Contents

Section I Introduction to Cellular and Molecular Biology

1. The Cellular System and the Code of Life 3

1.1 The Cellular Challenge 3

1.2 How Cells Meet the Challenge 4

1.3 Molecules in Cells 4

1.4 Intracellular Structures or Spaces 5

1.4.1 Nucleus 5

1.4.2 Cell Membrane 6

1.4.3 Cytoplasm 7

1.4.4 Endosome, Lysosome, and Peroxisome 8

1.4.5 Ribosome 9

1.4.6 Endoplasmic Reticulum (ER) 9

1.4.7 Golgi Apparatus 10

1.4.8 Cytoskeleton 10

1.4.9 Mitochondrion 10

1.4.10 Chloroplast 12

1.5 The Cell as a System 13

1.5.1 The Cellular System 13

1.5.2 Systems Biology of the Cell 13

1.5.3 How to Study the Cellular System 14

2. DNA Sequence: The Genome Base 17

2.1 The DNA Double Helix and Base Sequence 17

2.2 How DNA Molecules Replicate and Maintain Fidelity 18

2.3 How the Genetic Information Stored in DNA Is Transferred to Protein 20

2.4 The Genomic Landscape 21

2.4.1 The Minimal Genome 21

2.4.2 Genome Sizes 21

2.4.3 Protein-Coding Regions of the Genome 22

2.4.4 Noncoding Genomic Elements 23

2.5 DNA Packaging, Sequence Access, and DNA–Protein Interactions 25

2.5.1 DNA Packaging 25

2.5.2 Sequence Access 25

2.5.3 DNA–Protein Interactions 26

2.6 DNA Sequence Mutation and Polymorphism 27

2.7 Genome Evolution ... 28
2.8 Epigenome and DNA Methylation ... 29
2.9 Genome Sequencing and Disease Risk ... 30
2.9.1 Mendelian (Single-Gene) Diseases ... 31
2.9.2 Complex Diseases That Involve Multiple Genes ... 31
2.9.3 Diseases Caused by Genome Instability ... 32
2.9.4 Epigenomic/Epigenetic Diseases ... 32

3. RNA: The Transcribed Sequence ... 35
3.1 RNA as the Messenger ... 35
3.2 The Molecular Structure of RNA ... 35
3.3 Generation, Processing, and Turnover of RNA as a Messenger ... 36
3.3.1 DNA Template ... 37
3.3.2 Transcription of Prokaryotic Genes ... 37
3.3.3 Initial Transcription of Pre-mRNA from Eukaryotic Genes ... 38
3.3.4 Maturation of mRNA from Pre-mRNA ... 40
3.3.5 Transport and Localization ... 42
3.3.6 Stability and Decay ... 42
3.3.7 Major Steps of mRNA Transcript Level Regulation ... 43
3.4 RNA Is More Than a Messenger ... 44
3.4.1 Ribozyme ... 45
3.4.2 snRNA and snoRNA ... 46
3.4.3 RNA for Telomere Replication ... 46
3.4.4 RNAi and Small Noncoding RNAs ... 46
3.4.4.1 miRNA ... 47
3.4.4.2 siRNA ... 49
3.4.4.3 piRNA ... 49
3.4.5 Long Noncoding RNAs ... 50
3.4.6 Other Noncoding RNAs ... 50
3.5 The Cellular Transcriptional Landscape ... 51

Section II Introduction to Next-Generation Sequencing (NGS) and NGS Data Analysis

4. Next-Generation Sequencing (NGS) Technologies: Ins and Outs ... 55
4.1 How to Sequence DNA: From First Generation to the Next ... 55
4.2 A Typical NGS Experimental Workflow ... 58
4.3 Ins and Outs of Different NGS Platforms ... 60
4.3.1 Illumina Reversible Dye-Terminator Sequencing ... 61
4.3.1.1 Sequencing Principle ... 61
4.3.1.2 Implementation ... 61
4.3.1.3 Error Rate, Read Length, Data Output, and Run Time ... 64

4.3.2 Ion Torrent Semiconductor Sequencing 65
4.3.2.1 Sequencing Principle 65
4.3.2.2 Implementation 65
4.3.2.3 Error Rate, Read Length, Date Output, and Run Time 66
4.3.3 Pacific Biosciences Single Molecule Real-Time (SMRT) Sequencing 66
4.3.3.1 Sequencing Principle 66
4.3.3.2 Implementation 67
4.3.3.3 Error Rate, Read Length, Data Output, and Run Time 67
4.4 Biases and Other Adverse Factors That May Affect NGS Data Accuracy 69
4.4.1 Biases in Library Construction 69
4.4.2 Biases and Other Factors in Sequencing 70
4.5 Major Applications of NGS 71
4.5.1 Transcriptomic Profiling and Splicing Variant Detection (RNA-Seq) 71
4.5.2 Genetic Mutation and Variation Discovery 71
4.5.3 *De novo* Genome Assembly 71
4.5.4 Protein-DNA Interaction Analysis (ChIP-Seq) 71
4.5.5 Epigenomics and DNA Methylation Study (Methyl-Seq) ... 72
4.5.6 Metagenomics 72

5. Early-Stage Next-Generation Sequencing (NGS) Data Analysis: Common Steps 73
5.1 Base Calling, FASTQ File Format, and Base Quality Score 73
5.2 NGS Data Quality Control and Preprocessing 76
5.3 Reads Mapping 78
5.3.1 Mapping Approaches and Algorithms 78
5.3.2 Selection of Mapping Algorithms and Reference Genome Sequences 80
5.3.3 SAM/BAM as the Standard Mapping File Format 81
5.3.4 Mapping File Examination and Operation 83
5.4 Tertiary Analysis 86

6. Computing Needs for Next-Generation Sequencing (NGS) Data Management and Analysis 87
6.1 NGS Data Storage, Transfer, and Sharing 87
6.2 Computing Power Required for NGS Data Analysis 89
6.3 Software Needs for NGS Data Analysis 90
6.4 Bioinformatics Skills Required for NGS Data Analysis 92

Section III Application-Specific NGS Data Analysis

7. Transcriptomics by RNA-Seq ... 97
7.1 Principle of RNA-Seq ... 97
7.2 Experimental Design ... 98
7.2.1 Factorial Design ... 98
7.2.2 Replication and Randomization ... 98
7.2.3 Sample Preparation ... 99
7.2.4 Sequencing Strategy ... 100
7.3 RNA-Seq Data Analysis ... 101
7.3.1 Data Quality Control and Reads Mapping ... 101
7.3.2 RNA-Seq Data Normalization ... 103
7.3.3 Identification of Differentially Expressed Genes ... 105
7.3.4 Differential Splicing Analysis ... 107
7.3.5 Visualization of RNA-Seq Data ... 108
7.3.6 Functional Analysis of Identified Genes ... 108
7.4 RNA-Seq as a Discovery Tool ... 109

8. Small RNA Sequencing ... 111
8.1 Small RNA Next-Generation Sequencing (NGS) Data Generation and Upstream Processing ... 112
8.1.1 Data Generation ... 112
8.1.2 Preprocessing ... 113
8.1.3 Mapping ... 114
8.1.4 Identification of Known and Putative Small RNA Species ... 115
8.1.5 Normalization ... 115
8.2 Identification of Differentially Expressed Small RNAs ... 116
8.3 Functional Analysis of Identified Small RNAs ... 116

9. Genotyping and Genomic Variation Discovery by Whole Genome Resequencing ... 119
9.1 Data Preprocessing, Mapping, Realignment, and Recalibration... 120
9.2 Single Nucleotide Variant (SNV) and Indel Calling ... 121
9.2.1 SNV Calling ... 121
9.2.2 Identification of *de novo* Mutations ... 123
9.2.3 Indel Calling ... 124
9.2.4 Variant Calling from RNA-Seq Data ... 124
9.2.5 Variant Call Format (VCF) File ... 125
9.2.6 Evaluating VCF Results ... 126
9.3 Structural Variant (SV) Calling ... 126
9.3.1 Read-Pair-Based SV Calling ... 126
9.3.2 Breakpoint Determination ... 128
9.3.3 *De novo* Assembly-Based SV Detection ... 128

9.3.4 CNV Detection 128
9.3.5 Integrated SV Analysis 129
9.4 Annotation of Called Variants 129
9.5 Testing of Variant Association with Diseases or Traits 130

10. *De novo* Genome Assembly from Next-Generation Sequencing (NGS) Reads 131
10.1 Genomic Factors and Sequencing Strategies for *de novo* Assembly 132
10.1.1 Genomic Factors That Affect *de novo* Assembly 132
10.1.2 Sequencing Strategies for *de novo* Assembly 132
10.2 Assembly of Contigs 134
10.2.1 Sequence Data Preprocessing, Error Correction, and Assessment of Genome Characteristics 134
10.2.2 Contig Assembly Algorithms 136
10.3 Scaffolding 138
10.4 Assembly Quality Evaluation 139
10.5 Gap Closure 140
10.6 Limitations and Future Development 140

11. Mapping Protein–DNA Interactions with ChIP-Seq 143
11.1 Principle of ChIP-Seq 143
11.2 Experimental Design 145
11.2.1 Experimental Control 145
11.2.2 Sequencing Depth 145
11.2.3 Replication 146
11.3 Read Mapping, Peak Calling, and Peak Visualization 146
11.3.1 Data Quality Control and Read Mapping 146
11.3.2 Peak Calling 149
11.3.3 Peak Visualization 156
11.4 Differential Binding Analysis 156
11.5 Functional Analysis 159
11.6 Motif Analysis 159
11.7 Integrated ChIP-Seq Data Analysis 160

12. Epigenomics and DNA Methylation Analysis by Next-Generation Sequencing (NGS) 163
12.1 DNA Methylation Sequencing Strategies 163
12.1.1 Whole-Genome Bisulfite Sequencing (WGBS) 164
12.1.2 Reduced Representation Bisulfite Sequencing (RRBS) ... 165
12.1.3 Methylation Sequencing Based on Methylated DNA Enrichment 165
12.1.4 Differentiation of Cytosine Methylation from Demethylation Products in Bisulfite Sequencing 166
12.2 DNA Methylation Sequencing Data Analysis 166

12.2.1 Quality Control and Preprocessing 166
12.2.2 Read Mapping 167
12.2.3 Quantification of DNA Methylation 169
12.2.4 Visualization of DNA Methylation Data 170
12.3 Detection of Differentially Methylated Cytosines or Regions ... 172
12.4 Data Verification, Validation, and Interpretation 173

13. Metagenome Analysis by Next-Generation Sequencing (NGS) 175
13.1 Experimental Design and Sample Preparation 176
13.1.1 Metagenome Sample Collection 177
13.1.2 Metagenome Sample Processing 177
13.2 Sequencing Approaches 178
13.3 Overview of Whole-Genome Shotgun (WGS) Metagenome Sequencing Data Analysis 179
13.4 Sequencing Data Quality Control and Preprocessing 181
13.5 Taxonomic Characterization of a Microbial Community 181
13.5.1 Metagenome Assembly 181
13.5.2 Sequence Binning 182
13.5.3 Calling of Open Reading Frames (ORFs) and Other Genomic Elements from Metagenomic Sequences 184
13.5.4 Phylogenetic Gene Marker Analysis 184
13.6 Functional Characterization of a Microbial Community 185
13.6.1 Gene Function Annotation 185
13.6.2 Metabolic Pathway Reconstruction 185
13.7 Comparative Metagenomic Analysis 186
13.7.1 Metagenome Sequencing Data Normalization 186
13.7.2 Identification of Differentially Abundant Species or Operational Taxonomic Units (OTUs) 187
13.8 Integrated Metagenomics Data Analysis Pipelines 187
13.9 Metagenomics Data Repositories 188

Section IV The Changing Landscape of Next-Generation Sequencing Technologies and Data Analysis

14. What Is Next for Next-Generation Sequencing (NGS)? 191
14.1 The Changing Landscape of Next-Generation Sequencing (NGS) 191
14.2 Rapid Evolution and Growth of Bioinformatics Tools for High-Throughput Sequencing Data Analysis 193
14.3 Standardization and Streamlining of NGS Analytic Pipelines... 195
14.4 Parallel Computing 195
14.5 Cloud Computing 196

Appendix A: Common File Types Used in Next-Generation Sequencing (NGS) Data Analysis 199

Appendix B: Glossary 203

References 213

Index 237