### 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
Contents

10 Analysis of Protein–RNA Interactions with Single-Nucleotide Resolution Using iCLIP and Next-Generation Sequencing 153
Julian König, Nicholas J. McGlincy, and Jernej Ule
10.1 Introduction 153
10.2 Procedure Overview 154
10.3 Antibody and Library Preparation Quality Controls 155
10.4 Oligonucleotide Design 156
10.5 Recent Modifications of the iCLIP Protocol 158
10.6 Troubleshooting 158
10.7 Methods and Protocols 159
References 169

11 Massively Parallel Tag Sequencing Unveils the Complexity of Marine Protistan Communities in Oxygen-Depleted Habitats 171
Virginia Edgcomb and Thorsten Stoeck
11.1 Introduction 171
11.2 Cariaco Basin 173
11.3 Framvaren Fjord 176
11.4 Comparison of Cariaco Basin to Framvaren Fjord 177
11.5 Perspectives on Interpretation of Microbial Eukaryote 454 Data 179
References 182

12 Chromatin Interaction Analysis Using Paired-End Tag Sequencing (ChIA-PET) 185
Xiaoan Ruan and Yijun Ruan
12.1 Introduction 185
12.1.1 Development of the ChIA-PET Method 186
12.1.2 Applications of the ChIA-PET Method 187
12.1.3 Experimental Design of ChIA-PET Analysis 187
12.1.3.1 ChIP Sample Preparation 187
12.1.3.2 ChIA-PET Library Construction 189
12.1.3.3 ChIA-PET Library Sequencing and Mapping 190
12.1.3.4 Control Libraries 191
12.2 Methods and Protocols 192
12.2.1 Key Reagents and Consumables 192
12.2.2 Protocol 195
12.3 Timeline 206
12.4 Anticipated Results 207
12.4.1 Verification of Sonicated Chromatin DNA Size Range 207
12.4.2 ChIP Quality Control: Yield and Enrichment 207
12.4.3 ChIA-PET Library Quality Control 207
12.4.4 ChIA-PET Sequencing and Mapping Analysis 207
12.5 Perspectives 209
References 209

13 Tag-Seq: Next-Generation Tag Sequencing for Gene Expression Profiling 211
Sorana Morrissy, Yongjun Zhao, Allen Delaney, Jennifer Asano, Noreen Dhalla, Irene Li, Helen McDonald, Pawan Pandoh, Anna-Liisa Prabhu, Angela Tam, Martin Hirst, and Marco Marra
13.1 Introduction 211
13.2 Protocol Details 212
13.3 Protocol Overview and Timeline 213
13.4 Critical Parameters and Troubleshooting 214
13.5 Methods and Protocols 215
13.5.1 Basic Protocol 1: First- and Second-Strand cDNA Synthesis for Tag-Seq Library Construction 215
13.5.2 Basic Protocol 2: Tag Generation 219
13.5.3 Basic Protocol 3: PCR and Fragment Isolation 223
13.5.4 Basic Protocol 4: Preparing the Library for Illumina Sequencing 226
13.5.5 Alternate Protocol: Amplified Tag-Seq library construction (Tag-SeqLite) 227
13.5.6 Basic Protocol 5: Data Analysis 232
13.6 Applications 239
13.7 Perspectives 240
References 241

14 Isolation of Active Regulatory Elements from Eukaryotic Chromatin Using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) 243
Paul G. Giresi and Jason D. Lieb
14.1 Introduction 243
14.2 Methods and Protocols 245
14.2.1 FAIRE Procedure 245
14.2.2 Optimization of the FAIRE Procedure 245
14.2.3 Equipment and Reagents 246
14.2.4 Detection of FAIRE DNA 250
14.2.5 High-Throughput Sequencing 252
14.3 Applications 254
14.4 Perspectives 254
References 255

15 Identification of Nucleotide Variation in Genomes Using Next-Generation Sequencing 257
Hendrik-Jan Megens and Martien A.M. Groenen
15.1 Introduction 257
15.1.1 SNP Discovery and Nucleotide Variation Assessment 259
15.1.2 Sequence and Library Preparation Strategies 259
15.2 Methods 261
15.2.1 Preprocessing of Reads 262
15.2.1.1 FASTQ Format 262
15.2.1.2 FASTQ Format – Illumina Version 262
15.2.1.3 Illumina FASTQ to Sanger FASTQ 263
15.2.1.4 ABI SOLiD- and Roche 454-Specific Formats 264
15.2.1.5 Illumina SCARF or QSEQ to FASTQ 264
15.2.1.6 Quality Evaluation 265
15.2.1.7 Handling Adapter Sequences – Linkers and Barcodes 265
15.2.1.8 Quality Trimming 266
15.2.2 Mapping Reads to a Reference Genome 266
15.2.2.1 Making Alignments Using MOSAIK 267
15.2.2.2 Making Alignments Using BWA 268
15.2.3 Variant Calling 269
15.2.3.1 SAM Format 270
15.2.3.2 Variant Calling with SAMtools 273
15.2.3.3 Variant Calling with GATK 274
15.3 Notes 275
References 275

16 DGS (Ditag Genome Scanning) – A Restriction-Based Paired-End Sequencing Approach for Genome Structural Analysis 277
Jun Chen, Yeong C. Kim, and San Ming Wang
16.1 Introduction 277
16.2 Methods and Protocols 278
16.2.1 Cloning-Based DGS Protocol 278
16.2.2 Non-Cloning-Based DGS Protocol 281
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
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 21–22Mers 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
28.3.5 Building a Genome Browser 441
28.3.6 Launching a Portable Web Server 441
28.3.7 Configuring Track Views 441
28.3.8 Adding a New Track 442
28.3.8.1 FastaTrack 442
28.3.8.2 ReadTrack 442
28.3.8.3 WigTrack 443
28.3.8.4 Adding Keyword Search 443
28.3.9 Switching Views 443
28.3.10 Publishing Your Genome Browser 443
28.3.11 Manual Installation of UTGB Toolkit (Optional) 444
28.3.11.1 Windows 444
28.3.11.2 Mac OS X and Linux 444
28.3.12 Developing Your Own Tracks 444
28.4 Applications 444
28.4.1 Portable Web Server for Quickly Browsing Local Resources 444
28.4.2 Portable Database Engine 445
28.4.3 Web Application Development Framework 446
28.4.3.1 Server-Side Programming Support 446
28.4.3.2 Web Action 446
28.4.3.3 Database Connection 446
28.4.3.4 Object–Database Mapping 446
28.5 Perspectives 447
References 447

29 Beyond the Pipelines: Cloud Computing Facilitates Management, Distribution, Security, and Analysis of High-Speed Sequencer Data 449
Boris Umylny and Richard S.J. Weisburd
29.1 Introduction 449
29.2 Data Management 450
29.2.1 Data Quantity 450
29.2.2 HSSs 451
29.2.3 Data Analysis 452
29.2.4 Data Size 453
29.3 Distribution 454
29.3.1 Collaboration 454
29.3.2 Distribution of Data, Annotations, and Analysis Tools 455
29.4 Analysis 456
29.4.1 Integrating Data Repositories and Analytics 456
29.4.2 Integrating HSS Discovery Pipelines with Annotation Data 458
29.4.3 Integrating HSS and Traditional Analysis Algorithms 459
29.4.4 Cloud-Based Infrastructure 461
29.5 Security 462
29.6 Healthcare Data and Privacy Issues 464
29.7 Sample Evaluation of a Vendor Solution 465
29.8 Perspectives 465
References 467

30 Computational Methods for the Identification of MicroRNAs from Small RNA Sequencing Data 469
Eugene Berezikov
30.1 Introduction 469
30.2 Implementing the miR-Intess Pipeline 470
30.2.1 Preprocessing of Small RNA Sequencing Data 470
30.2.2 Mapping of Small RNA Reads to the Genome 471
30.2.3 Annotation 471
30.2.4 Identification of Hairpin Structures 472