Contents

Preface XIX
List of Contributors XXI

Part One Tag-Based Nucleic Acid Analysis 1

1 DeepSuperSAGE: High-Throughput Transcriptome Sequencing with Now- and Next-Generation Sequencing Technologies 3
Hideo Matsumura, Carlos Molina, Detlev H. Krüger, Ryohei Terauchi, and Günter Kahl
1.1 Introduction 3
1.2 Overview of the Protocols 5
1.2.1 Principle of the SuperSAGE Method 5
1.2.2 Power of the SuperSAGE Tag 5
1.2.3 Development of DeepSuperSAGE 6
1.2.4 Ditag-Based DeepSuperSAGE (for 454 Pyrosequencing) 7
1.2.5 Single-Tag-Based DeepSuperSAGE (HT-SuperSAGE) 8
1.3 Methods and Protocols 9
1.3.1 Linker or Adapter Preparation 9
1.3.2 RNA Samples 10
1.3.3 cDNA Synthesis and NlaIII Digestion 10
1.3.4 Tag Extraction from cDNA 10
1.3.5 Tag Extraction from cDNA 11
1.3.6 Purification of Linker–Tag Fragments 12
1.3.7 Ditag or Adapter–Tag Formation and Amplification 12
1.3.8 Preparation of Templates for Sequencing 14
1.4 Applications 14
1.4.1 Applications of DeepSuperSAGE in Combination with 454 Pyrosequencing 14
1.4.2 Practical Analysis of HT-SuperSAGE 18
1.5 Perspectives 19
References 20

2 DeepCAGE: Genome-Wide Mapping of Transcription Start Sites 23
Matthias Harbers, Mitchell S. Dushay, and Piero Carninci
2.1 Introduction 23
2.2 What is CAGE? 24
2.3 Why CAGE? 26
2.4 Methods and Protocols 28
2.4.1 Key Reagents and Consumables 28
2.4.2 Precautions 30
2.4.3 RNA Samples Used for DeepCAGE Library Preparation 30
2.4.4 DeepCAGE Library Preparation 32
2.5 Applications 43
2.6 Perspectives 44
References 45

3 Definition of Promotome–Transcriptome Architecture Using CAGEscan 47
Nicolas Bertin, Charles Plessy, Piero Carninci, and Matthias Harbers
3.1 Introduction 47
3.2 What is CAGEscan? 48
3.3 Why CAGEscan? 50
3.4 Methods and Protocols 51
3.4.1 Key Reagents and Consumables 51
3.4.2 Precautions 53
3.4.3 RNA Samples Used for CAGEscan Library Preparation 53
3.4.4 Considerations on Pooling CAGEscan Libraries 54
3.4.5 CAGEscan Library Preparation 54
3.5 Applications and Perspectives 59
References 61

4 RACE: New Applications of an Old Method to Connect Exons 63
Charles Plessy
4.1 Introduction 63
4.2 Deep-RACE 65
4.2.1 Choice of the Sequencer 65
4.2.2 Validation of Promoter Studies 65
4.2.3 Other Applications of Deep-RACE 66
4.2.4 Limitations of Deep-RACE 66
4.3 Methods Outline 67
4.3.1 Primer Design 67
4.3.2 Molecular Biology of Deep-RACE Library Preparation 67
4.3.3 Sequencing of Deep-RACE Libraries 68
4.3.4 Analysis 68
4.4 Perspectives 70
References 71

5 RNA-PET: Full-Length Transcript Analysis Using 5’- and 3’-Paired-End Tag Next-Generation Sequencing 73
Xiaoan Ruan and Yijun Ruan
5.1 Introduction 73
5.2 Methods and Protocols 75
5.2.1 Key Reagents and Consumables 75
5.2.2 Protocol 78
5.3 Applications 88
5.3.1 PET Sequencing with SOLiD 88
5.3.2 Mapping of the PETs 88
5.3.3 PET Clustering, Annotation, and Genome Browser Visualization 89
5.4 Perspectives 90
References 90

6 Stranded RNA-Seq: Strand-Specific Shotgun Sequencing of RNA 91
Alistair R.R. Forrest
6.1 Introduction 91
6.1.1 Before Starting 93
6.2 Methods and Protocols 93
6.2.1 Preface 93
6.2.2 Materials and Consumables 94
16.2.3 Computational Mapping Analysis of Experimental Ditags 282
16.3 Applications 283
16.3.1 Analyzing Normal Genome Structure 283
16.3.2 Identifying Somatic Rearrangements in Cancer Genomes 283
16.3.3 A Useful Tool to Study Family Germline Genetic Disorders 284
16.4 Perspectives 284

References 285

17 Next-Generation Sequencing of Bacterial Artificial Chromosome Clones for Next-Generation Physical Mapping 287

Robert Bogden, Keith Stormo, Jason Dobry, Amy Mraz, Quanzhou Tao, Michiel van Eijk, Jan van Oeveren, Marcel Prins, Jon Wittendorp, and Mark van Haaren

17.1 History of the Bacterial Artificial Chromosome Vector Systems 287
17.2 History of Physical Mapping 288
17.3 What is WGP? 289
17.4 Flow of a WGP Project 289
17.5 BAC Pooling Strategies 290
17.6 Methods and Protocols 291
17.6.1 BAC Library and Pooling Strategy 291
17.6.2 Sample Preparation for Illumina Sequencing 292
17.6.3 Illumina Sequencing 293
17.6.4 Deconvolution to Assign the BAC Address to Each Read 293
17.6.5 Contig Building 293
17.7 Applications 294
17.7.1 Results from Real WGP Projects Performed by the Authors 294
17.7.2 Reorganizing Project Funding and Sequencing Budgets 295
17.7.3 Unleash the Power of BAC Clones 296
17.8 Perspectives 296

References 297

18 HELP-Tagging: Tag-Based Genome-Wide Cytosine Methylation Profiling 299

Masako Suzuki and John M. Greally

18.1 Introduction 299
18.2 Genome-Wide DNA Methylation Analysis 299
18.3 What is HELP-Tagging? 300
18.3.1 When is HELP-Tagging the Preferred Cytosine Methylation Assay? 301
18.4 Methods and Protocols 301
18.4.1 Reagents, Materials, and Equipment 301
18.4.2 Buffers and Adapters for HELP-Tagging Library Preparation 302
18.4.3 Precautions 303
18.4.4 DNA Samples for HELP-Tagging Library Preparation 303
18.4.5 HELP-Tagging Library Preparation 304
18.4.6 Illumina Sequencing 307
18.5 Applications 308
18.6 Perspectives 308

References 309

19 Second-Generation Sequencing Library Preparation: In Vitro Tagmentation via Transposome Insertion 311

Fraz Syed

19.1 Introduction 311
19.2 Methods and Protocols 313
19.2.1 Materials 313
19.2.2 Methods 314
19.3 Perspectives 321

References 321
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>370</td>
</tr>
<tr>
<td>Total RNA-Seq Data Collection and Interpretation</td>
<td>378</td>
</tr>
<tr>
<td>Applications</td>
<td>380</td>
</tr>
<tr>
<td>References</td>
<td>381</td>
</tr>
</tbody>
</table>

**Part Three  Bioinformatics for Tag-Based Technologies  383**

<table>
<thead>
<tr>
<th>Chapter 24  Computational Infrastructure and Basic Data Analysis for Next-Generation Sequencing  385</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Sexton</td>
</tr>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Background</td>
</tr>
<tr>
<td>Getting Started with the Next-Generation Manufacturers</td>
</tr>
<tr>
<td>Infrastructure and Data Analysis</td>
</tr>
<tr>
<td>Computational Considerations</td>
</tr>
<tr>
<td>Data Dynamics</td>
</tr>
<tr>
<td>Software and Postanalysis</td>
</tr>
<tr>
<td>Staffing Requirements</td>
</tr>
<tr>
<td>Applications</td>
</tr>
<tr>
<td>Perspectives</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 25  CLC Bio Integrated Platform for Handling and Analysis of Tag Sequencing Data  393</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roald Forsberg, Søren Mønsted, and Anne-Mette Hein</td>
</tr>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Main Components and Features</td>
</tr>
<tr>
<td>Data Flow and Data Back-End</td>
</tr>
<tr>
<td>CLC Genomics Workbench</td>
</tr>
<tr>
<td>CLC Genomics Server</td>
</tr>
<tr>
<td>APIs</td>
</tr>
<tr>
<td>Acceleration of Analysis</td>
</tr>
<tr>
<td>Applications</td>
</tr>
<tr>
<td>First-Level Analysis</td>
</tr>
<tr>
<td>Import</td>
</tr>
<tr>
<td>Demultiplexing</td>
</tr>
<tr>
<td>Trim and Quality Control</td>
</tr>
<tr>
<td>Second and Third Levels – Application-Specific Steps</td>
</tr>
<tr>
<td>RNA-Seq</td>
</tr>
<tr>
<td>SmallRNA-Seq</td>
</tr>
<tr>
<td>Tag-Seq</td>
</tr>
<tr>
<td>ChIP-Seq</td>
</tr>
<tr>
<td>Fourth Level – Expression Analysis</td>
</tr>
<tr>
<td>Perspectives</td>
</tr>
<tr>
<td>References</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 26  Multidimensional Context of Sequence Tags: Biological Data Integration  407</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korbinian Grote and Thomas Werner</td>
</tr>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Methods and Strategies</td>
</tr>
<tr>
<td>Annotation Links Sequence Tags (Reads) to Biology</td>
</tr>
<tr>
<td>Application of the Methods and Strategies</td>
</tr>
<tr>
<td>Only Positive Results are Conclusive</td>
</tr>
<tr>
<td>Automatic Workflow: ChIP-seq of Peroxisome Proliferator-Activated Receptor</td>
</tr>
<tr>
<td>Perspectives</td>
</tr>
<tr>
<td>References</td>
</tr>
</tbody>
</table>
27 Experimental Design and Quality Control of Next-Generation Sequencing Experiments 417
Peter A.C. 't Hoen, Matthew S. Hestand, Judith M. Boer, Yuching Lai, Maarten van Iterson, Michiel van Galen, Henk P. Buermans, and Johan T. den Dunnen

27.1 Introduction 417
27.2 Choice of Platform 417
27.2.1 Read Length and Number of Reads 418
27.2.2 Single-End versus Paired-End Sequencing 419
27.2.3 Platform-Specific Advantages and Disadvantages 419
27.3 Sequencing Depth 420
27.3.1 Expression Profiling 420
27.3.2 ChIP-Seq: Relation Enrichment Factor and Sequencing Depth 421
27.3.3 Barcoding 422
27.4 Replicates, Randomization, and Statistical Testing 422
27.4.1 Technical and Biological Replicates 422
27.4.2 Technical Variability 423
27.4.3 Biological Replicates Increase Accuracy 424
27.4.4 Sample Size 424
27.4.5 Importance of Randomizing Samples 424
27.5 Experimental Controls 425
27.5.1 Spike-Ins 425
27.5.2 Negative Controls in ChIP-Seq Experiments 426
27.6 General Quality Assessment 427
27.6.1 Nucleotide Frequency Characteristics 428
27.6.2 Percentage Duplicate Reads 428
27.7 Platform-Specific Quality Scores 428
27.7.1 Sanger, Roche, Illumina, and SOLiD Quality Scores 429
27.7.2 Conversion and Visualization of Quality Scores 429
27.8 Quality Checks After Alignment 430
27.8.1 Percentage of Reads Aligned and Percentage in Repeat Regions 430
27.8.2 DeepSAGE: Percentage 21–22 Mers 430
27.8.3 RNA-Seq: Percentage Tags in Annotated Transcripts 430
27.8.4 miRNA Profiling: Percentage in Annotated miRNAs 430
27.8.5 ChIP-Seq: Enrichment 430
27.8.6 Correlation Measures 431
27.9 What Can Go Wrong 431
27.9.1 Sample Swaps 431
27.9.2 Contamination 431
27.10 Perspectives 432

References 432

28 UTGB Toolkit for Personalized Genome Browsers 435
Taro L. Saito, Jun Yoshimura, Budrul Ahsan, Atsushi Sasaki, Reginaldo Kurosh, and Shinichi Morishita

28.1 Introduction 435
28.2 Overview of the UTGB Toolkit 436
28.2.1 Availability of the UTGB Toolkit 438
28.3 Methods 438
28.3.1 Installation of the UTGB Toolkit 438
28.3.1.1 Prerequisites 438
28.3.1.2 Easy Installer 438
28.3.1.3 Mac OS X and Linux 438
28.3.1.4 Windows 439
28.3.2 Running the UTGB Toolkit 439
28.3.3 Viewing Help Messages 439
28.3.4 Creating a new UTGB Project 440