CONTENTS IN BRIEF

CONTRIBUTORS xix
PREFACE xxiii
ABOUT THE COMPANION WEBSITE xxv

PART I COMPUTING AND EXPERIMENTAL INFRASTRUCTURE FOR NGS 1

1 Cloud Computing for Next-Generation Sequencing Data Analysis 3
   Xuan Guo, Ning Yu, Bing Li, and Yi Pan

2 Introduction to the Analysis of Environmental Sequence Information Using Metapathways 25
   Niels W. Hanson, Kishori M. Konwar, Shang-Ju Wu, and Steven J. Hallam

3 Pooling Strategy for Massive Viral Sequencing 57
   Pavel Skums, Alexander Artyomenko, Olga Glebova, Sumathi Ramachandran, David S. Campo, Zoya Dimitrova, Ion I. Măndoiu, Alexander Zelikovsky, and Yury Khudyakov

4 Applications of High-Fidelity Sequencing Protocol to RNA Viruses 85
   Serghei Mangul, Nicholas C. Wu, Ekaterina Nenastyeva, Nicholas Mancuso, Alexander Zelikovsky, Ren Sun, and Eleazar Eskin
### PART II  GENOMICS AND EPIGENOMICS

5  **Scaffolding Algorithms**  
   *Igor Mandric, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky*  
   Page 107

6  **Genomic Variants Detection and Genotyping**  
   *Jorge Duitama*  
   Page 133

7  **Discovering and Genotyping Twilight Zone Deletions**  
   *Tobias Marschall and Alexander Schönhuth*  
   Page 149

8  **Computational Approaches for Finding Long Insertions and Deletions with NGS Data**  
   *Jin Zhang, Chong Chu, and Yufeng Wu*  
   Page 175

9  **Computational Approaches in Next-Generation Sequencing Data Analysis for Genome-Wide DNA Methylation Studies**  
   *Jeong-Hyeon Choi and Huidong Shi*  
   Page 197

10 **Bisulfite-Conversion-Based Methods for DNA Methylation Sequencing Data Analysis**  
   *Elena Harris and Stefano Lonardi*  
   Page 227

### PART III  TRANSCRIPTOMICS

11 **Computational Methods for Transcript Assembly from RNA-SEQ Reads**  
   *Stefan Canzar and Liliana Florea*  
   Page 247

12 **An Overview And Comparison of Tools for RNA-Seq Assembly**  
   *Rasiah Loganantharaj and Thomas A. Randall*  
   Page 269

13 **Computational Approaches for Studying Alternative Splicing in Nonmodel Organisms From RNA-SEQ Data**  
   *Sing-Hoi Sze*  
   Page 287

14 **Transcriptome Quantification and Differential Expression From NGS Data**  
   *Olga Glebova, Yvette Temate-Tiagueu, Adrian Caciula, Sahar Al Seesi, Alexander Artyomenko, Serghei Mangul, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky*  
   Page 301
CONTENTS IN BRIEF

PART IV  MICROBIOMICS

15  Error Correction of NGS Reads from Viral Populations  
Pavel Skums, Alexander Artyomenko, Olga Glebova, David S. Campo,  
Zoya Dimitrova, Alexander Zelikovsky, and Yury Khudyakov  

16  Probabilistic Viral Quasispecies Assembly  
Armin Töpfer and Niko Beerenwinkel  

17  Reconstruction of Infectious Bronchitis Virus Quasispecies  
from NGS Data  
Bassam Tork, Ekaterina Nenastyeva, Alexander Artyomenko,  
Nicholas Mancuso, Mazhar I. Khan, Rachel O’Neill, Ion I. Măndoiu,  
and Alexander Zelikovsky  

18  Microbiome Analysis: State of the Art and Future Trends  
Mitch Fernandez, Vanessa Aguiar-Pulido, Juan Riveros, Wenrui Huang,  
Jonathan Segal, Erliang Zeng, Michael Campos, Kalai Mathee,  
and Giri Narasimhan  

INDEX
CONTENTS

CONTRIBUTORS xix

PREFACE xxiii

ABOUT THE COMPANION WEBSITE xxv

PART I COMPUTING AND EXPERIMENTAL INFRASTRUCTURE FOR NGS 1

1 Cloud Computing for Next-Generation Sequencing Data Analysis 3
Xuan Guo, Ning Yu, Bing Li, and Yi Pan

1.1 Introduction, 3
1.2 Challenges for NGS Data Analysis, 4
1.3 Background For Cloud Computing and its Programming Models, 6
  1.3.1 Overview of Cloud Computing, 7
  1.3.2 Cloud Service Providers, 7
  1.3.3 Programming Models, 8
1.4 Cloud Computing Services for NGS Data Analysis, 13
  1.4.1 Hardware as a Service (HaaS), 13
  1.4.2 Platform as a Service (PaaS), 13
  1.4.3 Software as a Service (SaaS), 15
  1.4.4 Data as a Service (DaaS), 20
1.5 Conclusions and Future Directions, 20

References, 21
2 Introduction to the Analysis of Environmental Sequence Information Using Metapathways 25
Niels W. Hanson, Kishori M. Konwar, Shang-Ju Wu, and Steven J. Hallam

2.1 Introduction & Overview, 25
2.2 Background, 26
2.3 Metapathways Processes, 27
2.3.1 Open Reading Frame (ORF) Prediction, 29
2.3.2 Functional Annotation, 32
2.3.3 Analysis Modules, 33
2.3.4 ePGDB Construction, 38
2.4 Big Data Processing, 39
2.4.1 A Master–Worker Model for Grid Distribution, 39
2.4.2 GUI and Data Integration, 41
2.5 Downstream Analyses, 41
2.5.1 Large Table Comparisons, 43
2.5.2 Pathway Tools Cellular Overview, 43
2.5.3 Statistical Analysis with R, 45
2.5.4 Venn Diagram, 47
2.5.5 Clustering and Relating Samples by Pathways, 49
2.5.6 Faceting Variables with ggplot2, 49
2.6 Conclusions, 50

References, 50

3 Pooling Strategy for Massive Viral Sequencing 57
Pavel Skums, Alexander Artymenko, Olga Glebova, Sumathi Ramachandran, David S. Campo, Zoya Dimitrova, Ion I. Mândoiu, Alexander Zelikovsky, and Yury Khudyakov

3.1 Introduction, 57
3.2 Design of Pools for Big Viral Data, 60
3.2.1 Pool Design Optimization Formulation, 62
3.2.2 Greedy Heuristic for VSPD Problem, 63
3.2.3 The Tabu Search Heuristic for the OCBG Problem, 65
3.3 Deconvolution of Viral Samples from Pools, 68
3.3.1 Deconvolution Using Generalized Intersections and Differences of Pools, 68
3.3.2 Maximum Likelihood k-Clustering, 70
3.4 Performance of Pooling Methods on Simulated Data, 71
3.4.1 Performance of the Viral Sample Pool Design Algorithm, 71
3.4.2 Performance of the Pool Deconvolution Algorithm, 74
3.5 Experimental Validation of Pooling Strategy, 75
3.5.1 Experimental Pools and Sequencing, 75
3.5.2 Results, 77
3.6 Conclusion, 79
References, 81

4 Applications of High-Fidelity Sequencing Protocol to RNA Viruses 85
Serghei Mangul, Nicholas C. Wu, Ekaterina Nenastyeva, Nicholas Mancuso, Alexander Zelikovsky, Ren Sun, and Eleazar Eskin

4.1 Introduction, 85
4.2 High-Fidelity Sequencing Protocol, 86
4.3 Assembly of High-Fidelity Sequencing Data, 88
  4.3.1 Consensus Construction, 88
  4.3.2 Reads Mapping, 89
  4.3.3 Viral Genome Assembler (VGA), 89
  4.3.4 Viral Population Quantification, 91
4.4 Performance of VGA on Simulated Data, 93
4.5 Performance of Existing Viral Assemblers on Simulated Consensus Error-Corrected Reads, 98
4.6 Performance of VGA on Real Hiv Data, 99
  4.6.1 Validation of de novo Consensus, 99
4.7 Comparison of Alignment on Error-Corrected Reads, 100
4.8 Evaluating of Error Correction Tools Based on High-Fidelity Sequencing Reads, 101
Acknowledgment, 101
References, 102

PART II GENOMICS AND EPIGENOMICS 105

5 Scaffolding Algorithms 107
Igor Mandric, James Lindsay, Ion I. Măndoiu, and Alexander Zelikovsky

5.1 Scaffolding, 107
5.2 State-of-The-Art Scaffolding Tools, 108
  5.2.1 Sspace, 108
  5.2.2 OPERA, 109
  5.2.3 SOPRA, 110
  5.2.4 MIP, 110
  5.2.5 SCARPA, 111
5.3 Recent Scaffolding Tools, 111
  5.3.1 SILP2, 111
  5.3.2 ScaffMatch, 119
5.4 Scaffolding Software Evaluation, 124
  5.4.1 Data Sets, 124
  5.4.2 Quality Metrics, 124
  5.4.3 Evaluation and Comparison, 126
References, 129
6 Genomic Variants Detection and Genotyping 133
Jorge Duitama

6.1 Introduction, 133
6.2 Methods for Detection and Genotyping of SNPs and Small Indels, 135
   6.2.1 Description of the Problem, 135
   6.2.2 Bayesian Model, 136
   6.2.3 Common Issues Affecting Genotype Quality, 137
   6.2.4 Population Variability, 139
6.3 Methods for Detection and Genotyping of CNVs, 141
   6.3.1 Mean-Shift Approach for CNVs within a Sample, 141
   6.3.2 Identifying CNVs between Samples, 143
6.4 Putting Everything Together, 144
References, 145

7 Discovering and Genotyping Twilight Zone Deletions 149
Tobias Marschall and Alexander Schönhuth

7.1 Introduction, 149
   7.1.1 Twilight Zone Deletions, 151
7.2 Notation, 151
   7.2.1 Alignments, 152
   7.2.2 Gaps/Splits, 152
   7.2.3 Deletions, 152
7.3 Non-Twilight-Zone Deletion Discovery, 152
   7.3.1 Internal Segment Size-Based Approaches, 153
   7.3.2 Split-Read Mapping Approaches, 154
   7.3.3 Hybrid Approaches, 155
   7.3.4 The “Twilight Zone”: Definition, 156
7.4 Discovering “Twilight Zone” Deletions: New Solutions, 156
   7.4.1 CLEVER, 156
   7.4.2 Mate-Clever, 157
   7.4.3 Pindel, 157
7.5 Genotyping “Twilight Zone” Deletions, 158
   7.5.1 A Maximum Likelihood Approach under Read Alignment Uncertainty, 158
7.6 Results, 162
   7.6.1 Data Set, 162
   7.6.2 Tools, 163
   7.6.3 Discovery, 163
   7.6.4 Genotyping, 166
7.7 Discussion, 167
   7.7.1 HiSeq, 169
   7.7.2 MiSeq, 170
   7.7.3 Conclusion, 170
8 Computational Approaches for Finding Long Insertions and Deletions with NGS Data

Jin Zhang, Chong Chu, and Yufeng Wu

8.1 Background, 175
8.2 Methods, 177
8.2.1 Signatures of Long Indels in Sequence Reads, 177
8.2.2 Methods for Discovering Long Indels without Exact Breakpoints, 183
8.2.3 Methods for Discovering Long Indels with Exact Breakpoints, 185
8.2.4 Combined Approaches, 186
8.3 Applications, 191
8.3.1 Population SV Calling, 191
8.3.2 Cancer Genomics, 192
8.4 Conclusions and Future Directions, 193
Acknowledgment, 193
References, 193

9 Computational Approaches in Next-Generation Sequencing Data Analysis for Genome-Wide DNA Methylation Studies

Jeong-Hyeon Choi and Huidong Shi

9.1 Introduction, 197
9.2 Enrichment-Based Approaches, 201
9.2.1 Data Analysis Procedure, 201
9.2.2 Available Approaches, 205
9.3 Bisulfite Treatment-Based Approaches, 211
9.3.1 Data Analysis Procedure, 211
9.3.2 Available Approaches, 214
9.4 Conclusion, 221
References, 222

10 Bisulfite-Conversion-Based Methods for DNA Methylation Sequencing Data Analysis

Elena Harris and Stefano Lonardi

10.1 Introduction, 227
10.2 The Problem of Mapping BS-Treated Reads, 229
10.3 Algorithmic Approaches to the Problem of Mapping BS-Treated Reads, 231
10.4 Methylation Estimation, 234
10.5 Possible Biases in Estimation of Methylation Level, 234
10.6 Bisulfite Conversion Rate, 235
10.7 Reduced Representation Bisulfite Sequencing, 235
10.8 Accuracy as a Performance Measurement, 235
References, 241

PART III  TRANSCRIPTOMICS 245

11 Computational Methods for Transcript Assembly from RNA-SEQ Reads 247
Stefan Canzar and Liliana Florea

11.1 Introduction, 247
11.2 De Novo Assembly, 248
   11.2.1 Preprocessing of Reads, 249
   11.2.2 The De Bruijn Graph for RNA-seq Read Assembly, 250
   11.2.3 Contig Assembly, 252
   11.2.4 Filtering and Error Correction, 252
   11.2.5 Variations, 253
11.3 Genome-Based Assembly, 254
   11.3.1 Candidate Isoforms, 254
   11.3.2 Minimality, 258
   11.3.3 Accuracy, 261
   11.3.4 Completeness, 263
   11.3.5 Extensions, 264
11.4 Conclusions, 264
   Acknowledgment, 265
   References, 265

12 An Overview And Comparison of Tools for RNA-Seq Assembly 269
Rasiah Loganantharaj and Thomas A. Randall

12.1 Quality Assessment, 271
12.2 Experimental Considerations, 272
12.3 Assembly, 275
12.4 Experiment, 276
12.5 Comparison, 278
12.6 Results, 279
12.7 Summary and Conclusion, 280
   Acknowledgments, 284
   References, 284
13 Computational Approaches for Studying Alternative Splicing in Nonmodel Organisms From RNA-SEQ Data

Sing-Hoi Sze

13.1 Introduction, 287
  13.1.1 Alternative Splicing, 287
  13.1.2 Nonmodel Organisms, 288
  13.1.3 RNA-Seq Data, 288
13.2 Representation of Alternative Splicing, 289
  13.2.1 de Bruijn Graph, 289
  13.2.2 A Set of Transcripts, 291
  13.2.3 Splicing Graph, 292
13.3 Comparison to Model Organisms, 293
  13.3.1 A Set of Transcripts, 293
  13.3.2 Splicing Graph, 293
13.4 Accuracy of Algorithms, 293
  13.4.1 Assembly Results, 293
  13.4.2 mRNA BLAST Results, 295
  13.4.3 Alternative Splicing Junctions, 295
13.5 Discussion, 296
References, 297

14 Transcriptome Quantification and Differential Expression From NGS Data

Olga Glebova, Yvette Temate-Tiagueu, Adrian Caciula, Sahar Al Seesi, Alexander Artyomenko, Serghei Mangul, James Lindsay, Ion I. Mândoiu, and Alexander Zelikovsky

14.1 Introduction, 301
  14.1.1 Motivation and Problems Description, 302
  14.1.2 RNA-Seq Protocol, 303
14.2 Overview of the State-of-the-Art Methods, 304
  14.2.1 Quantification Methods, 304
  14.2.2 Differential Expression Methods, 305
14.3 Recent Algorithms, 307
  14.3.1 SimReg: Simulated Regression Method for Transcriptome Quantification, 307
  14.3.2 Differential Gene Expression Analysis: IsoDE, 311
14.4 Experimental Setup, 313
  14.4.1 Quantification Methods, 313
  14.4.2 Differential Expression Methods, 313
14.5 Evaluation, 316
  14.5.1 Transcriptome Quantification Methods Evaluation, 316
14.5.2 Differential Expression Methods Evaluation, 319
Acknowledgments, 326
References, 326

PART IV MICROBIOMICS

15 Error Correction of NGS Reads from Viral Populations
Pavel Skums, Alexander Artyomenko, Olga Glebova, David S. Campo, Zoya Dimitrova, Alexander Zelikovsky, and Yury Khudyakov

15.1 Next-Generation Sequencing of Heterogeneous Viral Populations and Sequencing Errors, 331
15.2 Methods and Algorithms for The Ngs Error Correction in Viral Data, 334
15.2.1 Clustering-Based Algorithms, 334
15.2.2 k-Mer-Based Algorithms, 339
15.2.3 Alignment-Based Algorithms, 345
15.3 Algorithm Comparison, 347
15.3.1 Benchmark Data, 347
15.3.2 Results and Discussion, 348
References, 350

16 Probabilistic Viral Quasispecies Assembly
Armin Töpfer and Niko Beerenwinkel

16.1 Intra-Host Virus Populations, 355
16.1.1 Viral Quasispecies, 356
16.1.2 Fitness, 357
16.1.3 HIV-1 as a Model System, 357
16.1.4 Recombination, 358
16.1.5 Clinical Implications, 359
16.1.6 Genotyping, 360
16.2 Next-Generation Sequencing for Viral Genomics, 360
16.2.1 Library Preparation, 360
16.2.2 Sequencing Approaches, 361
16.2.3 Specialized Viral Sequencing Methods, 363
16.2.4 Data Preprocessing and Read Alignment, 364
16.2.5 Spatial Scales of Viral Haplotype Reconstruction, 364
16.2.6 Quasispecies Assembly Performance, 365
16.3 Probabilistic Reconstruction Methods, 366
16.3.1 From Human to Viral Haplotype Reconstruction, 366
16.3.2 Viral Haplotype Inference Methods Overview, 369
16.3.3 Local Viral Haplotype Inference Approaches, 369
16.3.4 Quasispecies Assembly, 370
16.3.5 Recombinant Quasispecies Assembly, 370
16.4 Conclusion, 375
References, 376
17  Reconstruction of Infectious Bronchitis Virus Quasispecies from NGS Data 383

Bassam Tork, Ekaterina Nenastyeva, Alexander Artyomenko, Nicholas Mancuso, Mazhar I. Khan, Rachel O’Neill, Ion I. Măndoiu, and Alexander Zelikovsky

17.1 Introduction, 383
17.2 Background, 384
  17.2.1 Infectious Bronchitis Virus, 384
  17.2.2 High-Throughput Sequencing, 384
17.3 Methods, 385
  17.3.1 Compared Methods, 388
17.4 Results and Discussion, 388
  17.4.1 Data Sets, 388
  17.4.2 Validation of Error Correction Methods, 389
  17.4.3 Tuning, Comparison, and Validation of Methods for Quasispecies Reconstruction, 390

Acknowledgments, 397
References, 397

18  Microbiome Analysis: State of the Art and Future Trends 401

Mitch Fernandez, Vanessa Aguiar-Pulido, Juan Riveros, Wenrui Huang, Jonathan Segal, Erliang Zeng, Michael Campos, Kalai Mathee, and Giri Narasimhan

18.1 Introduction, 401
18.2 The Metagenomics Analysis Pipeline, 403
18.3 Data Limitations and Sources of Errors, 405
  18.3.1 Designing Degenerate Primers for Microbiome Work, 407
18.4 Diversity and Richness Measures, 407
18.5 Correlations and Association Rules, 409
18.6 Microbial Functional Profiles, 410
18.7 Microbial Social Interactions and Visualizations, 413
18.8 Bayesian Inferences, 418
18.9 Conclusion, 419

References, 420

INDEX 425