
Contents

Preface, xvii

Acknowledgments, xxi

Authors, xxiii

CHAPTER 1 ■ Introduction to RNA-seq	1
1.1 INTRODUCTION	1
1.2 ISOLATION OF RNAs	3
1.3 QUALITY CONTROL OF RNA	4
1.4 LIBRARY PREPARATION	6
1.5 MAJOR RNA-SEQ PLATFORMS	9
1.5.1 Illumina	9
1.5.2 SOLID	10
1.5.3 Roche 454	11
1.5.4 Ion Torrent	11
1.5.5 Pacific Biosciences	12
1.5.6 Nanopore Technologies	13
1.6 RNA-SEQ APPLICATIONS	14
1.6.1 Protein Coding Gene Structure	14
1.6.2 Novel Protein-Coding Genes	16
1.6.3 Quantifying and Comparing Gene Expression	16
1.6.4 Expression Quantitative Trait Loci (eQTL)	17
1.6.5 Single-Cell RNA-seq	18
1.6.6 Fusion Genes	18

1.6.7	Gene Variations	19
1.6.8	Long Noncoding RNAs	19
1.6.9	Small Noncoding RNAs (miRNA-seq)	20
1.6.10	Amplification Product Sequencing (Ampli-seq)	20
1.7	CHOOSING AN RNA-SEQ PLATFORM	21
1.7.1	Eight General Principles for Choosing an RNA-seq Platform and Mode of Sequencing	21
1.7.1.1	<i>Accuracy: How Accurate Must the Sequencing Be?</i>	21
1.7.1.2	<i>Reads: How Many Do I Need?</i>	22
1.7.1.3	<i>Length: How Long Must the Reads Be?</i>	23
1.7.1.4	<i>SR or PE: Single Read or Paired End?</i>	23
1.7.1.5	<i>RNA or DNA: Am I Sequencing RNA or DNA?</i>	23
1.7.1.6	<i>Material: How Much Sample Material Do I Have?</i>	24
1.7.1.7	<i>Costs: How Much Can I Spend?</i>	24
1.7.1.8	<i>Time: When Does the Work Need to Be Completed?</i>	24
1.7.2	Summary	25
	REFERENCES	25
CHAPTER 2	■ Introduction to RNA-seq Data Analysis	27
2.1	INTRODUCTION	27
2.2	DIFFERENTIAL EXPRESSION ANALYSIS WORKFLOW	30
2.2.1	Step 1: Quality Control of Reads	31
2.2.2	Step 2: Preprocessing of Reads	31
2.2.3	Step 3: Aligning Reads to a Reference Genome	31
2.2.4	Step 4: Genome-Guided Transcriptome Assembly	32
2.2.5	Step 5: Calculating Expression Levels	32
2.2.6	Step 6: Comparing Gene Expression between Conditions	33
2.2.7	Step 7: Visualization of Data in Genomic Context	33

2.3	DOWNSTREAM ANALYSIS	34
2.3.1	Gene Annotation	34
2.3.2	Gene Set Enrichment Analysis	34
2.4	AUTOMATED WORKFLOWS AND PIPELINES	35
2.5	HARDWARE REQUIREMENTS	35
2.6	FOLLOWING THE EXAMPLES IN THE BOOK	36
2.6.1	Using Command Line Tools and R	36
2.6.2	Using the Chipster Software	37
2.6.3	Example Data Sets	39
2.7	SUMMARY	40
	REFERENCES	40
CHAPTER 3 ■ Quality Control and Preprocessing		41
3.1	INTRODUCTION	41
3.2	SOFTWARE FOR QUALITY CONTROL AND PREPROCESSING	42
3.2.1	FastQC	42
3.2.2	PRINSEQ	43
3.2.3	Trimmomatic	44
3.3	READ QUALITY ISSUES	44
3.3.1	Base Quality	44
	3.3.1.1 <i>Filtering</i>	45
	3.3.1.2 <i>Trimming</i>	49
3.3.2	Ambiguous Bases	52
3.3.3	Adapters	54
3.3.4	Read Length	55
3.3.5	Sequence-Specific Bias and Mismatches Caused by Random Hexamer Priming	56
3.3.6	GC Content	57
3.3.7	Duplicates	57
3.3.8	Sequence Contamination	59
3.3.9	Low-Complexity Sequences and PolyA Tails	59

3.4 SUMMARY	60
REFERENCES	61
CHAPTER 4 ■ Aligning Reads to Reference	63
4.1 INTRODUCTION	63
4.2 ALIGNMENT PROGRAMS	64
4.2.1 Bowtie	64
4.2.2 TopHat	68
4.2.3 STAR	73
4.3 ALIGNMENT STATISTICS AND UTILITIES FOR MANIPULATING ALIGNMENT FILES	77
4.4 VISUALIZING READS IN GENOMIC CONTEXT	81
4.5 SUMMARY	82
REFERENCES	83
CHAPTER 5 ■ Transcriptome Assembly	85
5.1 INTRODUCTION	85
5.2 METHODS	87
5.2.1 Transcriptome Assembly Is Different from Genome Assembly	87
5.2.2 Complexity of Transcript Reconstruction	88
5.2.3 Assembly Process	89
5.2.4 de Bruijn Graph	90
5.2.5 Use of Abundance Information	91
5.3 DATA PREPROCESSING	92
5.3.1 Read Error Correction	93
5.3.2 Secer	93
5.4 MAPPING-BASED ASSEMBLY	95
5.4.1 Cufflinks	95
5.4.2 Scripture	97
5.5 <i>DE NOVO</i> ASSEMBLY	98
5.5.1 Velvet + Oases	98
5.5.2 Trinity	100

5.6 SUMMARY	104
REFERENCES	106
CHAPTER 6 ■ Quantitation and Annotation-Based Quality Control	109
6.1 INTRODUCTION	109
6.2 ANNOTATION-BASED QUALITY METRICS	110
6.2.1 Tools for Annotation-Based Quality Control	111
6.3 QUANTITATION OF GENE EXPRESSION	116
6.3.1 Counting Reads per Genes	117
6.3.1.1 <i>HTSeq</i>	117
6.3.2 Counting Reads per Transcripts	120
6.3.2.1 <i>Cufflinks</i>	122
6.3.2.2 <i>eXpress</i>	122
6.3.3 Counting Reads per Exons	126
6.4 SUMMARY	128
REFERENCES	129
CHAPTER 7 ■ RNA-seq Analysis Framework in R and Bioconductor	131
7.1 INTRODUCTION	131
7.1.1 Installing R and Add-on Packages	132
7.1.2 Using R	133
7.2 OVERVIEW OF THE BIOCONDUCTOR PACKAGES	134
7.2.1 Software Packages	134
7.2.2 Annotation Packages	134
7.2.3 Experiment Packages	135
7.3 DESCRIPTIVE FEATURES OF THE BIOCONDUCTOR PACKAGES	135
7.3.1 OOP Features in R	135
7.4 REPRESENTING GENES AND TRANSCRIPTS IN R	138
7.5 REPRESENTING GENOMES IN R	141
7.6 REPRESENTING SNPs IN R	143

7.7	FORGING NEW ANNOTATION PACKAGES	143
7.8	SUMMARY	146
	REFERENCES	146
CHAPTER 8 ■ Differential Expression Analysis		147
8.1	INTRODUCTION	147
8.2	TECHNICAL VS. BIOLOGICAL REPLICATES	148
8.3	STATISTICAL DISTRIBUTIONS IN RNA-SEQ DATA	149
8.3.1	Biological Replication, Count Distributions, and Choice of Software	150
8.4	NORMALIZATION	152
8.5	SOFTWARE USAGE EXAMPLES	154
8.5.1	Using Cuffdiff	154
8.5.2	Using Bioconductor Packages: DESeq, edgeR, limma	158
8.5.3	Linear Models, the Design Matrix, and the Contrast Matrix	158
8.5.3.1	<i>Design Matrix</i>	159
8.5.3.2	<i>Contrast Matrix</i>	160
8.5.4	Preparations Ahead of Differential Expression Analysis	161
8.5.4.1	<i>Starting from BAM Files</i>	162
8.5.4.2	<i>Starting from Individual Count Files</i>	162
8.5.4.3	<i>Starting from an Existing Count Table</i>	163
8.5.4.4	<i>Independent Filtering</i>	163
8.5.5	Code Example for DESeq(2)	163
8.5.6	Visualization	164
8.5.7	For Reference: Code Examples for Other Bioconductor Packages	168
8.5.8	Limma	169
8.5.9	SAMSeq (samr package)	170
8.5.10	edgeR	171

8.5.11 DESeq2 Code Example for a Multifactorial Experiment	171
8.5.12 For Reference: edgeR Code Example	174
8.5.13 Limma Code Example	175
8.6 SUMMARY	176
REFERENCES	177
CHAPTER 9 ■ Analysis of Differential Exon Usage	181
9.1 INTRODUCTION	181
9.2 PREPARING THE INPUT FILES FOR DEXSeq	183
9.3 READING DATA IN TO R	184
9.4 ACCESSING THE ExonCountSet OBJECT	185
9.5 NORMALIZATION AND ESTIMATION OF THE VARIANCE	187
9.6 TEST FOR DIFFERENTIAL EXON USAGE	190
9.7 VISUALIZATION	193
9.8 SUMMARY	198
REFERENCES	198
CHAPTER 10 ■ Annotating the Results	199
10.1 INTRODUCTION	199
10.2 RETRIEVING ADDITIONAL ANNOTATIONS	200
10.2.1 Using an Organism-Specific Annotation Package to Retrieve Annotations for Genes	201
10.2.2 Using BioMart to Retrieve Annotations for Genes	205
10.3 USING ANNOTATIONS FOR ONTOLOGICAL ANALYSIS OF GENE SETS	208
10.4 GENE SET ANALYSIS IN MORE DETAIL	210
10.4.1 Competitive Method Using GOstats Package	211
10.4.2 Self-Contained Method Using Globaltest Package	213
10.4.3 Length Bias Corrected Method	215
10.5 SUMMARY	216
REFERENCES	216

CHAPTER 11 ■ Visualization	217
11.1 INTRODUCTION	217
11.1.1 Image File Types	218
11.1.2 Image Resolution	218
11.1.3 Color Models	219
11.2 GRAPHICS IN R	219
11.2.1 Heatmap	220
11.2.2 Volcano Plot	224
11.2.3 MA Plot	226
11.2.4 Idiogram	228
11.2.5 Visualizing Gene and Transcript Structures	230
11.3 FINALIZING THE PLOTS	232
11.4 SUMMARY	234
REFERENCES	235
CHAPTER 12 ■ Small Noncoding RNAs	237
12.1 INTRODUCTION	237
12.2 MICRORNAs (miRNAs)	239
12.3 MICRORNA OFF-SET RNAs (moRNAs)	243
12.4 PIWI-ASSOCIATED RNAs (piRNAs)	243
12.5 ENDOGENOUS SILENCING RNAs (endo-siRNAs)	244
12.6 EXOGENOUS SILENCING RNAs (exo-siRNAs)	244
12.7 TRANSFER RNAs (tRNAs)	245
12.8 SMALL NUCLEOLAR RNAs (snoRNAs)	245
12.9 SMALL NUCLEAR RNAs (snRNAs)	245
12.10 ENHANCER-DERIVED RNAs (eRNA)	246
12.11 OTHER SMALL NONCODING RNAs	246
12.12 SEQUENCING METHODS FOR DISCOVERY OF SMALL NONCODING RNAs	248
12.12.1 microRNA-seq	248
12.12.2 CLIP-seq	251
12.12.3 Degradome-seq	254
12.12.4 Global Run-On Sequencing (GRO-seq)	254

12.13 SUMMARY	255
REFERENCES	255
CHAPTER 13 ■ Computational Analysis of Small Noncoding RNA Sequencing Data	259
13.1 INTRODUCTION	259
13.2 DISCOVERY OF SMALL RNAs—miRDeep2	260
13.2.1 GFF files	260
13.2.2 FASTA Files of Known miRNAs	263
13.2.3 Setting up the Run Environment	263
13.2.4 Running miRDeep2	266
13.2.4.1 <i>miRDeep2 Output</i>	266
13.3 miRANALYZER	268
13.3.1 Running miRanalyzer	271
13.4 miRNA TARGET ANALYSIS	271
13.4.1 Computational Prediction Methods	272
13.4.2 Artificial Intelligence Methods	274
13.4.3 Experimental Support-Based Methods	275
13.5 miRNA-SEQ AND mRNA-SEQ DATA INTEGRATION	276
13.6 SMALL RNA DATABASES AND RESOURCES	277
13.6.1 RNA-seq Reads of miRNAs in miRBase	277
13.6.2 Expression Atlas of miRNAs	279
13.6.3 Database for CLIP-seq and Degradome-seq Data	281
13.6.4 Databases for miRNAs and Disease	281
13.6.5 General Databases for the Research Community and Resources	282
13.6.6 miRNAblog	282
13.7 SUMMARY	284
REFERENCES	284
INDEX	287