

Computational Non-coding RNA Biology

Yun Zheng



Computational Non-coding RNA Biology

This page intentionally left blank

Computational Non-coding RNA Biology

Yun Zheng

Kunming University of Science and Technology
Institute of Primate Translational Medicine
Yunnan Key Laboratory of Primate Biomedical Research
Faculty of Information Engineering and Automation
Kunming, Yunnan, China



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2019 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalog record for this book is available from the British Library

ISBN: 978-0-12-814365-0

For information on all Academic Press publications
visit our website at <https://www.elsevier.com/books-and-journals>



Publisher: John Fedor
Acquisition Editor: Glyn Jones
Editorial Project Manager: Sandra Harron
Production Project Manager: Swapna Srinivasan
Designer: Victoria Pearson

Typeset by VTeX

This book is for my family.

Yun Zheng

This page intentionally left blank

Contents

About the Author	xv
Preface	xvii
Acknowledgments	xxi

PART 1 BACKGROUND

CHAPTER 1	Introduction to Non-coding RNAs and High Throughput Sequencing	3
1.1	Introduction to Different Classes of Non-coding RNAs	4
1.1.1	Introduction to microRNAs	4
1.1.2	Introduction to <i>trans</i> -acting siRNAs and Phased siRNAs	8
1.1.3	Introduction to Long Non-coding RNAs	10
1.1.4	Introduction to Lariat Originated Circular RNAs	13
1.1.5	Introduction to Back-spliced Circular RNAs	15
1.2	Introduction to High Throughput Sequencing Technologies	17
1.2.1	Introduction to RNA-seq Technologies	17
1.2.2	Introduction to sRNA-seq Technologies	19
1.2.3	Introduction to PAR-CLIP Sequencing Technology	19
1.2.4	Introduction to Degradome Sequencing Technology	19
1.3	Brief Introduction to the Software Used in the Book	20
1.3.1	The Java Platform	20
1.3.2	The JSmallRNA Package	20
1.3.3	The FastQC Program	21
1.3.4	The Vienna Package	21
1.3.5	BLAST	21
1.3.6	SOAP2	22
1.3.7	BOWTIE and BOWTIE2	22
1.3.8	The tcsh Shell	23
1.3.9	Cufflinks	23
1.3.10	SAMTools and BCFTools	23
1.3.11	BEDTools	24
1.3.12	Integrated Genomics Viewer	24
1.3.13	MATLAB	24
1.3.14	The R Environment and RStudio	24
1.3.15	The edgeR Package	25
1.3.16	The SRA Toolkit	25
1.4	File Formats of Sequences and Sequencing Profiles	25
1.4.1	The FASTA Format	25
1.4.2	Special FASTA Format for Processed Small RNA Profiles	25
1.4.3	The FASTQ Format for Raw Sequencing Profiles	26

1.5	File Formats for Gene Annotations	28
1.5.1	The GFF Format	28
1.5.2	The GTF Format	28
1.5.3	The BED Format	29
1.5.4	The bedGraph Format	30
1.6	Summary	30
	Glossary or Keywords	30
	Acknowledgments	31

PART 2 SMALL NCRNAS

CHAPTER 2	Identification of microRNAs From Small RNA Sequencing Profiles	35
2.1	A Schematic View of the Computational Analysis of Small RNA Profiles	36
2.1.1	Computational Analysis of Plant sRNA Profiles and Degradome Profiles	36
2.1.2	Computational Analysis for Animal sRNA Profiles and PAR-CLIP Profiles	36
2.2	The Biological Background for Identifying miRNAs	38
2.2.1	The Useful Characteristics for Identifying miRNAs	39
2.2.2	The Differences Between Animal and Plant Pre-miRNAs	39
2.2.3	Criteria for Annotating miRNAs	39
2.3	A General Pipeline for Processing Small RNA Sequencing Profiles	40
2.3.1	The Computational Pipeline for Processing Small RNA Sequencing Profiles	40
2.3.2	Detailed Commands to Fulfill the sRNA Tissue Frequency Pipeline	41
2.4	The Pipeline for Calculating Length Distributions of sRNAs	43
2.4.1	Motivation of Examining the Length Distributions of sRNAs in Sequencing Libraries	43
2.4.2	The Databases of Different Types of Molecules	43
2.4.3	Computational Pipeline for Calculating Length Distributions of sRNAs	43
2.4.4	Detailed Commands to Fulfill the Pipeline for Calculating sRNA Length Distributions	44
2.5	Calculating the Abundances of miRNAs in sRNA-seq Profiles	47
2.5.1	Computational Pipeline for Calculating the Abundances of miRNAs	47
2.5.2	Detailed Commands to Fulfill the Mature miRNA Frequency Pipeline	49
2.6	Identifying Precursors of Conserved miRNAs	50
2.6.1	Computational Pipeline for Identifying Precursors of Conserved miRNAs	50
2.6.2	Detailed Commands to Fulfill the Identification of the Conserved Pre-miRNA Pipeline	51
2.7	Identifying Novel Pre-miRNAs From sRNA-seq Profiles	55
2.7.1	Computational Pipeline for Identifying Novel Pre-miRNAs From sRNA-seq Profiles	55
2.7.2	Required Programs	55
2.7.3	Inputs and Outputs	55
2.7.4	Detailed Commands to Fulfill the Novel Pre-miRNA Identification Pipeline	56
2.8	Visualizing the Expression Levels of miRNAs in sRNA-seq Profiles	60
2.8.1	Computational Pipeline for Visualizing miRNA Expression Levels in sRNA-seq Profiles	60

2.8.2	Required Programs	61
2.8.3	Inputs and Outputs	61
2.8.4	Detailed Commands to Fulfill the Visualizing miRNA Pipeline	62
2.8.5	Processing of the MMFinder's Results	62
2.9	Analysis of the Expression Patterns of miRNAs	63
2.9.1	Computational Pipeline for Identifying Deregulated miRNAs in Different Sample Groups	63
2.9.2	Detailed Commands to Fulfill the Pipeline for Identifying Deregulated miRNAs	64
2.9.3	Clustering Analysis Using the Expression Profiles of miRNAs	67
2.9.4	Detailed Commands to Fulfill the Clustering Analysis for miRNA Profiles	69
2.9.5	Principle Component Analysis Using the Expression Profiles of miRNAs	70
2.9.6	Detailed Commands to Fulfill the PCA Analysis for miRNA Profiles	71
2.10	Analysis of miRNAs and Their Expression Patterns	73
2.10.1	Conserved and Novel miRNAs in <i>P. notoginseng</i>	73
2.10.2	The Expression Patterns of miRNAs in Watermelon	80
2.11	Summary	82
	Glossary or Keywords	82
	Acknowledgments	82
CHAPTER 3	Identification of TAS and PHAS Loci in Plants	83
3.1	Introduction of Secondary Small RNAs in Plants	83
3.2	Identification of TAS3 in Plants	84
3.2.1	Computational Pipeline for Identifying TAS3 Genes	84
3.2.2	Required Programs	85
3.2.3	Inputs and Outputs	85
3.2.4	Detailed Commands to Fulfill the TAS3 Identification Pipeline	85
3.2.5	Identifying TAS4 Loci	87
3.3	Visualizing the siRNAs Originating From TAS Loci	87
3.3.1	Computational Pipeline for Visualizing siRNAs Originating From TAS Loci	87
3.3.2	Detailed Commands to Fulfill the TAS Visualization Pipeline	87
3.4	Identification of PHAS in Plants Using sRNA-seq Profiles	89
3.4.1	Criteria for Identifying PHAS Loci From sRNA-seq Profiles	89
3.4.2	Computational Pipeline for Identifying PHAS Loci	89
3.4.3	Detailed Commands to Fulfill PHAS Loci Identification Pipeline	90
3.4.4	Processing of the Output Results of PhasiRNAFinderTree	92
3.5	Analysis of Results of Identified TAS and PHAS Loci	93
3.5.1	Data Sets Used	93
3.5.2	TAS3 Loci in <i>Panax notoginseng</i>	93
3.5.3	A TAS4 Locus in Chinese Sacred Lotus	95
3.5.4	<i>Panax notoginseng</i> PHAS Loci Encoding 21 nt phasiRNAs	96
3.5.5	24 nt PHAS Loci in <i>P. notoginseng</i>	102
3.5.6	PhasiRNA Targets in <i>P. notoginseng</i>	102
3.6	Summary	105

	Glossary or Keywords	105
	Acknowledgments	105
CHAPTER 4	Identification of Editing and Mutation Sites in miRNAs	107
4.1	Introduction of Editing and Mutation Sites in miRNAs	108
4.2	Identifying Mutation and Editing Sites in miRNAs	109
4.2.1	The MiRME Algorithm	109
4.2.2	Computational Pipeline for Identifying Mutation and Editing Sites in miRNAs	111
4.3	Detailed Commands to Fulfill the MiRME Pipeline	116
4.3.1	Generate Tissue Frequency File	116
4.3.2	Align Tissue Frequency File to Pre-miRNAs With BLASTN	117
4.3.3	Get the Unique Sequences That Were Mapped to the Pre-miRNAs	118
4.3.4	Align Unique Sequences That Were Mapped to Pre-miRNAs to the Genome	118
4.3.5	Append the Frequencies of Reads to the Alignments Generated From Bowtie	118
4.3.6	Use the Cross-mapping Correction Method to Calculate Weights of Mapped Reads	118
4.3.7	Analyze the Mutations and Editing in Pre-miRNAs With MiRME	118
4.3.8	The Output Files of MiRME	119
4.3.9	Processing of the MiRME Results	120
4.4	Auxiliary Tools in the MiRME Package	122
4.4.1	Compare MiRME Results to Reported SNPs in miRNAs	122
4.4.2	Compare MiRME Results to Reported Non-3' Editing Sites in miRNAs	122
4.4.3	Combine Results of Multiple Data Sets	122
4.5	Integrating Genome Sequencing Profiles to Differentiate Editing and Mutations	123
4.5.1	Computational Pipeline for Comparing Mutations to MiRME Identified Sites	123
4.5.2	Detailed Commands for Comparing Mutations to MiRME Identified Sites ..	124
4.5.3	Compare Mutation Sites to MiRME Results	127
4.6	Analysis of the MiRME Results	127
4.6.1	Data Sets Used	127
4.6.2	Analyzing Selected Samples and Combining Results of Different Samples ..	127
4.6.3	Overall Summary of the Detected M/E Sites	128
4.6.4	5' Editing Sites	129
4.6.5	A-to-I Editing Sites	129
4.6.6	Other Types of miRNA Editing Sites	131
4.6.7	Putative Small Insertions and Deletions in miRNAs	133
4.6.8	Detection of Known and Novel SNPs in miRNAs	134
4.6.9	Analyzing Non-canonical miRNA Editing Sites by Integrating Genome Sequencing Profiles	134
4.6.10	Comparisons With Related Works	137
4.6.11	Efficiencies of MiRME and Two Compared Methods	141
4.7	Summary	141
	Glossary or Keywords	141
	Acknowledgments	142

PART 3 MIRNA TARGETS

CHAPTER 5	Identification of miRNA Targets in Animals	145
5.1	Important Determinants for miRNA Target Recognition in Animals	145
5.2	Sequencing-based Methods for Identifying Animal miRNA Targets	147
5.3	Traditional miRNA Target Prediction Methods in Animals	147
5.3.1	The Criteria of Animal miRNA Target Prediction	147
5.3.2	How to Quantify Sequence-specific Determinants	148
5.3.3	The Hitsensor Algorithm for Animal miRNA Target Prediction	151
5.3.4	Evaluation of Methods	152
5.3.5	Computational Pipeline for Predicting Animal miRNA Targets With the Hitsensor Algorithm	153
5.3.6	Detailed Commands to Fulfill the Hitsensor Pipeline	153
5.3.7	Processing of the Results	155
5.3.8	Analysis of the Results of Hitsensor	155
5.4	Identifying Animal miRNA Targets Using PAR-CLIP	161
5.4.1	The MiCPAR Algorithm	161
5.4.2	Computational Pipeline for Predicting Animal miRNA Targets With the MiCPAR Algorithm	163
5.4.3	Detailed Commands to Fulfill the MiCPAR Pipeline	167
5.4.4	The Output Files of MiCPAR	169
5.4.5	Processing of the Results	170
5.5	Analysis of Identified miRNA Targets	171
5.5.1	Analysis of Reported miRNA Targets	171
5.5.2	Novel miRNA Targets	171
5.5.3	Non-coding RNAs as microRNA Targets	171
5.5.4	Analysis of miRNA Targets With Noncanonic Complementary Sites	171
5.6	Summary	176
	Glossary or Keywords	176
	Acknowledgments	176
CHAPTER 6	Identification of miRNA and siRNA Targets in Plants	177
6.1	The miRNA and siRNA Targets Recognition in Plants	177
6.2	Traditional miRNA Target Prediction Method in Plants	179
6.2.1	The Criteria of Plant miRNA Target Prediction	179
6.2.2	The Hitsensor Algorithm for Plant miRNA Target Prediction	179
6.2.3	The Computational Pipeline for Predicting Plant miRNA Targets With the Hitsensor Algorithm	179
6.2.4	Processing of the Hitsensor Results	181
6.3	Plant miRNA Target Prediction Using Degradome Sequencing Profiles	182
6.3.1	The SeqTar Algorithm	182
6.3.2	Identifying Plant miRNA Targets Using the SeqTar Algorithm	183
6.3.3	Processing the Output Results of SeqTar	188
6.4	Analysis of the Obtained Results of the SeqTar Pipeline	190
6.4.1	Validation of the Results From SeqTar	190

6.4.2	Conserved miRNAs Target Additional Members of Known Target Gene Families	190
6.4.3	Novel Targets of Conserved miRNAs and Experimental Validations	192
6.4.4	Identification of New Targets of Nonconserved miRNAs and siRNAs	193
6.4.5	The Combinatorial Regulation of miRNA Targets	196
6.4.6	Self- and Cross-repression of TAS/PPR Transcripts	196
6.4.7	Self-repression of miRNAs in <i>Arabidopsis</i>	202
6.4.8	Comparisons With CleaveLand	202
6.4.9	The False Discovery Rate of SeqTar	204
6.4.10	Efficiency of SeqTar	204
6.5	Summary	204
	Glossary or Keywords	205
	Acknowledgments	205

PART 4 LONG NCRNAS

CHAPTER 7	Identification of Long Non-coding RNAs From RNA-seq profiles	209
7.1	A Schematic View of the Computational Analysis for Sequencing Profiles of Long RNAs	209
7.2	A Brief Introduction to Long Non-coding RNAs	211
7.3	Identification and Quantification of lncRNAs From RNA-seq Profiles	211
7.3.1	Computational Pipeline for Identifying Putative lncRNAs With Cufflinks	211
7.3.2	Detailed Commands to Fulfill the Cufflinks Pipeline	212
7.3.3	Processing the Results	215
7.4	Computational Analysis of Structures of lncRNAs	215
7.4.1	Computational Command for Predicting the Structures of lncRNAs	215
7.5	Analyzing Coding Capacities of lncRNAs	216
7.5.1	Analyzing Coding Capacities With CPC2	216
7.5.2	Analyzing Coding Capacities With COME	216
7.5.3	Analyzing Coding Capacities With BLASTX	217
7.6	Analysis of the Identified lncRNA Candidates	217
7.6.1	Analysis of the Lengths and Structures of Putative lncRNAs	217
7.6.2	Classifying Putative lncRNAs Based on Their Genomic Contexts	218
7.6.3	Comparing the lncRNA Candidates to Reported lncRNAs	219
7.6.4	The Expression Patterns of the Putative lncRNAs	219
7.7	Summary	223
	Glossary or Keywords	223
CHAPTER 8	Identification of Intron Lariat RNAs and Intron Branch Points	225
8.1	Brief Introduction to Splicing and Lariat RNAs	225
8.2	Identification and Quantification of Lariat RNAs From RNA-seq Profiles	226
8.2.1	Computational Pipeline for Identifying Lariat RNAs	226
8.2.2	Detailed Commands to Fulfill Lariat RNA Identification Pipeline	227
8.2.3	Processing of the Results	228
8.2.4	Analysis of the Results of Lariat RNAs	228

8.3	Identification of Intron Branch Points	230
8.3.1	Computational Pipeline for Identifying Intron Branch Points	230
8.3.2	Detailed Commands to Fulfill Intron Branch Point Identification Pipeline . . .	231
8.3.3	Processing of the Results	232
8.3.4	Analysis of the Branch Points in <i>Arabidopsis thaliana</i>	232
8.4	Lariat RNAs Inhibits microRNA Biogenesis	234
8.4.1	Materials and Data Sets	234
8.4.2	DBR1 is Required for miRNA Accumulation in <i>Arabidopsis</i>	235
8.4.3	Pri-miRNAs Overaccumulate in the <i>dbr1-2</i> Mutant	237
8.4.4	The Binding of pri-miRNA With the Dicing Complex Is Impaired in <i>dbr1-2</i> .	237
8.4.5	Characterization of Lariat RNAs in <i>Arabidopsis</i>	239
8.4.6	Lariat RNAs Affect the Binding of the DCL1/HYL1 Complex With primiRNAs	239
8.4.7	Overexpression of Lariat RNAs Causes Reduced miRNA Accumulation . . .	243
8.5	Summary	245
	Glossary or Keywords	245
	Acknowledgments	245
CHAPTER 9	Identification of Back-spliced Circular RNAs	247
9.1	Brief Introduction to Back-splicing and Circular RNAs	247
9.2	Identifying circRNAs From RNA-seq Profiles	248
9.2.1	Considerations for Identifying circRNAs From RNA-seq Profiles	248
9.2.2	Computational Steps for Identifying circRNAs From RNA-seq Profiles	249
9.2.3	Detailed Commands to Fulfill the Pipeline for Identifying circRNAs	249
9.2.4	Processing of the Results	252
9.2.5	Analysis of the Identified circRNAs	252
9.3	Calculating the Expression Levels of circRNAs	255
9.3.1	Considerations When Calculating the Expression Levels of circRNAs	255
9.3.2	Computational Pipeline for Calculating the Expression Levels of circRNAs .	255
9.3.3	Detailed Commands to Fulfill the Pipeline for Quantifying circRNAs	256
9.3.4	Processing the Results	257
9.3.5	Analysis of the circRNA Expression Results	257
9.4	Analyzing the Repeat Elements in Introns Around circRNAs	259
9.4.1	Introduction of RepeatMasker	259
9.4.2	Analyzing Repeat Elements in Introns With RepeatMasker	260
9.4.3	Comparing circRNAs With Annotated Repeat Elements	260
9.4.4	Processing and Analysis of the Results	261
9.5	Identifying miRNA Binding Sites on circRNAs	261
9.5.1	Data Sets and Processing	261
9.5.2	Analysis of the Results	264
9.6	Summary	264
	Glossary or Keywords	264
Appendix A	Usage Guide of Web-based ncRNA Resources	267
A.1	Usage Guide of Web-based ncRNA Resources	267

A.2	UCSC Genome Browser	269
A.3	Visualization of ncRNAs With the Integrated Genomics Viewer	270
A.3.1	Visualizing the Expression Levels of ncRNAs With RNA-Seq Profiles	270
A.3.2	Visualizing the Mutations and Editing Sites in miRNAs With Sequencing Profiles	271
	Acknowledgments	274
Appendix B	Abbreviations and Acronyms	275
B.1	Abbreviations and Acronyms	275
	Bibliography	277
	Index	295