

# Contents

Preface, xiii

Author, xvii

|   |          |
|---|----------|
| <b>CHAPTER 1 ■ Sequencing and Raw Sequence Data Quality Control</b> | <b>1</b> |
| 1.1 NUCLEIC ACIDS   | 1        |
| 1.2 SEQUENCING  | 3        |
| 1.2.1 First-Generation Sequencing                                   | 3        |
| 1.2.2 Next-Generation Sequencing                                    | 4        |
| 1.2.2.1 Roche 454 Technology  | 5        |
| 1.2.2.2 Ion Torrent Technology                                      | 6        |
| 1.2.2.3 AB SOLiD Technology   | 6        |
| 1.2.2.4 Illumina Technology   | 7        |
| 1.2.3 Third-Generation Sequencing                                   | 8        |
| 1.2.3.1 PacBio Technology   | 9        |
| 1.2.3.2 Oxford Nanopore Technology                                  | 10       |
| 1.3 SEQUENCING DEPTH AND READ QUALITY                               | 11       |
| 1.3.1 Sequencing Depth  | 11       |
| 1.3.2 Base Call Quality   | 11       |
| 1.4 FASTQ FILES   | 13       |
| 1.5 FASTQ READ QUALITY ASSESSMENT                                   | 18       |
| 1.5.1 Basic Statistics  | 23       |
| 1.5.2 Per Base Sequence Quality                                     | 24       |
| 1.5.3 Per Tile Sequence Quality                                     | 25       |
| 1.5.4 Per Sequence Quality Scores                                   | 28       |
| 1.5.5 Per Base Sequence Content                                     | 28       |
| 1.5.6 Per Sequence GC Content                                       | 28       |
| 1.5.7 Per Base N Content  | 30       |

|   |  |           |
|---|--|-----------|
| 1.5.8   | Sequence Length Distribution                                     | 30        |
| 1.5.9   | Sequence Duplication Levels                                      | 31        |
| 1.5.10  | Overrepresented Sequences  | 31        |
| 1.5.11  | Adapter Content  | 32        |
| 1.5.12  | K-mer Content  | 33        |
| 1.6   | PREPROCESSING OF THE FASTQ READS                                 | 34        |
| 1.7   | SUMMARY  | 45        |
|   | REFERENCES   | 46        |
| <b>CHAPTER 2 ■ Mapping of Sequence Reads to the Reference Genomes</b> |  | <b>49</b> |
| 2.1   | INTRODUCTION TO SEQUENCE MAPPING                                 | 49        |
| 2.2   | READ MAPPING   | 55        |
| 2.2.1   | Trie   | 56        |
| 2.2.2   | Suffix Tree  | 56        |
| 2.2.3   | Suffix Arrays  | 57        |
| 2.2.4   | Burrows–Wheeler Transform  | 58        |
| 2.2.5   | FM-Index   | 62        |
| 2.3   | READ SEQUENCE ALIGNMENT AND ALIGNERS                             | 63        |
| 2.3.1   | SAM and BAM File Formats   | 65        |
| 2.3.2   | Read Aligners  | 70        |
|   | 2.3.2.1 <i>Burrows–Wheeler Aligner</i>                           | 71        |
|   | 2.3.2.2 <i>Bowtie2</i>   | 75        |
|   | 2.3.2.3 <i>STAR</i>  | 76        |
| 2.4   | MANIPULATING ALIGNMENTS IN SAM/BAM FILES                         | 79        |
| 2.4.1   | Samtools   | 79        |
|   | 2.4.1.1 <i>SAM/BAM Format Conversion</i>                         | 79        |
|   | 2.4.1.2 <i>Sorting Alignment</i>                                 | 80        |
|   | 2.4.1.3 <i>Indexing BAM File</i>                                 | 80        |
|   | 2.4.1.4 <i>Extracting Alignments of a Chromosome</i>             | 81        |
|   | 2.4.1.5 <i>Filtering and Counting Alignment in SAM/BAM Files</i> | 81        |
|   | 2.4.1.6 <i>Removing Duplicate Reads</i>                          | 82        |
|   | 2.4.1.7 <i>Descriptive Statistics</i>                            | 83        |
| 2.5   | REFERENCE-GUIDED GENOME ASSEMBLY                                 | 83        |
| 2.6   | SUMMARY  | 85        |
|   | REFERENCES   | 86        |

|   |            |
|---|------------|
| <b>CHAPTER 3 ■ De Novo Genome Assembly</b>                | <b>89</b>  |
| 3.1 INTRODUCTION TO DE NOVO GENOME ASSEMBLY               | 89         |
| 3.1.1 Greedy Algorithm                                    | 90         |
| 3.1.2 Overlap-Consensus Graphs                            | 90         |
| 3.1.3 De Bruijn Graphs                                    | 91         |
| 3.2 EXAMPLES OF DE NOVO ASSEMBLERS                        | 93         |
| 3.2.1 ABySS   | 93         |
| 3.2.2 SPAdes  | 97         |
| 3.3 GENOME ASSEMBLY QUALITY ASSESSMENT                    | 99         |
| 3.3.1 Statistical Assessment for Genome Assembly          | 100        |
| 3.3.2 Evolutionary Assessment for De Novo Genome Assembly | 103        |
| 3.4 SUMMARY   | 106        |
| REFERENCES  | 107        |
| <b>CHAPTER 4 ■ Variant Discovery</b>                      | <b>109</b> |
| 4.1 INTRODUCTION TO GENETIC VARIATIONS                    | 109        |
| 4.1.1 VCF File Format                                     | 110        |
| 4.1.2 Variant Calling and Analysis                        | 113        |
| 4.2 VARIANT CALLING PROGRAMS                              | 114        |
| 4.2.1 Consensus-Based Variant Callers                     | 114        |
| 4.2.1.1 <i>BCF Tools Variant Calling Pipeline</i>         | 115        |
| 4.2.2 Haplotype-Based Variant Callers                     | 125        |
| 4.2.2.1 <i>FreeBayes Variant Calling Pipeline</i>         | 127        |
| 4.2.2.2 <i>GATK Variant Calling Pipeline</i>              | 129        |
| 4.3 VISUALIZING VARIANTS                                  | 143        |
| 4.4 VARIANT ANNOTATION AND PRIORITIZATION                 | 143        |
| 4.4.1 SIFT  | 145        |
| 4.4.2 SnpEff  | 148        |
| 4.3.3 ANNOVAR   | 151        |
| 4.3.3.1 <i>Annotation Databases</i>                       | 153        |
| 4.3.3.2 <i>ANNOVAR Input Files</i>                        | 156        |
| 4.5 SUMMARY   | 160        |
| REFERENCES  | 161        |

|   |            |
|---|------------|
| <b>CHAPTER 5 ■ RNA-Seq Data Analysis</b>              | <b>163</b> |
| 5.1 INTRODUCTION TO RNA-SEQ                           | 163        |
| 5.2 RNA-SEQ APPLICATIONS                              | 165        |
| 5.3 RNA-SEQ DATA ANALYSIS WORKFLOW                    | 166        |
| 5.3.1 Acquiring RNA-Seq Data                          | 166        |
| 5.3.2 Read Mapping                                    | 167        |
| 5.3.3 Alignment Quality Assessment                    | 171        |
| 5.3.4 Quantification                                  | 172        |
| 5.3.5 Normalization                                   | 174        |
| 5.3.5.1 <i>RPKM and FPKM</i>                          | 174        |
| 5.3.5.2 <i>Transcripts per Million</i>                | 175        |
| 5.3.5.3 <i>Counts per Million Mapped Reads</i>        | 175        |
| 5.3.5.4 <i>Trimmed Mean of M-values</i>               | 175        |
| 5.3.5.5 <i>Relative Expression</i>                    | 176        |
| 5.3.5.6 <i>Upper Quartile</i>                         | 176        |
| 5.3.6 Differential Expression Analysis                | 176        |
| 5.3.7 Using EdgeR for Differential Analysis           | 180        |
| 5.3.7.1 <i>Data Preparation</i>                       | 181        |
| 5.3.7.2 <i>Annotation</i>                             | 183        |
| 5.3.7.3 <i>Design Matrix</i>                          | 184        |
| 5.3.7.4 <i>Filtering Low-Expressed Genes</i>          | 185        |
| 5.3.7.5 <i>Normalization</i>                          | 186        |
| 5.3.7.6 <i>Estimating Dispersions</i>                 | 186        |
| 5.3.7.7 <i>Exploring the Data</i>                     | 189        |
| 5.3.7.8 <i>Model Fitting</i>                          | 194        |
| 5.3.7.9 <i>Ontology and Pathways</i>                  | 202        |
| 5.3.8 Visualizing RNA-Seq Data                        | 204        |
| 5.3.8.1 <i>Visualizing Distribution with Boxplots</i> | 206        |
| 5.3.8.2 <i>Scatter Plot</i>                           | 207        |
| 5.3.8.3 <i>Mean-Average Plot (MA Plot)</i>            | 208        |
| 5.3.8.4 <i>Volcano Plots</i>                          | 209        |
| 5.4 SUMMARY   | 209        |
| REFERENCES  | 211        |

|  |            |
|--|------------|
| <b>CHAPTER 6 ■ Chromatin Immunoprecipitation Sequencing</b>                    | <b>213</b> |
| 6.1 INTRODUCTION TO CHROMATIN IMMUNOPRECIPITATION                              | 213        |
| 6.2 CHIP SEQUENCING  | 214        |
| 6.3 CHIP-SEQ ANALYSIS WORKFLOW   | 215        |
| 6.3.1 Downloading the Raw Data   | 217        |
| 6.3.2 Quality Control  | 218        |
| 6.3.3 ChIP-Seq and Input Read Mapping  | 219        |
| 6.3.4 ChIP-Seq Peak Calling with MACS3   | 223        |
| 6.3.5 Visualizing ChIP-Seq Enrichment in Genome Browser                        | 226        |
| 6.3.6 Visualizing Peaks Distribution   | 229        |
| 6.3.6.1 <i>ChIP-Seq Peaks' Coverage Plot</i>                                   | 230        |
| 6.3.6.2 <i>Distribution of Peaks in Transcription Start Site (TSS) Regions</i> | 233        |
| 6.3.6.3 <i>Profile of Peaks along Gene Regions</i>                             | 234        |
| 6.3.7 Peak Annotation  | 235        |
| 6.3.7.1 <i>Writing Annotations to Files</i>                                    | 237        |
| 6.3.8 ChIP-Seq Functional Analysis   | 239        |
| 6.3.9 Motif Discovery  | 243        |
| 6.4 SUMMARY  | 250        |
| REFERENCES   | 251        |
| <b>CHAPTER 7 ■ Targeted Gene Metagenomic Data Analysis</b>                     | <b>253</b> |
| 7.1. INTRODUCTION TO METAGENOMICS  | 253        |
| 7.2 ANALYSIS WORKFLOW  | 254        |
| 7.2.1 Raw Data Preprocessing   | 254        |
| 7.2.2 Metagenomic Features   | 255        |
| 7.2.2.1 <i>Clustering</i>  | 255        |
| 7.2.2.2 <i>Denoising</i>   | 256        |
| 7.2.3 Taxonomy Assignment  | 258        |
| 7.2.3.1 <i>Basic Local Alignment Search Tool</i>                               | 258        |
| 7.2.3.2 <i>VSEARCH</i>   | 259        |
| 7.2.3.3 <i>Ribosomal Database Project</i>                                      | 259        |
| 7.2.4 Construction of Phylogenetic Trees                                       | 260        |
| 7.2.5 Microbial Diversity Analysis   | 261        |
| 7.2.5.1 <i>Alpha Diversity Indices</i>   | 262        |
| 7.2.5.2 <i>Beta Diversity</i>  | 262        |

|  |   |            |
|--|---|------------|
| 7.3  | DATA ANALYSIS WITH QIIME2                             | 263        |
| 7.3.1  | QIIME2 Input Files                                    | 265        |
| 7.3.1.1  | <i>Importing Sequence Data</i>                        | 265        |
| 7.3.1.2  | <i>Metadata</i>                                       | 269        |
| 7.3.2  | Demultiplexing  | 269        |
| 7.3.3  | Downloading and Preparing the Example Data            | 271        |
| 7.3.3.1  | <i>Downloading the Raw Data</i>                       | 271        |
| 7.3.3.2  | <i>Creating the Sample Metadata File</i>              | 272        |
| 7.3.3.3  | <i>Importing Microbiome Yoga Data</i>                 | 274        |
| 7.3.4  | Raw Data Preprocessing                                | 275        |
| 7.3.4.1  | <i>Quality Assessment and Quality Control</i>         | 275        |
| 7.3.4.2  | <i>Clustering and Denoising</i>                       | 278        |
| 7.3.5  | Taxonomic Assignment with QIIME2                      | 289        |
| 7.3.5.1  | <i>Using Alignment-Based Classifiers</i>              | 289        |
| 7.3.5.2  | <i>Using Machine Learning Classifiers</i>             | 291        |
| 7.3.6  | Construction of Phylogenetic Tree                     | 297        |
| 7.3.6.1  | <i>De Novo Phylogenetic Tree</i>                      | 297        |
| 7.3.6.2  | <i>Fragment-Insertion Phylogenetic Tree</i>           | 298        |
| 7.3.7  | Alpha and Beta Diversity Analysis                     | 298        |
| 7.4  | SUMMARY   | 300        |
|  | REFERENCES  | 301        |
| <b>CHAPTER 8 ■ Shotgun Metagenomic Data Analysis</b> |   | <b>303</b> |
| 8.1  | INTRODUCTION  | 303        |
| 8.2  | SHOTGUN METAGENOMIC ANALYSIS WORKFLOW                 | 305        |
| 8.2.1  | Data Acquisition                                      | 305        |
| 8.2.2  | Quality Assessment and Processing                     | 305        |
| 8.2.3  | Removing Host DNA Reads                               | 306        |
| 8.2.3.1  | <i>Download Human Reference Genome</i>                | 306        |
| 8.2.3.2  | <i>Mapping Reads to the Reference Genome</i>          | 307        |
| 8.2.3.3  | <i>Converting SAM to BAM Format</i>                   | 307        |
| 8.2.3.4  | <i>Separating Metagenomic Reads in BAM Files</i>      | 307        |
| 8.2.3.5  | <i>Creating Paired-End FASTQ Files from BAM Files</i> | 308        |
| 8.2.4  | Assembly-Free Taxonomic Profiling                     | 310        |
| 8.2.4  | Assembly of Metagenomes                               | 315        |

|       |                                     |     |
|-------|-------------------------------------|-----|
| 8.2.5 | Assembly Evaluation                 | 317 |
| 8.2.6 | Mapping Reads to the Assemblies     | 318 |
| 8.2.7 | Binning                             | 321 |
| 8.2.8 | Bin Evaluation                      | 323 |
| 8.2.9 | Prediction of Protein-Coding Region | 324 |
| 8.3   | SUMMARY                             | 325 |
|       | REFERENCES                          | 326 |

## INDEX, 327