✉ <http://www.cdc.gov/nceh/lead/> (WebLink 21.10).

GIDEON (Global Infectious Disease and Epidemiology Network) is a commercial database of infectious diseases available at ✉ <http://www.gideononline.com> (WebLink 21.11).

## Disease and Genetic Background

While we have presented four disease categories so far (monogenic, complex, genomic, and environmental), these are interrelated categories. If four children who have the same highly elevated blood lead levels due to lead poisoning are examined, they may display entirely different responses. One might be aggressive, another intellectually disabled, another hyperactive, and another might appear unaffected. Four individuals exposed to the same pathogen might have different responses. It is likely that the genetic background has a key role in responses to environmental insults, as suggested by GWAS. Similarly, four children who have the identical single base pair mutation in the *ABCD1* gene might have entirely different severities of adrenoleukodystrophy, or the identical mutation in *MECP2* may cause very different forms of Rett syndrome. Modifier genes are likely to be involved (highlighting the concept that monogenic disorders may be caused primarily by the abnormal function of a single gene yet they always involve multiple genes), and environmental factors are certain to have large roles in genetic diseases.

There are other ways of classifying basic disease types. For example, particular ethnic groups or other discrete groups have high susceptibility to some genetic diseases. Examples include the following:

- Tay-Sachs disease is prevalent among Ashkenazi Jews.
- About 8% of the African-American population are carriers of a mutant *HBB* gene.
- Males rather than females are susceptible to Alport disease, male pattern baldness, and prostate cancer.
- Cystic fibrosis affects ~30,000 people in the United States with ~12 million carriers, and is the most common fatal genetic disease in that country. While it affects all groups, Caucasians of northern European ancestry are particularly susceptible.

## Mitochondrial Disease

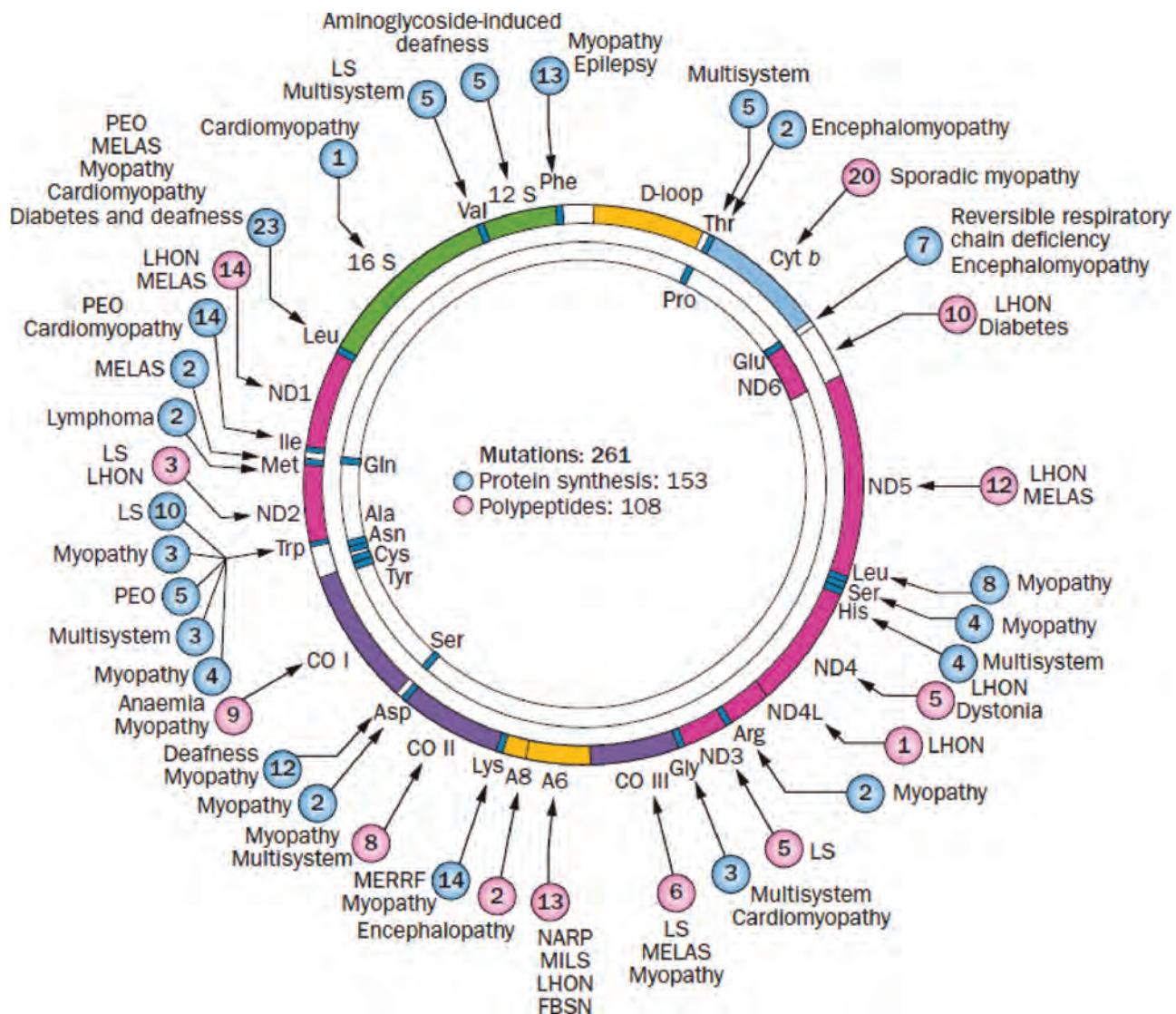
Most (~1500) mitochondrial proteins are the product of nuclear genes, and most mitochondrial diseases are caused by mutations in nuclear genes. Normally all mitochondrial genomes are the same, a condition called homoplasy. Pathogenic mutations may be heteroplasmic (having a mixture of normal and mutated genomes). We introduced NCBI tools to view the human mitochondrial genome in Chapter 15.

Another basis for classifying disease is according to tissue type organ system, or subcellular organelle. Eukaryotic cells are organized into organelles, such as the nucleus, endoplasmic reticulum, Golgi complex, peroxisome, lysosome, endosome, and mitochondrion. Each organelle serves a specialized function, gathering particular protein products to form enzymatic reactions necessary for cell survival, separating metabolic processes, and segregating harmful products. We have considered human disease from the perspective of genes and gene products. We can also examine disease in the context of the higher organizational level of organelles and pathways.

Consider the mitochondrion. This organelle was described as the site of respiration in the 1940s, and mitochondrial DNA was first reported by Nass and Nass (1963). It was not until 1988 that the first disease-causing mutations in mitochondria were described, however (Holt *et al.*, 1988; Wallace *et al.*, 1988a, b). Today, over 100 disease-causing point mutations have been described (reviewed in DiMauro and Schon, 2001; DiMauro *et al.*, 2013). The mitochondrial genome contains 37 genes, any of which can be associated with disease. **Figure 21.9** shows a morbidity map of the human mitochondrial genome.

Mitochondrial genetics differs from Mendelian genetics in three main ways (DiMauro and Schon, 2001; DiMauro *et al.*, 2013):

1. Mitochondrial DNA is maternally inherited. Mitochondria in the embryo are derived primarily from the ovum, while sperm mitochondria fail to enter the egg and are actively degraded. A woman having a mitochondrial DNA mutation may therefore transmit it to her children, but only her daughters will further transmit the mutation to their children.
2. While nuclear genes exist with two alleles (one maternal and one paternal), mitochondrial genes exist in hundreds or thousands of haploid copies per cell. (A typical mitochondrion contains about ten copies of the mitochondrial genome.)



**FIGURE 21.9** Morbidity map of the human mitochondrial genome. Colored sections represent protein-coding genes. These are seven subunits of complex I (ND; pink sections); one subunit of complex III (cyt b; light blue section); three subunits of cytochrome c oxidase (CO; purple sections); two subunits of ATP synthase (A6 and A8; yellow sections); 12 S and 16 S ribosomal RNA (green sections); and 22 transfer RNAs identified by three-letter codes for the corresponding amino acids (blue sections). Blue circles indicate diseases caused by mutations in genes that impair protein synthesis. Pink circles indicate diseases caused by mutations in genes that encode respiratory chain proteins. Numbers in circles represent the numbers of mutations reported at that site. Cyt b: cytochrome b; FBSN: familial bilateral striatal necrosis; LHON: Leber hereditary optic neuropathy; LS: Leigh syndrome; MELAS: mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF: myoclonus epilepsy with ragged-red fibres; MILS: maternally inherited Leigh syndrome; NARP: neuropathy, ataxia and retinitis pigmentosa; ND: NADH-dehydrogenase (complex I); PEO: progressive external ophthalmoplegia.

Source: DiMauro *et al.* (2013). Reproduced with permission from Macmillan Publishers.

An individual may harbor varying ratios of normal and mutated mitochondrial genomes. Some critical threshold of mutated mitochondrial genomes is required before a disease is manifested. As for nuclear DNA, mitochondrial DNA can therefore have somatic mutations that cause disease (Schon *et al.*, 2012).

3. As cells divide the proportion of mitochondria having mutated genomes can change, affecting the phenotypic expression of mitochondrial disorders. Clinically, mitochondrial disorders present at different times and in different regions of the body. An extremely broad variety of diseases are associated with mutations in mitochondrial DNA.

MITOMAP is online at  
 ⓘ <http://www.mitomap.org/>  
 (WebLink 21.12).

MITOMAP is a useful mitochondrial genome database (Ruiz-Pesini *et al.*, 2007). The site lists a broad variety of information on mutations and polymorphisms in mitochondrial genomes involving all known genetic mechanisms (inversions, insertions, deletions, etc.).

Next-generation sequencing has been employed to characterize variants in both the mitochondrial genome and in nuclear genes relevant to mitochondrial function (Vasta *et al.*, 2009), although Sanger sequencing also continues to be used (e.g., Tang *et al.*, 2013).

It is easy to analyze mitochondrial DNA variation in whole-exome sequence (WES) data. This may seem surprising since the basis of WES is selective capture or enrichment of nuclear-encoded exons using long oligonucleotides. However, there are so many copies of mitochondrial DNA that it is routinely, incidentally sequenced. Guo *et al.* (2013) developed MitoSeek, a package that extracts mitochondrial sequences from WES or whole-genome sequence data from BAM files, assembles the genome, and performs quality control (e.g., read depth, percent of base pairs covered, and base quality scores). We can invoke MitoSeek with the following command:

```
$ perl mitoSeek.pl -i /home/data/fshd216.bam -t 1 -d 5
```

MitoSeek is available at  
 ⓘ <https://github.com/riverlee/MitoSeek> (WebLink 21.13). It requires Perl scripts used by Circos (Box 19.2).

We invoke a perl script, use `-i` to specify the input BAM file location, use `-t` to define the type of BAM file (1 denotes whole exome, 2 is for whole-genome data, 3 is for RNA-seq data, and 4 is for mitochondrial DNA data) and specify `-d` for the minimum depth required to detect heteroplasmy. MitoSeek then reports heteroplasmy, somatic mutations (comparing allele counts between paired tumor/normal samples), relative copy number variation, and large structural variation.

## Somatic Mosaic Disease

Mosaicism is distinguished from chimerism in which cells having genetic differences are derived from separate ancestries, as may happen when an egg is fertilized by sperm from two different men.

Mosaicism is the occurrence of genetically distinct populations of cells within an organism (Yousoufian and Pyeritz, 2002; Lupski, 2013; Poduri *et al.*, 2013). Genetic changes may involve somatic cells such as skin or liver (somatic mosaicism), or they may involve germline cells (germline mosaicism, also called gonadal mosaicism). By some estimates the human body has  $\sim 10^{14}$  cells (e.g., Erickson, 2010) and, because errors during replication and cell division occur frequently, there is an appreciable amount of somatic mutation. We are therefore all mosaics. In many cases, somatic mosaicism is associated with disease. This idea was suggested by Sir Macfarlane Burnet (1959) who noted two examples of somatic mutation (in the fleece of sheep and in blood groups), and proposed mosaicism as an explanation for autoimmune disease.

Somatic mosaicism involving the skin, whether in the fleece of sheep or in human cutaneous disorders, is readily apparent. Rudolf Happle (1987) hypothesized that a set of disorders (such as McCune-Albright syndrome, Sturge-Weber syndrome, and Proteus syndrome) having a mosaic distribution of skin defects are each caused by a mutation in a gene that would be embryonic lethal as an inherited mutation in early development. Instead, the origin of these conditions is as a somatic variant.

Postzygotic, somatic, mosaic mutations have indeed been identified for the McCune-Albright syndrome (*GNAS* mutations; Weinstein *et al.*, 1991), the Proteus syndrome (*AKT* mutations; Lindhurst *et al.*, 2011), and other disorders. My lab reported that mutations in *GNAQ* encoding a G protein alpha subunit cause both the neurocutaneous Sturge-Weber syndrome and commonly occurring nonsyndromic port-wine stain birthmarks (Shirley *et al.*, 2013). Our approach was to obtain biopsies of affected parts of the body (e.g., a port-wine stain birthmark) and presumably unaffected regions (e.g., blood), and then to perform whole-genome sequencing of paired samples (from three individuals). After alignment to a reference genome and variant calling, the genotypes were compared using somatic variant callers. Matt Shirley, then a graduate student in my lab, employed

Somatic variation includes both single-nucleotide and copy number changes (Dumanski and Piotrowski, 2012). Pham *et al.* (2014) used high-resolution array comparative hybridization to evaluate >10,300 patients. They found somatic chromosomal mosaicism, resulting from postzygotic errors, in 57 cases (0.55% of the total).

Strelka (Saunders *et al.*, 2012); other prominent somatic variant callers include VarScan 2 (Koboldt *et al.*, 2012) and MuTect (Cibulskis *et al.*, 2013). By understanding the molecular defect – a G protein alpha subunit coupled to a seven transmembrane receptor is persistently activated – we hope we can modulate affected signaling pathways to offer treatment to patients.

We mentioned that mutations in *GNAQ* cause Sturge-Weber syndrome and port-wine stain birthmarks. These involve an R183Q mutation (an arginine is substituted by a glutamine). Somatic mutations in the same gene, causing the identical R183Q mutation, are also a cause of uveal melanoma and a pigmentation condition called blue nevi (Van Raamsdonk *et al.*, 2009). In the childhood disease the somatic mutation occurs before birth in an unknown cell type (perhaps endothelial cells). In uveal melanoma the somatic mutation occurs in adulthood, in melanocytes. The same mutation occurring in yet another cell type might have no clinical phenotype. Where in the body a mutation occurs, and at what stage of development, are critical considerations.

## Cancer: A Somatic Mosaic Disease

Cancer is a somatic mosaic disease, arising from a clone having somatic mutations and leading to malignant transformation (Chin *et al.*, 2011; Watson *et al.*, 2013). Cancer occurs when DNA mutations confer selective advantage to cells that proliferate, often uncontrollably (Varmus, 2006). Knudson (1971) introduced a two-hit hypothesis of cancer, suggesting that for dominantly inherited retinoblastoma one mutation is inherited through the germ cells while a second somatic mutation occurs; for a nonhereditary form of cancer two somatic mutations occur. There are six hallmarks of cancer, described by Hanahan and Weinberg (2011): proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, induction of angiogenesis, and inactivating invasion and metastasis.

There are >200 types of cancer and many disease mechanisms, and a growing number of key tumor suppressor genes and other oncogenic genes have been identified. Given the completion of the human genome project and the availability of improved sequencing capabilities, a human cancer genome project has been launched to catalog the DNA sequence of a variety of cancer genomes (Stratton, 2011).

One initiative is COSMIC (catalogue of somatic mutations in cancer; Forbes *et al.*, 2011). It includes information on nearly 1 million cancer samples, >1.6 million mutations, and various types of mutations (fusions, genomic rearrangements, and copy number variants). It also offers extensive literature annotation and a BioMart (Shepherd *et al.*, 2011). To explore *GNAQ*, you can do the following:

- Visit the main COSMIC website and obtain information such as a list of mutations in the gene.
- View the gene in a COSMIC web browser.
- Use the COSMIC BioMart to explore the many features.
- Use the main Ensembl human website, select BioMart, set the Dataset to *Homo sapiens* Somatic Short Variation (SNPs and indels). Under Filters select COSMIC as the variation source, and select the Ensembl Gene ID (ENSG00000156052, which is listed at the COSMIC or Ensembl sites). You can then select attributes of interest to learn more about the gene and associated mutations.
- In the main Ensembl human website you can perform a search for *GNAQ* and select the variation table. There you can access COSMIC database variants associated with the gene.

The website of the Cancer Genome Project at the Wellcome Trust Sanger Institute is <http://www.sanger.ac.uk/research/projects/cancergenome/> (WebLink 21.14) with links to many cancer resources. COSMIC is available at <http://cancer.sanger.ac.uk/cancergenome/projects/census/> (WebLink 21.15). For the COSMIC BioMart visit <http://www.sanger.ac.uk/genetics/CGP/cosmic/biomart/martview/> (WebLink 21.16).

The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are other major initiatives. Their goals are to analyze mutations in thousands of

The TCGA website is <http://cancergenome.nih.gov/> (WebLink 21.17). It is a US\$ 375 million sequencing project started in 2006 by the NHGRI. The National Cancer Institute (NCI) at the National Institutes of Health website is <http://www.cancer.gov/> (WebLink 21.18). Data for TCGA are stored at the UCSC Cancer Genomics Hub (CGHub) at <https://cghub.ucsc.edu/> (WebLink 21.19). Currently that site houses ~1,420,900 gigabytes of data (1.4 petabytes) organized by three dozens types of cancer. The ICGC website is <http://www.icgc.org/> (WebLink 21.20), and its data portal is <http://dcc.icgc.org/> (WebLink 21.21). Currently (March 2015) it lists ~13 million somatic mutations across 18 cancer sites.

The UCSC Cancer Genomics Browser is online at <https://genome-cancer.ucsc.edu/> (WebLink 21.22).

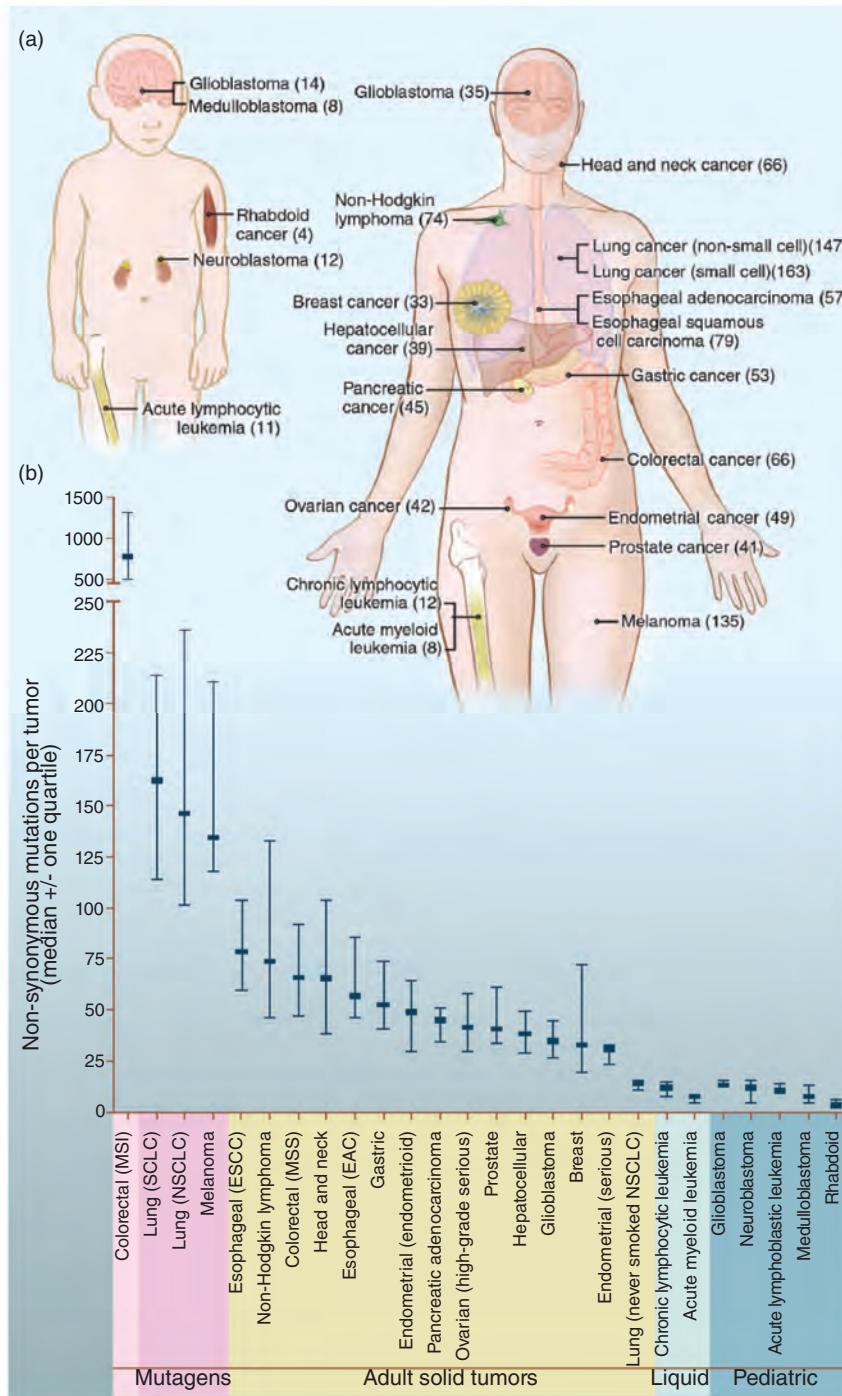
tumor samples to characterize genetic changes, as well as alterations in the transcriptome and epigenome.

The UCSC Cancer Genomics Browser is a resource offering extensive data on cancer, including from the TCGA project (Cline *et al.*, 2013; Goldman *et al.*, 2013). Features include the availability of large datasets with views by cancer subtype, chromosome location, clinical features, or genes and pathways of interest. Gemomic heatmaps can show regions of deletion and amplification, while a clinical heatmap shows samples (on the y axis) versus features such as tumor grade, histological type, and survival statistics (x axis). Kaplan–Meier plots can display percent survival versus time organized by user-selected groups such as patients receiving different treatments.

The landscape of cancer includes two types of mutations (Greenman *et al.*, 2007; Wood *et al.*, 2007; Vogelstein *et al.*, 2013). “Driver” mutations confer a selective growth advantage to cells, are implicated as causing the neoplastic process, and are positively selected for during tumorigenesis. “Passenger” mutations are retained by chance but confer no selective advantage and do not contribute to oncogenesis. A challenge is to identify driver mutations throughout the genome of a cancer cell and to distinguish them from passenger mutations. Driver mutations occurring at high frequency have been called mountains interspersed with many hills (corresponding to driver mutations that occur with lower frequency). The large number of infrequently mutated genes represented in the hills may be as important as the mountains, and may represent the relevant mutational signature of each cancer (e.g., Wood *et al.*, 2007). A goal is to relate such a molecular profile of a cancer to an appropriate therapy to eradicate the cancer.

The advent of next-generation sequencing has enabled deep cataloguing of many cancer types. Bert Vogelstein and colleagues summarized the number of somatic mutations in selected human cancers (Fig. 21.10a). They also described signaling pathways that are commonly affected. Some conclusions are (Vogelstein *et al.*, 2013):

- The rate of nonsynonymous mutations (each of which is predicted to alter a specified amino acid) varies greatly. Cancers that compromise DNA repair function can lead to thousands of nonsynonymous mutations per tumor (Fig. 21.10b). Cancer caused by mutagens such as tobacco or ultraviolet radiation from sunlight tend to cause ~100–200 nonsynonymous mutations per tumor. Lung cancers in smokers may therefore have ten times as many somatic mutations as lung cancers in nonsmokers.
- Cancers having relatively few mutations include pediatric tumors and leukemias (~10 per tumor). One reason is that some tumors acquire mutations over time (particularly tumors in self-renewing tissues).
- Metastatic cancer develops through somatic mutations that are acquired over a period of decades. Also, mutations in metastatic tumors were already present in many cells in primary tumors.
- Aneuploidy is common in cancer cells, including whole-chromosome or segmental copy number changes, inversions, and translocations. Translocations often result in the fusion of two genes to create an oncogene such as *BCR-ABL*. Chromosomal deletions are the most common form of aneuploidy in cancer, often deleting a tumor suppressor gene.
- It is challenging to determine whether a somatic mutation represents a driver or a passenger. A driver gene contains driver gene mutation(s) (Thiagalingam *et al.*, 1996), but it may also contain passenger mutations. Vogelstein *et al.* propose introducing the term “mut-driver genes” to denote driver gene mutations from “epi-driver genes” which are expressed aberrantly in tumors but are not frequently mutated.
- Oncogenes tend to have recurrent mutations at one or a few amino acid positions (as for *PIK3CA* and *IDH1*), while tumor suppressor genes tend to acquire truncating mutations along their length. Vogelstein *et al.* describe a “20/20 rule” in which a



**FIGURE 21.10** Somatic mutations in representative human cancers, based on genome-wide sequencing studies. (a) The genomes of adult (right) and pediatric (left) cancers are represented. Numbers in parentheses are the median number of nonsynonymous mutations per tumor. Redrawn from Vogelstein *et al.* (2013). Reproduced with permission from AAAS. (b) Median number of nonsynonymous substitutions per tumor. Horizontal bars indicate the 25% and 75% quartiles. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC: esophageal adenocarcinomas.

gene is classified as an oncogene if >20% of the recorded mutations are missense at recurrent positions. It is a tumor suppressor gene if >20% of its mutations are inactivating. Occasionally one gene (such as *NOTCH1*) may have different roles in different cancer types.

- Heterogeneity of cancer mutations may be observed: (1) among the cells of a single tumor; (2) among different metastatic lesions of one patient; (3) among the cells of a single metastatic lesion; and (4) among tumors of different patients.
- While the cancer genome appears extraordinarily complex, the vast majority of genetic variants are passengers that do not influence neoplasia. Vogelstein *et al.* list 138 key driver genes and divide their functional effects into a small number of cellular signaling pathways that confer a selective growth advantage. Functionally, driver mutations influence three cellular processes: cell fate, cell survival, and genome maintenance.

## DISEASE DATABASES

We next describe two major types of human disease database: (1) central databases such as OMIM, HGMD, and ClinVar provide great breadth in surveying thousands of diseases; and (2) thousands of locus-specific mutation databases provide great depth in reporting mutations associated with genes, with a focus on either one specific gene and/or one disease. Patrinos and Brookes (2005) and Thorisson *et al.* (2009) reviewed these two types of databases, emphasizing the great challenges associated with relating genotype to phenotype (that is, relating data on DNA mutations to clinical phenotypes).

Mendelian Inheritance in Man (MIM) was started in 1966 by Victor A. McKusick. The online version OMIM became integrated with NCBI in 1995, available at <http://www.ncbi.nlm.nih.gov/omim/> (WebLink 21.23) or through <http://www.omim.org> (WebLink 21.24). The scientific director of OMIM is Ada Hamosh of the Johns Hopkins Medical Institutions.

### OMIM: Central Bioinformatics Resource for Human Disease

OMIM® is a comprehensive database for human genes and genetic disorders, particularly rare (often monogenic) disorders having a genetic basis (McKusick, 2007; Amberger *et al.*, 2011). The OMIM database contains bibliographic entries for over 22,000 human diseases and relevant genes. A focus of OMIM is inherited genetic diseases. As indicated by its name, the OMIM database is concerned with Mendelian genetics. These are inherited traits that are transmitted between generations. There is relatively little information in the database about genetic mutations in complex disorders, or chromosomal disorders. Its focus is a comprehensive survey of single-gene disorders, with richly detailed descriptions as well as links to many database resources.

We can examine OMIM using sickle cell anemia and *HBB* as examples of a disease and a gene implicated in a disease. OMIM can be searched from the NCBI site, and is linked from NCBI Gene. Within the OMIM site, there is a search page that allows you to query a variety of fields including chromosome, map position, or clinical information. The result of a search for “beta globin” includes both the relevant gene (Fig. 21.11) and relevant diseases (e.g., sickle cell anemia and thalassemias).

We can next view the entry for beta globin (Fig. 21.12), with its OMIM identifier +141900. Each entry in OMIM is associated with a numbering system. There is a six-digit code in which the first digit indicates the mode of inheritance of the gene involved (Table 21.8). The beta globin entry is preceded by a plus sign to indicate that the entry contains the description of a gene of known sequence and a phenotype. The first number (1) indicates that this gene has an autosomal locus (and the entry was created by 1994). The entry includes bibliographic data such as available information on an animal model for globinopathies. OMIM entries link to a gene map, which provides a tabular listing of the cytogenetic position of disease loci. This gene map further links to the NCBI Map Viewer and to resources for the orthologous mouse gene. The OMIM morbid map also provides cytogenetic loci but is organized alphabetically.

Search: 'beta globin'

Results: 1 – 10 of 4,408 | Show top 100 | 1 2 3 4 5 6 7 8 9 10 Next Last

1 : + 141900. HEMOGLOBIN–BETA LOCUS; HBB  
METHEMOGLOBINEMIA, BETA-GLOBIN TYPE, INCLUDED  
Cytogenetic location: 11p15.4, Genomic coordinates (GRCh37): 11:5,246,695 - 5,248,300  
Matching terms: globin, beta

2 : # 141749. FETAL HEMOGLOBIN QUANTITATIVE TRAIT LOCUS 1; HBFQTL1  
DELTA-BETA THALASSEMIA, INCLUDED  
Cytogenetic locations: 11p15.4, 11p15.4, 11p15.4  
Matching terms: globin, beta

3 : # 603903. SICKLE CELL ANEMIA  
Cytogenetic location: 11p15.4  
Matching terms: globin, beta

4 : \* 142200. HEMOGLOBIN, GAMMA A; HBG1  
Cytogenetic location: 11p15.4, Genomic coordinates (GRCh37): 11:5,269,501 - 5,271,086  
Matching terms: globin, beta

5 : + 141800. HEMOGLOBIN–ALPHA LOCUS 1; HBA1  
METHEMOGLOBINEMIA, ALPHA-GLOBIN TYPE, INCLUDED  
Cytogenetic location: 16p13.3, Genomic coordinates (GRCh37): 16:226,678 - 227,519  
Matching terms: globin, beta

Gene Tests, Newborn Screening, Links

ICD+, Links

Newborn Screening, ICD+, Links

Links

Links

ICD+ for #603903 [x]

SNOMEDCT: 127040003, 417357006  
ICD10CM: D57, D57.1  
ICD9CM: 282.60, 282.6

**External Links for +141900 [x]**

- Genome
  - [Ensembl](#)
  - [NCBI Map Viewer](#)
  - [UCSC Genome Browser](#)
- DNA
  - [Ensembl](#)
  - [NCBI RefSeq](#)
  - [UCSC Genome Browser](#)
- Protein
  - [UniProt](#)
  - [HPRD](#)
- Gene Info
  - [BioGPS](#)
  - [Ensembl](#)
  - [GeneCards](#)
  - [Gene Ontology](#)
  - [KEGG](#)
  - [NCBI Gene](#)
  - [PharmGKB](#)
- Clinical Resources
  - [Clinical Trials](#)
  - [Gene Tests](#)
  - [Newborn Screening](#)
  - [GTR](#)
  - [GARD](#)
  - [Genetics Home Reference](#)
  - [NextGxDx](#)
- Variation
  - [ClinVar](#)
  - [Genetics Association DB](#)
  - [GWAS Central](#)
  - [HGVS](#)
  - [Locus Specific DBs](#)
  - [NHLBI EVS](#)
  - [1000 Genome](#)
- Animal Models
  - [NCBI HomoloGene](#)
  - [OMIA](#)
- Cell Lines
  - [Coriell](#)
- Cellular Pathways
  - [KEGG](#)
  - [Reactome](#)

**FIGURE 21.11** Online Mendelian Inheritance in Man (OMIM), accessible via the NCBI website, allows text searches by criteria such as author, gene identifier, or chromosome. A search of OMIM for “beta globin” produces results including entries on that gene, related globin genes, and diseases such as thalassemias and sickle cell anemia. The insets show links to external resources and to ICD clinical diagnostic categories.

Source: OMIM (<http://omim.org/>).

An important feature of OMIM entries is that many contain a list of allelic variants. Most of these represent disease-causing mutations. An example of several allelic variant entries is shown for *HBB* (Fig. 21.12). These allelic variants provide a glimpse of all the human genes that are known to contain disease-causing mutations. Allelic variants within OMIM are selected based on criteria such as being the first mutation to be discovered, having a high population frequency, or having an unusual pathogenetic mechanism. Some allelic variants in OMIM represent polymorphisms. These may be of particular interest if they show a positive correlation with common disorders. In the particular case of *HBB*, hundreds of allelic variants are included.

The current holdings of OMIM are summarized by chromosome (Table 21.9) and according to mode of inheritance (autosomal, X- or Y-linked, mitochondrial; Table 21.10). OMIM continues to be a crucial and comprehensive resource for information on the human genome. Many other disease databases incorporate OMIM identifiers to provide a common reference to disease-related genes.

**HGNC Approved Gene Symbol: HBB**Cytogenetic location: [11p15.4](#) Genomic coordinates (GRCh37): [11:5,246,695 – 5,248,300](#) (from NCBI)**Gene-Phenotype Relationships**

Location	Phenotype	Phenotype MIM number
11p15.4	Delta-beta thalassemia	<a href="#">141749</a>
	Erythremias, beta-	
	Heinz body anemias, beta-	<a href="#">140700</a>
	Hereditary persistence of fetal hemoglobin	<a href="#">141749</a>
	Methemoglobinemias, beta-	
	Sickle cell anemia	<a href="#">603903</a>
	Thalassemia-beta dominant inclusion-body	<a href="#">603902</a>
	Thalassemias, beta-	<a href="#">613985</a>
	[Malaria, resistance to]	<a href="#">611162</a>

**Clinical Synopsis****TEXT****Description**

The alpha (HBA1, [141800](#); HBA2, [141850](#)) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, HbA. Mutant beta globin that sickles causes sickle cell anemia ([603903](#)). Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. For clinical purposes, beta-thalassemia ([613985](#)) is divided into thalassemia major (transfusion dependent), thalassemia intermedia (of intermediate severity), and thalassemia minor (asymptomatic).

**Table of Contents for +141900**

Title
Gene-Phenotype Relationships
Text
Description
Gene Structure
Mapping
Gene Function
Biochemical Features
Molecular Genetics
Animal Model
History
Allelic Variants
Table View
Clinical Synopsis
See Also
References
Contributors
Creation Date
Edit History
External Links for Entry:
► Genome
► DNA
► Protein
► Gene Info
► Clinical Resources
► Variation
► Animal Models

**FIGURE 21.12** The OMIM entry for beta globin includes the OMIM identifier (+141900) and a variety of information, indexed on the sidebar, such as clinical features, a description of available animal models, and allelic variants.

Source: OMIM (<http://omim.org/>).

**TABLE 21.8** OMIM numbering system. OMIM number beginning 1 or 2 implies it entered the database before May 1994; OMIM number beginning 6 implies it was created after May 1994; + indicates a gene of known sequence and a phenotype; % indicates a confirmed mendelian phenotype (or phenotypic locus) for which the underlying molecular basis is not known; # indicates a descriptive entry (usually of phenotype); \* preceding entry indicates a gene of known sequence. For the AUTS1 entry, the number 1 indicates that this is the first listing of several autism susceptibility loci (e.g., AUTS2). Adapted from OMIM (<http://omim.org/help/faq>, accessed March 2014). Reproduced with permission from Johns Hopkins University.

OMIM no.	Phenotype	OMIM identifier	Disorder (example)	Chromosome number
1__	Autosomal dominant	+143100	Huntington disease	4p16.3
2__	Autosomal recessive	%209850	Autism, susceptibility to, (AUTS1)	7q
3__	X-linked loci or phenotypes	#312750	Rett syndrome	Xq28
4__	Y-linked loci or phenotypes	*480000	Sex-determining region Y	Yp11.3
5__	Mitochondrial loci or phenotypes	#556500	Parkinson disease	–
6__	Autosomal loci or phenotypes	#603903	Sickle cell anemia	–

**TABLE 21.9 Synopsis of OMIM human genes per chromosome. Total number of loci: 14,622. Adapted from OMIM (<http://omim.org/help/faq>, accessed March 2014).**  
Reproduced with permission from Johns Hopkins University.

Chromosome	Loci	Chromosome	Loci	Chromosome	Loci
1	1,445	9	553	17	838
2	919	10	545	18	212
3	782	11	886	19	912
4	565	12	770	20	371
5	659	13	273	21	154
6	865	14	474	22	355
7	692	15	436	X	807
8	516	16	593	Y	53

**TABLE 21.10 Current holdings of OMIM. Adapted from OMIM (<http://omim.org/help/faq>, accessed March 2014). Reproduced with permission from Johns Hopkins University.**

	Autosomal	X-linked	Y-linked	Mitochondrial	Total
* Gene with known sequence	13,752	672	48	35	14,507
+ Gene with known sequence and phenotype	100	2	0	2	104
# Phenotype description, molecular basis known	3,732	282	4	28	4,406
% Mendelian phenotype or locus, molecular basis unknown	1,577	135	5	0	1,717
Other, mainly phenotypes with suspected mendelian basis	1,745	115	2	0	1,862
Total	20,906	1,206	59	65	22,236

Earlier we introduced Genome Workbench, an NCBI tool to query Entrez databases. Select NCBI Gene and enter the query globin AND human[ORGN] (Fig. 21.13a). (Alternatively, search directly for the beta globin accession NM\_000518.) HBB appears in the results list; right-click to add it to a new project. You can then select a SNP table view and obtain a tabular list of genomic variants (with dbSNP identifiers), including those which have OMIM entries (Fig. 21.13b). Using Genome Workbench to obtain tabular outputs from NCBI Entrez is analogous to using the UCSC Table Browser to obtain tabular outputs from the UCSC Genome Browser. Note however that (at present) OMIM variant data can be viewed in the UCSC Genome Browser but are not available in the UCSC Table Browser.

Genome Workbench can be downloaded from <http://www.ncbi.nlm.nih.gov/tools/gbench/> (WebLink 21.25).

### Human Gene Mutation Database (HGMD)

The Human Gene Mutation Database (HGMD) is another major source of information on disease-associated mutations (Stenson *et al.*, 2012, 2014). The database is partly commercial (requiring payment for full access). George *et al.* (2008) compared OMIM and HGMD, noting differences in their approaches. For example, OMIM places emphasis on detailed descriptions of genes and disorders and their clinical phenotypes, while HGMD emphasizes more comprehensive cataloguing of mutations. In sequencing human genomes and exomes, it is common to filter variants based on whether they have been previously associated with disease; HGMD has emerged as a basic resource in many analysis pipelines.

HGMD is a project of David Cooper and colleagues at Cardiff University. It is available at <http://www.hgmd.cf.ac.uk/ac> (WebLink 21.26). There are ~115,000 mutation entries for public release and ~164,000 entries for commercial release (March, 2015).

(a) Genome Workbench query for human hemoglobin

The screenshot shows the 'Search View' window of the NCBI Genome Workbench. The search query 'hemoglobin AND human[ORGN]' has been entered. The results table lists 1241 items, with the first few rows shown below:

Label	Description	FASTA IDs	Taxonomic ID
NP_000509	hemoglobin subunit beta [Homo sapiens]	gi 4504349 ref NP_000509.1	9606
NP_000510	hemoglobin subunit delta [Homo sapiens]	gi 4504351 ref NP_000510.1	9606
NP_005321	hemoglobin subunit epsilon [Homo sapiens]	gi 4885393 ref NP_005321.1	9606
NP_000550	hemoglobin subunit gamma-1 [Homo sapiens]	gi 28302131 ref NP_000550.2	9606
NP_000175	hemoglobin subunit gamma-2 [Homo sapiens]	gi 6715607 ref NP_000175.1	9606
NP_001003938	hemoglobin subunit mu [Homo sapiens]	gi 51510893 ref NP_001003938.1	9606
NP_005322	hemoglobin subunit theta-1 [Homo sapiens]	gi 4885395 ref NP_005322.1	9606
NP_005323	hemoglobin subunit zeta [Homo sapiens]	gi 4885397 ref NP_005323.1	9606

(b) Genome Workbench SNP Table View

The screenshot shows the 'SNP Table View' for mRNA NM\_000518.4. The table includes columns for RefSNP ID, Alleles, Sequence, Location, Variation Class, Phenotype, and 1000 Genomes status. The first 18 rows of the table are listed below:

RS ID	Alleles	Sequence	Location	Variation Class	Phenotype	1000 Genomes
34305195	A/C	NM_000518.4	1	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
35352549	T/T	NM_000518.4	10	INDEL	From LSDB	Not present in 1000G
63750628	C/T	NM_000518.4	20	SNP	From LSDB	Not present in 1000G
34704828	C/T	NM_000518.4	22	SNP	From LSDB	Not present in 1000G
34135782	C/G	NM_000518.4	33	SNP	From LSDB	Not present in 1000G
113115948	C/T	NM_000518.4	39	SNP		Not present in 1000G
34196559	A/AAC	NM_000518.4	40	INDEL	From LSDB	Not present in 1000G
34563000	A/G	NM_000518.4	51	SNP	From LSDB	Not present in 1000G
33941849	A/C/G/T	NM_000518.4	52	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
33930702	A/C/G/T	NM_000518.4	53	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
33958358	A/G/T	NM_000518.4	54	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
33949930	A/C/G/T	NM_000518.4	55	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
35906307	C/T	NM_000518.4	57	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
33983205	A/C/G/T	NM_000518.4	58	SNP	From LSDB, OMIM/OMIA	Not present in 1000G
34058656	T/T	NM_000518.4	59	INDEL		Not present in 1000G
713040	A/C/G/T	NM_000518.4	59	SNP	From LSDB, OMIM/OMIA	1000G Phase 1, Has 1000G submission, Not present in 1000G
34126315	A/C/G	NM_000518.4	60	SNP	From LSDB	Not present in 1000G
33912272	C/G/T	NM_000518.4	66	SNP	From LSDB, OMIM/OMIA	1000G Phase 1, Has 1000G submission, Not present in 1000G

**FIGURE 21.13** Accessing OMIM allelic variants with NCBI Genome Workbench. (a) In Search View, the query hemoglobin AND human[ORGN] is entered. Results include RefSeq accessions for human hemoglobin proteins. By selecting hemoglobin subunit beta, it is added as a new project. (b) A SNP table view for the corresponding mRNA (accession NM\_000518.4) is launched by clicking the project on the left sidebar. The table includes columns listing RefSNPs (RS ID), alleles, location in the sequence, variation class (e.g., SNP, indel), phenotype, and whether each variant is observed in 1000 Genomes data. The phenotype entries include hyperlinks to OMIM, OMIA (Online Mendelian Inheritance in Animals), and locus-specific databases (LSDB).

Source: Genome Workbench, NCBI.

## ClinVar and Databases of Clinically Relevant Variants

The ClinVar database provides data on human variants and their relationship to disease (Landrum *et al.*, 2014). It further provides links to the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, and PubMed. GTR centralizes genetic test information that is volunteered by providers, for example listing where genetic testing can be performed for a particular condition. MedGen organizes human medical genetics information, for example providing several hundred entries on medical conditions relevant to a query for hemoglobin.

There are five categories of content in ClinVar (Landrum *et al.*, 2014): submitter, variation, phenotype, interpretation, and evidence. (1) Submissions are from organizations

and individuals. (2) Variation includes sequences at one location (single allele) or multiple alleles (e.g., compound heterozygotes in which two parents transmit different alleles at a single locus, sometimes causing a phenotypic change). Variants are cross-referenced to dbSNP and dbVar. (3) Phenotype may represent one concept or more and is annotated by MeSH term (see Fig. 21.4), OMIM number, MedGen identifier, or Human Phenotype Ontology (HPO; Robinson *et al.*, 2008). (4) Interpretation in ClinVar is submitter-driven and uses terms recommended by the American College of Medical Genetics and Genomics (ACMG). (5) Evidence typically consists of the number of individuals in which a given mutation was observed.

As an example of using ClinVar, enter a query for HBB[gene]. There are 720 results, and you can restrict these to single-nucleotide changes ( $n = 528$ ), pathogenic variants ( $n = 178$ ), or both ( $n = 120$ ). Several of these are shown in Figure 21.14a, as well as details on variant E7V (Fig. 21.14b). Those details include the HGVS nomenclature (e.g., NM\_000518.4:c.20A>T specifies the accession number with version number for the DNA sequence that has a coding (“c.”) change from A to T), genomic location, and allele frequency.

## GeneCards

GeneCards is a human gene compendium that includes a wealth of information on human disease genes (Stelzer *et al.*, 2011). GeneCards differs from OMIM in that it collects and integrates data from several dozen independent databases including OMIM, GenBank, UniGene, Ensembl, the University of California at Santa Cruz (UCSC), and the Munich Information Center for Protein Sequences (MIPS). Relative to OMIM, GeneCards uses relatively less descriptive text of human diseases and provides relatively more functional genomics data (George *et al.*, 2008).

## Integration of Disease Database Information at the UCSC Genome Browser

The UCSC Genome and Table Browser offers a convenient site to compare and contrast the contents of disease databases. We can view a 5000 base pair region encompassing the beta globin gene (Fig. 21.15). Although HGMD is overall far more comprehensive than OMIM and ClinVar, for genes such as *HBB* they represent comparable numbers of allelic variants, most of which overlap exons. Additional databases, described in the following, display allelic variants and copy number variants in that region.

## Locus-Specific Mutation Databases and LOVD

Central databases such as OMIM and HGMD attempt to comprehensively describe all disease-related genes without necessarily cataloguing every known allelic variant. In contrast, locus-specific mutation databases describe variations in a single gene (or sometimes in several genes) in depth (Samuels and Rouleau, 2011). Curators of these databases provide particular expertise on the genetic aspects of one specific gene, locus, or disease. The coverage of known mutations also tends to be far deeper in locus-specific databases as a group than in central databases (Scriver *et al.*, 1999). These two types of databases therefore serve complementary purposes.

A locus-specific mutation database is a repository for allelic variations. There are thousands of such databases. The essential components of a locus-specific database include the following (Scriver *et al.*, 1999, 2000; Claustres *et al.*, 2002; Cotton *et al.*, 2008):

- a unique identifier for each allele;
- information on the source of the data;
- the context of the allele; and
- information on the allele (e.g., its name, type, and nucleotide variation).

ClinVar is at NCBI (<http://www.ncbi.nlm.nih.gov/clinvar/>, WebLink 21.27). It currently includes ~75,000 variations from 18,700 genes. GTR is at <http://www.ncbi.nlm.nih.gov/gtr/> (WebLink 21.28), and MedGen is at <http://www.ncbi.nlm.nih.gov/medgen/> (WebLink 21.29). To learn more about the contents and fields of each database, use EDirect commands such as `$ einfo -db clinvar where -db` specifies the database of interest.

For the HPO see <http://www.human-phenotype-ontology.org/> (WebLink 21.30); ACMG is at <https://www.acmg.net/> (WebLink 21.31).

You can search ClinVar by disease with a command such as `autism[dis]`.

GeneCards, a project of Doron Lancet and colleagues at the Weizmann Institute, is available at <http://www.genecards.org/> (WebLink 21.32).

In the context of mutation databases, a mutation is defined as an allelic variant (Scriver *et al.*, 1999). The allele (or the unique sequence change) may be disease causing; such an allele tends to occur at low frequency. The allele may also be neutral, not having any apparent effect on phenotype.

(a) Tabular view of ClinVar results

	Gene	Variation	Freq	Phenotype	Clinical Significance	Review Status	Chr	Location (GRCh37 p10)
<a href="#">See details</a>	HBB	c.2T>C (p.Met1Thr)		Beta-thalassemia, lemontov type	Pathogenic	classified by single submitter	11	5248250
<a href="#">See details</a>	HBB	c.2T>C (p.Met1Thr)		beta0 <sup>+</sup> Thalassemia	Pathogenic	classified by single submitter	11	5248250
<a href="#">See details</a>	HBB	c.75T>A (p.Gly25=)		beta Thalassemia	Pathogenic	classified by single submitter	11	5248177

(b) Detailed view of a ClinVar entry for the E7V HBB variant

**HBB:c.20A>T (p.Glu7Val) AND Hb SS disease**

Clinical significance: Pathogenic (Last evaluated: Mar 14, 2013)  
 Review status:

Based on: 1 submission [Details]  
 Record status: current  
 Accession: RCV000016574.21

**Allele description**

Gene: HBB:hemoglobin, beta [Gene - OMIM]  
 Variant type: single nucleotide variant  
 Genomic location: Chr11:5248232 (on Assembly GRCh37)  
 Preferred name: HBB:c.20A>T (p.Glu7Val)  
 HGVS:  
   NC\_000011.9:g.5248232T>A  
   NG\_000007.3:g.70614A>T  
   NM\_000518.4:c.20A>T  
   NP\_000509.1:p.Glu7Val  
 Protein change: E6V, GLU6VAL  
 Links:  
   OMIM: [141900.0039](#); OMIM: [141900.0040](#); OMIM: [141900.0243](#); OMIM: [141900.0244](#); OMIM: [141900.0245](#); OMIM: [141900.0246](#); OMIM: [141900.0247](#); OMIM: [141900.0521](#); OMIM: [141900.0523](#); dbSNP: [334](#)

1000Genome: rs334  
 Allele Frequency: 0.0138, [GO-ESP](#)  
 Suspect: Not available

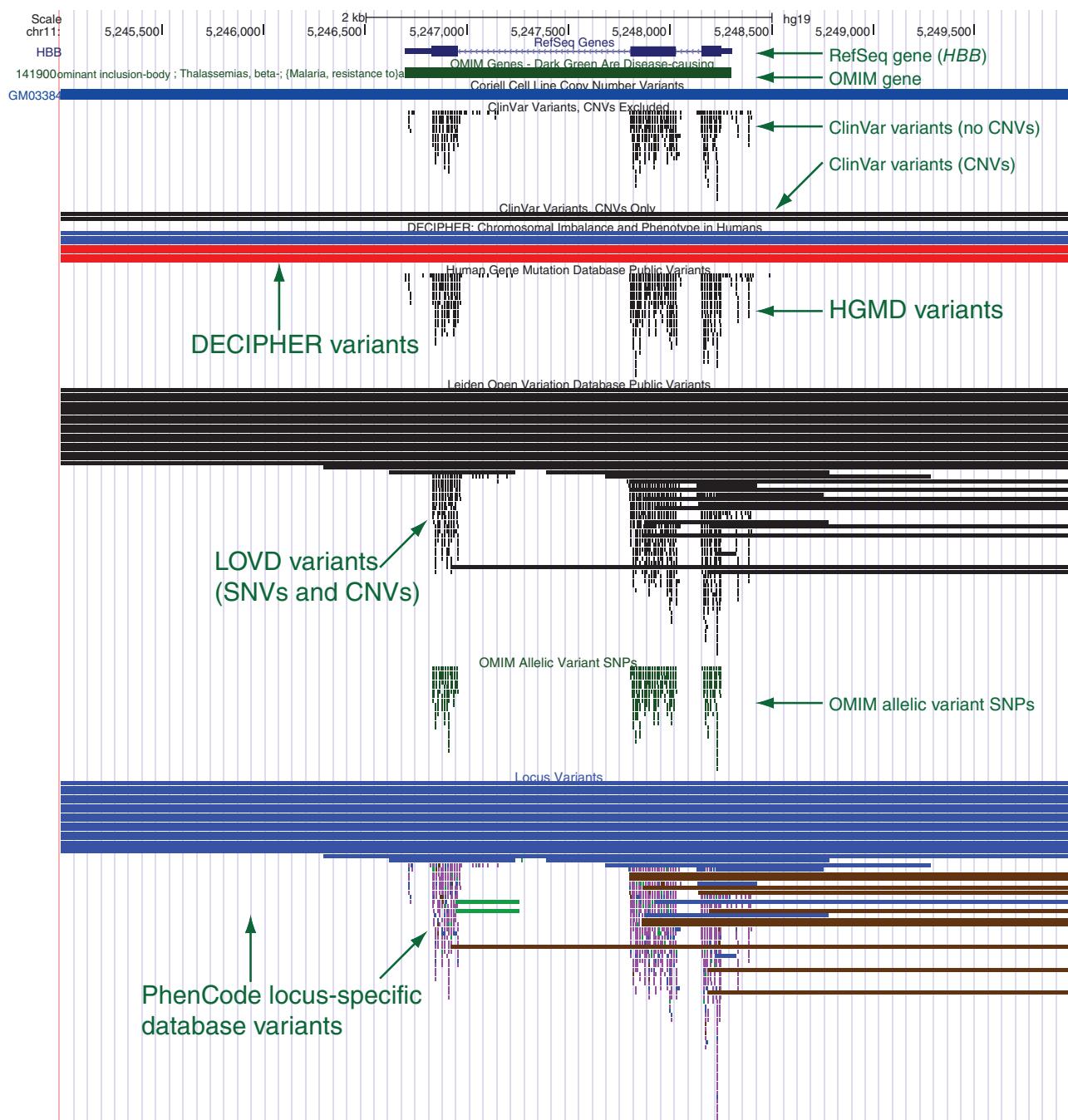
**Condition(s)**

Name: Hb SS disease  
 Synonyms: Hb SS disease; Hb SS disease; Hb SS disease  
 Identifiers: MedGen: [C0002895](#); OMIM: [603903](#); Orphanet: [232](#)  
 Age of onset: Variable  
 Prevalence: 1-5 / 10 000 [232](#)

**FIGURE 21.14** Output of a ClinVar query for the *HBB* gene. (a) There are 120 pathogenic, single-nucleotide variants of which three are shown here. (b) Details for a variant (E7V in which a wildtype glutamate at amino acid position 7 is substituted by a valine).

Source: ClinVar.

Mutation databases have an important role in gathering information about mutations, but there have not been uniform standards for their creation until recently. Claustres *et al.* (2002) surveyed 94 websites that encompassed 262 locus-specific databases; Cotton *et al.* (2008) noted over 700 such databases. Both studies noted great variability in the way data are collected, presented, linked, named, and updated. Scriver *et al.* (1999, 2000)



**FIGURE 21.15** UCSC Genome Browser includes tracks to display data from disease databases. A 5000 base pair region is shown (chr11:5,245,001–5,250,000) including *HBB* as shown by the RefSeq Genes track. The OMIM entry is shaded dark green, indicating it has disease-causing variants. HGMD, ClinVar, OMIM, and PhenCode entries are displayed at squish density, with similar profiles and with the majority of variants overlapping the exons (thick blue rectangles of the RefSeq track). Copy number variants (CNVs) are displayed in a separate ClinVar track, in the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) database, in the Coriell track displaying cell lines (and/or genomic DNA samples) available to the research community, and in the Leiden Open Variation Database (LOVD) which includes both single-nucleotide variants (SNVs) and CNVs.

Source: <http://genome.ucsc.edu>, courtesy of UCSC.

and Cotton *et al.* (2008) described guidelines for the content, structure, and deployment of mutation databases.

- There is now increased uniformity in naming alleles (Antonarakis, 1998; den Dunnen and Antonarakis, 2000). For example, the A of the ATG of the initiator Met codon is denoted nucleotide +1. Many such rules have been explicitly stated to allow uniform descriptions of mutations.
- Ethical guidelines have been described, such as the obligation of preserving the confidentiality of information (Knoppers and Laberge, 2000). Lowrance and Collins (2007) have reviewed issues of identifiability in genomic research.
- Generic software to build and analyze locus-specific databases has been provided, such as the Universal Mutation Database template (Bérroud *et al.*, 2005).

To see the Universal Mutation Database template of Bérroud *et al.*, visit <http://www.umd.be/> (WebLink 21.33).

HGVS databases are accessible at <http://www.hgvs.org/content/databases-tools> (WebLink 21.34); the Mitelman database is available at <http://cgap.nci.nih.gov/Chromosomes/Mitelman> (WebLink 21.35), and the OMIA website is <http://omia.angis.org.au/> (WebLink 21.36).

LOVD 3.0 is available online at <http://www.lovd.nl/3.0/home> (WebLink 21.37). It currently lists >22,000 genes.

HbVar is available at <http://globin.cse.psu.edu/hbvar/menu.html> (WebLink 21.38). It is a collaboration between investigators at Penn State University, INSERM Ccretel (France), and Boston University Medical Center.

A main point of entry to locus-specific databases is the Human Genome Variation Society (HGVS). This provides access to 1600 locus-specific mutation databases. Its major categories include: (1) locus-specific mutation databases, organized by HUGO approved gene symbols; (2) disease-centered central mutation databases, such as the Asthma Gene Database; (3) central mutation and SNP databases, such as OMIM, dbSNP, HGMD, and PharmGKB; (4) national and ethnic mutation databases, such as databases for diseases affecting Finns or Turks; (5) mitochondrial mutation databases, such as MITOMAP; (6) chromosomal variation databases, such as the Mitelman database of chromosome aberrations in cancer; (7) nonhuman mutation databases, such as OMIA (Online Mendelian Inheritance in Animals); and (8) clinical databases such as those of the National Organization for Rare Disorders (NORD).

The Leiden Open Variation Database (LOVD) has emerged as a platform supporting thousands of locus-specific databases (Fokkema *et al.*, 2011). This project provides software to establish locus-specific databases and curate data on individuals, phenotypes, and DNA sequencing variants following HGVS standards for nomenclature. LOVD provides access to Mutalyzer, a software package that confirms variant data are presented in a consistent standard.

As an example of a locus-specific database, we can examine HbVar (Giardine *et al.*, 2014). It can be accessed from searches in HGVS or LOVD, is linked from the NCBI Genome Workbench output of **Figure 21.13b**, and you can access it (and LOVD globin databases) at the bottom of its NCBI Gene page. The HbVar database is a useful resource for sequence variation associated with hemoglobinopathies, and is designed for both research purposes and clinical utility. The search page includes over a dozen fields that can be expanded to focus on a particular aspect of the globins such as those with particular physical properties (stability, chromatographic behavior, structural alterations) or functional properties (e.g., sickling of red blood cells, affinity of oxygen binding) or epidemiological aspects (ethnic background, frequency). There are currently over 6900 entries including categories such as entries involving hemoglobin variants (~980); thalassemia (~400 entries); the  $\alpha 1$ ,  $\alpha 2$ ,  $\beta$ ,  $\delta$ ,  $\text{A}\gamma$ , and  $\text{G}\gamma$  genes; and mutations involving insertions, deletions, substitutions, gene fusions, or altered stability or oxygen-binding properties.

## The PhenCode Project

Locus-specific mutation databases provide tremendous depth and breadth of information about one gene and/or disease. However, the information in these databases is usually separate from the wealth of information contained in major genome browsers. The PhenCode project connects data in locus-specific databases with genomic data from the UCSC Genome Browser (Giardine *et al.*, 2007), including the ENCODE project (described in Chapter 8). For a variety of locus-specific mutation databases, properties of interest can

be selected such as the type and location of the mutation. This information is then displayed as a custom track on the UCSC Genome Browser (Fig. 21.15, bottom panel). The significance of PhenCode is that it facilitates the exploration and discovery of genomic features associated with disease-causing mutations. For example, the genomic landscape could include ultraconserved elements in noncoding regions (Chapter 8) that are associated with disease, or repetitive elements that serve as substrates for recombination in deleted or duplicated regions.

The PhenCode website is  
🌐 <http://www.bx.psu.edu/phencode> (WebLink 21.39).

## Limitations of Disease Databases: The Growing Interpretive Gap

Databases reporting which alleles are associated with human disease have critical roles in the interpretation of the clinical significance of genomic variants. Data analysis pipelines for next-generation sequencing studies typically filter-exclude variants that are likely to be benign (neutral) because they appear in databases of apparently normal individuals. Such databases include dbSNP (although that contains a mixture of neutral and pathogenic SNPs) and the 1000 Genomes Project. Analysis pipelines typically filter-include variants that are likely to be pathogenic because they have been annotated as disease-associated. Many false positive results occur when a disease database contains entries that are actually neutral. For example, Bell *et al.* (2011) performed a targeted sequencing study involving severe, childhood, recessive diseases for which the causative mutation was likely to be extremely rare in the population. They reported that 74% of disease-associated calls were common polymorphisms with frequencies >5%. Also, 14 of 113 mutations that were annotated in the literature as disease mutations were incorrect.

Some of the challenges faced in assessing variants include the following (Cutting, 2014):

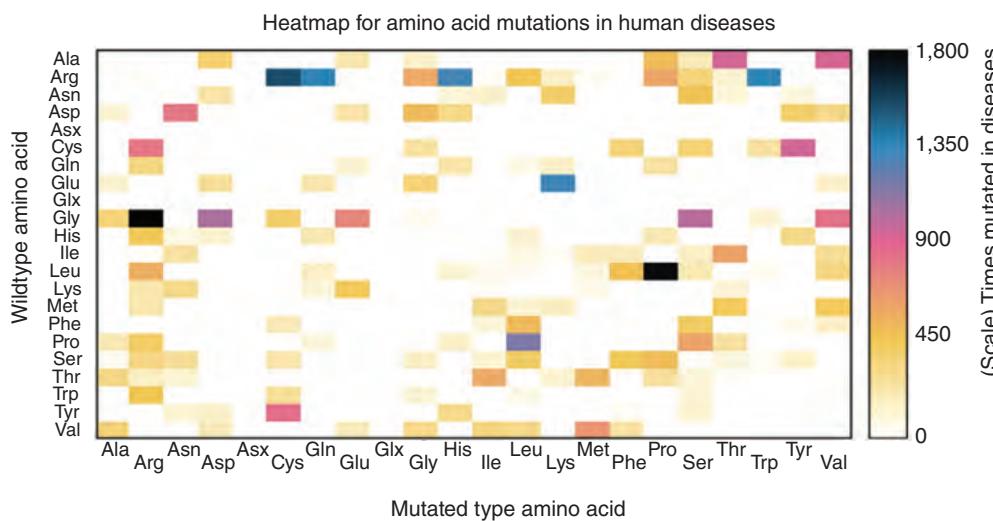
- For monogenic disorders, some variants in a disease-associated gene occur relatively frequently and their pathogenicity is established. For other rare variants, the clinical significance is unknown.
- For multigenic disorders, allelic heterogeneity makes the interpretation of the clinical significance of variants even more difficult.
- There is a large “interpretive gap” as increasing numbers of variants are identified, but their significance has not yet been assessed. Locus-specific databases are excellent repositories for the cataloguing of variants, but they also need associated clinical or phenotypic data.
- Databases such as the variants from the 1000 Genome Project are currently used to define neutral variants, but clinical and phenotypic data are not available for those individuals. Even if they are defined as “apparently normal,” all are susceptible to disease.

Efforts such as Gen2Phen are designed to broadly integrate human and model organism genetic variation databases into a federated network. Gen2Phen is establishing standards for data collection, storage, and sharing with a goal of facilitating genotype to phenotype studies (Webb *et al.*, 2011).

The Gen2Phen website is  
🌐 <http://www.gen2phen.org/> (WebLink 21.40).

## Human Disease Genes and Amino Acid Substitutions

The information in databases such as HGMD, OMIM, ClinVar, and central protein sequence repositories such as UniProt allows us to explore the amino acid substitutions that occur in human disease. Peterson *et al.* (2013) compiled substitutions in these databases and plotted a heatmap summarizing observed changes (Fig. 21.16). The three most common substitutions were leucine to proline, glycine to arginine, and arginine to



**FIGURE 21.16** Heat map of amino acid variants in human diseases. The observed frequencies of wildtype transitions to mutated variants that are implicated in human disease are shown. The variants are from OMIM, HGMD, UniProt/Swiss-Prot, and ClinVar. Redrawn from Peterson *et al.* (2013). Reproduced with permissions from Elsevier.

cysteine. From the BLOSUM62 matrix these have scores of  $-3$ ,  $-2$ , and  $0$  (Fig. 3.17) or, considering the substitutions of the PAM10 matrix involving closely related proteins, the scores are  $-10$ ,  $-13$ , and  $-11$ , respectively (Fig. 3.15).

These results are generally consistent with evolutionary approaches taken by Kumar and colleagues (Miller and Kumar, 2001; Miller *et al.*, 2003; Subramanian and Kumar, 2006). They combined data from locus-specific databases and alignments of metazoan orthologs (such as primates, rodents, fish, insects, and nematodes). There is general concordance between their results (Web Document 21.3) and those of Peterson *et al.* (2013). These analyses suggest that disease-associated changes tend to occur at conserved residues. Furthermore, the amino acid changes found in human disease do not commonly occur in comparisons between species.

## APPROACHES TO IDENTIFYING DISEASE-ASSOCIATED GENES AND LOCI

How can we determine the causes of diseases? There are many approaches to finding genes and loci that confer risk for the disease (Brunham and Hayden, 2013). By identifying such genes we may rationally develop treatments (or, ultimately, find cures). For example, phenylketonuria (PKU; OMIM +261600) is an inborn error of metabolism that results in intellectual disability and other symptoms. It is caused by a deficiency in phenylalanine hydroxylase activity. Knowing this, it is possible to screen newborns and, if PKU is found, to provide a diet lacking phenylalanine. PKU provides another example of the complexity of any disease. The enzyme phenylalanine hydroxylase is localized to the liver, and yet the symptoms of intellectual disability are neurological; if searching for the cause by studying brain tissue it would be challenging to discover any biochemical defects. Also, while mutation disrupting phenylalanine hydroxylase is overwhelmingly the major cause, it is not the only cause of PKU.

We next discuss several approaches that are used to identify disease-associated genes (or other genetic elements). Once a gene has been associated with a disease, it is further necessary to determine how susceptibility genes confer risk.

## Linkage Analysis

A genetic linkage map displays genetic information in reference to linkage groups (chromosomes) in a genome. The mapping units are centiMorgans, based on recombination frequency between polymorphic markers such as SNPs or microsatellites. (One cM equals one recombination event in 100 meioses; for the human genome, the recombination rate is typically 1–2 cM/Mb.)

In linkage studies, genetic markers are used to search for coinheritance of chromosomal regions within families, that is, polymorphic markers that flank a disease locus segregate with the disease in families. Two genes that are in proximity on a chromosome will usually cosegregate during meiosis. By following the pattern of transmission of a large set of markers in a large pedigree, linkage analysis can be used to localize a disease gene based on its linkage to a genetic marker locus. Huntington's disease (OMIM #143100), a progressive degenerative disorder, was the first autosomal disorder for which linkage analysis was used to identify the disease locus (reviewed in Gusella, 1989).

Linkage is usually successful for single-gene disease models rather than for complex traits. It also typically involves studies of large pedigrees. For Mendelian diseases the LOD score approach is used, providing a maximum likelihood estimate of the position of the disease locus (Ott, 2001; Szumilas, 2010). A LOD score of 3 implies that there is a 1 in 1000 chance that a given unlinked locus could have given rise to the observed cosegregation data. Many dozens of software packages are available for linkage analysis. Among the most widely used is Merlin (Multipoint Engine for Rapid Likelihood INference; Abecasis *et al.*, 2002).

Altschuler *et al.* (2008) reviewed genetic mapping by linkage and described these conclusions from studies of Mendelian disease genes:

- The “candidate gene” approach was inadequate because most disease genes could not have been predicted *a priori*.
- Mutations that cause disease often radically alter the function of encoded proteins.
- There is locus heterogeneity: there are often many disease-causing alleles within a gene, as we have seen in the example of *HBB* (e.g., Fig. 21.15). (There may also be locus heterogeneity involving distinct genes that cause a similar phenotype.)
- Mendelian diseases often display incomplete penetrance and variable expressivity.
- For common diseases linkage studies did not identify causal genes, consistent with a model in which common diseases are multigenic in origin.

## Genome-Wide Association Studies

While the genetic basis of over a thousand single-gene disorders has been found, it is far more difficult to identify the genetic causes of common human diseases that involve multiple genes. Part of the challenge is that a large number of genes may each make only a small contribution to the disease risk. Association studies provide an important approach (reviewed in Hirschhorn and Daly, 2005; McCarthy *et al.*, 2008; Pearson and Manolio, 2008). Genome-wide association studies (GWAS) provide a powerful approach that can rely on SNP microarrays (Chapter 8) having several hundred thousand to more than a million SNPs represented on a single array. There are two main experimental designs used in association studies (Laird and Lange, 2006). In family-based designs, markers are measured in affected individuals (probands) and unaffected individuals to identify differences in the frequency of variants (Ott *et al.*, 2011). In population-based designs, a large number of unrelated cases and controls are studied (typically hundreds or thousands in each group). Larger sample sizes offer increased statistical power.

As an example of a successful GWAS, Menzel *et al.* (2007) searched for variants associated with very high levels of fetal hemoglobin in adults. Fetal hemoglobin (HbF),

The Laboratory of Statistical Genetics at Rockefeller University, directed by Jürg Ott, offers a website listing dozens of software packages useful for linkage analysis. The Rockefeller website is <http://lab.rockefeller.edu/ott/> (WebLink 21.41). Merlin was developed by Gonçalo Abecasis and colleagues and is available at <http://www.sph.umich.edu/csg/abecasis/Merlin> (WebLink 21.42). Another popular software package, PLINK, was developed by Shaun Purcell and colleagues (2007) and is at <http://pngu.mgh.harvard.edu/~purcell/plink> (WebLink 21.43). We introduced PLINK in Chapter 20.

consisting of an  $\alpha_2\gamma_2$  tetramer, is normally expressed at high levels in early development but at negligible levels (<0.6% of total hemoglobin) in adults. Around 10–15% of adults have relatively high HbF levels (based on the presence of erythrocytes called F cells that contain measurable amounts of HbF). Such elevated HbF levels can be clinically beneficial because they improve the outcome of sickle cell disease and  $\beta$  thalassemia. Menzel *et al.* selected 179 unrelated individuals having very high or very low F cell levels, measured ~300,000 SNP genotypes for each individual, and identified major quantitative trait loci (QTLs). These included an expected variant in the beta globin gene cluster and two unexpected, independent QTLs overlapping a known oncogene, *BCL11A*, on chromosome 2p15. This exemplifies an ideal outcome for a GWAS: a relatively small number of subjects were genotyped and there was strong evidence of association for a gene not previously known to have any relationship to globin function. Functional studies could be pursued to further understand the mechanism of how *BCL11A* interacts with globins, and the possibility of clinical intervention in disease through manipulating *BCL11A* levels could be pursued.

More often the effect sizes in GWAS are relatively modest and, to achieve statistical power, larger sample sizes are employed. In one study of sexual dimorphism, genotype data from 270,000 individuals were analyzed (Randall *et al.*, 2013). In another study, Rietveld *et al.* (2013) performed a GWAS of >126,000 individuals to identify variants associated with educational attainment (based on years of education). They reported three SNPs with evidence of association; all had minuscule effect sizes (the largest having an effect size of 0.02%).

GWAS which succeed in identifying strong evidence of association often implicate intergenic regions far removed from protein-coding genes. Such loci could represent regulatory regions.

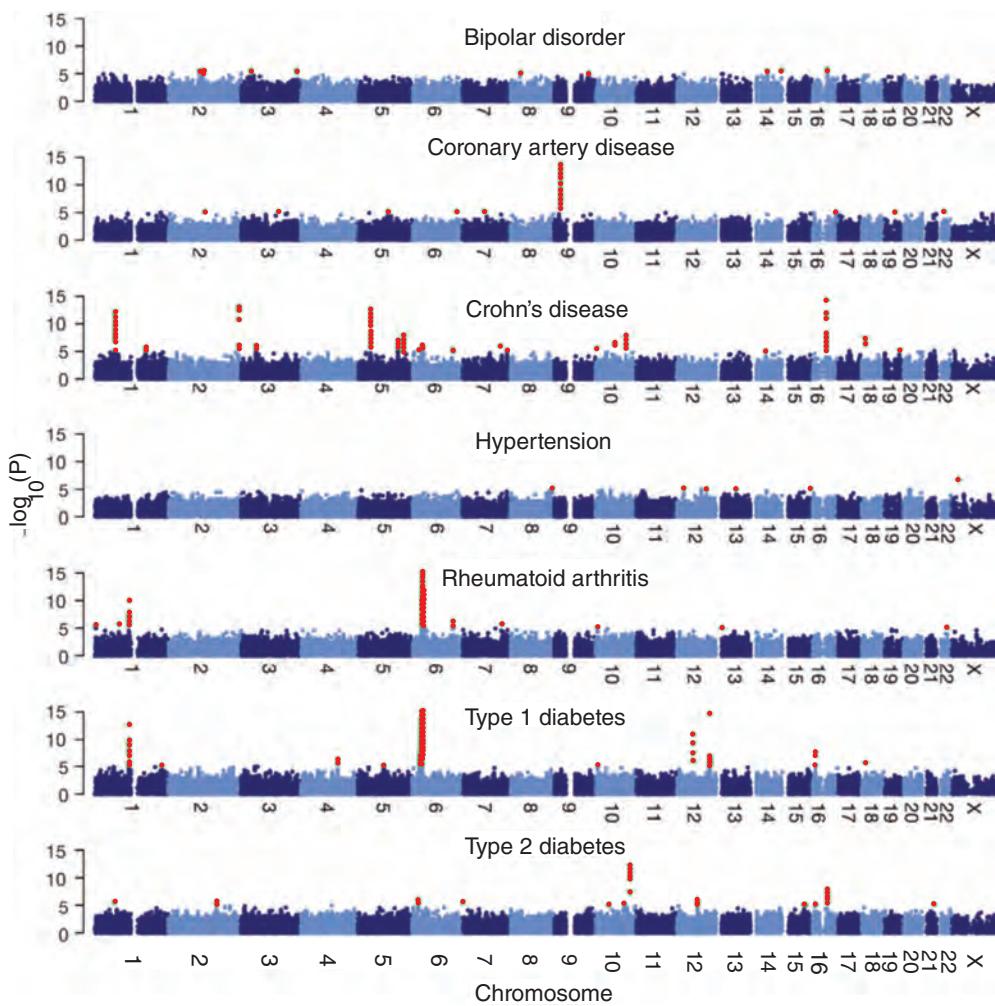
We can illustrate the genome-wide association approach with a large-scale study by the Wellcome Trust Case Control Consortium (2007) involving 50 research groups from the United Kingdom and >16,000 individuals (reviewed by Bowcock, 2007). Around 2000 affected individuals having one of seven common familial diseases – bipolar disorder, coronary artery disease, Crohn’s disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes – were studied. There were ~3000 control individuals. About 500,000 SNPs were measured for each individual, and the relationship between each SNP and the phenotypic trait (disease status) was measured. Twenty-four strong association signals were found for six of the seven diseases (Fig. 21.17). Many of these signals corresponded to previously characterized susceptibility loci, and many novel loci were also identified.

A key aspect of genome-wide association studies is that replication studies are required to confirm that positive signals are authentic. The NCI-NHGRI Working Group on Replication in Association Studies *et al.* (2007) has addressed many of the issues relevant to replication studies, emphasizing the need to eliminate false positive results that often occur. Proper experimental design is especially important, with efforts to assess phenotypes in a standard way and a need to account for biases such as population stratification.

There are several repositories of GWAS data. One is the Catalog of Published Genome-Wide Association Studies at the National Human Genome Research Institute (Hindorff *et al.*, 2009; Welter *et al.*, 2014). This includes an interactive diagram of the chromosomes, listing SNP-trait associations having  $p$  values  $<1\times10^{-5}$ .

The National Library of Medicine (NLM) offers the database of Genotype and Phenotype (dbGaP), a database of archived genome-wide association studies (Mailman *et al.*, 2007; Tryka *et al.*, 2014). dbGaP contains four types of data: (1) study documentation (e.g., protocols and data collection instruments); (2) phenotypic data (of individuals and as a summary); (3) genetic data (genotypes, pedigrees, mapping results); and

The NHGRI GWAS catalog is available at <http://www.genome.gov/gwastudies/> (WebLink 21.44). It is made in association with the European Bioinformatics Institute effort, online at <http://www.ebi.ac.uk/fgpt/gwas> (WebLink 21.45).



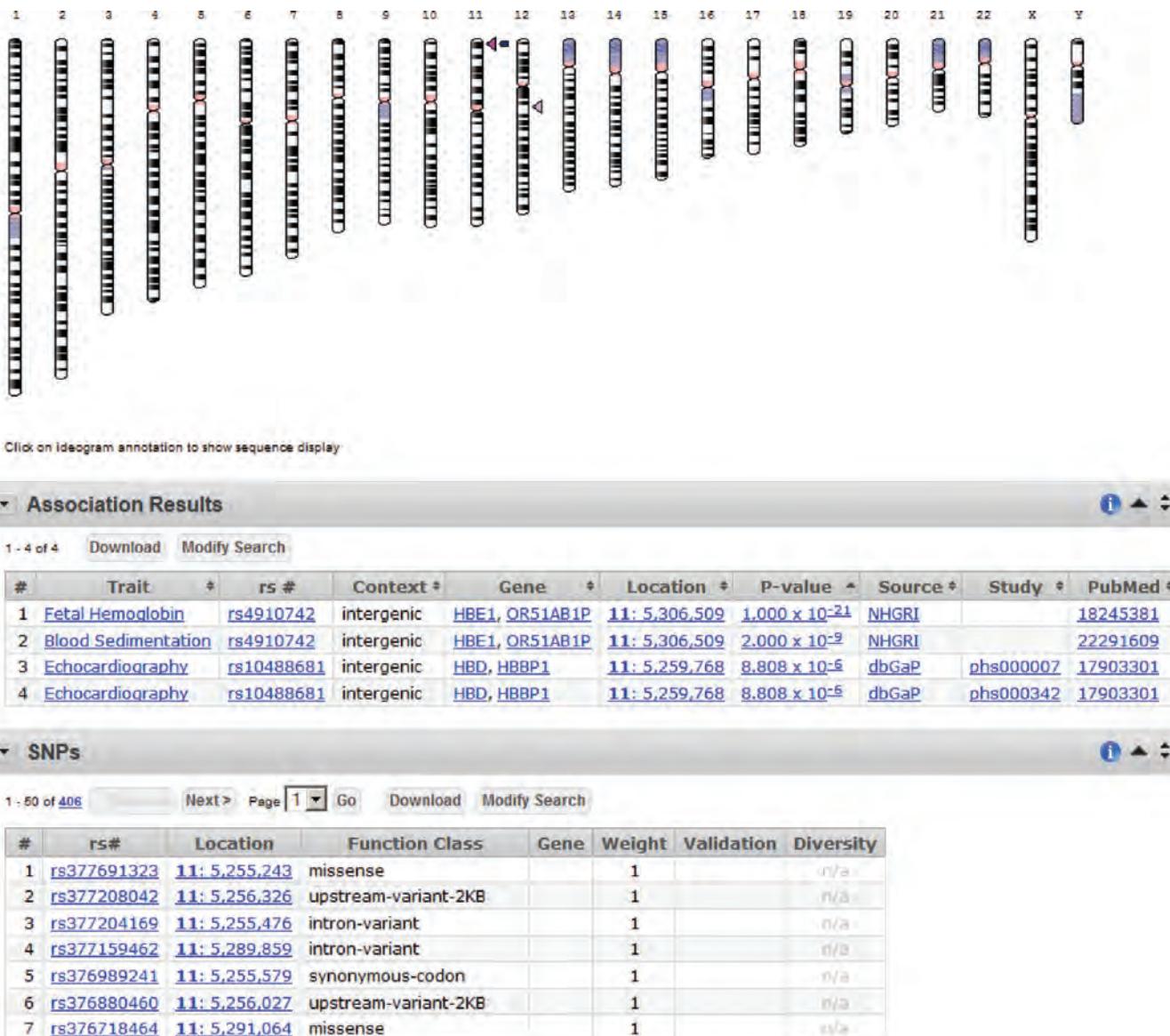
**FIGURE 21.17** Results of a genome-wide association study using 16,179 individuals to search for genes contributing to seven common familial disorders. For each of seven diseases, the  $y$  axis shows the  $-\log_{10} p$  value for SNPs that were positive for quality control criteria. The  $x$  axis shows the chromosomes.  $p$  values  $<1 \times 10^{-5}$  are highlighted in red. Panels are truncated at  $-\log_{10}(p \text{ value}) = 15$ . Redrawn from figure 4 of the Wellcome Trust Case Control Consortium (2007). Reproduced with permission from Macmillan Publishers.

(4) statistical results (e.g., linkage and association results). Open access is provided, as well as controlled access for situations in which permission from a committee is required to access information such as pedigrees or phenotypic data associated with genotype data.

If you have a gene of interest and you want to know whether it has been implicated in previous GWAS, one approach is to search the NHGRI catalog or dbGaP. You can also use the Phenotype-Genotype Integrator (PheGenI) tool at NCBI. Searches can begin with selection of phenotypes, location, gene(s), or SNPs. You can further restrict the search by  $p$  value of association and by SNP functional class (e.g., exon, intron, neighboring gene, or untranslated region). As an example, enter the text hbd and hbe1 for the delta and epsilon globin genes, *HBD* and *HBE1*. The output includes an ideogram (chromosome view) of SNPs showing significant evidence of association, association results (including one for the fetal hemoglobin trait with a  $p$  value of  $1 \times 10^{-21}$ ), and a list of SNPs including their functional class (Fig. 21.18). The output further displays expression quantitative trait locus (eQTL) data and relevant dbGaP studies.

dbGaP is available at  
<http://www.ncbi.nlm.nih.gov/gap> (WebLink 21.46).

PheGenI is available at  
<http://www.ncbi.nlm.nih.gov/gap/phegeni> (WebLink 21.47).



**FIGURE 21.18** The Phenotype and Genotype Integrator (PheGenI) tool at NCBI displays GWAS data from queries of traits, genes, SNPs, or genomic loci. Here a query with the gene symbols HBD and HBE1 results in an ideogram (top), association results (including fetal hemoglobin and blood sedimentation), and a list of SNPs.

Source: Phenotype and Genotype Integrator (PheGenI), NCBI.

### Identification of Chromosomal Abnormalities

The most common chromosomal aberrations in early development include the gain or loss of whole chromosomes. Such structural abnormalities may be detected by standard cytogenetic approaches such as karyotype analysis and fluorescence *in situ* hybridization (FISH). These techniques may also reveal commonly observed phenomena such as large-scale duplications, deletions, or rearrangements involving many millions of base pairs. One enhancement to FISH is spectral karyotyping/multiplex-FISH (SKY/M-FISH). This permits each chromosome to be depicted in a different color, facilitating the identification of abnormal karyotypes. In Chapter 8 we introduced array comparative genomic hybridization (aCGH), a form of genomic microarray using bacterial artificial chromosomes (BACs) that also represents an extension of FISH technology. NCBI offers a

SKY/M-FISH & CGH Database that includes tools to view SKY/M-FISH and aCGH data, particularly as ideograms of cancer datasets (Knutsen *et al.*, 2005).

Both genomic microarrays (aCGH) and SNP microarrays are used routinely to identify disease-associated chromosomal abnormalities. SNP microarrays offer higher resolution. (Currently, SNP arrays have ~one million markers per array, spaced several kilobases apart on average. Typical aCGH platforms have densely spaced oligonucleotide probes.) In addition to measuring copy number based on fluorescence intensity measurements, SNP technologies also permit estimates of genotypes which provides information about inheritance patterns and homozygosity. Both aCGH and SNP microarrays have been used to measure chromosomal variations in cancer, idiopathic intellectual disability, and a variety of other diseases.

The Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) is a web-based database that is a major resource for studying copy number imbalances in patients (Firth *et al.*, 2009). It can be searched by syndrome or karyotype, and includes data on copy number change in normal populations as well as patients. Data can be viewed through Ensembl or UCSC genome browsers (as shown in Fig. 21.15).

NCBI's SKY/M-FISH & CGH Database is available at  
🕒 <http://www.ncbi.nlm.nih.gov/sky/> (WebLink 21.48).

DECIPHER, led by Helen Firth, Nigel Carter and colleagues, is available at <http://decipher.sanger.ac.uk/> (WebLink 21.49).

## Human Genome Sequencing

We introduced high-throughput DNA sequencing in Chapter 9, and described genome sequencing projects in subsequent chapters. Another use of high-throughput sequencing is to resequence genomes, exomes, or targeted genomic regions in patients in order to define nucleotide differences that may be associated with disease.

As we explore the results of next-generation sequencing of patients, it is important to keep in mind how data analysis pipelines are only at their earliest stages of development. Gholson Lyon and colleagues compared multiple alignment and variant-calling pipelines (BWA-GATK, BWA-SAMtools, BWA-SNVer, GNUMAP, and SOAP; O'Rawe *et al.*, 2013). Single-nucleotide variant concordance across 15 exomes was ~57%, with 0.5–5.1% of variants unique to each pipeline. For three indel analysis pipelines, the concordance was only ~27%. These findings highlight the need for caution in interpreting variants in individual genomes, and the need for validation of findings.

### Genome Sequencing to Identify Monogenic Disorders

Exome sequencing has been particularly useful for identifying variants that cause monogenic disorders (Bamshad *et al.*, 2011). The majority of Mendelian diseases are caused primarily by mutations affecting the coding region of a gene. The yield of whole-exome sequencing has therefore been high: despite its focus on a small subset of the genome (typically ~60 megabases), the exome is enriched for functionally relevant loci. The main motivation to perform whole-exome rather than whole-genome sequencing is the cost of whole-genome sequencing that until recently has been substantially greater.

Targeted next-generation sequencing is a powerful approach to studying monogenic disorders. Stephen Kingsmore and colleagues used targeted sequencing as a preconception carrier screen for 448 severe, recessive, childhood diseases (Bell *et al.*, 2011). This approach targeted >7000 regions from 437 target genes. They observed an average carrier burden of 2.8 severe recessive substitutions, indels, or structural variants per genome (291 in 104 samples, primarily from individuals with severe, recessive disorders).

### Genome Sequencing to Solve Complex Disorders

Genome-wide association studies involving thousands of individuals with or without a given phenotype has been used to identify common alleles of small effect (as shown

in Fig. 21.6, lower right). GWAS has typically involved single-nucleotide polymorphism arrays; now whole-genome and whole-exome sequencing is being applied to complex diseases (Kilpinen and Barrett, 2013). While SNP arrays generate data on several hundred thousand (up to several million) variants, exome and genome sequencing permit broader variant discovery. Just as thousands of affected versus unaffected samples have been studied by GWAS in recent years, similarly large numbers of samples are currently being sequenced for complex disorders such as schizophrenia, bipolar disorder, and autism.

Recent studies on autism have employed father/mother/affected child trios (or sometimes quartets with two children). This design allows *de novo* mutations to be distinguished from inherited mutations. (A premise has been that *de novo* mutations are more likely to be relevant to the disease phenotype, although the possible presence of autism or subclinical features of autism in the parents is not typically assessed.) As an example, O’Roak *et al.* (2012) sequenced the exomes of 189 trios, most of whom lacked large *de novo* copy number variants. They reported 248 *de novo* variants (225 single-nucleotide variants, 17 indels, and 6 copy number variants); of the *de novo* events classified as severe, 33 of 120 (28%) were truncating. There were only two recurrent mutations (i.e., mutations occurring in the same gene in multiple affecteds), consistent with a model of extreme locus heterogeneity.

#### ***Research Versus Clinical Sequencing and Incidental Findings***

Whole-exome, whole-genome, and targeted next-generation sequencing can be performed in the context of research studies or clinical evaluations. (These purposes may also be combined.) In the United States, research studies must be approved by an Institutional Review Board (IRB) to confirm that appropriate procedures are in place. Informed consent must be obtained from the research participants; an informed consent document explains the risks and benefits of a study. For example, the risk of an exome study includes the potential loss of sequence data by the research team (e.g., if their computer server is breached) or the possible negative impact of learning that a family member has a disease-causing mutation.

Consider a research study involving whole-exome sequencing of a child with autism and his/her parents. The inclusion of the parents’ exomes is critical because it allows inherited variants to be distinguished from *de novo* variants. What procedure will be followed if a parent or child has a mutation in a cancer-causing gene? This possibility should be addressed as part of the informed consent process, and the IRB should review this procedure.

For clinical sequencing, the American College of Medical Genetics and Genomics (ACMG) issued recommendations for reporting incidental findings in exome and genome sequencing (Green *et al.*, 2013). They define a primary finding as “pathogenic alterations in a gene or genes that are relevant to the diagnostic indication for which the sequencing was ordered (e.g., a mutation in *MECP2* in a girl with loss of developmental milestones).” Incidental findings are unexpected positive findings. These are “the results of a deliberate search for pathogenic or likely pathogenic alterations in genes that are not apparently relevant to a diagnostic indication for which the sequencing test was ordered.” They produced a list of 56 genes (mutations which cause conditions such as cancer) for which they recommend results be returned (Table 21.11). In brief, the ACMG recommendations are as follows (with additional key details given in the ACMG paper):

1. Constitutional mutations found in the genes (listed in Table 21.11) should be reported by the laboratory to the ordering clinician.
2. Laboratories should seek and report only the types of variants in this list of genes.
3. It is the responsibility of the ordering clinician to provide pre- and post-test counseling to the patient.
4. These recommendations are focused on disorders caused by point mutations and small indels rather than structural variants, repeat expansions, or copy number variants.
5. The ACMG and others should refine and update the list of genes frequently.

**TABLE 21.11 Conditions, genes, and variants recommended by the ACMG for return of incidental findings in clinical sequencing. Abbreviations for inheritance: AD: autosomal dominant; SD: semidominant; XL: X-linked. Abbreviations for variants to report: KP: known pathogenic; EP: expected pathogenic (sequence variation is previously unreported and is expected to cause the disorder).**

Phenotype	MIM (disorder)	PMID (GeneReviews)	Typical age of onset	Gene	MIM (gene)	Inher.	Variants to report
Hereditary breast and ovarian cancer	604370, 612555	20301425	Adult	<i>BRCA1</i> <i>BRCA2</i>	113705 600185	AD	KP, EP
Li–Fraumeni syndrome	175200	20301488	Child/adult	<i>TP53</i>	191170	AD	KP, EP
Peutz–Jeghers syndrome	175200	20301443	Child/adult	<i>TK11</i>	602216	AD	KP, EP
Lynch syndrome	120435	20301390	Adult	<i>MLH1</i> <i>MSH2</i> <i>MSH6</i> <i>PMS2</i>	120436 609309 600678 600259	AD	KP, EP
Familial adenomatous polyposis MYH-associated polyposis; adenomas, multiple colorectal, FAP type 2; colorectal adenomatous polyposis, autosomal recessive, with pilomatricomas	175100 608456 132600	20301519 23035301	Child Adult	<i>APC</i> <i>MUTYH</i>	611731 604933	AD AR	KP, EP KP, EP
Von Hippel–Lindau syndrome	193300	20301636	Child/adult	<i>VHL</i>	608537	AD	KP, EP
Multiple endocrine neoplasia type 1	131100	20301710	Child/adult	<i>MEN1</i>	613733	AD	KP, EP
Multiple endocrine neoplasia type 2	171400 162300	20301434	Child/adult	<i>RET</i>	164761	AD	KP
Familial medullary thyroid cancer	1552401	20301434	Child/adult	<i>RET</i>	164761	AD	KP
<i>PTEN</i> hamartoma tumor syndrome	153480	20301661	Child/adult	<i>PTEN</i>	601728	AD	KP, EP
Retinoblastoma	180200	20301625	Child	<i>RB1</i>	614041	AD	KP, EP
Hereditary paraganglioma–pheochromocytoma syndrome	168000 (PGL1) 601650 (PGL2) 605373 (PGL3) 115310 (PGL4)	20301715	Child/adult	<i>SDHD</i> <i>SDHAF2</i> <i>SDHC</i> <i>SDHB</i>	602690 613019 602413 185470	AD	KP, EP KP KP, EP
Tuberous sclerosis complex	191100 613254	20301399	Child	<i>TSC1</i> <i>TSC2</i>	605284 191092	AD	KP, EP
WT1-related Wilms tumor	194070	20301471	Child	<i>WT1</i>	607102	AD	KP, EP
Neurofibromatosis type 2	101100	20301380	Child/adult	<i>NF2</i>	607379	AD	KP, EP
Ehlers–Danlos syndrome, vascular type	130050	20301667	Child/adult	<i>COL3A1</i>	120180	AD	KP, EP
Marfan syndrome, Loeys–Dietz syndromes, and familial thoracic aortic aneurysms and dissections	154700 609192 608967 610168 610380 613795 611788	20301510 20301312 20301299	Child/adult	<i>FBN1</i> <i>TGFBR1</i> <i>TGFBR2</i> <i>SMAD3</i> <i>ACTA2</i> <i>MYLK</i> <i>MYH11</i>	134797 190181 190182 603109 102620 600922 160745	AD	KP, EP

(Continued)

**TABLE 21.11 (continued)**

Phenotype	MIM (disorder)	PMID (GeneReviews)	Typical age of onset	Gene	MIM (gene)	Inher.	Variants to report
Hypertrophic cardiomyopathy, dilated cardiomyopathy	115197	20301725	Child/adult	MYBPC3	600958	AD	KP, EP
	192600			MYH7	160760		KP
	601094			TNNT2	191045		KP, EP
	613690			TNNI3	191044		KP
	115196			TPM1	191010		
	608751			MYL3	160790		
	612098			ACTC1	102540		
	600858			PRKAG2	602743		
	301500			GLA	300644	XL	KP, EP
	608758			MYL2	160781	AD	KP
	115200			LMNA	150330		KP, EP
Catecholaminergic polymorphic ventricular tachycardia	604772	—	—	RYR2	180902	AD	KP
Arrhythmogenic right- ventricular cardiomyopathy	609040	20301310	Child/adult	PKP2	602861	AD	KP, EP
	604400			DSP	125647		
	610476			DSC2	125645		
	607450			TMEM43	612048		KP
	610193			DSG2	125671		KP, EP
Romano–Ward long QT syndrome types 1, 2, and 3, Brugada syndrome	192500	20301308	Child/adult	KCNQ1	607542	AD	KP, EP
	613688			KCNH2	152427		
	603830			SCN5A	600163		
	601144						
Familial hypercholesterolemia	143890	No GeneReviews entry	Child/adult	LDLR	606945	SD	KP, EP
	603776			APOB	107730	SD	KP
				PCSK9	607786	AD	
Malignant hyperthermia susceptibility	145600	20301325	Child/adult	RYR1	180901	AD	KP
				CACMA1S	114208		

Source: Green *et al.* (2013). Reproduced with permission from Macmillan Publishers.

There are many ethical issues surrounding research and clinical sequencing technologies. For example, predictive testing for adult-onset diseases is performed by exome or genome sequencing of a child, and that information could impact the child, siblings, parents, and the entire family. For research-based sequencing it is common for researchers to explore as many variants as possible. These results are sometimes shared with study participants (depending on the details of the informed consent) even though the functional consequences of variation are always difficult to interpret.

#### *Disease-causing Variants in Apparently Normal Individuals*

How many disease-associated variants occur in healthy people? Specifically, how many variants that disrupt the function of protein-coding genes occur in apparently normal people? One might expect the answer to be extremely few, yet ~100 such variants may occur in each genome (MacArthur and Tyler-Smith, 2010).

The 1000 Genomes Project identified variants in apparently normal human populations (Chapter 20). Members of that consortium identified >2600 HGMD entries in the 1000 Genomes low-coverage pilot data (Xue *et al.*, 2012). Each individual harbored 281–515 missense mutations (40–85 of which were homozygous and predicted to be deleterious). Furthermore, each individual had 40–110 variants identified as disease-causing in the HGMD database. Of these variants 3–24 were homozygous, meaning that both chromosomal copies carried the deleterious variants. There are two perspectives on these findings. First, the number of deleterious alleles present in the genome of even apparently normal

individuals is quite high. Second, the databases such as HGMD that predict which variants are deleterious may include false positive entries. Xue *et al.* report that of 577 variants in the 1000 Genomes population that were classified as disease-causing mutations by HGMD, >90% were not predicted to be severely damaging by their analyses. Annotation of disease-associated variants therefore needs to be improved across all relevant databases.

Other studies involving the 1000 Genomes Project Consortium *et al.* (2010, 2012; MacArthur *et al.*, 2012) reached similar conclusions about the numbers of loss of function variants per individual (about 100) and the number of completely inactivated genes (~20 per person). MacArthur *et al.* noted that loss-of-function variants in healthy individuals may be categorized several ways:

- Severe recessive alleles may occur in the heterozygous state.
- Alleles that are not severe may still impact disease risk and phenotype.
- There is benign loss of function variation (perhaps an example would be the loss of an olfactory receptor gene).
- There are variants that do not appreciably disrupt gene function.
- Many variants represent sequencing and annotation artifacts.

MacArthur *et al.* identified, validated, and characterized many loss-of-function variants. HapMap individual NA12878 (whose genome has been extensively sequenced with multiple technologies) has 97 loss-of-function variants, including 26 that are known recessive disease-causing mutations (present in the heterozygous state) and 18 that are homozygous. For some variants, the loss of function was accompanied by a decreased level of RNA expression.

We described incidental findings above. How many clinically relevant incidental findings occur in apparently normal individuals? Dorshner *et al.* (2013) approached this problem by analyzing the exome sequences of 1000 adult individuals (half of European descent and half of African descent) with an emphasis on 114 medically actionable genes (i.e., genes for which variants are highly penetrant, pathological, and for which resulting medical recommendations could improve clinical outcomes in terms of morbidity and mortality). Their gene list included 52 of the 56 from the ACMG list described above.

- A total of 585 instances of 239 unique variants were found that were defined as disease-causing in HGMD. Most of those were false positive results, however; only 16 unique autosomal dominant variants were defined by Dorshner *et al.* as pathogenic or likely pathogenic.
- Pathogenic variants have low allele frequencies (typically <0.1% and in 15 cases out of 16 were observed just once in the cohort of 1000 individuals).
- Fewer pathogenic variants were identified in the individuals of African ancestry, perhaps reflecting the paucity of genetics literature on non-European populations.

## HUMAN DISEASE GENES IN MODEL ORGANISMS

The study of human disease genes and gene products in other organisms is of fundamental importance in our efforts to understand the pathophysiology of human disease. While mutations in genes cause many diseases, it is the aberrant protein product that has the proximal functional consequence on the cell and ultimately on the organism. Once a human disease gene is identified in a model organism, it can often be knocked out or otherwise manipulated. This allows the phenotypic consequences of specific mutations to be assessed. Earlier we introduced Rett syndrome. Over a dozen mouse models have been developed (Katz *et al.*, 2012), enabling Adrian Bird and colleagues to demonstrate phenotypic reversal of symptoms in adult mice (Guy *et al.*, 2007). Complementary studies in a *Drosophila* model confirmed anatomical and behavioral abnormalities (Cukier *et al.*, 2008). This work led to the identification of genetic modifiers relevant to the function of *MECP2*.

## Human Disease Orthologs in Nonvertebrate Species

A basic approach is to identify which known human disease genes have orthologs in model organisms. This is of interest even though the consequence of mutating that ortholog may differ. *Drosophila* has become an established model for overexpression of gain-of-function deleterious mutations that occur in genes that are orthologous to human disease genes (Chen and Crowther, 2012).

At the time that *C. elegans* was sequenced, about 65% of human disease genes had identifiable *C. elegans* orthologs (Ahringer, 1997).

Which human disease genes have orthologs in nonvertebrates? In an early comparative genomics study Rubin *et al.* (2000) analyzed the newly sequenced genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. They identified 289 genes that are mutated, altered, amplified, or deleted in human disease. Of these genes, 177 (61%) were found to have an ortholog in *Drosophila*. These data are displayed in Web Document 21.4 showing the presence of fly, worm, and yeast orthologs to human disease genes that are functionally categorized in cancer, neurological, cardiovascular, endocrine, and other disease types. Reiter *et al.* (2001) extended this study to 929 human disease genes in OMIM, 714 of which (77%) matched 548 *Drosophila* protein sequences (Web Document 21.5).

The cataloguing of human disease genes in model organisms is important in our efforts to establish functional assays for these genes. In addition to the results in *S. cerevisiae*, *D. melanogaster*, and *C. elegans*, similar descriptions have been made in other eukaryotes such as *Schizosaccharomyces pombe* (Wood *et al.*, 2002), *Arabidopsis* (Arabidopsis Genome Initiative, 2000), and the amoeba *Dictyostelium discoideum* (Eichinger *et al.*, 2005). For *S. pombe*, orthologs were identified both for human cancer genes (Table 21.12) and a

**TABLE 21.12** *Schizosaccharomyces pombe* genes related to human cancer genes. Score is the expect value from a BLAST search; a score of  $<1 \times 10^{-40}$  refers to a score between  $<1 \times 10^{-40}$  and  $1 \times 10^{-100}$ . Adapted from Wood *et al.* (2002), with permission from Macmillan Publishers.

Human cancer gene	Score	<i>S. pombe</i> gene/product	Systematic name
Xeroderma pigmentosum D; XPD	$<1 \times 10^{-100}$	rad15, rhp3	SPAC1D4.12
Xeroderma pigmentosum B; ERCC3	$<1 \times 10^{-100}$	rad25	SPAC17A5.06
Hereditary nonpolyposis colorectal cancer (HNPCC); MSH2	$<1 \times 10^{-100}$	rad16, rad10, rad20, swi9	SPBC24C6.12C
Xeroderma pigmentosum F; XPF	$<1 \times 10^{-100}$	cdc17	SPCC970.01
HNPCC; PMS2	$<1 \times 10^{-100}$	pms1	SPAC57A10.13C
HNPCC; MSH6	$<1 \times 10^{-100}$	msh6	SPAC19G12.02C
HNPCC; MSH3	$<1 \times 10^{-100}$	swi4	SPCC285.16C
HNPCC; MLH1	$<1 \times 10^{-100}$	mlh1	SPAC8F11.03
Haematological Chediak–Higashi syndrome; CHS1	$<1 \times 10^{-100}$	—	SPBC1703.4
Darier–White disease; SERCA	$<1 \times 10^{-100}$	Pgak	SPBC28E12.06C
Bloom syndrome; BLM	$<1 \times 10^{-100}$	Hus2, rqh1, rad12	SPBC31E1.02C
Ataxia telangiectasia; ATM	$<1 \times 10^{-100}$	Tel1	SPAC2G11.12
Xeroderma pigmentosum G; XPG	$<1 \times 10^{-40}$	rad13	SPBC3E7.08C
Tuberous sclerosis 2; TSC2	$<1 \times 10^{-40}$	—	SPAC630.13C
Immune bare lymphocyte; ABCB3	$<1 \times 10^{-40}$	—	SPBC9B6.09C
Downregulated in adenoma; DRA	$<1 \times 10^{-40}$	—	SPAC869.05C
Diamond–Blackfan anemia; RPS19	$<1 \times 10^{-40}$	rps19	SPBC649.02
Cockayne syndrome 1; CKN1	$<1 \times 10^{-40}$	—	SPBC577.09
RAS	$<1 \times 10^{-40}$	Ste5, ras1	SPAC17H9.09C
Cyclin-dependent kinase 4; CDK4	$<1 \times 10^{-40}$	Cdc2	SPBC11B10.09
CHK2 protein kinase	$<1 \times 10^{-40}$	Cds1	SPCC18B5.11C
AKT2	$<1 \times 10^{-40}$	Pck2, sts6, pkc1	SPBC12D12.04C

**TABLE 21.13** *Schizosaccharomyces pombe* genes related to human disease genes. Score is the expect value from a BLAST search. GNP: guanine nucleotide binding. Adapted from Wood *et al.* (2002), with permission from Macmillan Publishers.

Human cancer gene	Disease	Score	<i>S. pombe</i> gene/product
Wilson disease; <i>ATP7B</i>	Metabolic	<1×10 <sup>-100</sup>	P-type copper ATPase
Non-insulin-dependent diabetes; <i>PCSK1</i>	Metabolic	<1×10 <sup>-100</sup>	Krp1, kinesin related
Hyperinsulinism; <i>ABCC8</i>	Metabolic	<1×10 <sup>-100</sup>	ABC transporter
G6PD deficiency; <i>G6PD</i>	Metabolic	<1×10 <sup>-100</sup>	Zwf1 GP6 dehydrogenase
Citrullinemia type I; <i>ASS</i>	Metabolic	<1×10 <sup>-100</sup>	Arginosuccinate synthase
Wernicke–Korsakoff syndrome; <i>TKT</i>	Metabolic	<1×10 <sup>-40</sup>	Transketolase
Variegate porphyria; <i>PPOX</i>	Metabolic	<1×10 <sup>-40</sup>	Protoporphyrinogen oxidase
Maturity-onset diabetes of the young (MODY2); <i>GCK</i>	Metabolic	<1×10 <sup>-40</sup>	Hxk1, hexokinase
Gitelman's syndrome; <i>SLC12A3</i>	Metabolic	<1×10 <sup>-40</sup>	CCC Na-K-Cl transporter
Cystinuria type 1; <i>SLC3A1</i>	Metabolic	<1×10 <sup>-40</sup>	α-Glucosidase
Cystic fibrosis; <i>ABCC7</i>	Metabolic	<1×10 <sup>-40</sup>	ABC transporter
Bartter's syndrome; <i>SLC12A1</i>	Metabolic	<1×10 <sup>-40</sup>	CCC Na-K-Cl transporter
Menkes syndrome; <i>ATP7A</i>	Neurological	<1×10 <sup>-100</sup>	P-type copper ATPase
Deafness, hereditary; <i>MYO15</i>	Neurological	<1×10 <sup>-100</sup>	Myo51 class V myosin
Zellweger syndrome; <i>PEX1</i>	Neurological	<1×10 <sup>-40</sup>	AAA-family ATPase
Thomsen disease; <i>CLCN1</i>	Neurological	<1×10 <sup>-40</sup>	CIC chloride channel
Spinocerebellar ataxia type 6 (SCA6); <i>CACNA1A</i>	Neurological	<1×10 <sup>-40</sup>	VIC sodium channel
Myotonic dystrophy; <i>DM1</i>	Neurological	<1×10 <sup>-40</sup>	Orb6 Ser/Thr protein kinase
McCune–Albright syndrome; <i>GNAS1</i>	Neurological	<1×10 <sup>-40</sup>	Gpa1 GNP
Lowe's oculocerebrorenal syndrome; <i>OCRL</i>	Neurological	<1×10 <sup>-40</sup>	PIP phosphatase
Dents; <i>CLCN5</i>	Neurological	<1×10 <sup>-40</sup>	CIC chloride channel
Coffin–Lowry; <i>RPS6KA3</i>	Neurological	<1×10 <sup>-40</sup>	Ser/Thr protein kinase
Angelman; <i>UBE3A</i>	Neurological	<1×10 <sup>-40</sup>	Ubiquitin–protein Igase
Amyotrophic lateral sclerosis; <i>SOD1</i>	Neurological	<1×10 <sup>-40</sup>	Sod1, superoxide dismutase
Oguschi type 2; <i>RHKIN</i>	Neurological	<1×10 <sup>-40</sup>	Ser/Thr protein kinase
Familial cardiac myopathy; <i>MYH7</i>	Cardiac	<1×10 <sup>-100</sup>	Myo2, myosin II
Renal tubular acidosis; <i>ATP6B1</i>	Renal	<1×10 <sup>-100</sup>	V-type ATPase

variety of neurological, metabolic, and other disorders (Table 21.13). In *Dictyostelium*, which is intermediate in complexity between fungi and multicellular animals, many human disease orthologs were identified including nine that were absent from *S. pombe* and/or *S. cerevisiae*.

It is perhaps expected that human genes involved in cancer are also present in fungi; examples include genes encoding proteins involved in DNA damage and repair and the cell cycle. It might seem surprising that genes implicated in neurological disorders are present in single-celled fungi. However, the explanation may be that neurons are a particularly susceptible cell type with unique metabolic requirements. For example, most lysosomal disorders are caused by the loss of an enzyme that normally contributes to lysosomal function or to intracellular trafficking to lysosomes. Multiple organ systems are typically compromised, but neurological features such as intellectual disability are a common consequence of these disorders. The lysosome is a primary site for catabolism in the cell. The vacuole performs similar functions in fungi, and many human homologs of fungal vacuolar proteins have been identified.

## Human Disease Orthologs in Rodents

The mouse genome, reported by the Mouse Genome Sequencing Consortium *et al.* (2002), presents us with perhaps the most important animal model of human disease. A number of key resources are available:

You can access this mouse information at [http://www.rodentia.com/wmc/domain\\_genome.html#transgenics](http://www.rodentia.com/wmc/domain_genome.html#transgenics) (WebLink 21.50) and [http://www.rodentia.com/wmc/domain\\_mouse.html](http://www.rodentia.com/wmc/domain_mouse.html) (WebLink 21.51).

- The FANTOM database, part of the RIKEN Mouse Gene Encyclopedia Project, contains information on full-length mouse cDNA clones (Kawaji *et al.*, 2011).
- The Jackson Laboratory website offers a list of mouse/human gene homologs, including mouse models for human disease.
- High-efficiency mutagens such as *N*-ethyl-*N*-nitrosourea (ENU) or radiation have been applied to mice to generate models of human disease (see Chapter 14) (Probst and Justice, 2010; Stottmann and Beier, 2010).
- The Whole Mouse Catalog describes mouse models of human disease.

The sequencing of the mouse genome was achieved by both Celera Genomics and by a public consortium (Chapter 19). Celera sequenced the genomic DNA of several mouse strains and noted their differences in susceptibility to infectious disease (**Table 21.14**) and complex inherited disease (**Table 21.15**). Comparative genomic data will likely help explain why some mouse strains vary in their disease susceptibility.

The public consortium that sequenced the mouse genome reported that 687 human disease genes have clear orthologs in mouse (Mouse Genome Sequencing Consortium *et al.*, 2002). Surprisingly, for several dozen genes, the wildtype mouse gene sequence was identical to the sequence that is associated with disease in humans. These genes are

**TABLE 21.14 Infectious disease susceptibility of mouse strains.**

Infectious disease	Inbred mouse strain	
	A/J	C57BL/6J
Legionnaire's pneumonia	Susceptible	Resistant
Malaria	Susceptible	Resistant
Viral (MHV3) hepatitis	Resistant	Susceptible
Murine AIDS	Resistant	Susceptible

**TABLE 21.15 Common complex disease susceptibility of mouse strains.**

Complex disease	Inbred mouse strain	
	A/J	C57BL/6J
Arthritis	Susceptible	Resistant
Colon cancer	Susceptible	Resistant
Lung cancer	Susceptible	Resistant
Asthma	Susceptible	Resistant
Atherosclerosis	Resistant	Susceptible
Hypertension	Resistant	Susceptible
Type II diabetes	Resistant	Susceptible
Osteoporosis	Susceptible	Resistant
Obesity	Resistant	Susceptible

listed in **Table 21.16**. This suggests that, assuming the mouse does not have these diseases, any mouse model for these diseases must be used with caution. Conceivably, mice have modifying genes (or paralogous genes) not present in humans. Also, inbred strains of laboratory mice are exposed to different environmental stressors than mice in the wild, and their disease susceptibility could vary.

Sequencing of the genome of the Norway rat (Chapter 19; Rat Genome Sequencing Project Consortium, 2004) allowed the detailed comparison of human, mouse, and rat disease genes. Of 1112 well-characterized human disease genes from HGMD (described above), 76% have orthologs in rat. This is a higher percentage than for all rat versus all human genes (of which 46% have 1:1 orthologous matches). Only six human disease genes were found to lack rat orthologs. In general, the consortium concluded that human disease genes tend to be well conserved in rat, as also indicated by measurement of  $K_N/K_S$  ratios.

### Human Disease Orthologs in Primates

While the chimpanzee and human genomes are extremely closely related (Chimpanzee Sequencing and Analysis Consortium, 2005; Chapter 19), it is surprising that many common human disease variants correspond to the wildtype form allele in the chimpanzee;

**TABLE 21.16 Human disease-associated sequence variants for which wildtype mouse sequence matches diseased human sequence. Adapted from Mouse Genome Sequencing Consortium *et al.* (2002), with permission from Macmillan Publishers.**

Disease	OMIM	Mutation
Hirschsprung disease	142623	E251K
Leukencephaly with vanishing white matter	603896	R113H
Mucopolysaccharidosis type IVA	253000	R376Q
Breast cancer	113705	L892S
Breast cancer	600185	V211A, Q2421H
Parkinson's disease	601508	A53T
Tuberous sclerosis	605284	Q654E
Bardet–Biedl syndrome, type 6	209900	T57A
Mesothelioma	156240	N93S
Long QT syndrome 5	176261	V109I
Cystic fibrosis	602421	F87L, V754M
Porphyria variegata	176200	Q127H
Non-Hodgkin's lymphoma	605027	A25T, P183L
Severe combined immunodeficiency disease	102700	R142Q
Limb-girdle muscular dystrophy type 2D	254110	P30L
Long-chain acyl-CoA dehydrogenase deficiency	201460	Q333K
Usher syndrome type 1B	276902	G955S
Chronic nonspherocytic haemolytic anemia	206400	A295V
Mantle cell lymphoma	208900	N750K
Becker muscular dystrophy	300377	H2921R
Complete androgen insensitivity syndrome	300068	G491S
Prostate cancer	176807	P269S, S647N
Crohn's disease	266600	W157R

**TABLE 21.17 Human disease variants matching the wildtype chimpanzee allele.**  
 Variants are listed as benign variant, codon number, disease/chimpanzee variant.  
 Ancestral variants are inferred using primate outgroups. Frequency is of the disease allele in humans. *PON1* (Q192R) is polymorphic in chimpanzee.

Gene	Variant	Disease association	Ancestral	Frequency
<i>AIRE</i>	P252L	Autoimmune syndrome	Unresolved	0
<i>MKKS</i>	R518H	Bardet–Biedl syndrome	Wild type	0
<i>MLH1</i>	A441T	Colorectal cancer	Wild type	0
<i>MYOC</i>	Q48H	Glaucoma	Wild type	0
<i>OTC</i>	T125M	Hyperammonemia	Wild type	0
<i>PRSS1</i>	N29T	Pacreatitis	Disease	0
<i>ABCA1</i>	I883M	Coronary artery disease	Unresolved	0.136
<i>APOE</i>	C130R	Coronary artery disease and Alzheimer's disease	Disease	0.15
<i>DIO2</i>	T92A	Insulin resistance	Disease	0.35
<i>ENPP1</i>	K121Q	Insulin resistance	Disease	0.17
<i>GSTP1</i>	I105V	Oral cancer	Disease	0.348
<i>PON1</i>	I102V	Prostate cancer	Wild type	0.016
<i>PON1</i>	Q192R	Coronary artery disease	Disease	0.3
<i>PPARG</i>	A12P	Type 2 diabetes	Disease	0.85
<i>SLC2A2</i>	T110I	Type 2 diabetes	Disease	0.12
<i>UCP1</i>	A64T	Waist-to-hip ratio	Disease	0.12

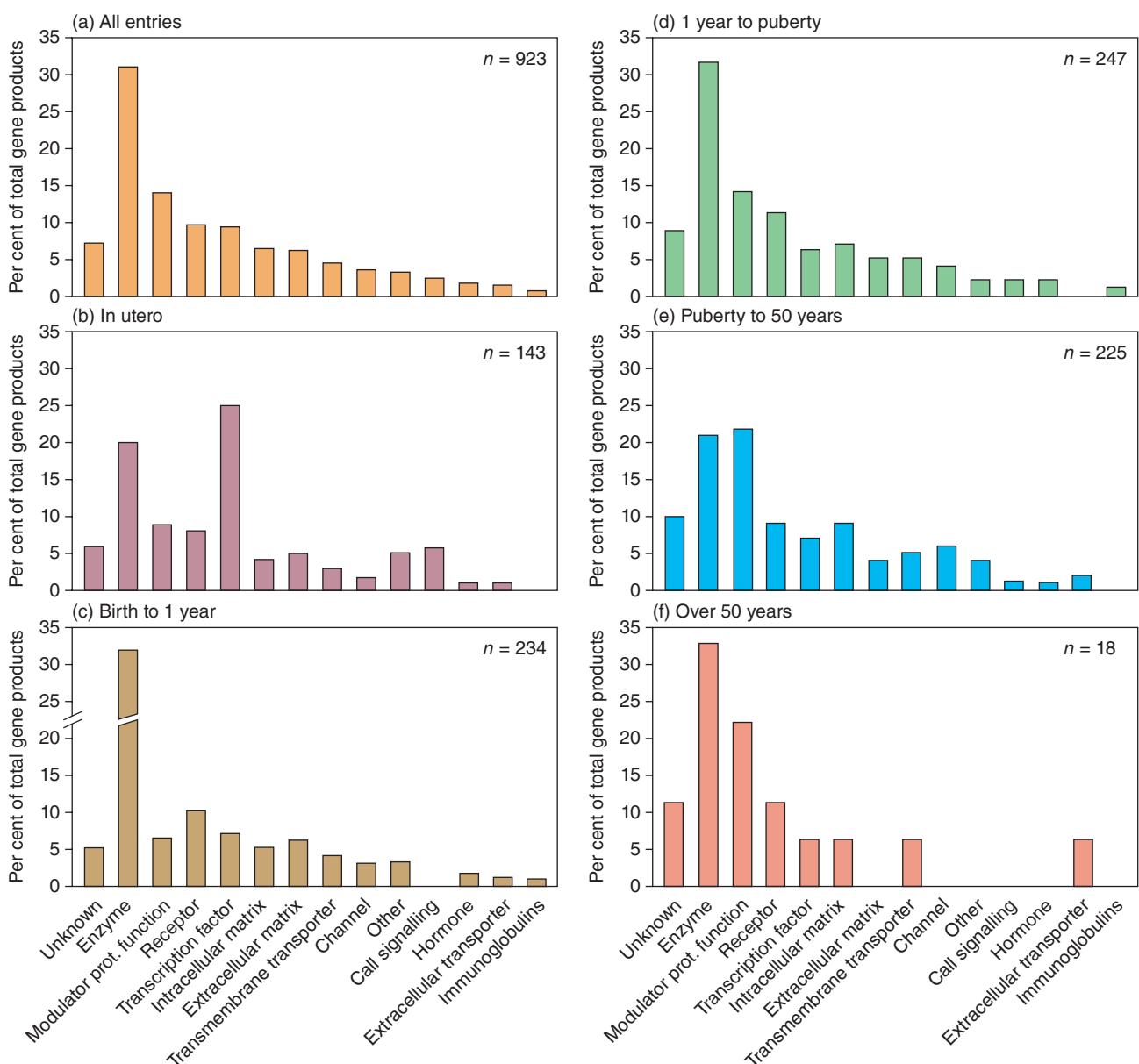
Source: Chimpanzee Sequencing and Analysis Consortium (2005). Reproduced with permission from Macmillan Publishers.

sixteen examples are presented in **Table 21.17**. In the gorilla, several genes also have wild-type sequences that correspond to human disease alleles (*GRN*, *TCAP*, and the globin *HBA1*; Scally *et al.*, 2012).

It is possible that not all of these mutations are true positive disease-associated alleles in humans. When a particular sequence occurs in chimpanzee, gorilla, and macaque, this indicates that it is an ancestral allele. Conceivably, specific changes in the human environment in the past several million years have made such ancestral sequences deleterious, such that an altered sequence in humans is adaptive. Other compensatory mutations may also be important in interpreting the findings. Similar results were reported by the Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) including 229 amino acid substitutions for which the amino acid identified as mutant in human corresponds to the wildtype allele in macaque, chimpanzee, and/or a reconstructed ancestral genome.

## FUNCTIONAL CLASSIFICATION OF DISEASE GENES

We conclude by considering the principles of human disease. The variety of human diseases is extraordinarily broad, yet the field of bioinformatics may provide insight into a logic of disease. One such attempt was by Jimenez-Sanchez *et al.* (2001), who analyzed 923 human genes that are associated with human disease. These genes primarily cause monogenic disorders. They classified each disease gene according to the function of its protein product (**Fig. 21.19a**). Enzymes represent the largest functional category and account for 31% of the total gene products associated with disease. In contrast, only 15% of positionally cloned disease genes encode enzymes. There may therefore be some



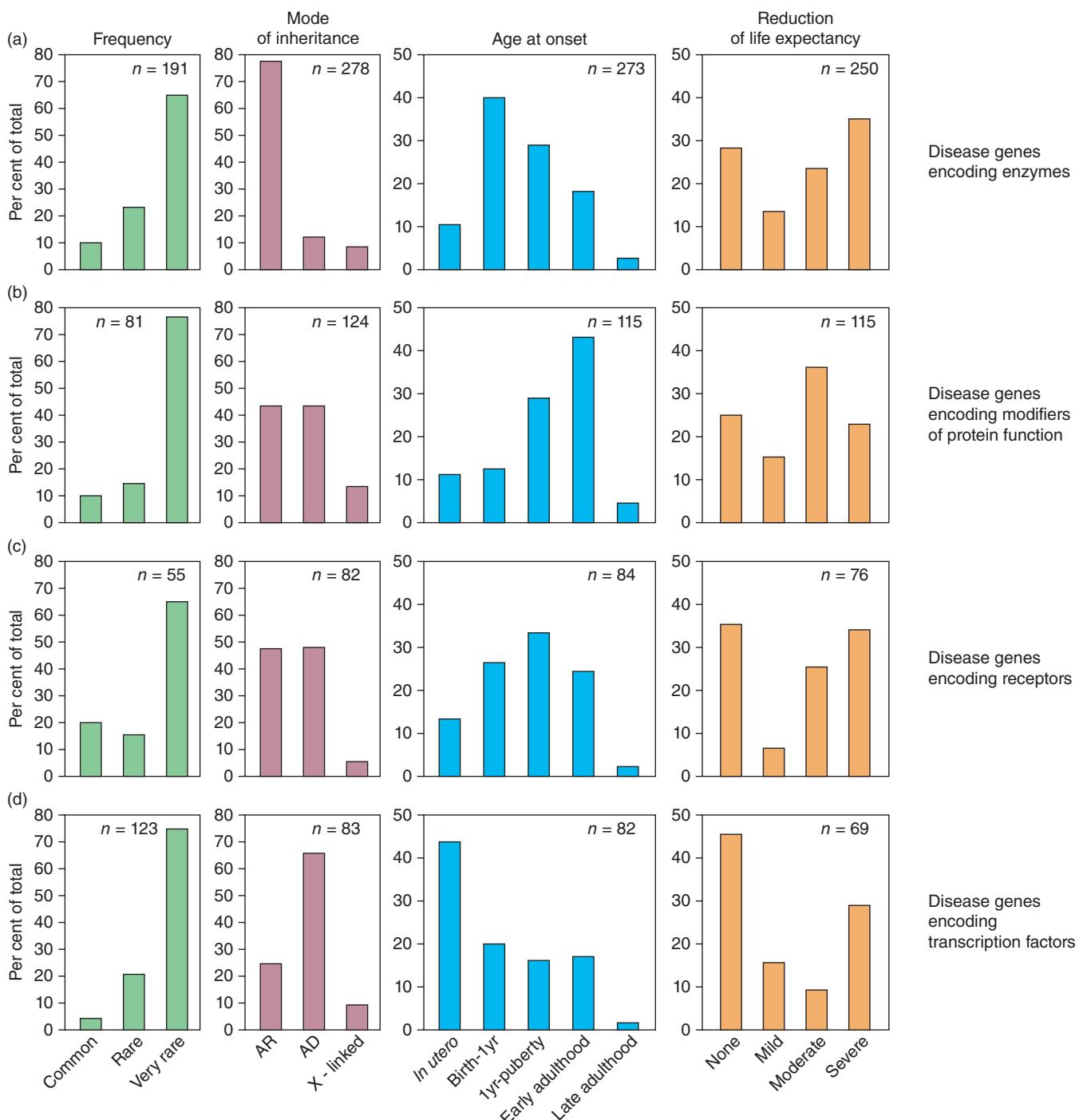
**FIGURE 21.19** The functions of the protein products of disease genes: (a) all genes ( $n = 923$ ); (b–f) disease genes listed according to the typical age of onset of the disease phenotype.

Source: Jimenez-Sanchez *et al.* (2001). Reproduced with permission from Macmillan Publishers.

historical bias toward our knowledge of disease-causing mutations that are based on enzymatic defects.

Jimenez-Sanchez *et al.* further analyzed the correlation between the function of a gene product and the age of disease onset (Fig. 20.19b–f). Genes encoding enzymes and transcription factors are especially likely to be involved in disease in utero, reflecting the importance of transcription factors in early development. Enzymes are particularly involved in disease up to puberty (Fig. 20.19b–d). The developing fetus has access to its mother's metabolic systems and thus may be viable even if it has a gene defect. After birth, such diseases are manifested. Disease genes encoding enzymes are less prevalent in diseases having a later onset in life (Fig. 21.19e).

All of the common diseases in this sample occur with only a very low frequency when analyzed for any of four functional categories of disease: frequency, mode of inheritance,



**FIGURE 21.20** The characteristics of diseases, organized by the function of the protein encoded by the disease gene. AR: autosomal recessive; AD: autosomal dominant; early adulthood: puberty to <50 years old; late adulthood: >50 years old.

Source: Jimenez-Sanchez *et al.* (2001). Reproduced with permission from Macmillan Publishers.

age of onset, and reduction of life expectancy (Fig. 21.20, leftmost column). This very low frequency reflects the population of disease genes that are currently available to study, that is, genes implicated in single-gene disorders. The mode of inheritance tends to be autosomal recessive, particularly for genes encoding enzymes. As also described in Figure 21.19, the age of onset tends to be: in utero for transcription factors; from birth to 1 year for genes encoding enzymes; between 1 year and puberty and into adulthood for receptors; and early adulthood for modifiers of protein function (such as proteins that stabilize, activate, or fold

other proteins). The severity of the disease, reflected in reduction of life expectancy, varies for diseases without a strong pattern based on functional categories.

These studies represent an early attempt to define a logic of disease. Such genomic-scale efforts will be enhanced when we have more information available on the genetic basis of complex disorders. Early whole-exome and whole-genome sequencing studies suggest that some complex disorders display extreme locus heterogeneity, involving mutations in many genes across and within individuals. Functional analyses may be combined using all the tools of bioinformatics and genomics to help elucidate the relationship between genotype and disease phenotype.

## PERSPECTIVE

There are several kinds of bioinformatics approaches to human disease:

- Human disease is a consequence of variation in DNA sequence. These variations are catalogued in databases of molecular sequences (such as GenBank, SRA, and ENA).
- Human disease databases have a major role in organizing information about disease genes. There are centralized databases, most notably OMIM, ClinVar, and HGMD, as well as locus-specific mutation databases.
- Functional genomics screens provide insight into the mechanisms of disease genes and disease processes.

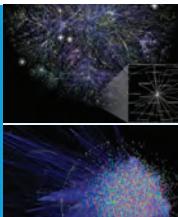
## PITFALLS

A fundamental gap in our understanding is how a genotype such as a mutated gene is related to a disease phenotype. We can approach disease from either end of the spectrum. Starting with a disease phenotype we can ask what genes, when mutated, might cause this disease? Starting with a gene, we can ask what disease occurs when this gene is mutated? However, connecting these two ends of the continuum has been nearly impossible. For the majority of diseases, the discovery of a disease gene has not yet led to the subsequent discovery of new treatment options or cures, or to an understanding of pathophysiology. A hope is that bioinformatics and functional genomics approaches may lead to an understanding of biochemical pathways that account for the molecular basis of pathophysiology. This could be accomplished by learning the function of disease-causing genes in model organisms or via high-throughput technologies such as RNA-seq that reveals the transcriptional response of susceptible cell types to the presence of a mutated gene.

The state of disease databases presents challenges for the field. As tens of millions of single-nucleotide variants are quickly being discovered through next-generation sequencing, how can we know which are neutral or pathogenic? Many diseases display allelic heterogeneity; many genes known to harbor pathogenic mutations also harbor benign variants; and there are many false positives (e.g., variants initially defined as pathogenic that are subsequently found to be neutral) and false negatives.

## ADVICE FOR STUDENTS

Select a disease and explore what is known about its mode of inheritance, clinical phenotype (e.g., by reviewing its OMIM entry), and implicated genes. Examine its variants in NCBI resources such as PheGenI and ClinVar, in locus-specific databases (if any), and HGMD. For genes that are associated with the disease, examine both paralogs and orthologs (e.g., beginning with HomoloGene). Do the phenotypes of a mouse, zebrafish, worm, fly, or yeast knockout illuminate the relevance of the gene to human disease? For a gene having allelic variants, systematically collect those variants (from OMIM, HGMD, or locus-specific databases). Do any of them have relatively high minor allele frequencies and, if so, does that suggest they are likely to be false positives?



# Discussion Questions

**[21-1]** Many neurological diseases such as Rett syndrome, vanishing white matter syndrome, and Huntington's disease have devastating consequences on brain function. For some of these diseases, the responsible genes have homologs in single-celled organisms such as fungi. Why do you think this is so?

**[21-2]** How have microarrays and next-generation sequencing been used to study human disease? Give some specific examples of progress that has been made.

## **COMPUTER LAB/PROBLEMS**

**[21-1]** How many inherited diseases have a known sequence associated with them? Visit OMIM and search for the number of genes having allelic variants. Use EDirect to search for the answer.

**[21-2]** Mutations in *MECP2* cause Rett syndrome.

(1) Explore this gene and this disease in OMIM. What is the phenotype of the disease? What chromosome is *MECP2* localized to? How many allelic variants are reported? Are mouse models available? (2) Explore *MECP2* at a locus-specific mutation database, RettBase. Compare the types of information you obtain from this resource versus OMIM. (3) Explore *MECP2* at dbSNP. Are there any SNPs that correspond to disease-associated substitutions? Do any SNPs alter the amino acid sequence? (4) Explore *MECP2* at the UCSC Genome Browser. Again, compare the types of information you obtain from this resource versus OMIM. (5) Use EDirect to produce a table of nonsynonymous variants identified in *MECP2*.

**[21-3]** Use BioMart at Ensembl to analyze any single cancer gene (such as *GNAQ*). In parallel, you can view the Variation Table for that gene to view its variants that are in dbSNP and COSMIC databases. For each variant it is possible to view the SIFT and PolyPhen scores (in the Variation Table they are color-coded red or green for deleterious or benign). We expect dbSNP entries to tend to be predicted to have neutral (benign) substitutions, while SIFT and PolyPhen predictions for COSMIC entries should tend to be deleterious. Tabulate the entries to determine if this is the case. Optionally, use the R package `biomaRt` in place of the BioMart web service.

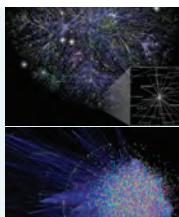
**21-4** What is the record of publications on the following diseases and conditions, organized by decade? Use EDirect and select mania, pertussis, diphtheria, schizophrenia, AIDS, SARS, and Ebola. You may select other diseases and time periods. You can use the following code (also available at the NCBI EDirect website).

```

$ for disease in mania pertussis diphtheria
schizophrenia AIDS SARS ebola
do
citations=`esearch -db pubmed -query "$disease
[TITLE]"`
current=`for (( yr = 2010; yr = 1860; yr -= 10 )) do
echo "$citations" |
efilter -mindate "$yr" -maxdate "$((yr+9))"
-datetype PDAT |
xtract -pattern ENTREZ_DIRECT -lbl "$((yr))s"
-element Count
done` 
heading=`echo -e "${disease:0:4}" | tr [a-z]
[A-Z]` 
current=`echo -e "YEARS\t$heading\n--\t--\
n$current" ` 
if [ -n "$result" ]
then
result=`join -t $'\t' <(echo "$result") <(echo
"$current")` 
else
result=$current
fi
done
echo "$result"
$ echo "$result"

```

YEARS	MANI	PERT	DIPH	SCHI	AIDS	SARS	EBOL
2010s	558	1154	405	12637	5800	375	587
2000s	1000	1966	890	17275	14117	2778	509
1990s	684	2660	1149	8113	23554	22	230
1980s	520	1746	780	4148	12351	5	46
1970s	194	698	749	3019	943	16	25
1960s	76	635	1152	2283	602	1	0
1950s	26	491	1224	1493	560	1	0
1940s	6	172	452	140	184	0	0
1930s	1	26	157	23	16	0	0
1920s	0	5	128	3	27	0	0
1910s	2	7	83	0	5	1	0
1900s	3	3	93	0	0	0	0
1890s	0	0	142	0	4	0	0
1880s	3	0	29	0	2	0	0
1870s	4	2	29	0	2	0	0
1860s	1	1	1	0	0	0	0



## Self-Test Quiz

**[21-1]** In humans, disorders that are inherited by simple Mendelian inheritance account for about what percentage of all human disease?

- (a) 1%;
- (b) 10%;
- (c) 50%; or
- (d) it is impossible to accurately measure the percentage.

**[21-2]** To a significant extent, susceptibility to a variety of infectious diseases is determined by variants of an individual's genes:

- (a) true; or
- (b) false.

**[21-3]** Which of the following best describes single-gene disorders? Each single-gene disorder:

- (a) Is caused by a mutation in a single gene; they represent a basic category of disease that is in contrast to complex disorders.
- (b) Is primarily caused by a mutation in a single gene, but the disease process always involve the contribution of many genes. They therefore represent a category of disease along a continuum with complex disorders.
- (c) Is primarily caused by a mutation in a single gene in which the mutation almost always introduces a synonymous substitution.
- (d) Is primarily caused by a mutation in a single gene in which the mutation almost always introduces a non-synonymous substitution.

**[21-4]** The United States population is ~320 million. How many people in the US have a rare disease?

- (a) 200,000;
- (b) 2 million;
- (c) 25 million; or
- (d) 100 million.

**[21-5]** Single-gene disorders tend to be:

- (a) rare in the general population, with an early onset in life;
- (b) common in the general population, with an early onset in life;

- (c) rare in the general population, with a late onset in life; or
- (d) common in the general population, with a late onset in life.

**[21-6]** Online Mendelian Inheritance in Man (OMIM) includes entries that focus on:

- (a) particular diseases;
- (b) particular genes;
- (c) either genes or diseases; or
- (d) complex chromosomal disorders.

**[21-7]** There are several thousand locus-specific databases. What information do they offer that is not available in central databases such as OMIM and GeneCards?

- (a) comprehensive descriptions of the gene implicated in a disease;
- (b) comprehensive lists of mutations associated with disease;
- (c) links to foundations and other organizations; or
- (d) links to chromosome maps displaying the disease-causing gene.

**[21-8]** You are interested in seeing a summary of the genome-wide association study (GWAS) results for a set of 10 genes. Which of the following resources is most useful?

- (a) HGMD;
- (b) NCBI GWAS;
- (c) OMIM; or
- (d) PheGenI.

**[21-9]** Human disease genes have orthologs in a variety of organisms including worms, insects, and fungi. For a number of human proteins that are implicated in disease, multiple sequence alignments with orthologous proteins have been made. These show that amino acid positions associated with disease-causing mutations in human proteins tend to be residues that are:

- (a) strongly conserved in other organisms;
- (b) sometimes conserved in other organisms;
- (c) poorly conserved in other organisms; or
- (d) only sometimes aligned with orthologous sequences.

## SUGGESTED READING

W. Gregory Feero, Alan Guttmacher, and Francis Collins provide an excellent primer on genomic medicine (Feero *et al.*, 2010). An essential resource for the study of human disease is *The Metabolic and Molecular Basis of Inherited Disease* (Scriver *et al.*, 2001). This four-volume tome has hundreds of chapters including introductions to disease from a variety of perspectives (e.g., Mendelian disorders, complex disorders, a logic of disease, mutation mechanisms, and animal models). A recommended introduction to disease is an essay by Barton Childs and David Valle (2000).

For an overview of cancer genomics see Vogelstein *et al.* (2013) as well as reviews by Chin *et al.* (2011) and Watson *et al.* (2013). Thorisson *et al.* (2009) introduce disease databases with an emphasis on genotype-phenotype correlations. Garry Cutting (2014) surveys the challenge of annotating genomic DNA variants and discusses the problem of the “interpretive gap.” David Altshuler, Mark Daly, and Eric Lander (2008) review genetic mapping in human disease, including linkage and association approaches. For an overview of GWAS see an article by Thomas Pearson and Teri Manolio (2008). Manolio *et al.* (2009) offer the important review “Finding the missing heritability of complex diseases” (see Fig. 21.6). See also Lupski *et al.* (2011) for a related article on the role of allelic variants of differing frequencies in human disease. For an overview of the fascinating topic of mitochondrial disease, see DiMauro *et al.* (2013).

## REFERENCES

- 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D. *et al.* 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073. PMID: 20981092.
- 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A. *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65. PMID: 23128226.
- Abecasis, G.R., Cherny, S.S., Cookson, W.O., Cardon, L.R. 2002. Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.
- Ahringer, J. 1997. Turn to the worm! *Current Opinion in Genetics and Development* **7**, 410–415.
- Altshuler, D., Daly, M.J., Lander, E.S. 2008. Genetic mapping in human disease. *Science* **322**(5903), 881–888. PMID: 18988837.
- Amberger, J., Bocchini, C., Hamosh, A. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human Mutation* **32**(5), 564–567. PMID: 21472891.
- Amir, R. E., Van den Veyver, I.B., Wan, M. *et al.* 1999. Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nature Genetics* **23**, 185–188. PMID: 10508514.
- Amir, R.E., Zoghbi, H.Y. 2000. Rett syndrome: Methyl-CpG-binding protein 2 mutations and phenotype–genotype correlations. *American Journal of Medical Genetics* **97**, 147–152.
- Antonarakis, S.E. 1998. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Human Mutation* **11**, 1–3.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.
- Bailey, A., Le Couteur, A., Gottesman, I. *et al.* 1995. Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychological Medicine* **25**, 63–77. PMID: 7792363.
- Bamshad, M.J., Ng, S.B., Bigham, A.W. *et al.* 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12**(11), 745–755. PMID: 21946919.
- Bauman, M.L., Kemper, T. L., Arin, D. M. 1995. Microscopic observations of the brain in Rett syndrome. *Neuropediatrics* **26**, 105–108.
- Beaudet, A.L., Scriver, C.R., Sly, W. S., Valle, D. 2001. Genetics, biochemistry, and molecular bases of variant human phenotypes. In *The Metabolic & Molecular Bases of Inherited Disease* (eds Scriver *et al.*), McGraw-Hill, New York, vol. 1, pp. 3–45.

- Bell, C.J., Dinwiddie, D.L., Miller, N.A. *et al.* 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science Translational Medicine* **3**(65), 65ra4. PMID: 21228398.
- Béroud, C., Hamroun, D., Collod-Béroud, G. *et al.* 2005. UMD (Universal Mutation Database): 2005 update. *Human Mutation* **26**(3), 184–191. PMID: 16086365.
- Bowcock, A.M. 2007. Guilt by association. *Nature* **447**, 645–646.
- Brunham, L.R., Hayden, M.R. 2013. Hunting human disease genes: lessons from the past, challenges for the future. *Human Genetics* **132**(6), 603–617. PMID: 23504071.
- Burnet, M. 1959. Auto-immune disease. II. Pathology of the immune response. *British Medical Journal* **2**(5154), 720–725. PMID: 13806211.
- Chapman, S.J., Hill, A.V. 2012. Human genetic susceptibility to infectious disease. *Nature Reviews Genetics* **13**(3), 175–188. PMID: 22310894.
- Chen, K.F., Crowther, D.C. 2012. Functional genomics in *Drosophila* models of human disease. *Briefings in Functional Genomics* **11**(5), 405–415. PMID: 22914042.
- Childs, B., Valle, D. 2000. *Genetics, Biology and Disease*. Annual Reviews, Palo Alto, CA, pp. 1–19.
- Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87.
- Chin, L., Hahn, W.C., Getz, G., Meyerson, M. 2011. Making sense of cancer genomic data. *Genes and Development* **25**(6), 534–555. PMID: 21406553.
- Cibulskis, K., Lawrence, M.S., Carter, S.L. *et al.* 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**(3), 213–219. PMID: 23396013.
- Claustres, M., Horaitis, O., Vanevski, M., Cotton, R. G. 2002. Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases. *Genome Research* **12**, 680–688.
- Cline, M.S., Craft, B., Swatloski, T. *et al.* 2013. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Science Reports* **3**, 2652. PMID: 24084870.
- Costa, T., Scriver, C.R., Childs, B. 1985. The effect of Mendelian disease on human health: a measurement. *American Journal of Medical Genetics* **21**(2), 231–242. PMID: 4014310.
- Cotton, R.G., Auerbach, A.D., Beckmann, J.S. *et al.* 2008. Recommendations for locus-specific databases and their curation. *Human Mutation* **29**, 2–5. PMID: 18157828.
- Cukier, H.N., Perez, A.M., Collins, A.L. *et al.* 2008. Genetic modifiers of MeCP2 function in *Drosophila*. *PLoS Genetics* **4**(9), e1000179. PMID: 18773074.
- Cutting, G.R. 2014. Annotating DNA variants is the next major goal for human genetics. *American Journal of Human Genetics* **94**(1), 5–10. PMID: 24387988.
- den Dunnen, J.T., Antonarakis, S. E. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation* **15**, 7–12.
- Denis, P.-S. 1842. *Études Chimiques, Physiologiques, et Médicales, Faites de 1835 à 1840, sur les Matières Albumineuses*. Imprimerie C.-F. Denis, Commercy.
- DiMauro, S., Schon, E. A. 2001. Mitochondrial DNA mutations in human disease. *American Journal of Medical Genetics* **106**, 18–26. PMID: 11579421.
- DiMauro, S., Schon, E.A., Carelli, V., Hirano, M. 2013. The clinical maze of mitochondrial neurology. *Nature Reviews Neurology* **9**(8), 429–444. PMID: 23835535.
- Dipple, K.M., McCabe, E. R. 2000. Modifier genes convert “simple” Mendelian disorders to complex traits. *Molecular Genetics and Metabolism* **71**, 43–50.
- Dipple, K. M., Phelan, J. K., McCabe, E. R. 2001. Consequences of complexity within biological networks: Robustness and health, or vulnerability and disease. *Molecular Genetics and Metabolism* **74**, 45–50.
- Dorschner, M.O., Amendola, L.M., Turner, E.H. *et al.* 2013. Actionable, pathogenic incidental findings in 1,000 participants’ exomes. *American Journal of Human Genetics* **93**(4), 631–640. PMID: 24055113.
- Dumanski, J.P., Piotrowski, A. 2012. Structural genetic variation in the context of somatic mosaicism. *Methods in Molecular Biology* **838**, 249–272. PMID: 22228016.
- Eichinger, L., Pachebat, J.A., Glöckner, G. *et al.* 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43–57. PMID: 15875012.

- Erickson, R.P. 2010. Somatic gene mutation and human disease other than cancer: an update. *Mutation Research* **705**(2), 96–106. PMID: 20399892.
- Feero, W.G., Guttmacher, A.E., Collins, F.S. 2010. Genomic medicine: an updated primer. *New England Journal of Medicine* **362**(21), 2001–2011. PMID: 20505179.
- Firth, H.V., Richards, S.M., Bevan, A.P. et al. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics* **84**(4), 524–533. PMID: 19344873.
- Fokkema, I.F., Taschner, P.E., Schaafsma, G.C. et al. 2011. LOVD v.2.0: the next generation in gene variant databases. *Human Mutation* **32**(5), 557–563. PMID: 21520333.
- Fombonne, E. 1999. The epidemiology of autism: A review. *Psychological Medicine* **29**, 769–786.
- Forbes, S.A., Bindal, N., Bamford, S. et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**(Database issue), D945–950. PMID: 20952405.
- Garrod, A. E. 1902. The incidence of alkaptonuria: A study in chemical individuality. *Lancet* **ii**, 1616–1620.
- Garrod, A.E. 1909. *Inborn errors of metabolism: The Croonian Lectures delivered before the Royal College of Physicians of London, in June, 1908*. Frowde, Hodder and Stoughton, London.
- Garrod, A.E. 1931. *Inborn factors in disease: An essay*. Clarendon Press, Oxford.
- George, R.A., Smith, T.D., Callaghan, S. et al. 2008. General mutation databases: analysis and review. *Journal of Medical Genetics* **45**(2), 65–70.
- Giardine, B., Riemer, C., Hefferon, T. et al. 2007. PhenCode: connecting ENCODE data with mutations and phenotype. *Human Mutation* **28**, 554–562. PMID: 17326095.
- Giardine, B., Borg, J., Viennas, E. et al. 2014. Updates of the HbVar database of human hemoglobin variants and thalassemia mutations. *Nucleic Acids Research* **42**(Database issue), D1063–1069. PMID: 24137000.
- Gillberg, C., Wing, L. 1999. Autism: not an extremely rare disorder. *Acta Psychiatrica Scandinavica* **99**(6), 399–406. PMID: 10408260.
- Goldman, M., Craft, B., Swatloski, T. et al. 2013. The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Research* **41**(Database issue), D949–954. PMID: 23109555.
- Green, R.C., Berg, J.S., Grody, W.W. et al. 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in Medicine* **15**(7), 565–574. PMID: 23788249.
- Greenman, C., Stephens, P., Smith, R. et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158. PMID: 17344846.
- Guo, Y., Li, J., Li, C.I., Shyr, Y., Samuels, D.C. 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**(9), 1210–1211. PMID: 23471301.
- Gusella, J.F. 1989. Location cloning strategy for characterizing genetic defects in Huntington's disease and Alzheimer's disease. *FASEB Journal* **3**, 2036–2041.
- Guy, J., Gan, J., Selfridge, J., Cobb, S., Bird, A. 2007. Reversal of neurological defects in a mouse model of Rett syndrome. *Science* **315**(5815), 1143–1147. PMID: 17289941.
- Hagberg, B., Aicardi, J., Dias, K., Ramos, O. 1983. A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. *Annals of Neurology* **14**, 471–479.
- Hammer, S., Dorrani, N., Dragich, J., Kudo, S., Schanen, C. 2002. The phenotypic consequences of MECP2 mutations extend beyond Rett syndrome. *Mental Retardation and Developmental Disabilities Research Reviews* **8**, 94–98.
- Hanahan, D., Weinberg, R.A. 2011. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674.
- Happle, R. 1987. Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. *Journal of the American Academy of Dermatology* **16**(4), 899–906. PMID: 3033033.
- Hindorff, L.A., Sethupathy, P., Junkins, H.A. et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, USA* **106**(23), 9362–9367. PMID: 19474294.

- Hirschhorn, J. N., Daly, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.
- Holt, I. J., Harding, A. E., Morgan-Hughes, J. A. 1988. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* **331**, 717–719.
- Human Genome Structural Variation Working Group *et al.* 2007. Completing the map of human genetic variation. *Nature* **447**, 161–165.
- Jimenez-Sanchez, G., Childs, B., Valle, D. 2001. Human disease genes. *Nature* **409**, 853–855.
- Kanner, L. 1943. Autistic disturbances of affective contact. *The Nervous Child* **2**, 217–250.
- Katz, D.M., Berger-Sweeney, J.E., Eubanks, J.H. *et al.* 2012. Preclinical research in Rett syndrome: setting the foundation for translational success. *Disease Models and Mechanisms* **5**(6), 733–745. PMID: 23115203.
- Kawaiji, H., Severin, J., Lizio, M. *et al.* 2011. Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Research* **39**(Database issue), D856–860. PMID: 21075797.
- Kilpinen, H., Barrett, J.C. 2013. How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics* **29**(1), 23–30. PMID: 23103023.
- Knoppers, B. M., Laberge, C. M. 2000. Ethical guideposts for allelic variation databases. *Human Mutation* **15**, 30–35.
- Knudson, A.G. Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences, USA* **68**, 820–823.
- Knutsen, T., Gobu, V., Knaus, R. *et al.* 2005. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer* **44**, 52–64.
- Koboldt, D.C., Zhang, Q., Larson, D.E. *et al.* 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**(3), 568–576. PMID: 22300766.
- Laird, N. M., Lange, C. 2006. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**, 385–394.
- Landrum, M.J., Lee, J.M., Riley, G.R. *et al.* 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**(Database issue), D980–985. PMID: 24234437.
- Lindhurst, M.J., Sapp, J.C., Teer, J.K. *et al.* 2011. A mosaic activating mutation in *AKT1* associated with the Proteus syndrome. *New England Journal of Medicine* **365**(7), 611–619. PMID: 21793738.
- Lowrance, W.W., Collins, F.S. 2007. Identifiability in genomic research. *Science* **317**, 600–602.
- Lupski, J. R. 1998. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics* **14**, 417–422.
- Lupski, J.R. 2013. Genetics. Genome mosaicism: one human, multiple genomes. *Science* **341**(6144), 358–359. PMID: 23888031.
- Lupski, J. R., Stankiewicz, P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics* **1**, e49.
- Lupski, J.R., Belmont, J.W., Boerwinkle, E., Gibbs, R.A. 2011. Clan genomics and the complex architecture of human disease. *Cell* **147**(1), 32–43. PMID: 21962505.
- MacArthur, D.G., Tyler-Smith, C. 2010. Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics* **19**(R2), R125–130. PMID: 20805107.
- MacArthur, D.G., Balasubramanian, S., Frankish, A. *et al.* 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**(6070), 823–828. PMID: 22344438.
- Mailman, M. D., Feolo, M., Jin, Y. *et al.* 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* **39**, 1181–1186.
- Manolio, T.A., Collins, F.S., Cox, N.J. *et al.* 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753. PMID: 19812666.
- Mathers, C.D., Loncar, D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine* **3**, e442.

- McCarthy, M.I., Abecasis, G.R., Cardon, L.R. *et al.* 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**(5), 356–369. PMID: 18398418.
- McGraw, C.M., Samaco, R.C., Zoghbi, H.Y. 2011. Adult neural function requires MeCP2. *Science* **333**(6039), 186. PMID: 21636743.
- McKusick, V.A. 2007. Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics* **80**, 588–604.
- Menzel, S., Garner, C., Gut, I. *et al.* 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nature Genetics* **39**(10), 1197–1199. PMID: 17767159.
- Miller, M. P., Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics* **10**, 2319–2328.
- Miller, M.P., Parker, J.D., Rissing, S.W., Kumar, S. 2003. Quantifying the intragenic distribution of human disease mutations. *Annals of Human Genetics* **67**, 567–579.
- Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K. *et al.* 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562. PMID: 12466850.
- Murray, C. J. L., Lopez, A. D. (eds) 1996. *The Global Burden of Disease*. Harvard University Press, Cambridge.
- Murray, C.J., Vos, T., Lozano, R. *et al.* 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**(9859), 2197–2223. PMID: 23245608.
- Nan, X., Campoy, F. J., Bird, A. 1997. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88**, 471–481.
- Nass, S., Nass, M. M. K. 1963. Intramitochondrial fibers with DNA characteristics. *Journal of Cell Biology* **19**, 613–629.
- NCI-NHGRI Working Group on Replication in Association Studies *et al.* 2007. Replicating genotype–phenotype associations. *Nature* **447**, 655–660.
- Olsson, I., Steffenburg, S., Gillberg, C. 1988. Epilepsy in autism and autisticlike conditions. A population-based study. *Archives of Neurology* **45**, 666–668.
- O’Rawe, J., Jiang, T., Sun, G. *et al.* 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* **5**(3), 28. PMID: 23537139.
- O’Roak, B.J., Vives, L., Girirajan, S. *et al.* 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**(7397), 246–250. PMID: 22495309.
- Ott, J. 2001. Major strengths and weaknesses of the lod score method. *Advances in Genetics* **42**, 125–132.
- Ott, J., Kamatani, Y., Lathrop, M. 2011. Family-based designs for genome-wide association studies. *Nature Reviews Genetics* **12**(7), 465–474. PMID: 21629274.
- Patrinos, G.P., Brookes, A.J. 2005. DNA, diseases and databases: disastrously deficient. *Trends in Genetics* **21**, 333–338.
- Pauling, L., Itano, H. A., Singer, S. J., Wells, I. C. 1949. Sickle cell anemia, a molecular disease. *Science* **110**, 543–548.
- Pearson, T.A., Manolio, T.A. 2008. How to interpret a genome-wide association study. *JAMA* **299**(11), 1335–1344. PMID: 18349094.
- Peterson, T.A., Doughty, E., Kann, M.G. 2013. Towards precision medicine: advances in computational approaches for the analysis of human variants. *Journal of Molecular Biology* **425**(21), 4047–4063. PMID: 23962656.
- Pham, J., Shaw, C., Pursley, A. *et al.* 2014. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10,362 consecutive cases. *European Journal of Human Genetics* **22**(8), 969–978. PMID: 24398791.
- Poduri, A., Evrony, G.D., Cai, X., Walsh, C.A. 2013. Somatic mutation, genomic variation, and neurological disease. *Science* **341**(6141), 1237758. PMID: 23828942.
- Probst, F.J., Justice, M.J. 2010. Mouse mutagenesis with the chemical supermutagen ENU. *Methods in Enzymology* **477**, 297–312. PMID: 20699147.

- Purcell, S., Neale, B., Todd-Brown, K. *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**(3), 559–575. PMID: 17701901.
- Ramocki, M.B., Tavyev, Y.J., Peters, S.U. 2010. The *MECP2* duplication syndrome. *American Journal of Medical Genetics A* **152A**(5), 1079–1088. PMID: 20425814.
- Randall, J.C., Winkler, T.W., Kutalik, Z. *et al.* 2013. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genetics* **9**(6), e1003500. PMID: 23754948.
- Rapin, I. 1997. Autism. *New England Journal of Medicine* **337**, 97–104.
- Rapin, I., Katzman, R. 1998. Neurobiology of autism. *Annals of Neurology* **43**(1), 7–14. PMID: 9450763.
- Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.
- Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., Bier, E. 2001. A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. *Genome Research* **11**, 1114–1125.
- Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234.
- Rietveld, C.A., Medland, S.E., Derringer, J. *et al.* 2013. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**(6139), 1467–1471. PMID: 23722424.
- Robinson, P.N., Köhler, S., Bauer, S. *et al.* 2008. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics* **83**(5), 610–615. PMID: 18950739.
- Rossi, P. G., Parmeggiani, A., Bach, V., Santucci, M., Visconti, P. 1995. EEG features and epilepsy in patients with autism. *Brain Development* **17**, 169–174.
- Rubin, G.M., Yandell, M.D., Wortman, J.R. *et al.* 2000. Comparative genomics of the eukaryotes. *Science* **287**, 2204–2215.
- Ruiz-Pesini, E., Lott, M.T., Procaccio, V. *et al.* 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Research* **35**, D823–828.
- Samuels, M.E., Rouleau, G.A. 2011. The case for locus-specific databases. *Nature Reviews Genetics* **12**(6), 378–379. PMID: 21540879.
- Saunders, C.T., Wong, W.S., Swamy, S. *et al.* 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**(14), 1811–1817. PMID: 22581179.
- Scally, A., Dutheil, J.Y., Hillier, L.W. *et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**(7388), 169–175. PMID: 22398555.
- Schon, E.A., DiMauro, S., Hirano, M. 2012. Human mitochondrial DNA: roles of inherited and somatic mutations. *Nature Reviews Genetics* **13**(12), 878–890. PMID: 23154810.
- Scriver, C.R., Childs, B. 1989. *Garrod's Inborn Factors in Disease*. New York, Oxford University Press.
- Scriver, C.R., Nowacki, P. M., Lehvaslaiho, H. 1999. Guidelines and recommendations for content, structure, and deployment of mutation databases. *Human Mutation* **13**, 344–350.
- Scriver, C.R., Nowacki, P. M., Lehvaslaiho, H. 2000. Guidelines and recommendations for content, structure, and deployment of mutation databases: II. Journey in progress. *Human Mutation* **15**, 13–15.
- Scriver, C.R., Beaudet, A., Sly, W., Valle, D. (eds). 2001. *The Metabolic and Molecular Basis of Inherited Disease*. McGraw-Hill, New York.
- Shepherd, R., Forbes, S.A., Beare, D. *et al.* 2011. Data mining using the Catalogue of Somatic Mutations in Cancer BioMart. *Database (Oxford)* **2011**, bar018. PMID: 21609966.
- Shirley, M.D., Tang, H., Gallione, C.J. *et al.* 2013. Sturge-Weber syndrome and port-wine stains caused by somatic mutation in *GNAQ*. *New England Journal of Medicine* **368**(21), 1971–1979. PMID: 23656586.
- Smalley, S.L., Asarnow, R. F., Spence, M. A. 1988. Autism and genetics. A decade of research. *Archives of General Psychiatry* **45**, 953–961.
- Stankiewicz, P., Lupski, J. R. 2002. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* **18**, 74–82.

- Stelzer, G., Dalah, I., Stein, T.I. *et al.* 2011. In-silico human genomics with GeneCards. *Human Genetics* **5**(6), 709–717. PMID: 22155609.
- Stenson, P.D., Ball, E.V., Mort, M. *et al.* 2012. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics Chapter 1, Unit1.13.* PMID: 22948725.
- Stenson, P.D., Mort, M., Ball, E.V. *et al.* 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics* **133**(1), 1–9. PMID: 24077912.
- Stottmann, R.W., Beier, D.R. 2010. Using ENU mutagenesis for phenotype-driven analysis of the mouse. *Methods in Enzymology* **477**, 329–348. PMID: 20699149.
- Stratton, M.R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science* **331**(6024), 1553–1558. PMID: 21436442.
- Subramanian, S., Kumar, S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* **7**, 306.
- Szatmari, P., Jones, M. B., Zwaigenbaum, L., MacLean, J. E. 1998. Genetics of autism: Overview and new directions. *Journal of Autism and Development Disorders* **28**, 351–368.
- Szumilas, M. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* **19**(3), 227–229. PMID: 20842279.
- Tang, S., Wang, J., Zhang, V.W. *et al.* 2013. Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects. *Human Mutation* **34**(6), 882–893. PMID: 23463613.
- Thiagalingam, S., Lengauer, C., Leach, F.S. *et al.* 1996. Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nature Genetics* **13**(3), 343–346. PMID: 8673134.
- Thomas, C.L. (ed.) 1997. *Taber's Cyclopedic Medical Dictionary*. F. A. Davis Company, Philadelphia.
- Thorisson, G.A., Muilu, J., Brookes, A.J. 2009. Genotype–phenotype databases: challenges and solutions for the post-genomic era. *Nature Reviews Genetics* **10**(1), 9–18. PMID: 19065136.
- Todd, J.A. 2001. Multifactorial diseases: Ancient gene polymorphism at quantitative trait loci and a legacy of survival during our evolution. In *The Metabolic & Molecular Bases of Inherited Disease* (eds C.Scrivner *et al.*), McGraw-Hill, New York, vol. 1, pp. 193–201.
- Tryka, K.A., Hao, L., Sturcke, A. *et al.* 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research* **42**(Database issue), D975–979. PMID: 24297256.
- Turner, M., Barnby, G., Bailey, A. 2000. Genetic clues to the biological basis of autism. *Molecular Medicine Today* **6**, 238–244.
- van Echten-Arends, J., Mastenbroek, S., Sikkema-Raddatz, B. *et al.* 2011. Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Human Reproduction Update* **17**(5), 620–627. PMID: 21531753.
- Van Raamsdonk, C.D., Bezrookove, V., Green, G. *et al.* 2009. Frequent somatic mutations of *GNAQ* in uveal melanoma and blue naevi. *Nature* **457**(7229), 599–602. PMID: 19078957.
- Varmus, H. 2006. The new era in cancer research. *Science* **312**, 1162–1165.
- Vasta, V., Ng, S.B., Turner, E.H., Shendure, J., Hahn, S.H. 2009. Next generation sequence analysis for mitochondrial disorders. *Genome Medicine* **1**(10), 100. PMID: 19852779.
- Vogelstein, B., Papadopoulos, N., Velculescu, V.E. *et al.* 2013. Cancer genome landscapes. *Science* **339**(6127), 1546–1558. PMID: 23539594
- Volkmar, F. 1998. Recently diagnosed with autism, autism or not. *Journal of Autism and Development Disorders* **28**, 269–270.
- Volkmar, F. R., Nelson, D. S. 1990. Seizure disorders in autism. *Journal of the American Academy of Child and Adolescent Psychiatry* **29**, 127–129.
- Vouillaire, L., Slater, H., Williamson, R., Wilton, L. 2000. Chromosome analysis of blastomeres from human embryos by using comparative genomic hybridization. *Human Genetics* **106**, 210–217.
- Wallace, D.C., Singh, G., Lott, M. T. *et al.* 1988a. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242**, 1427–1430.

- Wallace, D.C., Zheng, X. X., Lott, M. T. *et al.* 1988b. Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease: *Cell* **55**, 601–610.
- Watson, I.R., Takahashi, K., Futreal, P.A., Chin, L. 2013. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics* **14**(10), 703–718. PMID: 24022702.
- Webb, A.J., Thorisson, G.A., Brookes, A.J. 2011. GEN2PHEN Consortium. An informatics project and online “Knowledge Centre” supporting modern genotype-to-phenotype research. *Human Mutation* **32**(5), 543–550. PMID: 21438073.
- Weinstein, L.S., Shenker, A., Gejman, P.V. *et al.* 1991. Activating mutations of the stimulatory G protein in the McCune-Albright syndrome. *New England Journal of Medicine* **325**(24), 1688–1695. PMID: 1944469.
- Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
- Wells, D., Delhanty, J. D. 2000. Comprehensive chromosomal analysis of human preimplantation embryos using whole genome amplification and single cell comparative genomic hybridization. *Molecular Human Reproduction* **6**, 1055–1062.
- Welter, D., Macarthur, J., Morales, J. *et al.* 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–1006. PMID: 24316577.
- Wood, L.D., Parsons, D.W., Jones, S. *et al.* 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113. PMID: 17932254.
- Wood, V., Gwilliam, R., Rajandream, M.A. *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880. PMID: 11859360.
- Xue, Y., Chen, Y., Ayub, Q. *et al.* 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics* **91**(6), 1022–1032. PMID: 23217326.
- Youssoufian, H., Pyeritz, R.E. 2002. Mechanisms and consequences of somatic mosaicism in humans. *Nature Reviews Genetics* **3**(10), 748–758. PMID: 12360233.
- Zeev, B. B., Yaron, Y., Schanen, N. C. *et al.* 2002. Rett syndrome: Clinical manifestations in males with *MECP2* mutations. *Journal of Child Neurology* **17**, 20–24.
- Zuckerlandl, E., Pauling, L. 1962. Molecular disease, evolution, and genic heterogeneity. In *Horizons in Biochemistry* (eds M.Kasha and B.Pullman), Albert Szent-Gyorgyi Dedicatory Volume, Academic Press, New York.



# Glossary

This glossary is combined from six web-based glossaries and each entry is marked accordingly: (1) the National Center for Biotechnology Information (NCBI BLAST); (2) NCBI genome; (3) the Oak Ridge National Laboratory (ORNL); (4) the talking glossary at the National Human Genome Research Institute (NHGRI); (5) the SMART database; and (6) the protein folds glossary from the Structural Classification of Proteins website (SCOP) (these entries are modified). The glossaries are available online.

## A

**Accession number** An accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented but the accession number will remain constant.

**Additive genetic effects** When the combined effects of alleles at different loci are equal to the sum of their individual effects (ORNL).

**Adenine (A)** A nitrogenous base, one member of the base pair AT (adenine–thymine). *See also:* base pair (ORNL).

**AGP** A file that describes how primary sequences can be assembled to make a nonredundant, contiguous sequence. The sequence being assembled may be a contig or a chromosome. This file describes the portion of the component sequence used in the contig, in addition to the location on the contig of the component sequence (NCBI).

**Algorithm** A fixed procedure embodied in a computer program (NCBI BLAST).

**Alignment** (a) The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology (NCBI BLAST). (b) Representation of a prediction of the amino acids in tertiary structures of homologs that overlay in three dimensions. Alignments held by SMART are mostly based on published observations (see domain annotations for details) but are updated and edited manually (SMART).

**All alpha** A class that has the number of secondary structures in the domain or common core described as 3-, 4-, 5-, 6-, or multihelical (SCOP).

**All beta** A class that includes two major fold groups: sandwiches and barrels. The sandwich folds are made of two  $\beta$  sheets which are usually twisted and packed so their strands are aligned. The barrel fold are made of a single  $\beta$  sheet that twists and coils upon itself so that, in most cases, the first strand in the  $\beta$  sheet hydrogen bonds to the last strand. The strand directions in the two opposite sides of a barrel fold are roughly orthogonal. Orthogonal packing of sheets is also seen in a few special cases of sandwich folds (SCOP).

**Allele** (a) Alternative form of a genetic locus; a single allele for each locus is inherited from each parent (e.g., at a locus for eye color the allele might result in blue or brown

- <http://www.ncbi.nlm.nih.gov/books/NBK62051/> (NCBI BLAST glossary by Drs Jan Fassler and Peter Cooper)
- <http://www.ncbi.nlm.nih.gov/projects/genome/glossary.shtml> (NCBI)
- [http://web.ornl.gov/sci/techresources/Human\\_Genome/glossary.shtml](http://web.ornl.gov/sci/techresources/Human_Genome/glossary.shtml) (ORNL)
- <http://www.genome.gov/glossary.cfm> (NHGRI)
- [http://smart.embl-heidelberg.de/help/smart\\_glossary.shtml](http://smart.embl-heidelberg.de/help/smart_glossary.shtml) (SMART)
- <http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html> (SCOP)

eyes; ORNL). (b) One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant form) may be expressed more than another form (the recessive form; NHGRI).

**Allelic series** A collection of distinct mutations that affect a single locus. Often, these different mutations will produce different phenotypes, thus providing a powerful genetic tool for the dissection of gene function.

**Allogeneic** Variation in alleles among members of the same species (ORNL).

**Alternative splicing** Different ways of combining a gene's exons to make variants of the complete protein (ORNL).

**Amino acid** Any of a class of 20 molecules that are combined to form proteins in living things. The sequence of amino acids in a protein and hence protein function is determined by the genetic code (ORNL).

**Amplification** An increase in the number of copies of a specific DNA fragment; can be *in vivo* or *in vitro*. *See also:* cloning (ORNL).

**Animal model** *See:* model organisms (ORNL).

**Annotation** (a) Adding pertinent information such as gene coded for, amino acid sequence, or other commentary to the database entry of raw sequence of DNA bases. *See also:* bioinformatics (ORNL). (b) Adding biological information to genome sequence. This is a very complex task, and the process for performing this is rapidly evolving. Several groups are performing automated computational annotation of several genomes. Features that are added to the genome often include gene models, SNPs, and STSs (NCBI).

**Anticipation** Each generation of offspring has increased severity of a genetic disorder; for example, a grandchild may have earlier onset and more severe symptoms than the parent, who had earlier onset than the grandparent. *See also:* additive genetic effects, complex trait (ORNL).

**Antisense** Nucleic acid that has a sequence exactly opposite to an mRNA molecule made by the body; binds to the mRNA molecule to prevent a protein from being made. *See also:* transcription (ORNL).

**Apoptosis** Programmed cell death, the body's normal method of disposing of damaged, unwanted, or unneeded cells (ORNL).

**Array (of hairpins)** An assemble of  $\alpha$  helices that cannot be described as a bundle or a folded leaf (SCOP).

**Array comparative genome hybridization (aCGH)** A technique involving the competitive hybridization of "test" and "reference" DNA probes to target genomic (or cDNA clones) immobilized on a microarray. Most often used for the detection of copy number variation (CNV), aCGH also has applications in gene annotation and diagnostics (NCBI).

**Arrayed library** Individual primary recombinant clones (hosted in phage, cosmid, YAC, or other vector) that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific gene or genomic region of interest. *See also:* library, genomic library, gene chip technology (ORNL).

**Assembly** Putting sequenced fragments of DNA into their correct chromosomal positions (ORNL).

**Autoradiography** A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis (ORNL).

**Autosomal dominant** A gene on one of the non-sex chromosomes that is always expressed, even if only one copy is present. The chance of passing the gene to offspring is 50% for each pregnancy. *See also:* autosome, dominant, gene (ORNL).

**Autosome** A chromosome not involved in sex determination. The diploid human genome consists of a total of 46 chromosomes: 22 pairs of autosomes and 1 pair of sex chromosomes (the X and Y chromosomes). *See also:* sex chromosome (ORNL).

## B

**Backcross** A cross between an animal that is heterozygous for alleles obtained from two parental strains and a second animal from one of those parental strains. Also used to describe the breeding protocol of an outcross followed by a backcross. *See also:* model organisms (ORNL).

**Bacterial artificial chromosome (BAC)** (a) A vector used to clone DNA fragments (100–300 kb insert size; average 150 kb) in *Escherichia coli* cells. Based on naturally occurring F-factor plasmid found in the bacterium *E. coli*. *See also:* cloning vector (ORNL). (b) Large segments of DNA, 100,000–200,000 bases, from another species cloned into bacteria. Once the foreign DNA has been cloned into the host bacteria, many copies of it can be made (NHGRI).

**BAC end sequence** The ends of BACs are sequenced and the clone association information is retained. In this way, BAC clones that do not have insert sequence can be integrated with other BAC clones, or with WGS assemblies (NCBI).

**Bacteriophage** *See also:* phage (ORNL).

**Barrel** Structures are usually closed by main-chain hydrogen bonds between the first and last strands of the  $\beta$  sheet. In this case it is defined by the two integer numbers: the number of strand in the  $\beta$  sheet  $n$ , and a measure of the extent to which the strands in the sheet are staggered, the shear number  $S$  (SCOP).

**Base** One of the molecules that form DNA and RNA molecules. *See also:* nucleotide, base pair, base sequence (ORNL).

**Base pair (bp)** Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs (ORNL).

**Base sequence** The order of nucleotide bases in a DNA molecule; determines the structure of proteins encoded by that DNA (ORNL).

**Base sequence analysis** A method, sometimes automated, for determining the base sequence (ORNL).

**Behavioral genetics** The study of genes that may influence behavior (ORNL).

**Beta ( $\beta$ ) sheet** Can be antiparallel (i.e., the strand direction in any two adjacent strands are antiparallel), parallel (all strands are parallel to each other), and mixed (there is at least one strand that is parallel to one of its two neighbors and antiparallel to the other) (SCOP).

**Bioinformatics** (a) The merger of biotechnology and information technology with the goal of revealing new insights and principles in biology (NCBI BLAST). (b) The science of managing and analyzing biological data using advanced computing techniques. Especially important in analyzing genomic research data (ORNL).

**Bioremediation** The use of biological organisms such as plants or microbes to aid in removing hazardous substances from an area (ORNL).

**Biotechnology** A set of biological techniques developed through basic research and now applied to research and product development. In particular, biotechnology refers to the use by industry of recombinant DNA, cell fusion, and new bioprocessing techniques (ORNL).

**Birth defect** Any harmful trait, physical or biochemical, present at birth, whether a result of a genetic mutation or some other nongenetic factor. *See also:* congenital, gene, mutation, syndrome (ORNL).

**Bit score** (a) The value  $S'$  is derived from the raw alignment score  $S$  in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches (NCBI BLAST). (b) Alignment scores are reported by HMMer and BLAST as bit scores. The likelihood that the query sequence is a *bona fide* homolog of the database sequence is compared to the likelihood that the sequence was instead generated by a “random” model. Taking the logarithm (to base 2) of this likelihood ratio gives the bits score (SMART).

**BLAST** (a) Basic Local Alignment Search Tool. A sequence comparison algorithm optimized for speed used to search sequence databases for optimal local alignments to a query. The initial search is performed for a word of length  $W$  that scores at least  $T$  when compared to the query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of  $S$ . The  $T$  parameter dictates the speed and sensitivity of the search. For additional details, see one of the BLAST tutorials (Query or BLAST) or the narrative guide to BLAST (NCBI BLAST). (b) A computer program that identifies homologous (similar) genes in different organisms, such as human, fruit fly, or nematode (ORNL).

**BLAT** A hashing algorithm developed by Jim Kent to allow rapid searching of large amounts of genome sequence. A hashing algorithm divides the database into words of a prescribed size (often 12–14 bases). The locations of these words are stored in memory. The query sequence is scanned for exact matches to words stored in memory. These types of algorithms tend to be very fast and effective for closely related sequences, but fail as sequences diverge. In addition to nucleotide BLAT, translated BLAT allows for comparison of protein sequences. This sequence aligner also allows for accurate alignment of transcribed sequences by looking at splice site information (NCBI).

**BLOSUM** Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment in order to avoid overweighting closely related family members (NCBI BLAST).

**Bundle** An array of  $\alpha$  helices each oriented roughly along the same (bundle) axis. It may have twist, either left-handed if each helix makes a positive angle with the bundle axis or right-handed if each helix makes a negative angle with the bundle axis (SCOP).

## C

**Cancer** Diseases in which abnormal cells divide and grow unchecked. Cancer can spread from its original site to other parts of the body and can be fatal. *See also:* hereditary cancer, sporadic cancer (ORNL).

**Candidate gene** A gene located in a chromosome region suspected of being involved in a disease. *See also:* positional cloning, protein (ORNL).

**Capillary array** Gel-filled silica capillaries used to separate fragments for DNA sequencing. The small diameter of the capillaries permit the application of higher electric fields, providing high-speed, high-throughput separations that are significantly faster than traditional slab gels (ORNL).

**Carcinogen** Something which causes cancer to occur by causing changes in a cell's DNA. *See also:* mutagen (ORNL).

**Carrier** An individual who possesses an unexpressed, recessive trait (ORNL).

**cDNA library** A collection of DNA sequences that code for genes. The sequences are generated in the laboratory from mRNA sequences. *See also:* messenger RNA (ORNL).

**CDS** Coding sequence. This is the portion of an mRNA or genomic sequence that encodes for a protein sequence (NCBI).

**Cell** The basic unit of any living organism that carries on the biochemical processes of life. *See also:* genome, nucleus (ORNL).

**Centimorgan (cM)** A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, one centimorgan is equivalent, on average, to one million base pairs. *See also:* megabase (ORNL).

**Centromere** A specialized chromosome region to which spindle fibers attach during cell division (ORNL).

**Chimera (plural chimaera)** An organism that contains cells or tissues with a different genotype. These can be mutated cells of the host organism or cells from a different organism or species (ORNL).

**ChIP/chip** The hybridization of ChIP purified DNA to microarrays containing genomic DNA sequences to achieve genome-wide identification of protein-DNA interactions (NCBI).

**ChIP/seq** A technique involving size selection, high-throughput sequencing (typically using next-generation sequencing technologies that produce millions of reads in a run) and mapping of ChIP purified DNA onto a reference genome to achieve genome-wide identification of protein-DNA interactions (NCBI).

**Chloroplast chromosome** Circular DNA found in the photosynthesizing organelle (chloroplast) of plants instead of the cell nucleus, where most genetic material is located (ORNL).

**Chromatin immunoprecipitation (ChIP)** A method for identifying protein-DNA interactions. Genomic DNA and associated proteins are cross-linked, sheared, and immunoprecipitated with antibodies that recognize specific DNA proteins. Purified DNA fragments are then assayed by various techniques to determine the association of specific sequences with the protein of interest (NCBI).

**Chromosomal deletion** The loss of part of a chromosome's DNA (ORNL).

**Chromosomal inversion** Chromosome segments that have been turned 180°. The gene sequence for the segment is reversed with respect to the rest of the chromosome (ORNL).

**Chromosome** The self-replicating genetic structure of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins (ORNL).

**Chromosome painting** Attachment of certain fluorescent dyes to targeted parts of the chromosome. Used as a diagnostic for particular diseases, for example, types of leukemia (ORNL).

**Chromosome region p** A designation for the short arm of a chromosome (ORNL).

**Chromosome region q** A designation for the long arm of a chromosome (ORNL).

**Clone** An exact copy made of biological material such as a DNA segment (e.g., a gene or other region), a whole cell, or a complete organism (ORNL).

**Clone bank** *See:* genomic library (ORNL).

**Cloning** Using specialized DNA technology to produce multiple, exact copies of a single gene or other segment of DNA to obtain enough material for further study. This process, used by researchers in the Human Genome Project, is referred to as cloning DNA. The resulting cloned (copied) collections of DNA molecules are called clone libraries. A second type of cloning exploits the natural process of cell division to make many copies of an entire cell. The genetic makeup of these cloned cells, called a cell line, is identical to the original cell. A third type of cloning produces complete, genetically identical animals such as the famous Scottish sheep, Dolly. *See also:* cloning vector (ORNL).

**Cloning vector** DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where the DNA can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes; vectors are often recombinant molecules containing DNA sequences from several sources (ORNL).

**Closed, Partly Opened, and Opened** For all-alpha structures, the extent to which the hydrophobic core is screened by the  $\alpha$  helices comprising the structure. *Opened* means that there is space for at least one more helix to be easily attached to the core (SCOP).

**Code** *See:* genetic code (ORNL).

**Codominance** Situation in which two different alleles for a genetic trait are both expressed. *See also:* autosomal dominant, recessive gene (ORNL).

**Codon** *See:* genetic code (ORNL).

**Coisogenic or congenic** Nearly identical strains of an organism which vary at only a single locus (ORNL).

**Comparative genomics** The study of human genetics by comparisons with model organisms such as mice, the fruit fly, and the bacterium *Escherichia coli* (ORNL).

**Complementary DNA (cDNA)** DNA that is synthesized in the laboratory from a messenger RNA template (ORNL).

**Complementary sequence** Nucleic acid-base sequence that can form a double-stranded structure with another DNA fragment by following base-pairing rules (A pairs with T and C with G). The complementary sequence to GTAC, for example, is CATG (ORNL).

**Complex trait** Trait that has a genetic component that does not follow strict Mendelian inheritance. May involve the interaction of two or more genes or gene-environment interactions. *See also:* Mendelian inheritance, additive genetic effects (ORNL).

**Computational biology** *See:* bioinformatics (ORNL).

**Confidentiality** In genetics, the expectation that genetic material and the information gained from testing that material will not be available without the donor's consent (ORNL).

**Congenital** Any trait present at birth, whether the result of a genetic or nongenetic factor. *See also:* birth defect (ORNL).

**Conservation** Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physicochemical properties of the original residue (NCBI BLAST).

**Conserved sequence** A base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution (ORNL).

**Contig** (a) Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome (ORNL). (b) Short for contiguous sequence. When two sequences overlap at their ends (known as a "dove-tail"). The sequences can be collapsed into a single, nonredundant sequence (NCBI).

**Contig map** A map depicting the relative order of a linked library of overlapping clones representing a complete chromosomal segment (ORNL).

**Copy number variation** Large-scale structural changes in DNA that vary from individual to individual. These include insertions, deletions, duplications, and complex multi-site variants that range from kilobases to megabases in size. CNV can influence gene expression, phenotypic variation, and gene dosage. In certain instances it may be associated with developmental disorders, cause disease, or confer susceptibility to complex disease traits (NCBI).

**Cosmid** Artificially constructed cloning vector containing the *cos* gene of phage lambda. Cosmids can be packaged in lambda phage particles for infection into *Escherichia coli*; this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in plasmid vectors (ORNL).

**Crossing over** The breaking during meiosis of one maternal and one paternal chromosome, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of alleles between chromosomes. *See also:* recombination (ORNL).

**Cross-over** Connection that links secondary structures at the opposite ends of the structural core and goes across the surface of the domain (SCOP).

**Cytogenetics** The study of the physical appearance of chromosomes. *See also:* karyotype (ORNL).

**Cytological band** An area of the chromosome that stains differently from areas around it. *See also:* cytological map (ORNL).

**Cytological map** A type of chromosome map whereby genes are located on the basis of cytological findings obtained with the aid of chromosome mutations (ORNL).

**Cytoplasmic trait** A genetic characteristic in which the genes are found outside the nucleus, in chloroplasts or mitochondria. Results in offspring inheriting genetic material from only one parent (ORNL).

**Cytoplasmic (uniparental) inheritance** *See:* cytoplasmic trait (ORNL).

**Cytosine (C)** A nitrogenous base, one member of the base pair GC (guanine and cytosine) in DNA. *See also:* base pair, nucleotide (ORNL).

## D

**Data warehouse** A collection of databases, data tables, and mechanisms to access the data on a single subject (ORNL).

**Deletion** A loss of part of the DNA from a chromosome; can lead to a disease or abnormality. *See also:* chromosome, mutation (ORNL).

**Deletion map** A description of a specific chromosome that uses defined mutations – specific deleted areas in the genome – as “biochemical signposts” or markers for specific areas (ORNL).

**Deoxyribonucleotide** *See:* nucleotide (ORNL).

**Deoxyribose** A type of sugar that is one component of DNA (deoxyribonucleic acid) (ORNL).

**Diploid** A full set of genetic material consisting of paired chromosomes, one from each parental set. Most animal cells except the gametes have a diploid set of chromosomes. The diploid human genome has 46 chromosomes. *See also:* haploid (ORNL).

**Directed evolution** A laboratory process used on isolated molecules or microbes to cause mutations and identify subsequent adaptations to novel environments (ORNL).

**Directed mutagenesis** Alteration of DNA at a specific site and its reinsertion into an organism to study any effects of the change (ORNL).

**Directed sequencing** Successively sequencing DNA from adjacent stretches of chromosome (ORNL).

**Disease-associated genes** Alleles carrying particular DNA sequences associated with the presence of disease (ORNL).

**DNA (deoxyribonucleic acid)** The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs only form between A and T and between G and C; the base sequence of each single strand can therefore be deduced from that of its partner (ORNL).

**DNA bank** A service that stores DNA extracted from blood samples or other human tissue (ORNL).

**DNA probe** *See:* probe (ORNL).

**DNA repair genes** Genes encoding proteins that correct errors in DNA sequencing (ORNL).

**DNA replication** The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other eukaryotes, replication occurs in the cell nucleus (ORNL).

**DNA sequence** The relative order of base pairs, whether in a DNA fragment, gene, chromosome, or an entire genome. *See also:* base sequence analysis (ORNL).

**Domain** (a) A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function (NCBI BLAST). (b) A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function (ORNL). (c) Conserved structural entities with distinctive secondary structure content and an hydrophobic core. In small disulfide-rich and Zn<sup>2+</sup>- or Ca<sup>2+</sup>-binding domains, the hydrophobic core may be provided by cystines and metal ions, respectively. Homologous domains with common functions usually show sequence similarities (SMART).

**Domain composition** Proteins with the same domain composition have at least one copy of each of the domains of the query (SMART).

**Domain organization** Proteins having all the domains as the query in the same order (additional domains are allowed) (SMART).

**Dominant** An allele that is almost always expressed, even if only one copy is present. *See also:* gene, genome (ORNL).

**Double helix** The twisted-ladder shape that two linear strands of DNA assume when complementary nucleotides on opposing strands bond together (ORNL).

**Draft sequence** (a) The sequence generated by the Human Genome Project that, while incomplete, offers a virtual road map to an estimated 95% of all human genes. Draft sequence data are mostly in the form of 10,000 bp-sized fragments whose approximate chromosomal locations are known. *See also:* sequencing, finished DNA sequence, working draft DNA sequence (ORNL). (b) This term has had several definitions, but generally refers to a sequence that is not yet finished but is of generally high quality. In terms of clone-based projects, draft sequence refers to a project in which greater than 90% of the bases are of high quality. This means that a clone project will have several fragments connected by Ns. Often, the order and orientation of these fragments is unknown. However, these sequences, in conjunction with other data, are a useful substrate for genome assembly and annotation (NCBI).

**DUST** A program for filtering low-complexity regions from nucleic acid sequences (NCBI BLAST).

**E**

**E value** (a) Expectation value. The number of different alignments with scores equivalent to or better than  $S$  that are expected to occur in a database search by chance. The lower the  $E$  value, the more significant the score (NCBI BLAST). (b) This represents the number of sequences with a score greater than or equal to  $X$  expected absolutely by chance. The  $E$  value connects the score ( $X$ ) of an alignment between a user-supplied sequence and a database sequence, generated by any algorithm, with the number of alignments with similar or greater scores that would be expected from a search of a random-sequence database of equivalent size. Since version 2.0,  $E$  values are calculated using hidden Markov models, leading to more accurate estimates than before (SMART).

**Electrophoresis** A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids (ORNL).

**Electroporation** A process using high-voltage current to make cell membranes permeable to allow the introduction of new DNA; commonly used in recombinant DNA technology. *See also:* transfection (ORNL).

**Embryonic stem (ES) cells** An embryonic cell that can replicate indefinitely, transform into other types of cells, and serve as a continuous source of new cells (ORNL).

**Endonuclease** *See:* restriction enzyme (ORNL).

**Enzyme** A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction (ORNL).

**Epistasis** One gene interferes with or prevents the expression of another gene located at a different locus (ORNL).

**Escherichia coli** Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory (ORNL).

**Eugenics** The study of improving a species by artificial selection; usually refers to the selective breeding of humans (ORNL).

**Eukaryote** Cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue-green algae. *See also:* prokaryote, chromosome (ORNL).

**Evolutionarily conserved** *See:* conserved sequence (ORNL).

**Exogenous DNA** DNA originating outside an organism that has been introduced into the organism (ORNL).

**Exon** The protein-coding DNA sequence of a gene. *See also:* intron (ORNL).

**Exonuclease** An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate (ORNL).

**Expressed gene** *See:* gene expression (ORNL).

**Expressed sequence tag (EST)** (a) A short strand of DNA that is part of a cDNA molecule and can act as identifier of a gene. Used in locating and mapping genes. *See also:* cDNA, sequence-tagged site (ORNL). (b) These are single-pass sequences of cDNA clones. Databases of EST sequences are highly redundant but quite useful for gene identification. There are many efforts to cluster EST sequences to remove the redundancy and low-quality sequences (NCBI).

**F**

**FASTA** (a) The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called “words.” Initially, the scores of segments in which there are multiple word hits are calculated (“init1”). Later, the scores of several segments may be summed to generate an “initn” score. An optimized alignment that includes gaps is shown in the output as “opt.” The sensitivity and speed of the search are inversely related and controlled by the “k-tup” variable, which specifies the size of a word (NCBI BLAST). (b) An output format for nucleic acid or protein sequences.

**Filial generation (F1, F2)** Each generation of offspring in a breeding program, designated F1, F2, et. (ORNL).

**Filtering** Also known as masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores. *See also:* SEG and DUST (NCBI BLAST).

**Fingerprinting** (a) In genetics, the identification of multiple specific alleles on a person’s DNA to produce a unique identifier for that person. *See also:* forensics (ORNL). (b) The pattern of bands produced by a clone when restricted by a particular enzyme, such as *Hind*III. Clones that are related will have fingerprint bands in common. The more bands in common, the greater the degree of overlap (NCBI).

**Finished DNA sequence** High-quality, low-error, gap-free DNA sequence of the human genome. Achieving this ultimate 2003 Human Genome Project (HGP) goal requires additional sequencing to close gaps, reduce ambiguities, and allow for only a single error every 10,000 bases, the agreed-upon standard for HGP finished sequence. *See also:* sequencing, draft sequence (ORNL).

**Flow cytometry** Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., chromosomes) passing in a narrow stream through a laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet (ORNL).

**Flow karyotyping** Use of flow cytometry to analyze and separate chromosomes according to their DNA content (ORNL).

**Fluorescence *in situ* hybridization (FISH)** A physical mapping approach that uses fluorescein tags to detect hybridization of probes with metaphase chromosomes and with the less-condensed somatic interphase chromatin (ORNL).

**Folded leaf** A layer of  $\alpha$  helices wrapped around a single hydrophobic core but not with the simple geometry of a bundle (SCOP).

**Forensics** The use of DNA for identification. Some examples of DNA use are to establish paternity in child support cases, establish the presence of a suspect at a crime scene, and identify accident victims (ORNL).

**Fosmid** A cloning system based on the *E. coli* F factor. These clones have an average insert size of 40 kb, with a very small standard deviation (NCBI).

**Fraternal twin** Siblings born at the same time as the result of fertilization of two ova by two sperm. They share the same genetic relationship to each other as any other siblings. *See also:* identical twin (ORNL).

**Full gene sequence** The complete order of bases in a gene. This order determines which protein a gene will produce (ORNL).

**Functional genomics** The study of genes, their resulting proteins, and the role played by the proteins in the body’s biochemical processes (ORNL).

**G**

**Gamete** Mature male or female reproductive cell (sperm or ovum) with a haploid set of chromosomes (23 for humans) (ORNL).

**Gap** (a) A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, the introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment (NCBI BLAST). (b) A position in an alignment that represents a deletion within one sequence relative to another. Gap penalties are requirements for alignment algorithms in order to reduce excessively gapped regions. Gaps in alignments represent insertions that usually occur in protruding loops or beta-bulges within protein structures (SMART).

**GC-rich area** Many DNA sequences carry long stretches of repeated G and C, which often indicates a gene-rich region (ORNL).

**Gel electrophoresis** *See:* electrophoresis (ORNL).

**Gene** The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule). *See also:* gene expression (ORNL).

**Gene amplification** Repeated copying of a piece of DNA; a characteristic of tumor cells. *See also:* gene, oncogene (ORNL).

**Gene chip technology** Development of cDNA microarrays from a large number of genes. Used to monitor and measure changes in gene expression for each gene represented on the chip (ORNL).

**Gene expression** The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein, and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs) (ORNL).

**Gene family** Group of closely related genes that make similar products (ORNL).

**Gene library** *See:* genomic library (ORNL).

**Gene mapping** Determination of the relative positions of genes on a DNA molecule (chromosome or plasmid) and of the distance, in linkage units or physical units, between them (ORNL).

**Gene pool** All the variations of genes in a species. *See also:* allele, gene, polymorphism (ORNL).

**Gene prediction** Predictions of possible genes made by a computer program based on how well a stretch of DNA sequence matches known gene sequences (ORNL).

**Gene product** The biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing alleles (ORNL).

**Gene targeting** This is a specific type of transgenesis that targets a particular gene. If a mutated copy of a gene is electroporated into a cell, the inserted DNA will find the endogenous copy of itself and recombination will occur with some frequency (1–25%). If this event occurs in embryonic stem cells, cells carrying the new copy of the gene can be used to generate embryos that can be assessed for the phenotypic consequences of the mutation. This technique is used frequently in mice to study loss-of-function mutations (NCBI).

**Gene testing** *See:* genetic testing, genetic screening (ORNL).

**Gene therapy** An experimental procedure aimed at replacing, manipulating, or supplementing nonfunctional or malfunctioning genes with healthy genes. *See also:* gene, inherit, somatic cell gene therapy, germ line gene therapy (ORNL).

**Gene trapping** This strategy uses transgenesis to introduce DNA carrying a reporter gene (*lacZ* or GFP) flanked by various genomic signals (splice donor or acceptor sites, promoters, etc.). Expression of the reporter gene indicates that the DNA has integrated into a region of the genome containing a gene. The gene that has been trapped can be recovered using the DNA sequences associated with the reporter construct. Often, the introduction of the gene-trapping vector inactivates the gene into which it was introduced (NCBI).

**Genetic code** The sequence of nucleotides, coded in triplets (codons) along the mRNA, that determines the sequence of amino acids in protein synthesis. A gene's DNA sequence can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence (ORNL).

**Genetic counseling** Provides patients and their families with education and information about genetic-related conditions and helps them to make informed decisions (ORNL).

**Genetic discrimination** Prejudice against those who have or are likely to develop an inherited disorder (ORNL).

**Genetic engineering** Altering the genetic material of cells or organisms to enable them to make new substances or perform new functions (ORNL).

**Genetic engineering technology** *See:* recombinant DNA technology (ORNL).

**Genetic illness** Sickness, physical disability, or other disorder resulting from the inheritance of one or more deleterious alleles (ORNL).

**Genetic informatics** *See:* bioinformatics (ORNL).

**Genetic map** *See:* linkage map (ORNL).

**Genetic marker** A gene or other identifiable portion of DNA whose inheritance can be followed. *See also:* chromosome, DNA, gene, inherit (ORNL).

**Genetic material** *See:* genome (ORNL).

**Genetic mosaic** An organism in which different cells contain different genetic sequence. This can be the result of a mutation during development or fusion of embryos at an early developmental stage (ORNL).

**Genetic polymorphism** Difference in DNA sequence among individuals, groups, or populations (e.g., genes for blue eyes versus brown eyes) (ORNL).

**Genetic predisposition** Susceptibility to a genetic disease. May or may not result in actual development of the disease (ORNL).

**Genetic screening** Testing a group of people to identify individuals at high risk of having or passing on a specific genetic disorder (ORNL).

**Genetic testing** Analyzing an individual's genetic material to determine predisposition to a particular health condition or to confirm a diagnosis of genetic disease (ORNL).

**Genetics** The study of inheritance patterns of specific traits (ORNL).

**Gene transfer** Incorporation of new DNA into an organism's cells, usually by a vector such as a modified virus. Used in gene therapy. *See also:* mutation, gene therapy, vector (ORNL).

**Genome** All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs (ORNL).

**Genome project** Research and technology development effort aimed at mapping and sequencing the genome of human beings and certain model organisms. *See also:* Human Genome Initiative (ORNL).

**Genomic library** A collection of clones made from a set of randomly generated overlapping DNA fragments that represent the entire genome of an organism. *See also:* library, arrayed library (ORNL).

**Genomics** The study of genes and their function (ORNL).

**Genomic sequence** *See:* DNA (ORNL).

**Genotype** The genetic constitution of an organism, as distinguished from its physical appearance (its phenotype) (ORNL).

**Germ cell** Sperm and egg cells and their precursors. Germ cells are haploid and have only one set of chromosomes (23 in all), while all other cells have two copies (46 in all) (ORNL).

**Germ line** The continuation of a set of genetic information from one generation to the next. *See also:* inherit (ORNL).

**Germ line gene therapy** An experimental process of inserting genes into germ cells or fertilized eggs to cause a genetic change that can be passed on to offspring. May be used to alleviate effects associated with a genetic disease. *See also:* genomics, somatic cell gene therapy (ORNL).

**Germ line genetic mutation** *See:* mutation (ORNL).

**Global alignment** The alignment of two nucleic acid or protein sequences over their entire length (NCBI BLAST).

**Greek key** A topology for a small number of  $\beta$ -sheet strands in which some interstrand connections go across the end of a barrel or, in a sandwich fold, between  $\beta$  sheets (SCOP).

**Guanine (G)** A nitrogenous base, one member of the base pair GC (guanine and cytosine) in DNA. *See also:* base pair, nucleotide (ORNL).

## H

**H** The relative entropy of the target and background residue frequencies  $H$  can be thought of as a measure of the average information (in bits) available per position that distinguishes an alignment from chance. At high values of  $H$  short alignments can be distinguished by chance, whereas at lower  $H$  values a longer alignment may be necessary (NCBI BLAST).

**Haploid** A single set of chromosomes (half the full set of genetic material) present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. *See also:* diploid (ORNL).

**Haplotype** (a) A way of denoting the collective genotype of a number of closely linked loci on a chromosome (ORNL). (b) A set of closely linked genetic markers present on one chromosome that tend to be inherited together. A haplotype may also refer to a set of single-nucleotide polymorphisms (SNPs) on a single chromatid that are statistically associated with one another (NCBI).

**Hemizygous** Having only one copy of a particular gene. For example, in humans, males are hemizygous for genes found on the Y chromosome (ORNL).

**Heredity cancer** Cancer that occurs due to the inheritance of an altered gene within a family. *See also:* sporadic cancer (ORNL).

**Heterozygosity** The presence of different alleles at one or more loci on homologous chromosomes (ORNL).

**Heterozygote** *See:* heterozygosity (ORNL).

**Highly conserved sequence** DNA sequence that is very similar across several different types of organisms. *See also:* gene, mutation (ORNL).

**High-throughput sequencing** A fast method of determining the order of bases in DNA. *See also:* sequencing (ORNL).

**Homeobox** A short stretch of nucleotides whose base sequence is virtually identical in all the genes that contain it. Homeoboxes have been found in many organisms from fruit flies to human beings. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development (ORNL).

**Homolog** A member of a chromosome pair in diploid organisms or a gene that has the same origin and functions in two or more species (ORNL).

**Homologous chromosome** Chromosome containing the same linear gene sequences as another, each derived from one parent (ORNL).

**Homologous recombination** Swapping of DNA fragments between paired chromosomes (ORNL).

**Homology** (a) Similarity attributed to descent from a common ancestor (NCBI BLAST). (b) Similarity in DNA or protein sequences between individuals of the same species or among different species (ORNL). (c) Evolutionary descent from a common ancestor due to gene duplication (SMART).

**Homozygote** An organism that has two identical alleles of a gene. *See also:* heterozygote (ORNL).

**Homozygous** *See:* homozygote (ORNL).

**HSP** High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search (NCBI BLAST).

**Human gene therapy** *See:* gene therapy (ORNL).

**Human Genome Initiative** Collective name for several projects begun in 1986 by the US Department of Energy (DOE) to create an ordered set of DNA segments from known chromosomal locations, develop new computational methods for analyzing genetic map and DNA sequence data, and develop new techniques and instruments for detecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The joint national effort, led by the DOE and National Institutes of Health, is known as the Human Genome Project (ORNL).

**Human Genome Project (HGP)** Formerly titled Human Genome Initiative. *See also:* Human Genome Initiative (ORNL).

**Hybrid** The offspring of genetically different parents. *See also:* heterozygote (ORNL).

**Hybridization** The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule (ORNL).

## I

**Identical twin** Twins produced by the division of a single zygote, with identical genotypes. *See also:* fraternal twin (ORNL).

**Identity** The extent to which two (nucleotide or amino acid) sequences are invariant (NCBI BLAST).

**Immunotherapy** Using the immune system to treat disease, for example, in the development of vaccines. May also refer to the therapy of diseases caused by the immune system. *See also:* cancer (ORNL).

**Imprinting** A phenomenon in which the disease phenotype depends on which parent passed on the disease gene. For instance, both Prader-Willi and Angelman syndromes are inherited when the same part of chromosome 15 is missing. When the father's complement of 15 is missing, the child has Prader-Willi, but when the mother's complement of 15 is missing, the child has Angelman syndrome (ORNL).

**Independent assortment** During meiosis each of the two copies of a gene is distributed to the germ cells independently of the distribution of other genes. *See also:* linkage (ORNL).

**Informatics** *See:* bioinformatics (ORNL).

**Informed consent** An individual willingly agrees to participate in an activity after first being advised of the risks and benefits. *See also:* privacy (ORNL).

**Inherit** In genetics, to receive genetic material from parents through biological processes (ORNL).

**Inherited** *See:* inherit (ORNL).

**Insertion** A chromosome abnormality in which a piece of DNA is incorporated into a gene and thereby disrupts the gene's normal function. *See also:* chromosome, DNA, gene, mutation (ORNL).

**Insertional mutation** *See:* insertion (ORNL).

**In situ hybridization** Use of a DNA or RNA probe to detect the presence of the complementary DNA sequence in cloned bacterial or cultured eukaryotic cells (ORNL).

**Intellectual property rights** Patents, copyrights, and trademarks. *See also:* patent (ORNL).

**Interference** One cross-over event inhibits the chances of another cross-over event. Also known as positive interference. Negative interference increases the chance of a second cross-over. *See also:* crossing over (ORNL).

**Interphase** The period in the cell cycle when DNA is replicated in the nucleus; followed by mitosis (ORNL).

**Intracellular domains** Domain families that are most prevalent in proteins within the cytoplasm (SMART).

**Intron** DNA sequence that interrupts the protein-coding sequence of a gene; an intron is transcribed into RNA but is cut out of the message before it is translated into protein. *See also:* exon (ORNL).

**In vitro** Studies performed outside a living organism, such as in a laboratory (ORNL).

**In vivo** Studies carried out in living organisms (ORNL).

**Isoenzyme** An enzyme performing the same function as another enzyme but having a different set of amino acids. The two enzymes may function at different speeds (ORNL).

## J

**Jelly roll** A variant of Greek-key topology with both ends of a sandwich or a barrel fold being crossed by two interstrand connections. *See also:* Greek key (SCOP).

**Junk DNA** Stretches of DNA that do not code for genes; most of the genome consists of so-called junk DNA which may have regulatory and other functions. Also called non-coding DNA (ORNL).

## K

**K** A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for search space size. The value *K* is used in converting a raw score (*S*) to a bit score (*S'*) (NCBI BLAST).

**Karyotype** A photomicrograph of an individual's chromosomes arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution physical mapping to correlate gross chromosomal abnormalities with the characteristics of specific diseases (ORNL).

**Kilobase (kb)** Unit of length for DNA fragments equal to 1000 nucleotides (ORNL).

**Knockout** Deactivation of specific genes; used in laboratory organisms to study gene function. *See also:* gene, locus, model organisms (ORNL).

**L**

**Lambda ( $\lambda$ )** A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for a scoring system. The value  $\lambda$  is used in converting a raw score ( $S$ ) to a bit score ( $S'$ ) (NCBI BLAST).

**Library** An unordered collection of clones (i.e., cloned DNA from a particular organism) whose relationship to each other can be established by physical mapping. *See also:* genomic library, arrayed library (ORNL).

**Linkage** The proximity of two or more markers (e.g., genes, restriction fragment length polymorphism markers) on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together (ORNL).

**Linkage disequilibrium** Where alleles occur together more often than can be accounted for by chance. Indicates that the two alleles are physically close on the DNA strand. *See also:* Mendelian inheritance (ORNL).

**Linkage map** A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans (cM) (ORNL).

**Local alignment** The alignment of some portion of two nucleic acid or protein sequences (NCBI BLAST).

**Localization** Numbers of domains that are thought from SwissProt annotations to be present in different cellular compartments (cytoplasm, extracellular space, nucleus, and membrane associated) are shown in annotation pages (SMART).

**Localize** Determination of the original position (locus) of a gene or other marker on a chromosome (ORNL).

**Locus (plural loci)** The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. The use of locus is sometimes restricted to mean expressed DNA regions. *See also:* gene expression (ORNL).

**Long-range restriction mapping** Restriction enzymes are proteins that cut DNA at precise locations. Restriction maps depict the chromosomal positions of restriction enzyme cutting sites. These are used as biochemical “signposts” or markers of specific areas along the chromosomes. The map will detail the positions where the DNA molecule is cut by particular restriction enzymes (ORNL).

**Low-complexity region (LCR)** Regions of biased composition including homopolymeric runs, short-period repeats, and more subtle overrepresentation of one or a few residues. The SEG program is used to mask or filter LCRs in amino acid queries. The DUST program is used to mask or filter LCRs in nucleic acid queries (NCBI BLAST).

**M**

**Macrorestriction map** Map depicting the order of and distance between sites at which restriction enzymes cleave chromosomes (ORNL).

**Mapping** *See:* gene mapping, linkage map, physical map (ORNL).

**Mapping population** The group of related organisms used in constructing a genetic map (ORNL).

**Marker** *See:* genetic marker (ORNL).

**Masking** Also known as filtering. The removal of repeated or low-complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence (NCBI BLAST).

**Mass spectrometer** An instrument used to identify chemicals in a substance by their mass and charge (ORNL).

**Mate pair** The sequence obtained from opposite ends of a particular clone are referred to as mate pairs. Knowing that two sequences are derived from the same clone allows these sequences to be linked, even if the full insert of the clone is unavailable. This is key to WGS assemblies (NCBI).

**Meander** A simple topology of a  $\beta$  sheet where any two consecutive strands are adjacent and antiparallel (SCOP).

**Megabase (Mb)** Unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM. *See also:* centimorgan (ORNL).

**Meiosis** The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes. *See also:* mitosis (ORNL).

**Mendelian inheritance** One method in which genetic traits are passed from parents to offspring. Named for Gregor Mendel, who first studied and recognized the existence of genes and this method of inheritance. *See also:* autosomal dominant, recessive gene, sex linked (ORNL).

**Messenger RNA (mRNA)** RNA that serves as a template for protein synthesis. *See also:* genetic code (ORNL).

**Metaphase** A stage in mitosis or meiosis during which the chromosomes are aligned along the equatorial plane of the cell (ORNL).

**Microarray** Sets of miniaturized chemical reaction areas that may also be used to test DNA fragments, antibodies, or proteins (ORNL).

**Microbial genetics** The study of genes and gene function in bacteria, archaea, and other microorganisms. Often used in research in the fields of bioremediation, alternative energy, and disease prevention. *See also:* model organisms, biotechnology, bioremediation (ORNL).

**Microinjection** A technique for introducing a solution of DNA into a cell using a fine microcapillary pipet (ORNL).

**Mitochondrial DNA** The genetic material found in mitochondria, the organelles that generate energy for the cell. Not inherited in the same fashion as nucleic DNA. *See also:* cell, DNA, genome, nucleus (ORNL).

**Mitosis** The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell. *See also:* meiosis (ORNL).

**Modeling** The use of statistical analysis, computer analysis, or model organisms to predict outcomes of research (ORNL).

**Model organisms** A laboratory animal or other organism useful for research (ORNL).

**Molecular biology** The study of the structure, function, and makeup of biologically important molecules (ORNL).

**Molecular farming** The development of transgenic animals to produce human proteins for medical use (ORNL).

**Molecular genetics** The study of macromolecules important in biological inheritance (ORNL).

**Molecular medicine** The treatment of injury or disease at the molecular level. Examples include the use of DNA-based diagnostic tests or medicine derived from DNA sequence information (ORNL).

**Monogenic disorder** A disorder caused by mutation of a single gene. *See also:* mutation, polygenic disorder (ORNL).

**Monogenic inheritance** *See:* monogenic disorder (ORNL).

**Monosomy** Possessing only one copy of a particular chromosome instead of the normal two copies. *See also:* cell, chromosome, gene expression, trisomy (ORNL).

**Morbid map** A diagram showing the chromosomal location of genes associated with disease (ORNL).

**Motif** (a) A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains (NCBI BLAST). (b) Sequence motifs are short conserved regions of polypeptides. Sets of sequence motifs need not necessarily represent homologs (SMART).

**Mouse model** *See:* model organisms (ORNL).

**Multifactorial or multigenic disorder** *See:* polygenic disorder (ORNL).

**Multiple sequence alignment** An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. ClustalW is one of the most widely used multiple sequence alignment programs (NCBI BLAST).

**Multiplexing** A laboratory approach that performs multiple sets of reactions in parallel (simultaneously); greatly increasing speed and throughput (ORNL).

**Murine** Organism in the genus *Mus*. A rat or mouse (ORNL).

**Mutagen** An agent that causes a permanent genetic change in a cell. Does not include changes occurring during normal genetic recombination (ORNL).

**Mutagenicity** The capacity of a chemical or physical agent to cause permanent genetic alterations. *See also:* somatic cell genetic mutation (ORNL).

**Mutation** (a) Any heritable change in DNA sequence. *See also:* polymorphism (ORNL). (b) A sequence variation that deviates from the reference, or “wildtype”, sequence. This variation can be a SNP, an insertion of sequence, or a deletion of sequence. There can be a great deal of sequence variation between individuals in a population. For example, different humans may have as many as 1 base pair difference every 1000 bp. In practice, mutations are distinguished from variation because they have phenotypic consequences. Mutations in the Pax6 gene that lead to a loss of the function of that gene lead to the *eyeless* mutation in flies, the Small eye mutation in mice, and aniridia in humans (NCBI).

## N

**N50** The contig/scaffold length at which half of the bases in a given assembly reside. This provides a measure of continuity. For instance, a scaffold N50 of 15 Mb means that at least half of the bases in the assembly are in a contig that is at least 15 Mb (NCBI).

**Nitrogenous base** A nitrogen-containing molecule having the chemical properties of a base. DNA contains the nitrogenous bases adenine (A), guanine (G), cytosine (C), and thymine (T). *See also:* DNA (ORNL).

**Northern blot** A gel-based laboratory procedure that locates mRNA sequences on a gel that are complementary to a piece of DNA used as a probe. *See also:* DNA, library (ORNL).

**Nuclear transfer** A laboratory procedure in which a cell’s nucleus is removed and placed into an oocyte with its own nucleus removed so the genetic information from the donor nucleus controls the resulting cell. Such cells can be induced to form embryos. This process was used to create the cloned sheep Dolly. *See also:* cloning (ORNL).

**Nucleic acid** A large molecule composed of nucleotide subunits. *See also:* DNA (ORNL).

**Nucleolar organizing region** A part of the chromosome containing rRNA genes (ORNL).

**Nucleotide** A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. *See also:* DNA, base pair, RNA (ORNL).

**Nucleus** The cellular organelle in eukaryotes that contains most of the genetic material (ORNL).

## O

**Oligo** *See:* oligonucleotide (ORNL).

**Oligogenic** A phenotypic trait produced by two or more genes working together. *See also:* polygenic disorder (ORNL).

**Oligonucleotide** A molecule usually composed of 25 or fewer nucleotides; used as a DNA synthesis primer. *See also:* nucleotide (ORNL).

**Oncogene** A gene, one or more forms of which are associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth (ORNL).

**Open reading frame (ORF)** The sequence of DNA or RNA located between the start-code sequence (initiation codon) and the stop-code sequence (termination codon) (ORNL).

**Operon** A set of genes transcribed under the control of an operator gene (ORNL).

**Optimal alignment** An alignment of two sequences with the highest possible score (NCBI BLAST).

**ORF** *See:* open reading frame (SMART).

**Orthologous** Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function (NCBI BLAST).

**Overlapping clones** *See:* genomic library (ORNL).

## P

**p value** The probability of an alignment occurring with the score in question or better. The *p* value is calculated by relating the observed alignment score *S* to the expected distribution of high-scoring segment pair scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant *p* values will be those close to zero. The *p* and *E* values are different ways of representing the significance of the alignment (NCBI BLAST).

**P1-derived artificial chromosome (PAC)** One type of vector used to clone DNA fragments (insert size 100–300 kb; average 150 kb) in *Escherichia coli* cells. Based on bacteriophage (a virus) P1 genome. *See also:* cloning vector (ORNL).

**PAM** Point accepted mutation. A unit used to quantify the amount of evolutionary change in a protein sequence. The amount of evolution which will change, on average, 1% of amino acids in a protein sequence is 1.0 PAM units. A PAM(*x*) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (*x*) of evolutionary divergence (NCBI BLAST).

**Paralogous** Homologous sequences within a single species that arose by gene duplication (NCBI BLAST).

**Patent** In genetics, conferring the right or title to genes, gene variations, or identifiable portions of sequenced genetic material of an individual or organization. *See also:* gene (ORNL).

**Pedigree** A family tree diagram that shows how a particular genetic trait or disease has been inherited. *See also:* inherit (ORNL).

**Penetrance** The probability of a gene or genetic trait being expressed. “Complete” penetrance means the gene or genes for a trait are expressed in the whole population that has the genes. “Incomplete” penetrance means the genetic trait is expressed in only part of the population. The percent penetrance may also change with the age range of the population (ORNL).

**Peptide** Two or more amino acids joined by a bond called a “peptide bond.” *See also:* polypeptide (ORNL).

**Phage** A virus for which the natural host is a bacterial cell (ORNL).

**Pharmacogenomics** The study of the interaction of an individual’s genetic makeup and response to a drug (ORNL).

**Phenocopy** A trait not caused by inheritance of a gene but that appears to be identical to a genetic trait (ORNL).

**Phenotype** (a) The physical characteristics of an organism or the presence of a disease that may or may not be genetic. *See also:* genotype (ORNL). (b) An observable characteristic displayed by an organism. These characteristics can be controlled by genes, by the environment, or a combination of both. The characteristic can be directly observable, such as having brown eyes. In some cases, the phenotype will be measurable, such as having high blood pressure (NCBI).

**Physical map** A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns on the 24 different chromosomes; the highest-resolution map is the complete nucleotide sequence of the chromosomes (ORNL).

**Plasmid** Autonomously replicating extrachromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors (ORNL).

**Pleiotropy** One gene that causes many different physical traits such as multiple disease symptoms (ORNL).

**Pluripotency** The potential of a cell to develop into more than one type of mature cell, depending on environment (ORNL).

**Polygenic disorder** Genetic disorder resulting from the combined action of alleles of more than one gene (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; hereditary patterns are therefore usually more complex than those of single-gene disorders. *See also:* single-gene disorder (ORNL).

**Polymerase chain reaction (PCR)** A method for amplifying a DNA base sequence using a heat-stable polymerase and two 20-base primers, one complementary to the (+) strand at one end of the sequence to be amplified and one complementary to the (–) strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR can also be used to detect the existence of the defined sequence in a DNA sample (ORNL).

**Polymerase, DNA or RNA** Enzyme that catalyzes the synthesis of nucleic acids on pre-existing nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides (ORNL).

**Polymorphism** Difference in DNA sequence among individuals that may underlie differences in health. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic linkage analysis. *See also:* mutation (ORNL).

**Polypeptide** A protein or part of a protein made of a chain of amino acids joined by a peptide bond (ORNL).

**Population genetics** The study of variation in genes among a group of individuals (ORNL).

**Positional cloning** A technique used to identify genes, usually those that are associated with diseases, based on their location on a chromosome (ORNL).

**Primer** Short pre-existing polynucleotide chain to which new deoxyribonucleotides can be added by DNA polymerase (ORNL).

**Privacy** In genetics, the right of people to restrict access to their genetic information (ORNL).

**Probe** Single-stranded DNA or RNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect the complementary base sequence by hybridization (ORNL).

**Profile** (a) A table that lists the frequencies of each amino acid in each position of protein sequence. Frequencies are calculated from multiple alignments of sequences containing a domain of interest. *See also:* PSSM (NCBI BLAST). (b) A table of position-specific scores and gap penalties, representing a homologous family, that may be used to search sequence databases. In ClustalW-derived profiles those sequences that are more distantly related are assigned higher weights (SMART).

**Prokaryote** Cell or organism lacking a membrane-bound, structurally discrete nucleus and other subcellular compartments. Bacteria are examples of prokaryotes. *See also:* chromosome, eukaryote (ORNL).

**Promoter** A DNA site to which RNA polymerase will bind and initiate transcription (ORNL).

**Pronucleus** The nucleus of a sperm or egg prior to fertilization. *See also:* nucleus, transgenic (ORNL).

**Protein** A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs; each protein has unique functions. Examples are hormones, enzymes, and antibodies (ORNL).

**Proteome** Proteins expressed by a cell or organ at a particular time and under specific conditions (ORNL).

**Proteomics** Systematic analysis of protein expression of normal and diseased tissues that involves the separation, identification, and characterization of all of the proteins in an organism (NCBI BLAST).

**Pseudogene** A sequence of DNA similar to a gene but nonfunctional; probably the remnant of a once-functional gene that accumulated mutations (ORNL).

**PSI-BLAST** Position-specific iterative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired, with new sequences found in each cycle used to refine the profile (NCBI BLAST).

**PSSM** Position-specific scoring matrix. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence. *See also:* profile (NCBI BLAST).

**Purine** A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine. *See also:* base pair (ORNL).

**Pyrimidine** A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil. *See also:* base pair (ORNL).

## Q

**Query** The input sequence (or other type of search term) with which all of the entries in a database are to be compared (NCBI BLAST).

## R

**Radiation hybrid** A hybrid cell containing small fragments of irradiated human chromosomes. Maps of irradiation sites on chromosomes for the human, rat, mouse, and other genomes provide important markers, allowing the construction of very precise sequence-tagged site maps indispensable to studying multifactorial diseases. *See also:* sequence-tagged site (ORNL).

**Rare-cutter enzyme** *See:* restriction enzyme cutting site (ORNL).

**Raw score** The score of an alignment  $S$  calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table. Gap scores are typically calculated as the sum of  $G$ , the gap opening penalty, and  $L$ , the gap extension penalty. For a gap of length  $n$ , the gap cost would be  $G + Ln$ . The choice of gap costs  $G$  and  $L$  is empirical, but it is customary to choose a high value for  $G$  (10–15) and a low value for  $L$  (1–2). *See also:* PAM, BLOSUM (NCBI BLAST).

**Recessive gene** A gene which will only be expressed if there are two identical copies or, for a male, if one copy is present on the X chromosome (ORNL).

**Reciprocal translocation** When a pair of chromosomes exchange exactly the same length and area of DNA. Results in a shuffling of genes (ORNL).

**Recombinant clone** Clone containing recombinant DNA molecules. *See also:* recombinant DNA technology (ORNL).

**Recombinant DNA molecules** A combination of DNA molecules of different origin that are joined using recombinant DNA technologies (ORNL).

**Recombinant DNA technology** Procedure used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome (ORNL).

**Recombination** The process by which progeny derives a combination of genes different from that of either parent. In higher organisms, this can occur by crossing over. *See also:* crossing over, mutation (ORNL).

**RefSeq (Reference Sequence)** The goal of the RefSeq project is to produce a reference sequence for all naturally occurring molecules from the central dogma (DNA, RNA, Protein) (NCBI).

**Regulatory region or sequence** A DNA base sequence that controls gene expression (ORNL).

**Repetitive DNA** Sequences of varying lengths that occur in multiple copies in the genome; it represents much of the human genome (ORNL).

**Reporter gene** *See:* marker (ORNL).

**Resolution** Degree of molecular detail on a physical map of DNA, ranging from low to high (ORNL).

**Restriction enzyme cutting site** A specific nucleotide sequence of DNA at which a particular restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs); others much less frequently (rare cutter; e.g., every 10,000 bp) (ORNL).

**Restriction enzyme, endonuclease** A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. Bacteria contain over 400 such enzymes that recognize and cut more than 100 different DNA sequences. *See also:* restriction enzyme cutting site (ORNL).

**Restriction fragment length polymorphism (RFLP)** Variation between individuals in DNA fragment sizes cut by specific restriction enzymes; polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps. RFLPs are usually caused by mutation at a cutting site. *See also:* marker, polymorphism (ORNL).

**Retroviral infection** The presence of retroviral vectors, such as some viruses, which use their recombinant DNA to insert their genetic material into the chromosomes of the host's cells. The virus is then propagated by the host cell (ORNL).

**Reverse transcriptase** An enzyme used by retroviruses to form a complementary DNA sequence (cDNA) from their RNA. The resulting DNA is then inserted into the chromosome of the host cell (ORNL).

**Ribonucleotide** *See:* nucleotide (ORNL).

**Ribose** The five-carbon sugar that serves as a component of RNA. *See also:* ribonucleic acid, deoxyribose (ORNL).

**Ribosomal RNA (rRNA)** A class of RNA found in the ribosomes of cells (ORNL).

**Ribosomes** Small cellular components composed of specialized ribosomal RNA and protein; site of protein synthesis. *See also:* RNA (ORNL).

**Risk communication** In genetics, a process in which a genetic counselor or other medical professional interprets genetic test results and advises patients of the consequences for them and their offspring (ORNL).

**RNA (ribonucleic acid)** A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose (ORNL).

## S

**Sanger sequencing** A widely used method of determining the order of bases in DNA. *See also:* sequencing, shotgun sequencing (ORNL).

**Satellite** A chromosomal segment that branches off from the rest of the chromosome but is still connected by a thin filament or stalk (ORNL).

**Scaffold** In genomic mapping, a series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence (ORNL).

**Seed alignment** Alignment that contains only one of each pair of homologs that are represented in a ClustalW-derived phylogenetic tree linked by a branch of length less than a distance of 0.2 (SMART).

**SEG** A program for filtering low-complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment. SEG filtering is performed by default in the blastp subroutine of BLAST 2.0 (NCBI BLAST).

**Segmental duplication** A region of genomic DNA ranging from 1 to 400 kb that may be found at more than one site in the genome. Segmental duplications often share >90% sequence identity. See also Copy Number Variation (CNV) (NCBI).

**Segregation** The normal biological process whereby the two pieces of a chromosome pair are separated during meiosis and randomly distributed to the germ cells (ORNL).

**Sequence** *See:* base sequence (ORNL).

**Sequence assembly** A process whereby the order of multiple sequenced DNA fragments is determined (ORNL).

**Sequence-tagged site (STS)** Short (200–500 bp) DNA sequence that has a single occurrence in the human genome and whose location and base sequence are known. Detectable by polymerase chain reaction, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories, and serve as landmarks on the developing physical map of the human genome. Expressed sequence tags (ESTs) are STSs derived from cDNAs (ORNL).

**Sequencing** Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein (ORNL).

**Sequencing technology** The instrumentation and procedures used to determine the order of nucleotides in DNA (ORNL).

**Sex chromosome** The X or Y chromosome in human beings that determines the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a karyotype. *See also:* autosome (ORNL).

**Sex linked** Traits or diseases associated with the X or Y chromosome; generally seen in males. *See also:* gene, mutation, sex chromosome (ORNL).

**Shotgun method** Sequencing method that involves randomly sequenced cloned pieces of the genome, with no prior knowledge of where the piece originally came from. This can be contrasted with “directed” strategies, in which pieces of DNA from known chromosomal locations are sequenced. Because there are advantages to both strategies, researchers use both random (or shotgun) and directed strategies in combination to sequence the human genome. *See also:* library, genomic library (ORNL).

**Similarity** The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST, similarity refers to a positive matrix score (NCBI BLAST).

**Single-gene disorder** Hereditary disorder caused by a mutant allele of a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). *See also:* polygenic disorders (ORNL).

**Single-nucleotide polymorphism (SNP)** (a) DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. *See also:* mutation, polymorphism, single-gene disorder (ORNL). (b) A single base difference found when comparing the same DNA sequence from two different individuals (NCBI).

**Somatic cell** Any cell in the body except gametes and their precursors. *See also:* gamete (ORNL).

**Somatic cell gene therapy** Incorporating new genetic material into cells for therapeutic purposes. The new genetic material cannot be passed to offspring. *See also:* gene therapy (ORNL).

**Somatic cell genetic mutation** A change in the genetic structure that is neither inherited nor passed to offspring. Also called acquired mutations. *See also:* germ line genetic mutation (ORNL).

**Southern blotting** Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific base sequences by radiolabeled complementary probes (ORNL).

**Spectral karyotype (SKY)** A graphic of all an organism's chromosomes, each labeled with a different color. Useful for identifying chromosomal abnormalities. *See also:* chromosome (ORNL).

**Splice site** Location in the DNA sequence where RNA removes the noncoding areas to form a continuous gene transcript for translation into a protein (ORNL).

**Sporadic cancer** Cancer that occurs randomly and is not inherited from parents. Caused by DNA changes in one cell that grows and divides, spreading throughout the body. *See also:* hereditary cancer (ORNL).

**SSAHA** A hashing algorithm developed for rapid searching of large amounts of genome sequence. This program is similar to BLAT but does not use splice information to align mRNA sequences, nor can it perform translated searches (NCBI).

**Stem cell** Undifferentiated, primitive cells in the bone marrow that have the ability both to multiply and to differentiate into specific blood cells (ORNL).

**Structural genomics** The effort to determine the three-dimensional structures of large numbers of proteins using both experimental techniques and computer simulation (ORNL).

**STS (sequence tag site)** In general, short sequences (200–500 bp) are produced throughout a genome. Oligonucleotide primers are generated such that this sequence can be amplified using PCR to produce a discrete band when analyzed by electrophoresis. STS markers can be polymorphic or monomorphic. They are critical to integrating nonsequence-based maps (such as genetic or radiation hybrid) with sequence-based maps (NCBI).

**Substitution** (a) The presence of a nonidentical amino acid at a given position in an alignment. If the aligned residues have similar physicochemical properties, the substitution is said to be “conservative” (NCBI BLAST). (b) In genetics, a type of mutation due to replacement of one nucleotide in a DNA sequence by another nucleotide or replacement of one amino acid in a protein by another amino acid. *See also:* mutation (ORNL).

**Substitution matrix** A substitution matrix containing values proportional to the probability that amino acid  $i$  mutates into amino acid  $j$  for all pairs of amino acids. Such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occurring through a period of evolution (NCBI BLAST).

**Supercontig (scaffold)** A supercontig is formed when an association can be made between two contigs that have no sequence overlap. This commonly occurs using information obtained from paired plasmid ends. For example, both ends of a BAC clone are sequenced. It can be inferred that these two sequences are approximately 150–200 kb apart (based on the average size of a BAC). If the sequence from one end is found in a particular sequence contig, and the sequence from the other end is found in a different sequence contig, the two sequence contigs are said to be linked. In general, it is useful to have end sequences from more than one clone to provide evidence for linkage (NCBI).

**Suppressor gene** A gene that can suppress the action of another gene (ORNL).

**Syndrome** The group or recognizable pattern of symptoms or abnormalities that indicate a particular trait or disease (ORNL).

**Syngeneic** Genetically identical members of the same species (ORNL).

**Synteny** Genes occurring in the same order on chromosomes of different species. *See also:* linkage, conserved sequence (ORNL).

**T**

**Tandem repeat sequences** Multiple copies of the same base sequence on a chromosome; used as markers in physical mapping. *See also:* physical map (ORNL).

**Targeted mutagenesis** Deliberate change in the genetic structure directed at a specific site on the chromosome. Used in research to determine the targeted region's function. *See also:* mutation, polymorphism (ORNL).

**Technology transfer** The process of transferring scientific findings from research laboratories to the commercial sector (ORNL).

**Telomerase** The enzyme that directs the replication of telomeres (ORNL).

**Telomere** The end of a chromosome. This specialized structure is involved in the replication and stability of linear DNA molecules. *See also:* DNA replication (ORNL).

**Teratogenic** Substances such as chemicals or radiation that causes abnormal development of an embryo. *See also:* mutatgen (ORNL).

**Thymine (T)** A nitrogenous base, one member of the base pair AT (adenine–thymine). *See also:* base pair, nucleotide (ORNL).

**Toxicogenomics** The study of how genomes respond to environmental stressors or toxicants. Combines genome-wide mRNA expression profiling with protein expression patterns using bioinformatics to understand the role of gene–environment interactions in disease and dysfunction (ORNL).

**Transcription** The synthesis of an RNA copy from a sequence of DNA (a gene); the first step in gene expression. *See also:* translation (ORNL).

**Transcription factor** A protein that binds to regulatory regions and helps control gene expression (ORNL).

**Transcriptome** The full complement of activated genes, mRNAs, or transcripts in a particular tissue at a particular time (ORNL).

**Transfection** The introduction of foreign DNA into a host cell. *See also:* cloning vector, gene therapy (ORNL).

**Transfer RNA (tRNA)** A class of RNA having structures with triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. The role of tRNAs in protein synthesis is to bond with amino acids and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA (ORNL).

**Transformation** A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its genome (ORNL).

**Transgenic** An experimentally produced organism in which DNA has been artificially introduced and incorporated into the organism's germ line. *See also:* cell, DNA, gene, nucleus, germ line (ORNL).

**Translation** The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. *See also:* transcription (ORNL).

**Translocation** A mutation in which a large segment of one chromosome breaks off and attaches to another chromosome. *See also:* mutation (ORNL).

**Transposable element** A class of DNA sequences that can move from one chromosomal site to another (ORNL).

**Trisomy** Possessing three copies of a particular chromosome instead of the normal two copies. *See also:* cell, gene, gene expression, chromosome (ORNL).

**U**

**Unitary matrix** Also known as identity matrix. A scoring system in which only identical characters receive a positive score (NCBI BLAST).

**Up and down** The simplest topology for a helical bundle or folded leaf, in which consecutive helices are adjacent and antiparallel; it is approximately equivalent to the meander topology of a  $\beta$  sheet (SCOP).

**Uracil** A nitrogenous base normally found in RNA but not DNA; it is capable of forming a base pair with adenine. *See also:* base pair, nucleotide (ORNL).

## V

**Vector** *See:* cloning vector (ORNL).

**Virus** A noncellular biological entity that can reproduce only within a host cell. Viruses consist of nucleic acid covered by protein; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus. *See also:* cloning vector (ORNL).

## W

**Western blot** A technique used to identify and locate proteins based on their ability to bind to specific antibodies. *See also:* DNA, Northern blot, protein, RNA, Southern blotting (ORNL).

**Whole-genome shotgun sequencing (WGS)** A sequencing method by which an entire genome is cut into chunks of discrete sizes (usually 2,10, 50 and 150 kb) and cloned into an appropriate vector. The ends of these clones are sequenced. The two ends from the same clone are referred to as mate pairs. The distance between two mate pairs can be inferred if the library size is known and should have a narrow window of deviation (NCBI).

**Wildtype** The form of an organism that occurs most frequently in nature (ORNL).

**Working draft DNA sequence** *See:* Draft DNA sequence (ORNL).

## X

**X chromosome** One of the two sex chromosomes, X and Y. *See also:* Y chromosome, sex chromosome (ORNL).

**Xenograft** Tissue or organs from an individual of one species transplanted into or grafted onto an organism of another species, genus, or family. A common example is the use of pig heart valves in humans (ORNL).

## Y

**Y chromosome** One of the two sex chromosomes, X and Y. *See also:* X chromosome, sex chromosome (ORNL).

**Yeast artificial chromosome (YAC)** Constructed from yeast DNA, it is a vector used to clone large DNA fragments. *See also:* cloning vector, cosmid (ORNL).

## Z

**Zinc-finger protein** A secondary feature of some proteins containing a zinc atom; a DNA-binding protein (ORNL).



# Self-Test Quiz: Solutions

[2-1] e	[4-10] c	[8-8] d
[2-2] e		[8-9] a
[2-3] c	[5-1] b	[8-10] c
[2-4] a	[5-2] b	
[2-5] a	[5-3] c	[9-1] b
[2-6] a	[5-4] b	[9-2] c
[2-7] c	[5-5] a	[9-3] c
[2-8] d	[5-6] a	[9-4] d
[2-9] c	[5-7] a	[9-5] b
	[5-8] b	[9-6] b
[3-1] asparagine N glutamine Q tryptophan W tyrosine Y phenylalanine F	[5-9] d	[9-7] a [9-8] d [9-9] c
	[6-1] b	
	[6-2] b	
	[6-3] c	[10-1] a
[3-2] a	[6-4] d	[10-2] d
[3-3] d	[6-5] d	[10-3] c
[3-4] c	[6-6] a	[10-4] c
[3-5] d	[6-7] a	[10-5] d
[3-6] a	[6-8] c	[10-6] c
[3-7] c		[10-7] a
[3-8] false	[7-1] d	[10-8] b
[3-9] c	[7-2] b	[10-9] c
[3-10] d	[7-3] c	
	[7-4] a	[11-1] c
[4-1] d	[7-5] b	[11-2] d
[4-2] c	[7-6] a	[11-3] a
[4-3] a	[7-7] b	[11-4] b
[4-4] BLASTP d BLASTN a BLASTX c TBLASTN b TBLASTX e	[7-8] a [7-9] c [8-1] c [8-2] c	[11-5] d [11-6] d [11-7] a [11-8] d [11-9] a
[4-5] c	[8-3] b	
[4-6] a	[8-4] c	[12-1] a
[4-7] a	[8-5] d	[12-2] c
[4-8] b	[8-6] d	[12-3] b
[4-9] b	[8-7] a	[12-4] c

- |          |          |          |
|----------|----------|----------|
| [12-5] c | [15-6] d | [19-1] a |
| [12-6] b | [15-7] c | [19-2] d |
| [12-7] d | [15-8] c | [19-3] a |
| [12-8] b | [16-1] c | [19-4] d |
| [13-1] a | [16-2] b | [19-5] a |
| [13-2] c | [16-3] d | [19-6] b |
| [13-3] d | [16-4] c | [19-7] a |
| [13-4] c | [16-5] a | [19-8] c |
| [13-5] d | [16-6] d | [19-9] c |
| [13-6] c | [16-7] b | [20-1] c |
| [13-7] a | [16-8] d | [20-2] c |
| [13-8] b | [17-1] c | [20-3] a |
| [13-9] d | [17-2] c | [20-4] a |
| [14-1] d | [17-3] a | [20-5] b |
| [14-2] a | [17-4] c | [20-6] b |
| [14-3] b | [17-5] d | [20-7] d |
| [14-4] c | [17-6] a | [20-8] d |
| [14-5] a | [18-1] c | [20-9] d |
| [14-6] e | [18-2] c | [21-1] a |
| [14-7] d | [18-3] c | [21-2] a |
| [14-8] a | [18-4] b | [21-3] b |
| [14-9] c | [18-5] a | [21-4] c |
| [15-1] c | [18-6] a | [21-5] a |
| [15-2] a | [18-7] c | [21-6] c |
| [15-3] d | [18-8] b | [21-7] b |
| [15-4] d | [18-9] a | [21-8] d |
| [15-5] b |          | [21-9] a |

# Author Index

- Abecasis, Gonçalo, 1047  
Aebersold, Ruedi, 47  
Albert, Istvan, 15  
Allen, Jonathan, 339  
Altman, Russ, 3  
Altman, Sidney, 436  
Altschul, Stephen, 121, 162  
Altshuler, David, 1066  
Anfinsen, Christian, 68, 589  
Aristotle, 701, 708  
Avery, Oswald T., 433, 1022  
Axelrod, Julius, 1022
- Bairoch, Amos, 227  
Baker, David, 621, 624  
Baltimore, David, 759  
Bateman, Alex, 200, 472  
Baxevanis, Andy, 53, 719  
Beadle, George, 433, 874  
Bedel, Joseph, 162  
Beccari, Iacopo Bartolomeo, 18  
Beijerinck, Martinus, 754  
Berg, Paul, 248  
Bernal, John D., 754  
Bernardi, Giorgio, 969  
Berzelius, Jöns Jacob, 18, 539  
Bird, Adrian, 1055  
Birney, Ewan, 425  
Blakeslee, Albert, 634  
Blaser, Martin, 837  
Blattner, Frederick, 805  
Boeke, Jef, 657  
Bork, Peer, 224, 573, 704, 806, 837  
Botstein, David, 660  
Bourne, Philip, 628  
Bouton, Christopher, 528, 668  
Brazma, Alvis, 465  
Brenner, Steven, 116, 612  
Brenner, Sydney, 919  
Bridges, C. B., 645  
Brown, Donald, 353  
Brown, Patrick, 478, 660
- Bull, James, 704  
Bullard, James, 362  
Burnet, Sir MacFarlane, 755, 1032  
Byrne, Kevin, 867
- Cagniard-Latour, Baron Charles, 849  
Capecchi, Mario, 650  
Carter, Nigel, 1051  
Cech, Thomas, 436  
Chandonia, John-Marc, 612  
Chatton, Edouard, 702  
Childs, Barton, 1066  
Chothia, Cyrus, 116, 179, 602, 612  
Churchill, Gary, 181  
Claverie, Jean-Michel, 783  
Clint, 939  
Coghlan, Avril, 818  
Colantuoni, Carlo, 488  
Collado-Vides, Julio, 814  
Collins, Francis, 1066  
Cooper, David, 1039  
Corey, Robert, 594  
Crick, Francis, 433-434  
Crowfoot, Dorothy, 754  
Cummings, E.E., 635  
Cutting, Garry, 1066  
Cuvier, Baron Georges, 956
- Daly, Mark, 1066  
Darlington, Charles, 306  
Darwin, Charles, 72, 245, 254, 699  
Davis, Ron, 872  
Dayhoff, Margaret, 19, 31, 69, 79, 113,  
    116, 705, 739  
Denis, Prosper-Sylvain, 1010  
DePristo, Mark, 382  
Dickerson, Richard, 251  
Dobzhansky, Theodosius, 245  
Doolittle, R.F., 116, 205, 208, 238  
Doolittle, W. F., 322  
Dowell, Robin, 913  
Deutsch, Eric, 47
- Dujon, Bernard, 880  
Dunker, Keith, 622  
Durbin, Richard, 47  
Durinck, Steffen, 362  
Dyson, Jane, 622
- Eddy, Sean, 185, 322  
Edgar, Robert, 160, 218  
Edman, Pehr, 543  
Edwards, David, 745, 837  
Edwards, Robert, 727  
Eichler, Evan, 967  
Eisen, Jonathan, 809  
Eisen, Michael, 533  
Evans, Sir Martin, 650  
Eve, 985
- Feero, W. Gregory, 1066  
Felsenstein, Joseph, 282, 297  
Feng, Da-Fei, 205, 208, 238  
Fenn, John, 548  
Field, Dawn, 699  
Fire, Andrew, 447  
Firth, Helen, 1051  
Fisher, Emil, 591  
Fitch, Walter M., 71  
Flieck, Paul, 425  
Fox, Naomi, 612  
Franklin, Rosalind, 754  
Fraser, Claire, 837
- Garrod, Sir Archibald, 1014-1015  
Gerstein, Mark, 328  
Gilbert, Don, 272  
Gilbert, Walter, 248  
Gish, Warren, 121, 170  
Golub, Todd, 517  
Gordon, Jeffrey, 789  
Green, Eric, 16, 308, 719, 961  
Green, Phil, 326  
Gregory, T. Ryan, 311  
Griffith, Frederick, 433

- Gu, Jenny, 628  
 Guigó, Roderic, 307  
 Gumbel, Emil, 120, 142  
 Guttmacher, Alan, 1066  
 Guyer, Mark, 16
- Haeckel, Ernst, 246, 698  
 Hamosh, Ada, 1036  
 Happle, Rudolph, 1032  
 Hartwell, Leland, 875  
 Haussler, David, 181, 961  
 Henikoff, Jorja A., 79  
 Henikoff, Steven, 79  
 Henle, Jakob, 765  
 Heringa, Jaap, 220  
 Hermjakob, Henning, 47  
 Higgins, Desmond, 220  
 Hillis, David, 704  
 Hodgkin, Dorothy Crawford, 754  
 Hofmeister, Franz, 591  
 Holley, Robert, 435  
 Holt, Kathryn, 745, 837  
 Holmes, Edward, 766, 789  
 Hoppe-Seyler, Felix, 376  
 Horvitz, H. Robert, 919  
 Hubbard, Tim, 116, 612  
 Huebner, Robert, 765  
 Huelsenbeck, John, 282  
 Hunt, Timothy, 875
- Irizarry, Rafael, 488
- Johannsen, Wilhelm, 433  
 Jones, David, 613  
 Jongeneel, C.V., 116, 163  
 Junier, Thomas, 104  
 Jurka, Jerzy, 326  
 Justice, Monica, 665
- Kabsch, Wolfgang, 597  
 Karplus, Kevin, 161  
 Kendrew, John, 204, 250, 588, 590  
 Kent, Jim, 413  
 Khorana, Har, 435  
 Kimura, Motoo, 258  
 Kingsmore, Stephen, 1051  
 Kluyver, Albert Jan, 705  
 Knight, Rob, 837  
 Kobilka, Brian, 599, 600  
 Koch, Robert, 765  
 Kohara, Yuji, 928  
 Koonin, Eugene, 573, 704, 798, 805,  
     811, 944  
 Korf, Ian, 162, 740  
 Kornberg, Arthur, 433  
 Kossel, Albrecht, 376, 433
- Krogh, Anders, 181, 184, 820  
 Krzywinski, Martin, 902  
 Kumar, Sudhir, 70, 282, 298  
 Kyrides, Nikos, 710
- Lamarck, Jean-Baptiste, 702  
 Lancet, Doron, 1041  
 Lander, Eric, 517, 1066  
 Landsman, David, 53  
 Laveran, Charles Louis Alphonse, 895  
 Lederberg, Joshua, 874  
 Leek, Jeff, 480, 519  
 Leeuwenhoek, Anton von, 796, 849, 891  
 Leuckart, Karl, 917  
 Lefkowitz, Robert, 600  
 Levinthal, Cyrus, 598  
 Levitt, Michael, 628  
 Lewis, Edward B., 645, 919  
 Liebig, Justus, 6  
 Linnaeus, Carl, 702, 705  
 Lipman, David, 121  
 Lowe, Todd, 439, 440, 441  
 Lucretius, 635  
 Lucy, 707  
 Lupski, James R., 1000  
 Lynch, Michael, 944  
 Lyon, Gholson, 408, 1051
- Maddison, David R., 710  
 Mann, Gustav, 538  
 Manolio, Teri, 1066  
 Mao, Rong, 482  
 Marchionni, Luigi, 519  
 Margoliash, Emanuel, 250  
 Markov, Anrei Andreyevich, 181  
 Martens, Lennart, 47  
 Martin, William, 704, 798  
 Mayer, Adolf, 754  
 McClintock, Barbara, 325, 1022  
 McKusick, Victor A., 32, 1036  
 Mello, Craig, 447  
 Mesirov, Jill, 529  
 Meyer, Edgar F., 1997  
 Miescher, Johann Friedrich, 376, 379  
 Miller, Webb, 121, 191  
 Minden, Jonathan, 547  
 Morgan, Thomas Hunt, 351, 645, 919  
 Mulder, Gerardus Johannes, 6, 166, 539  
 Muller, Hermann J., 645, 665  
 Murzin, Alexey, 612  
 Myers, Gene, 121
- Nei, Masatoshi, 282, 354  
 Neuvéglise, Cécile, 853  
 Nirenberg, Marshall, 435  
 Noble, William, 44
- Notredame, Cédric, 220  
 Novick, Peter, 649  
 Nurse, Sir Paul, 875  
 Nüsslein-Volhard, Christiane, 645, 919  
 Nuttall, George, 244
- Ochoa, Severo, 433  
 Ohno, Susumu, 307, 348, 349, 861, 880  
 Oliver, Stephen, 847  
 Olsen, Gary, 433  
 Orengo, Christine, 613  
 Ott, Jürg, 1047  
 Owen, Richard, 72, 797
- Pääbo, Svante, 725-726  
 Pace, Norman, 702, 798  
 Pagni, Marco, 104, 116, 163  
 Pandey, Akhilesh, 542  
 Pauling, Linus, 113, 250, 594, 1011,  
     1022  
 Pearson, Thomas, 1066  
 Pearson, William, 104, 107, 115, 162  
 Perutz, Max, 123, 250, 588, 590  
 Pevzner, Pavel, 351, 377, 396  
 Piccard, Jules, 376  
 Posada, David, 280  
 Prusiner, Stanley, 760  
 Purcell, Shaun, 1047
- Quinlan, Aaron, 414
- Rambaut, Andrew, 293  
 Reese, Martin, 307  
 Roberson, Eli, 992  
 Roberts, Richard J., 451  
 Ronquist, Fredrik, 282  
 Ross, Ronald, 895  
 Rothman, James E., 643  
 Rubin, Edward, 837  
 Ruska, Helmut, 754
- Salzberg, Steven, 47, 339, 400, 783, 821  
 Sander, Christian, 597  
 Sanger, Frederick, 68, 248, 379, 543,  
     711  
 Schekman, Randy, 643, 649, 877  
 Schmidt, Heiko, 282, 290  
 Schuster-Böckler, Benjamin, 200  
 Schwann, Theodor, 849  
 Searls, David, 13  
 Sharp, Phillip A., 451  
 Shirley, Matt, 1032  
 Simpson, George Gaylord, 757  
 Smit, Arian, 326  
 Smith, Hamilton O., 732  
 Smithies, Oliver, 650

- Smyth, Gordon, 508  
Snyder, Michael, 382  
Speed, Terry, 488  
Spiegelman, Sol, 478  
Sprinzl, Mathias, 441  
Stein, Lincoln, 913  
Strimmer, Korbinian, 282, 290  
Sturtevant, A. H., 351, 645  
Südhoff, Thomas, 643  
Sueoka, Noboru, 737  
Sulston, John E., 919  
Swofford, David, 282  
  
Tamura, Koichiro, 282, 298  
Tanaka, Koichi, 548  
Tatum, Edward, 433, 874  
Tatusova, Tatiana, 758
- Taubenberger, Jeffery, 774  
Thornton, Janet, 613  
Treangen, Todd, 400  
Trent, Jeffrey, 478  
Tringe, Susannah, 837  
Tumpey, Terrence, 774  
Tuppy, Hans, 543  
  
Valle, David, 1066  
Vassilenko, Konstantin, 441  
Venter, J. Craig, 718, 723, 732, 764, 998,  
    1001  
Vogelstein, Bert, 1034  
von Haeseler, Arndt, 282, 290  
  
Wagner, Andreas, 864  
Wang, Jun, 837
- Washburn, Michael, 539  
Watson, James, 204, 433-434, 998,  
    1000  
Weismann, August, 637  
White, Owen, 821  
Wieschaus, Eric F., 645, 919  
Wilkins, Maurice, 434, 754  
Woese, Carl, 433, 702, 809  
Wolfe, Kenneth, 862-863, 867  
Wright, Peter, 622  
Wu, Ray, 379  
Wu, Zhijin, 488  
  
Yandell, Mark, 162, 420  
  
Zhogbi, Huda, 638  
Zuckerkandl, Emil, 113, 250, 1011



# Subject Index

- Ab initio* prediction, 621  
ABI SOLiD, 385  
Accepted point mutations (PAM), 79, 80, 81  
alternatives to, 91–94  
mutation probability matrix, 82–84  
PAM250 and other PAM matrices, 84–88  
practical usefulness of PAM matrices, 91  
Accession numbers, 34–36  
Accuracy, definition, 490  
Acquired Immune Deficiency Syndrome (AIDS), 766–770  
Acrocentric chromosomes, 348  
Adenine, 435  
Adenosine 5' phosphosulfate (APS), 385  
Adenosine triphosphate (ATP), 564  
Advanced database searching, 167–168  
advice for students, 198  
BLAST-like alignment tools for rapid DNA searches, 186–194  
domain enhanced lookup time accelerated BLAST (DELTA-BLAST), 177–178, 197–198, 226  
finding distantly related proteins, 171–181  
hidden Markov models (HMMs), 181–186  
next-generation sequencing (NGS) alignment to reference genome, 194–197  
organism-specific BLAST sites, 168–170  
pattern-hit initiated BLAST (PHI-BLAST), 179–181  
perspective, 197  
pitfalls, 197–198  
position-specific iterated BLAST (PSI-BLAST), 171–177, 197–198  
profile searches, 181–186  
specialized BLAST sites, 168–171  
specialized BLAST-related algorithms, 170–171  
web resources, 198  
Affinity chromatography, 673, 675–676  
Agglomerative hierarchical clustering, 512–513  
Alanine (Ala), 76  
Algorithms, 77  
Alkaptonuria, 1014  
Allele frequencies, 1020–1021  
*Allium cepa*, 311  
Alzheimer's disease, 624  
Amino acids abbreviations, 76  
codons of DNA, 83  
conformational preferences, 597  
corrected number of changes per 100 residues, 252  
frequency of, 79, 81  
peptide bonds, 593  
phylogenetic approach to alignment, 80  
polypeptides, 591–594  
relative mutability, 80–82  
step matrix, 270, 271  
structures, 76  
substitution, 272–281  
substitution in disease, 1022, 1035, 1045–1046, 1060, 1064  
substitution rates, 253  
*Amoeba dubia*, 311  
*Amoeba proteus*, 311  
*Amphiuma means*, 311  
aneuploidy, 1025  
ANNOVAR program, 419  
*Anopheles gambiae*, 921–922  
Apicomplexans  
    *Babesia bovis*, 898  
    *Cryptosporidium hominis*, 898  
    *Hammondia hammondi*, 899  
    *Theileria annulata*, 898  
    *Toxoplasma gondii*, 898–899  
*Arabidopsis thaliana*, 28, 643, 910–913  
ARB project, 444  
Archaea, 7, 24, 700 classification, 798–811  
genome size and geometry, 801–805  
genomes, 797–798  
lifestyle, 805–808  
molecular sequences, 810–811  
phylogenetic diversity, 810  
protein-encoding genes, 804  
ribosomal RNA sequences, 809–810  
Archaeal genome analysis, 814–830 challenges, 825  
finding genes, 819–825  
gene annotation, 825–827  
lateral gene transfer (LGT), 827–830  
nucleotide composition, 817–819  
Archaeal genome comparisons species with closely related strains determined, 831  
Arginine (Arg), 76  
Array comparative genomic hybridization (aCGH), 356, 357  
Arthropods, 920  
Asexual reproduction, 309  
*Ascomycetes*, 869  
Asparagine (Asn), 76  
Aspartic acid (Asp), 76  
*Aspergillus*, 871  
Association of Biomolecular Resource Facilities (ABRF), 542–543, 551, 565  
Autism, 1024–1025  
*Babesia bovis*, 898  
Bacteria, 7, 24, 700 classification, 798–811  
genome size and geometry, 801–805  
genomes, 797–798  
human disease relevance, 808–809  
lifestyle, 805–808  
molecular sequences, 810–811

- Bacteria (*continued*)  
 morphological criteria, 800–801  
 phylogenetic diversity, 810  
 protein-encoding genes, 804  
 ribosomal RNA sequences, 809–810  
 vaccine-preventable disease, 808  
 Bacterial artificial chromosome (BAC), 355, 398  
 Bacterial genome analysis, 814–830  
   challenges, 825  
   finding genes, 819–825  
   gene annotation, 825–827  
   lateral gene transfer (LGT), 827–830  
   nucleotide composition, 817–819  
 Bacterial genome comparisons, 830  
   advice for students, 835  
   MUMmer, 833–834  
   perspective, 834–835  
   pitfalls, 835  
   species with closely related strains  
   determined, 831  
 TaxPlot, 830–833  
   web resources, 835  
 Bacteriophages, 711–712  
 Basic Local Alignment Search Tool (BLAST), 31–32, 103–104  
   advice for students, 160  
   algorithm parts, 138–141  
   algorithm schematic, 139  
   bit scores, 143  
   dotplots, 104–105  
   E value, 142–143  
   E values and *p* values, 143–144  
   extreme value distribution, 142  
   formatting parameters, 132–134  
   gene discovery, 155–159  
   gene identification, 338  
   graphic summary, 134, 149, 153, 156  
   handling too few results, 150–151  
   handling too many results, 150  
   human beta globin example, 136–138  
   introduction, 121–123  
   Karlin–Altschul statistics, 142  
   list of alignments, 147, 149, 153  
   local alignment search strategy, 138–144  
   multidomain protein search, 151–155  
   optional search parameters, 127–131  
   organism-specific sites, 168–170  
   output, 134  
   *p* value, 143–144  
   pairwise alignments, 147, 149  
   perspective, 159  
   pitfalls, 160  
   raw scores, 143  
   result significance, 146–150  
   search concepts, 145  
   search overview, 151  
   search principles, 146–151  
   search statistics, 141–142  
   search steps, 124–138  
   search strategies, 145–155  
   search summary, 133  
   selecting appropriate program, 124–126  
   selecting database, 126–127  
   specialized sites, 168–171  
   specifying sequence of interest, 124  
   stand-alone version, 135–138  
   taxonomy report, 154  
   TBLASTN tool, 70  
   threshold value, effect of, 140  
   web resources, 160  
*Basidiomycetes*, 870  
 Bayesian inference tree-building  
   method, 290–293, 294  
 BED files, 56–57, 327, 361, 414–416, 424  
 BEDtools, 413–417  
 Beta globin gene (*HBB*) *see* Globins  
 Bifurcating rooted trees, 263  
 Binomial coefficient, 289  
 BioConductor, 11  
 BioGrid network map, 677  
 Bioinformatics  
   definitions, 3–4  
   other informatics disciplines, 15  
   overview, 5–8  
   software *see* Software for bioinformatics  
 Biological databases *see* Databases  
 BioMart project, 20, 54, 314–320  
 Birth-and-death evolution model, 353, 354  
 Bit score, 75, 91–92, 132, 134, 137, 143, 175, 438  
 BLAST-like alignment tools for rapid DNA searches, 186–194  
 benchmarking to assess genomic alignment performance, 187–188  
 BLAST-like tool (BLAT), 192, , 193  
 BLASTZ, 188–191  
 discontinuous MegaBLAST, 191  
 Enredo, 191  
 Limited Area Global Alignment of Nucleotides (LAGAN), 192–194  
 MegaBLAST, 191–192  
 PatternHunter, 188, 189  
 Pecan, 191  
 Sequence Search and Alignment by Hashing Algorithm (SSAHA2), 194  
 BLAST-like tool (BLAT), 192, 193  
 BLAST-related algorithms, 170  
   *see also* Domain enhanced lookup  
   time accelerated BLAST (DELTA-BLAST); Pattern-hit initiated BLAST (PHI-BLAST); Position-specific iterated BLAST (PSI-BLAST)  
 European Bioinformatics Institute (EBI), 170  
 National Center for Biotechnology Information (NCBI), 170  
 next-generation sequencing (NGS) data, 170–171  
 WU-BLAST 2.0, 170  
 BLASTZ, 188–191  
 Block substitution matrix (BLOSUM), 91–94  
*Boa constrictor*, 311  
 BodyMap-2, 459  
*Bos taurus*, 25, 28, 718, 736, 744, 934  
 Bowtie-2 program, 196–197  
 Branches of phylogenetic trees, 259–262  
 Burrows–Wheeler Transform (BWT)  
   alignment, 196–197  
 Butterflies, 922–923  
 BWA program, 196–197  
 C value, 310–312, 322–323  
   protein-coding gene paradox, 342  
*Caenorhabditis elegans*, 311, 643–644, 918–919  
 Cancer, 718, 721, 760, 762, 1033  
*Candida albicans*, 175, 871–872  
*Canis familiaris*, 311, 934  
 Cap analysis gene expression (CAGE), 467  
*Carcharias obscurus*, 311  
 Caseins, 85, 86  
 Catalogue of somatic mutations in cancer (COSMIC), 1033  
 CATH database, 613–615  
   search results, 615  
 CCAAT box, 336  
 CEL files, 496  
   CEL definition file (CDF), 505  
   input, 506  
 Central bioinformatics resources, 31–34  
 Centroid axes, 500  
 Chagas’ disease, 892  
 Charge coupled device (CCD) camera, 385  
 Chemical mutagenesis, 665  
 Chi-squared analysis, 256  
*Chlorophyta*, 908–910  
 Chloroplast genomes, 714–715  
 Chloroplasts, 703, 714–715, 723, 743, 902

- Chromalveolates: *Plasmodium falciparum*, 895–898
- Chromatin diminution, 349
- Chromatin immunoprecipitation sequencing (ChIP-seq), 345
- Chromosomal abnormalities, 349, 351–353, 355–359, 1050–1051
- Chromosomal aneuploidy frequency, 1025
- Chromosome territories, 314
- Chromosomes
- analysis by ENCODE project, 320–321, 323–324
  - analysis in genome browsers, 314
  - analysis using BioMart and biomaRt, 314–317
  - deletions, 349, 352
  - duplications, 351, 353
  - fission, 348
  - fusion, 348
  - gene density, 336
  - gene family models, 353–354
  - human, 313, 959–963, 979–986
  - inactivation, 349
  - inversions, 351
  - measuring change, 355–359
  - single nucleotide polymorphisms (SNPs), 354–355
  - translocation, 348
  - yeast, 854–860
- Ciliophora, 899
- Paramecium tetraurelia*, 899–901
  - Tetrahymena thermophila*, 902
- Ciona intestinalis*, 925–926
- Circular binary segmentation (CBS) method, 356
- Cis-regulatory modules (CRMs), 342
- Clades, 262
- Cladograms, 262
- Classical structure biology, 601
- ClinVar database, 1041
- ClustalW program, 208, 209–214
- Clusters
- agglomerative hierarchical clustering, 512–513
  - divisive hierarchical clustering, 512–513
  - partition methods, 516–517
  - relatedness, 515
  - self-organizing maps, 517
- Clusters of Orthologous Groups (COGs) database, 573
- Escherichia coli*, 639
- Coding sequence (CDS), 40
- Codons of DNA, 83
- COILS program, 559, 563
- Co-immunoprecipitation, 672
- Column score (CS), 223
- Command line software tools, 10, 11–12, 317–320, 824
- NCBI access, 42–49
- Comparative genome hybridization (CGH), 355
- Comparative genome hybridization, array (aCGH), 356, 357
- Comparative modeling, 618–619
- Complete Genomics self-assembly DNA nanoarrays, 387
- Complex disorders, 1024–1025
- Concerted evolution model, 353–354
- Confidentiality, 390, 1044
- Consensus Coding Sequence (CCDS) project, 37
- Consent, informed, 390
- Conservative substitutions, 74
- Conserved Domain Database (CDD), 177, 226
- Consistency-based alignment, 218–220
- Contigs, 734
- mapping, 734–735
- Copy number variation (CNVs), 356, 1029
- Coscinodiscus asteromphalus*, 311
- Covariance models, 439, 441
- CpG islands, 342
- CRAM file format, 406–408
- Creutzfeldt–Jakob disease, 624
- Critical Assessment of Genome Interpretation (CAGI), 668
- Critical Assessment of Protein Function Annotation (CAFA), 672
- Critical Assessment of Techniques for Protein Structure Prediction (CASP), 621–622, 623
- Cross-linking of proteins, 673
- Cryptococcus neoformans*, 872–873
  - Cryptosporidium hominis*, 898
- CuffLinks sample protocol, 523–524
- assembling transcripts, 525
  - differential expression, 525–526
- CummeRbund sample protocol, 526–527, 528
- Curation of databases, 227
- Cyclic reversible termination, 382–384
- Cysteine (Cys), 76
- Cystic fibrosis, 622–624
- Cystic fibrosis transmembrane regulator (CFTR), 624
- Cytosine, 435
- Dali Domain Dictionary, 615–616
- search results, 616
- Danio rerio*, 25, 28, 645, 926, 928
- Data ownership, 390
- Data storage, 421
- Database Referencing of Array Genes Online (DRAGON) database, 528–529
- Databases, 19–20
- access examples, 52–54
  - access to information, 34–37
  - access via gene resource at NCBI, 38–42
  - accessing large data sets, 54–58
  - advice for students, 60
  - central bioinformatics resources, 31–34
  - centralized DNA sequences, 20–24
  - Conserved Domain Database (CDD), 226
  - contents of DNA, RNA, and protein databases, 24–31
  - curation, manual versus automated, 227
  - genome browsers, 49–52
  - genomic regulatory factors, 342–345
  - integrated multiple sequence alignment resources, 226–227
  - literature access, 58–59
  - multiple sequence alignments, 222–227
  - perspective, 59–60
  - pitfalls, 60
  - protein databases, 29–31
  - Protein Family database of profile HMMs (Pfam), 223–224, 225, 226
  - RNA databases, 27–29
  - Simple Modular Architecture Research Tool (SMART), 224
  - web resources, 60
- Dayhoff Model of protein scoring, 79
- step 1 – accepted point mutations (PAM), 79, 80, 81
  - step 2 – frequency of amino acids, 79, 81
  - step 3 – relative mutability of amino acids, 80–82
  - step 4 – mutation probability matrix for 1 PAM evolutionary distance, 82–84
  - step 5 – PAM250 and other PAM matrices, 84–88
  - step 6 – relatedness odds matrix, 88
  - step 7 – log-odds scoring matrix, 89–91
- Dayhoff's protein superfamilies, 77
- DbEST database, 455

- De Bruijn graphs, 395–396, 398  
 DeepView software, 593, 594  
 Degree of divergence, 274  
 Deletions, 349, 409–410  
   chromosomes, 352  
   mutations, 78  
 Descriptive statistics for microarray data  
   analysis, 511  
   classification of genes or samples,  
 517–519, 520  
   clustering strategies, 517  
   confusion matrix, 521  
   data visualization methods, 518  
   hierarchical cluster analysis, 511–516  
   *k*-means clustering, 516–517  
   multidimensional scaling (MDS)  
     compared with principal component  
       analysis (PCA), 517  
   self-organizing maps (SOMs), 517, 518  
*Dictyostelium discoideum*, 916–917  
 Dideoxynucleotide sequencing, 19, 380  
 Difference gel electrophoresis (DIGE), 547  
 Digital Differential Display (DDD) tool,  
   456, 457  
 Diploid cells, 309, 348  
 Direct protein sequencing, 543  
   Edman degradation, 544  
 Disability-adjusted life years (DALYs),  
   1016–1017  
 Discontinuous MegaBLAST, 191  
 Distance measures compared with  
   similarity measures, 213  
 Divergent evolution model, 353  
 Divisive hierarchical clustering, 512–513  
 DNA  
   chromosomal variation, 347–355  
   codons, 83  
   functional genomics, 668  
   interspersed repeats, 325–326  
   noncoding and repetitive sequences,  
 323–325, 327  
   processed pseudogenes, 326–331  
   rDNA, 444  
   segmental duplications, 331–333  
   simple sequence repeats, 331  
   structure, 434  
   tandemly repeated sequences, 333–334  
   transcription, 434  
   transposon-derived repeats, 325–326  
 DNA Database of Japan (DDBJ), 21, 22  
   data types, 26–27  
   organisms, 24–25  
 DNA next-generation sequencing (NGS)  
   analysis, 387–421  
   experimental design, 389–390  
   FASTQ files, 391–394  
   genome assembly, 394–398  
   interpretation biological significance of  
    variants, 417–421  
   overview, 387–389  
   SAM/BAM format files and SAMtools,  
 402–408  
   sample preparation, 390  
   sequence generation, 390–391  
   storing data, 421  
   variant calling, 408–410  
   VCF format and VCFtools, 410–413  
   visualizing and tabulating data,  
 413–417  
   workflow chart, 388, 389  
 DNA next-generation sequencing (NGS)  
   technologies, 379, 382  
 ABI SOLiD, 385  
   compared with Sanger sequencing, 382  
 Complete Genomics self-assembly  
   DNA nanoarrays, 387  
   cyclic reversible termination, 382–384  
   decline in costs, 383  
   Illumina, 382–384  
   Ion Torrent, 387  
   ligation sequencing, 385  
   Pacific Biosciences DNA sequencing,  
 387  
   pyrosequencing, 384–385, 386  
 DNA sequence databases, 20–24  
   content, 24–31  
   data types, 26–27  
   genomic DNA, 27  
   growth, 22  
   range of file sizes, 23–24  
   scales of basepairs, 23  
 DNA trees, 268–270  
 DNase I, 345  
 Documentation, 14  
 Domain enhanced lookup time  
   accelerated BLAST (DELTA-  
     BLAST), 177–178, 226  
   assessing performance, 179  
   pitfalls, 197–198  
 Domains, 552–559  
   definition, 553, 554  
   *Homo sapiens*, 554  
   methyl-binding domains, 557  
   multidomain proteins, 556–557  
   yeast, 853  
 Dotplots, 104–105, 783–785, 832–834  
 Down syndrome, 313, 349, 480, 1020,  
   1025  
*Drosophila* genome database (FlyBase),  
   566  
*Drosophila melanogaster*, 311, 645,  
   919–921  
 Duplication of chromosomes, 351, 353  
*Dysidea crawshagi*, 311  
*E* value, 142–143  
   relation to *p* values, 143–144  
 Ebola virus, 775–776  
 EcoCyc database, 825–826  
 Edge branch of phylogenetic trees,  
   259–262  
 EDirect, 21, 42, 44–49, 137, 160, 236,  
   541, 578, 744, 878, 1002, 1018–1019  
 Edman degradation, 544  
 ELANDv2 program, 196  
 Electron crystallography, 600  
*Encephalitozoon cuniculi*, 873  
 ENCODE Genome Annotation  
   Assessment Project (EGASP),  
 339–340  
 ENCODE project, 447, 467  
   chromosome analysis, 320–321  
   critiques of, 322–323  
   functional elements catalog, 323  
   gene definition, 335–336  
 Enredo program, 191, 231, 232  
 Ensembl, 11, 32–33, 34, 50–52  
   BioMart, 316  
   chromosomes, 314, 315  
   genomic multiple sequence alignments,  
 231, 232  
   Human Genome Project, 959–961  
 Ensembl BLAST, 168–170  
 Ensembl Genomes, 710  
 Entrez, 20, 31  
   command-line access, 45, 46  
   usage tips, 32  
 Environmentally caused disease, 1029  
 Enzyme Commission (EC) numbers, 38,  
   567, 573–574  
 Epicellular bacteria, 806  
 Epstein–Barr virus (EBV), 762  
 Equilibrium dialysis, 673  
 Errors  
   DNA alignment, 400  
   DNA sequencing error rates, 379, 382,  
 385, 391, 393  
   genome annotation, 826–827  
   genome assembly, 397–398  
   Microsoft Excel, 42, 485  
   protein structure modeling, 619  
   spelling, 827  
   variant calling, 404, 420  
   web-based versus command-line  
 analysis, 320  
*Erysiphe cichoracearum*, 311  
*Escherichia coli*, 639–640  
   genome, 715–716

- phylogenetic relationships of strains, 815
- Ethical considerations, 390
- N*-Ethyl-*N*-nitrosourea (ENU), 665
- Euclidean distance, 513
- Euglenozoa, 893
- Eukaryotes, 7, 24, 700
- first chromosome, 715
  - first genome, 715
  - first organellar genome, 712–714
  - gene annotation, 738–742
  - genomes, 228
  - origins, 704
  - phylogeny, 888
  - ribosomal DNA, 444
- Eukaryotic chromosome, 308
- advice to students, 360
  - algorithms for finding genes, 338
  - chromosomal variation in individual genomes, 349–355
  - comparison of eukaryotic DNA, 346–347
  - deletions, 352
  - differences compared with Bacteria and Archaea, 308–309
  - duplications, 351, 353
  - dynamic nature, 347–349
  - EGASP competition, 339–340
  - features, 309, 310–323
  - gene content, 334–342
  - gene family models, 353–354
  - genes, finding, 336–339
  - genomes, 310
  - inversions, 351
  - measuring change, 355–359
  - organization, 312–314
  - perspective, 359
  - pitfalls, 359–360
  - protein-coding gene paradox, 342
  - protein-coding gene study resources, 340–342
  - regulatory regions, 342–346
  - repetitive DNA content, 323–324
  - single nucleotide polymorphisms (SNPs), 354–355
  - structural variations, 351–354
  - variation in chromosomal DNA, 347–355
  - web resources, 360
  - whole-genome duplication, 347–349
- Eukaryotic genomes
- advice for students, 941
  - bioinformatics, 889
  - biological principles, 889
  - cataloguing information, 888–889
  - Chromalveolates, 895–906
- comparative genomics, 889
- databases, 913
- introduction, 887–889
- metazoans, 916–940
- perspective, 940–941
- pitfalls, 941
- plant genomes, 906–916
- protozoans, 890–892
- sequence analysis, 889
- unicellular pathogens, 892–895
- web resources, 942
- European Bioinformatics Institute (EBI), 11, 21, 32–34
- BLAST-related algorithms, 170
  - sequence similarity search tools, 171
- European Molecular Biology Laboratory (EMBL)-Bank Nucleotide Sequence Database (EMBL-BANK), 20–21, 22, 32
- data types, 26–27
  - organisms, 24–25
- European Molecular Biology Open Software Suite (EMBOSS), 100
- European Nucleotide Archive (ENA), 14, 393, 465
- Evolution of life, 78–79
- Evolutionary distance, 82–84
- Evolutionary divergence between sequences, 275
- Exons, 336, 337
- Expert Protein Analysis System (ExPASy), 546, 562
- Expressed Sequence Tags (ESTs), 28, 122, 337, 454, 455
- Expression quantitative trait loci (eQTLs), 459, 468, 469
- Expresso program, 220–221
- Extracellular bacteria, 806
- Extreme value distribution, 120, 141–142
- Extremophilic microbes, 806
- Facultatively intracellular bacteria, 806
- False discovery rate (FDR), 495, 502
- FASTA format, 40–41, 73, 169, 381
- FASTA local alignment algorithm, 103–104
- FASTG files, 394
- FASTQ, 733–734
- FASTQ files, 391–392
- finding files, 392–393
  - format, 387
  - format conversion, 392
  - Illumina 1.3+ format, 391
  - quality assessment, 393–394
  - quality scores, 392
  - Sanger format, 391
- Solexa format, 391
- viewing files, 393
- Feng–Doolittle method for progressive sequence alignment, 208–209, 210, 214, 218
- File formats, 57
- Find-a-gene project, 156–159
- Fisher's exact test, 457, 458
- Fission of chromosomes, 348
- Fluorescence *in situ* hybridization (FISH), 355
- Fluorescence resonance energy transfer (FRET), 673
- Fold recognition, 619–620
- Forbidden mutations, 863–864
- Forward genetics, 649
- chemical mutagenesis, 665
  - compared with reverse genetics, 665–666
- Fragile sites, 349
- Frequency of amino acids, 79, 81
- Frogs, 929
- Functional annotation of microarray data, 528–529
- Functional Annotation of the Mouse (FANTOM) project, 459
- Functional genomics, 5
- advice for students, 686
  - approaches to function, 666, 667
  - bioinformatics resources, 682–685
  - central dogma, 666–670
  - definitions of function, 666–667
  - DNA, 668
  - forward and reverse proteomics, 671
  - integrating information, 668
  - introduction, 635–638
  - model organisms, 638–647
  - pairwise interactions and protein networks, 678–682
  - perspective, 685
  - pitfalls, 686
  - protein–protein interactions, 672–678
  - proteins, 672
  - reverse and forward genetics, 648–666
  - RNA, 668–670
  - web resources, 686
- Fungal genome analysis, 869–870
- Aspergillus*, 871
  - Candida albicans*, 871–872
  - Cryptococcus neoformans*, 872–873
  - Encephalitozoon cuniculi*, 873
  - Neurospora crassa*, 873–874
  - other genomes, 876
  - Phanerochaete chrysosporium*, 875
  - projects using *Ascomycetes*, 869
  - projects using *Basidiomycetes*, 870
  - Saccharomyces pombe*, 875–876

- Fungal genomes, 847–848  
advice for students, 877  
perspective, 876  
pitfalls, 877  
web resources, 877
- Fungi, 24  
description and classification, 848–849  
pathogens, 876  
taxonomy, 849
- Fusion of chromosomes, 348
- G protein-coupled receptor (GPCR), 599, 624
- Galaxy program: genomic multiple sequence alignments, 229–231
- Galaxy program, 20, 57–58
- Galius domesticus*, 311
- Gamma distribution, 278
- Gaussian distribution, 107
- GCRMA normalization, 488–489
- GenBank, 11, 20, 22, 52–53  
data types, 26–27  
organisms, 24–25
- GENCODE project, 328–329, 339, 340–342, 447
- Gene discovery, 155–159
- Gene expression, 5–6, 450–466
- Gene Expression Omnibus (GEO), 14, 465
- Gene Ontology (GO) Consortium, 551, 552, 566–570  
controlled vocabulary, 569  
evidence codes, 568  
participating databases, 567
- Gene Set Enrichment Analysis (GSEA), 529
- Gene silencing, 662–664
- Gene symbols, official, 27, 317–320
- Gene trapping, 657–660  
strategies, 659
- Gene trees, 266–268
- GeneCards, 1041
- General linear models, 504
- Generic Model Organism Project (GMOD), 913
- Genes  
classification, 517–519  
definition, 334–336  
eukaryotic chromosomes, 334–342  
identification, 335  
models for creating families, 353–354  
single nucleotide polymorphisms (SNPs), 354–355
- Genetic code, 83
- Genetic footprinting, 661
- Genome, 5
- Genome analysis projects  
advice for students, 743  
ancient DNA, 725–727  
annotation, 737–738  
annotation in eukaryotes, 738–742  
applications of sequencing, 720  
assembly, 730–733  
challenges, 735–737  
FASTQ, 733–734  
HTGS archive, 730  
introduction, 720–727  
large-scale projects, 721–722  
mapping contigs, 734–735  
metagenomics projects, 727  
perspective, 742  
pitfalls, 742–743  
resequencing projects, 725  
role of comparative genomics, 724–725  
selection criteria, 722–724  
sequence completed, 735  
sequencing centers, 728  
trace archive, 728–730
- Genome Analysis Toolkit (GATK), 388, 401, 403–404
- Genome Assembly Gold-standard Evaluation (GAGE), 397
- Genome assembly using NGS data, 394–396  
completion standards, 398  
performance evaluation, 396–398  
software, 395
- Genome browsers, 49–52
- Genome Portal of DOE, 710
- Genome Reference Consortium (GRC), 49, 395
- Genome sequence alignment, 399–400  
Genome Analysis Toolkit (GATK), 401  
repetitive DNA, 400–401  
software, 399, 1039–1040
- Genome Survey Sequences (GSSs), 27
- Genome Workbench software, 417, 418, 855–856, 1039–1040
- Genomes On Line Database (GOLD), 710
- Genomes  
Archaea, 797–798  
Bacteria, 797–798  
bioinformatics aspects, 701  
biological principles, 701  
cataloguing comparative genomic information, 701  
cataloguing genomic information, 701  
chromosomal variation in individual genomes, 349–355  
compared, 310
- currently sequenced, 709  
features of eukaryotic genomes, 310–323  
human disease relevance, 701  
introduction, 700–710  
life on Earth, 705  
molecular sequences, 705–709  
multicellular organism, 716  
organization of eukaryotic chromosomes, 312–314  
size variation, 310–312  
sizes, 309–310, 801  
systematics, 701–704  
taxonomy, 709–710  
web resources, 710  
whole-genome duplication, 347–349
- Genome-sequencing projects, 711  
1976–1978, 711–712  
1981, 712–714  
1986, 714–715  
1992, 715  
1996, 715  
1997, 715–716  
1998, 716  
1999, 716  
2000, 716  
2001, 716–717  
2002, 717  
2003, 717  
2004, 717–718  
2005, 718  
2006, 718  
2007, 718  
2008, 718  
2009, 718  
2011, 719  
2012, 719  
2013, 719  
2014, 719  
2015, 720  
chronology, 711
- Genome-wide association studies (GWAS), 468, 643, 1047–1050, 1052
- Genomic disorders, 1025–1028  
chromosomal aneuploidy frequency, 1025  
molecular mechanisms, 1028
- Genomic DNA databases, 27, 227–234  
assessment of whole-genome alignment methods, 231–234
- Ensembl program, 231, 232
- Galaxy program, 229–231
- UCSC Genome Bowser, 229, 230
- Genomic Evolutionary Rate Profiling (GERP), 229

- Genomic multiple sequence alignments, 227–234  
assessment of whole-genome alignment methods, 231–234  
Ensembl program, 231, 232  
Galaxy program, 229–231  
UCSC Genome Bowser, 229, 230
- Genomic promoter regions software, 343
- Genomic regulatory factors  
databases, 342–345  
nonconserved elements, 346  
ultraconserved elements, 345
- Genotype and phenotype, 637–638
- Genotype-Tissue Expression (GTEx) project, 459
- GENSCAN program, 338
- GEO dataset analysis using R  
analyses, 504–505  
CEL file input, 506  
identifying differentially expressed genes, 508–510  
microarray analysis and reproducibility, 510–511  
microarray analysis plots, 507  
RMA normalization, 506–508
- GEO2R resource at NCBI, 482  
corrections for multiple comparisons, 494–495  
data normalization, 486–488  
R scripts, 482–485  
robust multiarray analysis (RMA)  
normalization, 488–490  
statistical tests, 490–494
- Giant viruses, 782–783
- Giardia lamblia*, 891–892, 942
- Gibbs free energy, 590
- GLEAN gene model combiner, 339
- GLIMMER gene finding algorithm, 820–825
- Global alignment algorithms, 96  
Needleman and Wunsch algorithm, 96–100  
statistical significance, 106–108  
websites and URLs, 110
- Globins, 6–8  
*see also* Myoglobin  
beta globin gene, 20, 650–653  
beta globin gene mutants, 654  
beta globin BLAST example, 136–138
- Dayhoff subfamilies, 248  
hemoglobin, 6, 7  
phylogeny tree, 247  
structural biology, 591  
substitution frequencies, 111
- Glutamic acid (Glu), 76
- Glutamine (Gln), 76
- Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), 85
- Glycine max*, 28, 910, 915
- Green algae, 908–910
- Green fluorescent protein (GFP), 645
- Guanine, 435
- GXW motif, 590
- Haemophilus influenzae*, 19
- Hamming distance, 274, 277
- Hammonia hammondi*, 899
- Haploid cells, 309, 348  
*C* value, 310–312
- HaplotypeCaller, 408
- HapMap browser, 988
- HapMap project, 354, 717
- Hash tables, alignment based upon, 194–196
- Hemiascomycetes, 865  
functional element identification, 868–869  
whole-genome duplication, 866–868
- Hemoglobin *see* Globins
- Hepatitis A, 761
- Hepatitis B virus (HBV), 761, 762
- Hepatitis C virus (HCV), 258, 762
- Herpesvirus, 776–780
- Heterokonta, 905
- Heuristic algorithms, 77
- HGNC *see* HUGO Gene Nomenclature Committee
- Hidden Markov models (HMMs), 181–186, 566  
HHMMER program, 184–186, 187  
Protein Family database of profile HMMs (Pfam), 223–224, 225, 226
- Hierarchical cluster analysis, 511–516
- Highly homologous sequences, 75
- High-throughput gene expression data acquisition, 462
- High-Throughput Genomic Sequences (HTGS), 27
- H-Invitational Database, 459
- Histidine (His), 76
- Histones, 52–53
- HIV-1 pol, 53–54  
BLAST multidomain protein search, 151–155
- HHMMER program, 184–186, 187, 223, 224
- Homogenitase 1,2-dioxygenase (HGD), 1014
- Homo sapiens*, 25, 28, 311  
chromosome sequenced, 716  
chromosomes, 313
- draft sequence of genome, 716–717  
protein domains, 554  
variation, 647
- HomoloGene, 270  
compared with NCBI Gene, 42
- Homology  
definition, 70–74  
evolution of life, 78–79  
history, 72  
modeling, 618–619
- Honeybee, 921, 923
- Hordeum vulgare*, 25
- Hubbard plots, 623
- HUGO Gene Nomenclature Committee, 20, 27, 54, 60, 61, 62, 314, 316–317, 319
- Human chromosomes, 313, 979–981  
Group A, 981–982  
Group B, 982–983  
Group C, 983  
Group D, 983  
Group E, 984  
Group F, 984  
Group G, 984  
groups, 980
- Human disease-associated genes and loci, 1046  
chromosomal abnormalities, 1050–1051  
genome-wide association studies (GWAS), 1047–1050  
linkage analysis, 1047
- Human disease categories, 1020–1036  
allele frequencies and effects sizes, 1020–1021  
cancer, 1033  
complex disorders, 1024  
environmentally caused disease, 1029  
genetic background, 1030  
genomic disorders, 1025  
mitochondrial disease, 1030–1032  
monogenetic disorders, 1021–1024  
somatic mosaic disease, 1032–1033, 1035
- Human disease databases, 1036  
amino acid substitutions, 1045–1046  
ClinVar, 1040–1041  
GeneCards, 1041  
Human Disease Mutation Database (HGMD), 1039  
Integration of Disease Database, 1041  
limitations, 1045  
Locus-Specific Mutation Databases, 1041–1044

- Online Mendelian Inheritance in Man (OMIM), 1036–1039  
 PhenCode project, 1044–1045
- Human disease gene functional classification, 1060–1063  
 disease characteristics, 1062  
 protein products, 1061
- Human disease genes in model organisms, 1055  
 nonvertebrate species, 1056–1057  
 primates, 1059–1060  
 rodents, 1058–1059
- Human Disease Mutation Database (HGMD), 1039
- Human diseases  
 advice for students, 1063  
 perspective, 1063  
 pitfalls, 1063
- Human Gene Mutation Database (HGMD), 417, 420
- Human genetic disease  
 bioinformatics perspective, 1012–1014, 1014  
 causes of death in USA, 1016  
 classification, 1015–1017  
 consequence of DNA variation, 1011–1019  
 Garrod's view of disease, 1014–1015  
 ICD classification system, 1018  
 MeSH terms, 1017–1019  
 mutation mechanisms, 1013  
 projected global deaths, 1016
- Human genome  
 advice for students, 1001  
 gateway to access, 959–964  
 introduction, 957–958  
 perspective, 999–1000  
 pitfalls, 1000–1001  
 statistics, 961
- Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), 27, 317
- Human Genome Project, 964  
 assemblies, 966–968  
 background, 964  
 broad genomic landscape, 968–970  
 centromeres, 974  
 comparative proteome analysis, 975–977  
 complexity of proteome, 978–979  
 CpG islands, 969–970  
 exons, 977  
 gene characteristics, 976  
 gene content, 974–979  
 genetic and physical distance compared, 970–971, 973
- global statistics, 968  
 goals, 965  
 introns, 977  
 issues addressed, 967  
 long-range variation in GC content, 969  
 main conclusions, 958–959  
 noncoding genes, 976  
 noncoding RNAs, 975  
 paralogous genes, 979  
 protein-coding genes, 975  
 repeat content, 971–974  
 segmental duplications, 973–974  
 simple sequence repeats, 973  
 simple sequence repeats, 974  
 strategic issues, 966  
 transposon-derived repeats, 972–973
- Human genome sequencing, 1051–1055  
 complex disorders, 1051–1052  
 conditions, genes, and variants, 1053–1054  
 disease-causing variants in otherwise healthy people, 1054–1055  
 incidental findings, 1052–1054  
 monogenic disorders, 1051  
 research versus clinical sequencing, 1052–1054
- Human genome variation, 986  
 1000 Genomes Project, 995–998  
 haplotype phasing, 996  
 major conclusion of HepMap project, 994  
 sequencing individual genomes, 998–999  
 SNPs, haplotypes, and HapMap, 986–988, 989  
 viewing and analyzing, 988–993  
 viewing and analyzing, 990
- Human Immunodeficiency Virus (HIV-1), 765–770
- Human leukocyte antigen (HLA), 258
- Human microbiome, 811–814  
 bacterial taxa, 813–814  
 fungi, 870
- Human Microbiome Project (HMP), 708, 710, 721, 727, 811–814
- Human mitochondrial genome, 985–986  
 haplogroups, 986
- Human papillomavirus (HPV), 762
- Human Protein Reference Database (HPRD), 542, 565
- Human Proteome Organization (HUPO), 542
- Human T-lymphotropic virus-I (HTLV-I), 762
- Huntingdon disease, 624  
 Hydropathy index, 565
- Identity, definition, 75  
 Illumina, 382–384  
 Indels, 408  
 Influenza, 761  
 Influenza Genome Sequencing Project (IGSP), 773  
 Influenza virus, 771–774  
 genes, 772
- Informed consent, 390, 1052, 1054
- Insertion mutations, 78  
 Insulin, 248–250
- Integrated Microbial Genomes (IMG) website, 816
- Integration of Disease Database, 1041
- Integrative Genomics Viewer (IGV), 407, 988, 990
- International Cancer Genome Consortium (ICGC), 1033
- International Committee of Taxonomy of Viruses (ICTV), 756–758
- International Gene Trap Consortium (IGTC), 658
- International Human Genome Sequencing Consortium (IHGSC), 957
- International Mouse Phenotyping Consortium (IMPC), 647
- International Nucleotide Sequence Database Collaboration (INSDC), 21
- Interpolated context model (ICM), 822–823
- Interpolated Markov models (IMMs), 820–821
- InterPro database, 226–227
- Interspersed duplication, 409, 410
- Introns, 336
- Inversion of chromosomes, 351, 409, 410
- Ion Torrent, 387
- IProClass database, 226–227
- Isoelectric focusing, 544
- Isoleucine (Ile), 76
- Iterative alignment, 214–218
- JIGSAW program, 339
- Jukes–Cantor correction, 277
- Kaposi's sarcoma herpesvirus (KSHV), 762
- Kappa caseins, 85, 86
- Karlin–Altschul statistics, 142
- Kazusa mammalian cDNA set (KIAA), 459
- KDEL sequence, 570
- Kimura two-parameter model, 277

- Klenow fragment, 379  
*K*-means clustering algorithm, 516–517  
 Knockout Mouse Project (KOMP), 647, 653  
 Kyoto Encyclopedia of Genes and Genomes (KEGG), 682–685
- Lactase-phlorizin hydrolase (LPH), 468  
 Last universal cellular ancestor (LUCA), 766  
 Lateral gene transfer (LGT), 827–830  
 Laurasia, 705, 706, 777, 778  
*Leishmania*, 894–895  
 Leucine (Leu), 76  
 Levinthal’s paradox, 598  
 Life on Earth, 705  
   geological history, 706, 707–708  
   evolution of, 78–79  
 Ligation sequencing, 385  
 Likelihood mapping, 290  
 Likelihood ratio test, 279  
*Lilium formosanum*, 311  
 Limited Area Global Alignment of Nucleotides (LAGAN), 192–194  
   algorithm, 193  
 Linkage analysis, 1047  
 Linux, 11  
   commands, 43–44  
 Literature access, 58–59  
 Local alignment, 74  
 Local alignment algorithms, 96  
   dotplots, 104–105  
   FASTA and BLAST, 103–104  
   Smith and Waterman algorithm, 101–103  
   statistical significance, 108  
   websites and URLs, 110  
 Locus Reference Genomic (LRG), 37  
 Locus-Specific Mutation Databases, 1041–1044  
*Locusta migratoria*, 311  
 Logarithms in base-2, 499  
 Log-expectation (LE) score, 219  
 Log-odds ratio, 92  
 Log-odds scoring matrix, 89–91  
 Long interspersed elements (LINEs), 326, 327  
 Long-branch attraction, 287–289  
 Los Alamos National Laboratory (LANL)  
   HIV databases, 769, 770  
 Lymphoblastoid cell lines (LCLs), 468  
 Lysine (Lys), 76
- MA plots, 498  
 Major histocompatibility complex (MHC), 50
- Malaria, 895–898  
 Mammalian Gene Collection (MGC), 459  
 Mann–Whitney test, 493  
 MAQ program, 195  
 Markov chains, 181, 183, 292–293  
 Markov models, 181  
   *see also* hidden Markov models (HMMs)  
 MASCOT® software, 551  
 Mass spectrometry (MS), 547–551  
   matrix-assisted laser desorption ionization (MALDI), 548  
   matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF), 548, 551, 675  
   protein complexes, 675–676  
   tandem affinity purification mass spectrometry (TAP), 676  
   triple quadrupole (QQQ), 548  
 Mass Spectrometry protein sequence Data Base (MSDB), 548  
 Massive Open Online Courses (MOOCs), 13  
 Matrix multiplication method, 87  
 Matrix-assisted laser desorption ionization (MALDI) mass spectrometry, 548  
 Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, 548, 551, 675  
 Maximal unique matches (MUM), 784  
 Maximum likelihood methods, 289–90, 291  
 Maximum parsimony methods, 287–289  
 Mean (average), 107, 491  
 Measles, 761  
 Measles virus, 774–775  
   proteins, 775  
 Medical Subject Heading (MeSH) terms, 1017–1019  
 MEDLINE, 31, 58  
 MEGA software, 259  
 MegaBLAST, 191–192  
 Meiotic errors, 348  
 Mendelian genomic disorders, 1026  
 Merkel cell polyomavirus (MCV), 762  
 Messenger RNA (mRNA), 434, 450  
   full-length cDNA projects, 459  
   gene expression analysis in cDNA libraries, 455–458  
   gene expression studies, 450–452  
   low- and high-throughput studies, 452–455  
 measuring gene expression across the body, 459  
 RNA export, 451, 452  
 RNA processing, 451, 452  
 RNA surveillance, 451, 452  
 transcription, 450–451, 452
- Metadata, 14  
 Metaphase, 312  
 Metazoans, 24, 916, 917–918  
   *Anopheles gambiae*, 921–922  
   butterflies, 922–923  
   *Caenorhabditis elegans*, 918–919  
   *Ciona intestinalis*, 925–926  
   *Dictyostelium discoideum*, 916–917  
   *Drosophila melanogaster*, 919–921  
   fish genomes, 926–929  
   frogs, 929  
   honeybee, 923  
   insect genomes, 923–924  
   mammalian genomes, 933–934  
   mouse, 934–937  
   opposum, 931–933  
   platypus, 931–933  
   primate genomes, 937–940  
   rat, 934–937  
   reptiles, 929–931  
   sea urchin, 924–925  
   silkworms, 922–923  
 Methionine (Met), 76  
 MicroArray Quality Control (MAQC) project, 495, 511  
 Microarray RNA expression data analysis, 479–482  
   advice for students, 531  
   descriptive statistics, 511–519  
   functional annotation, 528–529  
   GEO dataset analysis using R, 504–511  
   GEO2R at NCBI, 482–495  
   Partek Genomics Suite, 495–504  
   perspective, 529–530  
   pitfalls, 530–531  
 Microarray RNA expression measurement, 460–466  
   biological confirmation, 465  
   data acquisition, 464–465  
   data analysis, 465  
   databases, 465  
   experimental design, 461, 494  
   further analysis, 466  
   probe preparation, 464  
   radioactive probes, 463  
   sample preparation, 461–464  
 Microbe, 700  
 Microorganism, 700  
 MicroRNA (miRNA), 445–447

Minimum Information About a Microarray Experiment (MIAME), 465  
 MiRBase, 446  
 Mitochondrial disease, 1030–1032  
 Mitochondrial Eve, 985  
 MitoSeek software, 421–422  
 Mitotic errors, 348  
 Mobile-element insertions, 409, 410  
 ModelTest program, 279  
 Molecular barcodes, 653–657  
 Molecular clock hypothesis, 250–254  
 amino acid substitution rates, 253  
 corrected number of amino acid changes per 100 residues, 252  
 nucleotide substitution rates, 254  
 Molecular evolution, 245–246  
 Molecular Genetics Evolutionary Analysis (MEGA) program, 282  
 bootstrap procedure, 295  
 maximum parsimony method, 287  
 Molecular Modeling Database (MMDB), 32, 608  
 Molecular phylogeny and evolution advice for students, 296  
 five stages of phylogenetic analysis, 270–295  
 goals, 246–247  
 historical background, 247–250  
 molecular clock hypothesis, 250–254  
 neutral theory of molecular evolution, 258–259  
 perspective, 295  
 phylogenetic trees, 259–266  
 pitfalls, 295–296  
 positive and negative selection, 254–258  
 Tajima's relative rate test, 255, 256  
 types of phylogenetic trees, 266–270  
 web resources, 297  
 Monogenic disorders, 1021–1024 examples, 1022  
 Monte Carlo Markov Chain (MCMC) command, 292–293  
 Mosquito-borne human diseases, 922  
 Moss, 916  
 Motifs, 552–559  
 characteristic of proteins, 557–559  
 definition, 553  
 Mouse Genome Informatics Database (MGD/GXD), 566, 646  
 Mouse knockouts, 650–653  
 MrBayes program, 282, 292, 294  
 MSD, 32

Multidimensional scaling (MDS), 518 compared with principal component analysis (PCA), 517  
 Multiparent Advanced Generation Inter-Cross (MAGIC), 913  
 Multiple Alignment using Fast Fourier Transform (MAFFT) program, 214–218  
 Multiple sequence alignment advice for students, 235  
 algorithm assessment, 207–208  
 benchmarking, 207–208  
 benchmarking studies, 221–222  
 consistency-based alignment, 218–220  
 databases, 222–227  
 definition, 206–207  
 evaluation, 223  
 exact approaches, 208  
 genomic regions, 227–234  
 introduction, 205–208  
 iterative alignment, 214–218  
 main approaches, 208–221  
 perspective, 234  
 pitfalls, 234  
 practical strategies, 207  
 progressive sequence alignment, 209–214  
 structure-based alignment, 220–221  
 typical uses, 207  
 Multiple Sequence Comparison by Log Expectation (MUSCLE) program, 215–218  
 profile-profile alignment, 219  
 MUMmer program, 346–347, 833–834 virus genomes, 783–785  
 Mumps, 761  
*Mus musculus*, 25, 28, 646–647, 935–937  
 knockout mice, 650–653  
 MUSCLE program, 208, 215–219  
 Mutagenic Insertion and Chromosome Engineering Resource (MICER), 658  
 Mutation probability matrix for 1 PAM evolutionary distance, 82–84  
 Mutations, 78, 269  
 forbidden, 863–864  
 mechanisms, 1013  
 transition substitutions, 270  
 transversion substitutions, 270  
*Mycobacterium tuberculosis*, 101  
 Myoglobin, 6, 7, 70–75, 77–80, 94, 96, 106–108, 111, 114, 247, 248, 250, 253, 258, 260, 263, 266, 272, 273, 275, 276, 280, 283, 285, 289, 291, 294, 295, 297, 588, 590, 593–596, 603–613, 615, 616, 626, 678

National Center for Biotechnology Information (NCBI), 11, 21  
 access via gene resource, 38–42  
 BLAST-related algorithms, 170  
 command-line access, 42–49  
 command-line access to Entrez databases, 45  
 compared with HomoloGene, 42  
 compared with UniGene, 41–42  
 Conserved Domain Database (CDD), 177, 226  
 EDirect access, 45–49  
 Ensembl, 50–52  
 genome browsers, 49–52  
 Genomes, 710  
 GEO2 for RNA gene expression analysis, 482–495  
 Human Genome Project, 959  
 introduction, 31–32  
 Map Viewer, 52, 314  
 PDB access, 606–609  
 National Human Genome Research Institute (NHGRI), 398  
 Human Genome Project, 961–963  
 National Library of Medicine (NLM), 31, 58  
 Needleman and Wunsch global alignment algorithm, 96–100  
 Negative selection, 254–258  
 Neighbor-joining tree p-distance correction, 276  
 Poisson correction, 276, 280  
*Neurospora crassa*, 311, 873–874  
 Next-generation sequencing (NGS), 19, 359, 378–379, 1051  
 advice for students, 423  
 alignment to reference genome, 194–197  
 BLAST-related algorithms, 170–171  
 Burrows–Wheeler Transform (BWT) alignment, 196–197  
 DNA sequencing analysis, 387–421  
 DNA sequencing technologies, 379–387  
 hash table alignment, 194–196  
 perspective, 422–423  
 pitfalls, 423  
 short read alignment strategies, 195  
 specialized applications, 421–422  
 web resources, 424  
*Nicotiana tabacum*, 311, 908  
 Nodes of phylogenetic trees, 259  
 Noncoding RNA, 436  
 long noncoding RNA (lncRNA), 447–448

- microRNA (miRNA), 445–447  
 other noncoding RNAs, 448  
*Rfam* database, 436–438  
 ribosomal RNA (rRNA), 441–444  
 short interfering RNA (RNAi), 447  
 small nuclear RNA (snRNA), 445  
 small nucleolar RNA (snoRNA), 445  
 splicosomal RNAs, 445  
 transfer RNA (tRNA), 438–441  
 UCSC Genome and Table Browser, 448–449  
 Nonconserved elements, 346  
 Nonparametric bootstrapping, 293–295  
 Nonparametric tests, 488, 493  
 Normal distribution, 107  
 Novel sequence insertions, 409, 410  
 Novel species, 348  
 Nuclear magnetic resonance (NMR) spectroscopy, 600  
 Nucleomorphs, 902–904  
 Nucleotides  
   Limited Area Global Alignment of Nucleotides (LAGAN), 192–194  
   sequence databases searchable via BLAST, 127  
   step matrix, 270, 271  
   substitution rates, 254  
   transition substitutions, 270  
   transversion substitutions, 270  
 Null hypothesis, 106  
 Obligately intracellular and parasitic bacteria, 807–808  
 Obligately intracellular and symbiotic bacteria, 806–807  
 Odds ratio, 89  
 Odorant-binding protein (OBP), 572–573  
 One-based counting, 57, 58  
 Online Mendelian Inheritance in Man (OMIM), 32, 417, 419, 420, 1021, 1036–1039  
   numbering system, 1038  
 Oömycetes, 905  
 Open reading frames (ORFs), 336, 447  
   *Saccharomyces cerevisiae*, 640  
 Open Regulatory annotation (ORegAnno) database, 342–343  
 Operational taxonomic unit (OTU), 259–264  
   microarray data, 514  
   number of rooted and unrooted trees, 263, 265  
   tree-building by neighbor-joining (NJ), 285–287  
 Opposum, 931–933  
 Organism-specific BLAST sites, 168–170  
 Ortheus program, 231, 232, 233  
 Orthologous proteins or genes, 70–71, 319, 552, 568, 571, 573, 575, 642, 645, 650, 673, 680, 685, 866, 868, 874  
*Oryza sativa*, 311, 913–914  
 Oxytocin, 250  
*P* value, 491–492  
 Pacific Biosciences DNA sequencing, 387  
 Pairwise sequence alignment, 69–79  
   advice for students, 112  
   gaps, 78  
   global and local alignment algorithms, 96–106  
   homology and evolution of life, 78–79  
   limits of detection, 94–96  
   perspective, 110–112  
   pitfalls, 112  
   scoring matrices, 79–96  
   statistical significance, 106–110  
   web resources, 112  
 Pairwise Sequence Comparison (PASC) tool, 780–782  
*Pan troglodytes*, 91  
 Pangaea, 705, 706, 778  
 Parabasala, 891  
 Paralogous proteins or genes, 70–71, 73, 320, 552, 642, 643, 656, 679, 864, 866, 867  
*Paramecium aurelia*, 311  
*Paramecium caudatum*, 311  
*Paramecium tetraurelia*, 899–901  
 Parametric tests, 488  
*Parascaris equorum*, 311  
 Parkinson disease, 624  
 Parsimony analysis, 287–289  
   maximum parsimony principle, 288  
 Partek Genomics Suite, 495–496  
   ANOVA, 501–504  
   data analysis, exploratory, 498–501  
   data import, 496  
   log<sub>2</sub> transformed microarray data, 498  
   MA plots, 498  
   principal components analysis (PCA), 498–501  
   quality control, 496–497  
   sample histogram, 498  
   sample information, adding, 497  
   scatter plots, 498  
   *t*-test, 503–504  
 PartTree program, 215  
 Pattern-hit initiated BLAST (PHI-BLAST), 179–181  
   choosing a pattern, 180  
 PatternHunter, 188  
   nonconsecutive seeds, 189  
 P-distance correction, 276  
 Pearson correlation coefficient, 510  
 Pecan program, 191, 231, 232  
 Peptide bonds, 593  
 Percent identity, 108–109  
 Percent similarity, 75  
 Perfect match (PM) probesets, 488  
 Perl, 11, 627, 728, 1032  
 Phaeophyceae, 906  
*Phanerochaete chrysosporium*, 875  
 PHAST package, 229  
 PHD program, 596  
 PhenCode project, 1044–1045  
 Phenotype and genotype, 637–638  
 Phenylalanine (Phe), 76  
 Phosphorylation, 564–565  
 PHRED scores, 391, 392  
 Phylogenetic analysis, 270–295  
   amino acid substitution, 272–281  
   DNA models, 272–281  
   multiple sequence alignment, 271–272  
   pitfalls, 295–296  
   sequence acquisition, 270–271  
   tree-building methods, 281–293  
 Phylogenetic Analysis Using Parsimony (PAUP) program, 282  
 Phylogenetic trees, 257–259  
   DNA trees, 268–270  
   enumeration, 263–266  
   globins, 247, 248, 260, 276, 280, 291, 294  
   number of rooted and unrooted trees, 263, 266  
   protein-based trees, 268–270  
   RNA trees, 268–270  
   roots, 262–263  
   search strategies, 263–266  
   species trees versus gene/protein trees, 266–268  
   topologies and branch length, 259–262, 286  
   types, 266–270  
 Phylogenetic trees of species  
   bacteria, 802  
   comparative, 725  
   *E. coli*, 815  
   eukaryotes, 849, 888, 907  
   fish, 927  
   fungi, 849, 850  
   global tree of life, 7, 703, 812, 873  
   herpesviruses, 777, 778

- Phylogenetic trees of species (*continued*)  
 lentiviruses, 767  
 mammals, 932  
 metazoans (animals), 917  
 origin of life, 704  
 plants, 908  
 primates, 938  
 reptiles, 930
- Phylogenies, inconsistent, 890
- Phylogeny, 246
- Phylogeny Inference Package (PHYLIP)  
 program, 282
- Phylogenograms, 262
- Pinus resinosa*, 311
- PipMaker software, 346–347
- Plant genomes, 906–908  
*Arabidopsis thaliana*, 910–913  
*Chlorophyta*, 908–910  
 detecting ancient whole-genome  
 duplications, 912  
 evolution of plants, 907  
 giant genomes, 915  
 land plants, 915  
 moss, 916  
*Oryza sativa*, 913–914  
 phylogeny, 908  
*Populus trichocarpa*, 914  
 tiny genomes, 915  
*Vitis vinifera*, 915
- Plasmodium falciparum*, 895–898
- Plastids, 902
- Platypus, 931–933
- PLINK, 992
- Ploidy *see* Polyploidy
- Point-and-click web-based software,  
 10–11
- Poisson correction, 276
- Poisson distribution, 274
- Poliomyelitis, 761
- Polyacrylamide gel electrophoresis  
 (PAGE), 543–547
- Polymerase chain reaction (PCR), 384,  
 660
- Polymorphism Phenotyping-2 (PolyPhen)  
 software, 417–420
- Polypeptides, 540  
 amino acids, 591–594  
 phi and psi angles, 593
- Polyploidy, 347, 348, 863  
 plants, 911
- Populus trichocarpa*, 914
- Position-specific iterated BLAST (PSI-BLAST), 171–177  
 assessing performance, 179  
 corruption problem, 177  
 hits from human beta globin, 174, 175
- homologous matches, 176  
 matrix view, 173  
 pitfalls, 197–198  
 reverse position-specific BLAST (RPS-BLAST), 177, 178, 226  
 target frequencies, 173
- Position-specific scoring matrix (PSSM),  
 172–173, 564
- Positive selection, 254–258
- Posterior probability, 219, 282, 289,  
 291–293, 295
- Post-translational modifications to  
 proteins, 560, 561
- Precision, definition, 490
- Primary structure of proteins, 591–594
- Principal components analysis (PCA),  
 498–501  
 axes, 501  
 compared with multidimensional  
 scaling (MDS), 517  
 plots, 500
- Privacy, 390
- Probability matrix, 82–84
- Probability that base not sequenced, 405
- ProbCons program, 215, 217  
 consistency-based alignment, 219–220
- Profile Alignment (PRALINE) program,  
 215–218  
 structure-based alignment, 220–221
- Profile searches, 181–186
- Profile sum-of-pairs (PSP) scoring  
 method, 219
- Profile–profile alignment, 219
- Progenote, 700
- Programming, learning, 13–14
- Programs, 77
- Progressive sequence alignment,  
 209–214
- Prokaryote, 700, 702
- Proline (Pro), 76
- PROSITE database, 226, 558–559
- Protein alignment, 70
- Protein analysis, 539–540, 543  
 direct protein sequencing, 543  
 mass spectrometry, 547–551  
 polyacrylamide gel electrophoresis  
 (PAGE), 543–547
- Protein complexes, 675–676
- Protein Data Bank (PDB), 30, 32, 590,  
 602–617  
 accessing entries, 606–609  
 CATH database, 613–615  
 classification, 603  
 comparison of resources, 617  
 Dali Domain Dictionary, 615–616  
 molecule types, 604
- number of searchable structures, 604  
 protein folding, 609–610  
 search results, 605
- Structural Classification of Proteins  
 (SCOP) database, 610–613
- structure retrieval, 609  
 visualization tools, 606, 607
- Protein databases, 29–31  
 content, 24–31
- Protein homologs, 832, 833
- Protein Information Resource (PIR), 30
- Protein networks, 678–680  
 accuracy, 680  
 data, 680  
 experimental organism, 680–681  
 map categories, 681–682  
 pathway variation, 681
- Protein Research Foundation (PRF), 30
- Protein structure  
 advice for students, 625  
 disease, 622–625  
 high-resolution structures, 599  
 overview, 589–591  
 perspective, 625  
 pitfalls, 625  
 primary structure, 591–594  
 principles, 591–602  
 protein-folding, 598–600  
 quaternary structure, 592  
 secondary structure, 592, 594–598  
 sequence and structure, 590  
 structural genomics, 600–601  
 target selection, 602  
 tertiary structure, 592, 598–600
- Protein Structure Initiative (PSI),  
 601–602
- Protein structure prediction, 617–618  
*ab initio* prediction, 621  
 accuracy, 620  
 approaches, 618  
 fold recognition, 619–620  
 homology modeling, 618–619  
 progress assessment, 621–622  
 websites, 620
- Protein trees, 266–268
- Protein-based trees, 268–270
- Protein–protein interactions, 672–680  
 accuracy, 680  
 data, 680  
 databases, 676–678  
 experimental organism, 680–681  
 map categories, 681–682  
 pathway variation, 681
- Proteins  
 amino acid relative abundance, 76  
 coding gene study resources, 340–342

- databases, 540–542  
*Dayhoff's superfamilies*, 77  
 finding distantly related proteins, 171–181  
 functional genomics, 672  
 functions, 570–573, 574, 575  
 homology, 70–74  
 interactions, 672–678  
 intrinsically disordered, 622  
 localization, 570  
 methyl-binding domains, 557  
 modular nature, 552–559  
 motif characteristic, 557–559  
 multidomain proteins, 556–557  
 multiple copies of distinct domains, 555  
 orthologous, 70–71  
 overview, 552  
 pairwise alignment, 72–74  
 pairwise interactions, 678–682  
 paralogous, 70–71, 73  
 perspectives, 551–573  
 physical properties, 559–566  
 post-translational modifications, 560, 561  
 Protein Family database of profile HMMs (Pfam), 223–224, 225, 226  
 sequence databases searchable via BLAST, 126  
 sharing common domains, 556  
**Proteomics**, 539–540  
 accuracy of prediction programs, 564–565  
 advice for students, 574–575  
 Association of Biomolecular Resource Facilities (ABRF), 542–543  
 community standards, 542  
 definitions, 553  
 domains and motifs, 552–559  
 forward and reverse, 671  
 function of proteins, 570–573, 574, 575  
 gene ontology websites, 568  
 localization of proteins, 570  
 perspective, 573–574  
 phosphorylation, 564–565  
 physical properties, 559–566  
 pitfalls, 574  
 transmembrane regions, 565–566  
 Gene Ontology (GO) Consortium, 566–570  
 web resources, 576–578  
**Proteomics Identifications (PRIDE)**  
 database, 549–550  
*Protopterus aethiopicus*, 311  
**Protozoans**, 890  
*Giardia lamblia*, 891–892  
*Trichomonas*, 890–891
- ProtTest program, 280–281  
**Pseudogene Decoration Resource (psiDR)**, 329  
**Pseudogenes**, 326–331  
**PubMed**, 31, 59  
**Pyrosequencing**, 384–385, 386  
**Python**, 11, 112, 114
- Quaternary structure of proteins, 592
- R programming language, 482–485  
 biomaRt, 317  
 seqinr, 818  
 RStudio, 504  
 Ramachandran plots, 594, 596  
 Random insertional metagenesis, 657–660  
**Rapid Annotations using Subsystems (RAST) server**, 825, 827  
*Rattus norvegicus*, 25, 311, 935, 937  
**Receiver operating characteristic (ROC) curves**, 356, 421  
**Redundancy of coverage**, 405  
**Reference sequence (RefSeq) project**, 36–37, 340–342  
 RefSeqGene, 37  
**Regulatory factor databases**, 342–345  
**Relatedness odds matrix**, 88  
**Relative entropy**, 109–110  
**Relative mutability of amino acids**, 80–82  
**RepeatMasker software**, 325, 326, 329  
**Reproducible research in bioinformatics**, 14  
**Reptiles**, 929–931  
**Rett syndrome (RTT)**, 555, 1023  
**Reverse genetics**, 649  
 compared with forward genetics, 665–666  
 gene silencing, 662–664  
 $\beta$ -globin gene, 650–653  
 insertional mutagenesis, 660–662  
 molecular barcodes, 653–657  
 mouse knockouts, 650–653  
 random insertional mutagenesis, 657–660  
 techniques, 658  
 yeast knockouts, 653–657  
**Reverse position-specific BLAST (RPS-BLAST)**, 177, 178, 226  
**Reverse transcription polymerase chain reaction (RT-PCR)**, 452–453  
**Rfam database**, 436–438  
**Ribosomal Database Project (RDB)**, 444
- Ribosomal RNA (rRNA)**, 435, 441–444  
 gene numbers in selected organisms, 443  
 major forms in bacteria and eukaryotes, 443
- Risk allele frequencies, 1021
- RNA**  
 advice to students, 470  
 disrupting, 662–664  
 functional genomics, 668–670  
 gene expression analysis method overview, 481  
 genome-wide gene expression, 460–466  
 introduction, 433–436  
 messenger RNA (mRNA), 450–459  
 noncoding RNA, 436–449  
 perspective, 469–470  
 pitfalls, 470  
 structure, 434  
 transcription of DNA, 435  
 web resources, 470
- RNA analysis interpretation, 466  
 relationship between DNA, mRNA, and protein levels, 466–467  
 transcription, 467–468
- RNA databases, 27–29  
 content, 24–31
- RNA-inducing silencing complex (RISC), 447, 662
- RNA interference (RNAi), 662
- RNA-seq expression data analysis, 479–482, 519–523  
 advice for students, 531  
 CuffLinks sample protocol, 523–524  
 CuffLinks to assemble transcripts, 525  
 CuffLinks to determine differential expression, 525–526  
 CummeRbund sample protocol, 526–527, 528  
 perspective, 529–530  
 pitfalls, 530–531
- RNA-seq Genome Annotation Assessment Project (RGASP), 527–528
- TopHat reference genome, 524–525
- TopHat sample protocol, 523–524  
 workflow chart, 522
- RNA-seq expression measurement, 460–466  
 biological confirmation, 465  
 data acquisition, 464–465  
 data analysis, 465  
 databases, 465  
 experimental design, 461

- RNA-seq expression measurement  
(*continued*)  
further analysis, 466  
probe preparation, 464  
sample preparation, 461–464
- RNA-seq Genome Annotation  
Assessment Project (RGASP), 527–528
- RNA trees, 268–270
- Robertsonian translocation, 348
- Robust multiarray analysis (RMA), 488  
normalization for accuracy and precision, 488–490  
normalization for CEL files, 506–508
- Root mean square deviation (RMSD), 221, 611, 616, 619, 621, 623
- Roots of phylogenetic trees, 262–263
- Rosetta method, 621
- Rotavirus, 761
- Rubella, 761
- Saccharomyces cerevisiae*, 23, 311, 640–643, 849, 850  
chromosomes, 854–860  
features, 851–854  
gene duplication, 860–865  
gene nomenclature, 856  
genome duplication, 860–865  
genome sequencing, 851  
insertional mutagenesis, 660–662  
introns, 853  
knockout genes, 653–657  
multiple yeast genome browsers, 857  
NCBI Genome Workbench, 855  
NCBI Map Viewer, 854  
polyploidy, 863  
protein domains, 853  
proteins, 852  
secretion proteins, 642
- Saccharomyces* Genome Database (SGD), 566, 640–643, 856
- Saccharomyces pombe*, 875–876
- SAM/BAM format files and SAMtools, 402–405  
CRAM file format, 406–408  
file anatomy, 403, 404  
finding files, 405–406  
mandatory fields, 402  
read depth calculation, 405  
viewing files, 406
- Sanger sequencing, 19, 35, 379–382  
compared with next-generation sequencing (NGS) technologies, 382
- dideoxynucleotide sequencing, 380
- genomic DNA, 381  
quality scores, 381
- Scatter plots, 498
- Scoring matrices for proteins, 79  
accepted point mutations (PAM), 79, 80, 81  
block substitution matrix (BLOSUM), 91–94  
frequency of amino acids, 79, 81  
limits of detection, 94–96  
log-odds scoring matrix, 89–91  
mutation probability matrix for 1 PAM evolutionary distance, 82–84  
PAM250 and other PAM matrices, 84–88  
practical usefulness of PAM matrices, 91  
relatedness odds matrix, 88  
relative mutability of amino acids, 80–82
- Sea urchin, 924–925
- Secondary structure of proteins, 225, 592, 594–598  
DSSP codes, 598  
web servers for prediction, 597
- Segmental duplications in DNA, 331–333
- Selected reaction monitoring (SRM), 548
- Selection, 254–258
- Self-organizing maps (SOMs), 517, 518
- Sequence Read Archive (SRA), 14, 23, 465
- Sequence Search and Alignment by Hashing Algorithm (SSAHA2), 194
- Sequence-tagged sites (STSs), 27
- Serine (Ser), 76
- Sexual reproduction, 309
- Short interfering RNA (RNAi), 447
- Short read alignment strategies, 195
- Sickle cell anemia, 624, 1022–1023
- Silkworms, 922–923
- Simian immunodeficiency virus (SIV), 766–767
- Similarity  
compared with distance measures, 213  
definition, 74–75  
percent similarity, 75
- Simple Modular Architecture Research Tool (SMART) database, 224
- Simple sequence repeats in DNA, 331
- Single nucleotide polymorphisms (SNPs), 354–355, 408  
microarrays, 356–358
- Single nucleotide variants (SNVs), 408
- Slime molds, 916–917
- Small nuclear RNA (snRNA), 445
- Small nucleolar RNA (snoRNA), 445
- Small subunits (SSUs), 707
- Small-angle X-ray scattering, 600
- Smallpox, 761
- Smith and Waterman local alignment algorithm, 101–103, 122  
rapid heuristic versions, 103–104
- SNARE protein, 641, 864
- Sodium dodecyl sulfate (SDS), 544
- Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), 544–545, 547
- Software for bioinformatics, 10–14  
bridging the two cultures, 12–13  
command-line, 10, 11–12  
programming, learning, 13–14  
reproducible research, 14  
web-based, 10–11
- Solanum lycopersicum*, 25
- Solexa quality score, 391
- Somatic mosaic disease, 318, 349, 378, 422, 1032–1033, 1035
- Sorting Tolerant from Intolerant (SIFT) software, 417–420
- Species trees, 266–268
- Splitting of chromosomes, 348
- SRA Toolkit, 392–393
- Standard deviation, 107
- Standard genetic code, 83
- Statistical significance of pairwise alignments, 106  
global alignments, 106–108  
local alignments, 108  
percent identity, 108–109  
relative entropy, 109–110
- Statistics  
average, 491  
binomial coefficient, 289  
bit score, 75, 91–92, 132, 134, 137, 143, 175, 438  
Chi-squared analysis, 256  
covariance models, 439, 441  
*E* value, 142–143  
false discovery rate (FDR), 495, 502  
Fisher's exact test, 457, 458  
gamma distribution, 278  
Gaussian distribution, 107  
logarithms in base-2, 499  
log-expectation (LE) score, 219  
log-odds ratio, 92  
log-odds scoring matrix, 89–91  
Mann–Whitney test, 493  
mean (average), 107, 491  
multidimensional scaling (MDS), 518
- MDS compared with principal components analysis (PCA), 517
- nonparametric bootstrapping, 293–295
- nonparametric tests, 488, 493
- normal distribution, 107
- null hypothesis, 106

- odds ratio, 89  
*p* value, 491–492  
parametric tests, 488  
p-distance correction, 276  
Pearson correlation coefficient, 510  
Poisson correction, 276  
Poisson distribution, 274  
precision, definition, 490  
probability matrix, 82–84  
probability that base not sequenced, 405  
receiver operating characteristic (ROC) curves, 356, 421  
standard deviation, 107  
variance, 107, 491  
Wilcoxon test, 493  
Z scores, 107–108  
Step matrices, 270, 271  
Stochastic context-free grammar (SCFG), 439, 441  
Stramenopila, 904–906  
*Strongylocentrotus purpuratus*, 311, 924  
Structural biology, 591  
Structural Biology Knowledgebase (SBKB), 602  
Structural Classification of Proteins (SCOP) database, 610–613  
release notes, 612  
Structural Classification of Proteins-extended (SCOPe) database, 613  
Structural genomics, 600–601  
Structural polymorphisms, 1027  
Structural variants, identifying, 409–410  
Structure-based alignment, 220–221  
Substitution mutations, 78  
Sum-of-pairs scores (SPS), 223  
Surface plasmon resonance, 673  
*Sus scrofa*, 25, 28, 934  
SWISS-PROT, 30, 31, 32  
Synonymous Nonsynonymous Analysis Program (SNAP), 258  
Synteny, 346  
Systematics, 701–704  
Tajima’s relative rate test, 255, 256  
Tandem affinity purification mass spectrometry (TAP-MS), 676  
Tandem duplication, 409, 410  
Tandemly repeated sequences in DNA, 333–334  
Target frequencies, 90  
TATA box, 336  
Taxonomy, 709–710  
Taxonomy Browser, 52  
Taxons (taxa), 259–262  
TaxPlot, 830–833  
TBLASTN tool, 70  
T-COFFEE program, 215, 217  
consistency-based alignment, 220  
Expresso program, 220–221  
structure-based alignment, 220–221  
Telomeres, 312, 333, 348, 855, 860, 868, 894  
Template-free modeling, 621  
Tertiary structure of proteins, 592, 598–600  
Thalassemias, 1023  
The *Arabidopsis* Information Resource (TAIR), 643, 644, 911, 913  
The Cancer Genome Atlas (TCGA), 1033–1034  
*Theileria annulata*, 898  
Threaded Blockset Aligner (TBA), 229  
Threading, 619–620  
Threonine (Thr), 76  
*Thuja occidentalis*, 311  
Thymine, 435  
TMHMM program, 566  
Tool makers, 15  
Tool users, 15  
TopHat sample protocol, 523–524  
reference genome, 524–525  
*Toxoplasma gondii*, 898–899  
Transcription activator-like effector nucleases (TALENs), 664  
Transcriptome, 5  
Transcripts of uncertain coding potential (TUCP), 448  
Transcript-specific variance, 491  
Transfer RNA (tRNA), 434–435, 438–441  
identification, 440  
Transition substitutions, 270  
Translocation, 348  
Transposon tagging, 662  
Transposons, 325–326  
Transversion substitutions, 270  
Tree bisection reconnection (TBR)  
approach, 266  
Tree of life, 7, 246, 812, 828  
global, 703  
inferred tree, 246  
molecular sequences, 705–709  
nomenclature, 700  
true tree, 246  
Tree-building methods, 281–282  
Bayesian inference method, 290–293, 294  
branch lengths, 286  
distance-based, 282–287  
evaluating trees, 293–295  
maximum likelihood, 289–90  
maximum parsimony, 287–289  
neighbor-joining (NJ), 285–287  
unweighted pair group method of arithmetic averages (UPGMA), 283–285  
TREE-PUZZLE program, 282  
maximum likelihood method, 289–290, 291  
TrEMBL, 32, 540  
*Trichomonas*, 890–891  
Triple quadrupole mass spectrometry (QQQ), 548  
Triploid cells, 348  
Trisomy 21 (Down syndrome), 313, 349, 480  
*Triticum aestivum*, 25, 28, 908, 915  
tRNAscan-SE, 439–441, 825  
*Truturus cristatus*, 311  
*Trypanosoma*, 892–894  
Tryptophan (Trp), 76  
T-test statistic, 491, 492–493  
Tyrosine (Tyr), 76  
UCNEbase database, 345  
Ultraconserved elements, 345  
UniBuild  
human cluster sizes, 455  
human cluster sizes, ten largest, 456  
nonhuman cluster sizes, ten largest, 456  
Unicellular pathogens, 892–895  
*Leishmania*, 894–895  
*Trypanosoma*, 892–894  
UniGene, 28  
compared with NCBI Gene, 41–42  
phyla and organisms, 29  
UniMes database, 540  
UniParc database, 540  
Uniparental disomy, 349  
UniProt, 29, 31, 540  
UniProtKB database, 540, 548, 550, 553  
UniRef database, 540  
University of California at Santa Cruz (UCSC), 11  
Genes, 340–342  
Genome Bioinformatics site, 321  
Genome Browser, 50, 190, 229, 230, 314, 710  
Human Genome Project, 961  
Table Brower, 54–56, 229, 327, 328  
Unix, 11, 12, 15, 42–47, 379  
3'-Untranslated region (3'UTR), 451  
5'-Untranslated region (5'UTR), 451  
Unweighted pair group method of arithmetic averages (UPGMA), 211, 218, 282, 513  
distance-based method, 283–285  
Uracil, 435

- VAAST software, 420–421  
Vaccine-preventable bacterial disease, 808  
Valine (Val), 76  
Variance, 107, 491  
Variant call format (VCF) files and  
  VCFTools, 388, 407, 408, 410–413  
  file columns, 411  
  file description, 412  
  finding and viewing files, 413  
  globin VCF, 996–998  
Variant Effect Predictor (VEP) program,  
  419  
Variant surface glycoprotein (VSG), 895  
Variants, biological significance of,  
  417–421  
Varicella, 761  
VAST program, 611  
Vector alignment search tool (VAST), 32  
Velvet software, 734  
Vertebrate Genome Annotation (VEGA)  
  project, 37, 314  
Vertebrate Multiz Alignment and  
  Conservation, 229  
*Vitis vinifera*, 915  
Viral diseases, 761  
Viridiplantae, 24  
Virus genomes, 763  
  advice for students, 786  
  bioinformatics approaches, 765–766
- International Committee of Taxonomy  
  of Viruses (ICTV), 756–758  
introduction, 755–758  
metagenomics and diversity,  
  764–765  
MUMmer comparisons, 783–785  
Pairwise Sequence Comparison  
  (PASC) tool, 780–782  
perspectives, 785–786  
pitfalls, 786  
web resources, 786–787
- Viruses, 24, 704, 1029  
cancer-causing, 762  
classification, 758–765  
diversity and evolution, 762–764  
ebola virus, 775–776  
genome size, 758–760, 761  
genomes, 711–712  
giant viruses, 782–783  
herpesvirus, 776–780  
Human Immunodeficiency Virus  
  (HIV-1), 765–770  
influenza virus, 771–774  
measles virus, 774–775  
morphology, 758  
nucleic acid composition, 758, 760  
VISTA program, 346–347
- Web-based software tools, 10–11
- Wellcome Trust Sanger Institute (WTSI),  
  168, 170  
Human Genome Project, 964  
Whole-genome duplication,  
  347–349  
  graphical representation, 902  
Whole-genome shotgun (WGS) strategy,  
  22, 395  
Wilcoxon test, 493  
WoLF PSORT program, 570, 571  
WU-BLAST 2.0, 170
- X chromosome dosage compensation,  
  349, 983  
*Xenopus laevis*, 311, 927, 929  
*Xenopus tropicalis*, 28  
X-ray crystallography, 598, 599  
X-ray free-electron lasers (XFELs),  
  599
- Yeast knockouts, 653–657  
Yeast two-hybrid system, 673–675
- Z scores, 107–108  
*Zea mays*, 25, 28, 910  
Zebrafish Information Network (ZFIN),  
  645
- Zero-based counting, 57, 58  
Zinc fingers, 664

## **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.