Preface

AIMS AND AUDIENCE

Non-coding RNAs (ncRNAs) have become critical molecules in almost all biological processes and human diseases. With the development of high throughput sequencing technologies, our knowledge of the scale of transcriptome has fundamentally changed. While only 2% of the human genome are protein coding genes, 60%–75% of the human genome is transcribed into different types of RNAs. Consequently, in the last few years many new ncRNAs have come into our view.

In early years, the studies of ncRNAs were performed in a low-throughput way. Today, a single study could easily generate even several terabyte (TB) of data with hundreds of RNA-seq profiles. This large amount of data naturally needs novel ways, especially computational methods, of accurate and efficient analysis. Although some algorithms have been developed for the identification and quantification of some ncRNAs, these algorithms or methods are distributed in many sources and are often specialized for one type of ncRNA. Currently, it still lacks a unified source that clearly and comprehensively shows the detailed computational steps for identifying and characterizing the ever-expanding repository of ncRNAs from the profuse sequencing data.

To fill the gap between the large amount of sequencing data and the accurate annotations of ncRNAs, this book aims to provide practical computational methods or pipelines and skills for performing diverse analyses of different types of ncRNAs, mainly microRNA (miRNA), small interfering RNA (siRNA), long non-coding RNA (lncRNA), and circular RNA (circRNA), using diverse high throughput sequencing profiles. More emphasis is given to the *identification and quantification of ncRNAs*, including miRNA, *trans*-acting siRNA, phased siRNA, long non-coding RNA, circular intronic RNA (or lariat RNA), and circular RNA originating from back-spliced exons. The functional mechanisms of most miRNAs and some siRNAs have been clearly characterized. This book includes two chapters that introduce how to identify miRNA targets in both animals and plants, with and without the high throughput sequencing based technologies. The mutation and editing sites in miRNAs may lead to radical changes in either the expression levels of miRNAs or in the targets of miRNAs, which may cause severe human diseases. Thus, the book includes a chapter to show how to identify mutation and editing sites in miRNAs from small RNA high throughput sequencing profiles, and to show how to integrate the genome sequencing profile of the same sample to distinguish the mutation and editing sites identified in the sRNA-seq profile. The book also shows how to identify deregulated ncRNAs, to perform clustering, and to perform Principle Component Analysis using the expression levels of different types of ncRNAs. For more downstream analyses, after ncRNAs have been identified and quantified, readers can choose from many other works about machine learning and data mining.

The field of computational ncRNAs is in a phase of fast development and evolution. The current version of the book only discusses the topics related to miRNA, tasiRNA/phasiRNA, lncRNA, lariat RNA, and circRNA. We do not include the topics of traditional ncRNAs, such as tRNA and rRNA, because methods for these ncRNAs have already been established. We do not include the topics of other recently identified ncRNAs, such as piRNA and tRNA-derived RNA fragments, due to their limited audience or the immature mechanisms of their biogenesis or functions.

The potential audience of the book includes, but is not limited to, wet-lab biologists, medical doctors, computer scientists, and computational biologists.

OUTLINE OF THE CONTENT

The book consists of four main parts.

The first part, which has only one chapter, is background knowledge. This first chapter starts with introductions of five types of non-coding RNAs (ncRNAs) that will be discussed in the book: microRNA (miRNA); *trans*-acting small interfering RNA (tasiRNA) and phased small interfering RNA (phasiRNA); long non-coding RNA (lncRNA); lariat originated circular RNA (lariat RNA); and back-spliced circular RNA (circRNA). We mainly talk about the biogenesis, functions, computational considerations, and related online resources (if available) of these ncRNAs. We then introduce some high throughput sequencing technologies for the identification and quantification of ncRNAs, including RNA sequencing (RNA-seq), and variants of RNA-seq. Two sequencing technologies for identifying miRNA targets in animals and plants, PAR-CLIP and degradome (or PARE) sequencing, respectively, are also introduced. Next, we briefly introduce the software that is used throughout the book. Finally, we introduce several file formats of sequences and sequencing profiles and for gene annotations.

The second part of the book is about small RNA and is organized in three chapters, i.e., Chapters 2–4. The second chapter starts with an overview of computational analysis for small RNA profiles, as well as PAR-CLIP and degradome sequencing profiles, and links these analyses to different pipelines in the book. We introduce nine computational pipelines for processing of sRNA-seq profile, calculating the length distributions of sRNA-seq profiles, calculating the frequencies of mature miRNAs, identifying conserved miRNAs, identifying new miRNAs, visualizing miRNA expression levels, identifying deregulated miRNAs in different sample groups, performing clustering analysis using miRNA expression profiles, and performing Principle Component Analysis using miRNA expression profiles. We then show some results of miRNAs in *Panax notoginseng* and watermelon.

The third chapter is about the tasiRNAs and phasiRNAs found in plants. We introduce three pipelines for identifying a highly conserved TAS locus, TAS3, in land plants, for visualizing siRNA originated from TAS loci and for identifying PHAS and phasiRNAs from sRNA-seq profiles. We then show results of TAS3 in *Panax notoginseng*, TAS4 in the Chinese sacred lotus, and PHAS and phasiRNAs in *Panax notoginseng*.

In the fourth chapter, we first introduce a pipeline for identifying mutation and editing sites in miRNAs from sRNA-seq profiles. When the genome sequencing profile and sRNA-seq profile of the same sample are available at the same time, it is feasible to distinguish the mutation and editing sites by comparing the genome sequencing and sRNA-seq profiles, which is introduced in the second pipeline in the fourth chapter. We then show some results of identified mutation and editing sites in miRNAs.

The third part of the book is about miRNA targets. The functions of miRNAs are largely determined by their targets. There are two chapters in this part, Chapters 5 and 6, which are dedicated to identifying miRNA targets in animals and plants, respectively.

In the fifth chapter, we introduce two computational pipelines for identifying animal miRNA targets, with and without PAR-CLIP sequencing profiles. We then show some results found by these two pipelines, including reported targets, novel targets, lincRNAs targeted by miRNAs, and targets of

non-canonic complementary miRNA sites. Recent evidence show that some circRNAs are targeted by miRNAs. We demonstrate the miRNA:circRNA interactions in Chapter 9.

The sixth chapter is for identifying plant miRNA targets. We introduce pipelines for identifying plant miRNA targets, with and without degradome sequencing profiles. Then we show the results for the identified miRNA and siRNA targets in both *Arabidopsis thaliana* and rice.

The fourth part of the book is about long non-coding RNAs. There are three chapters in this part, Chapters 7, 8, and 9, dedicated to lncRNAs, lariat RNAs, and circRNAs, respectively.

The seventh chapter starts with a schematic view of computational tasks of lncRNAs and links these tasks to different pipelines in the book. We then introduce how to use the Cufflinks pipeline to identify putative lncRNAs. Next, we introduce how to analyze the secondary structures and coding capacities of the lncRNA candidates. We also show the results of identified lncRNAs and their expression patterns.

In the eighth chapter, we introduce two pipelines for identifying lariat RNAs and for identifying branch points in introns from RNA-seq profiles. We then show some results of lariat RNAs and branch points of introns in *Arabidopsis thaliana*. We also include a section to show that some lariat RNAs in *Arabidopsis thaliana* could inhibit miRNA biogenesis by attracting the miRNA generation complex.

The ninth chapter introduces two pipelines for identifying circRNAs and quantifying circRNAs. We show some results of circRNAs and their expression levels by analyzing four human RNA-seq profiles. Then, we introduce how to analyze the repeat elements in introns adjacent to circRNAs with RepeatMasker since some circRNAs are formed through the pairing of *Alu* elements around circRNAs. Some of the identified circRNAs and their neighboring *Alu* elements are also shown in detail. We also introduce how to identify miRNA complementary sites on circRNAs with the MiCPAR pipeline introduced in Chapter 5. Our results suggest that a widely reported circRNA, CDR1as, could be targeted by another miRNA, miR-1180-3p, in addition to miR-7-5p.

Two appendixes are given after the main chapters. The first, Appendix A, gives a brief usage guide to the online resources of ncRNAs. Some online resources of miRNAs, lncRNAs, and circRNAs are introduced in the first chapter. This appendix gives some advice and suggestions when using these resources. Then, we introduce the features related to ncRNAs in the UCSC Genome Browser. We also introduce how to visualize the expression levels of ncRNAs and visualize the miRNA mutation and editing sites with the Integrated Genomics Viewer. The second, Appendix B, lists the abbreviations and/or acronyms and their full descriptions.

For each of the computational pipelines presented in the book, we first list the main steps in a box, then introduce the required programs of the pipeline, i.e., the inputs and outputs, followed by detailed commands of the whole pipeline. Some processing that is necessary after obtaining the results is also introduced.

Many pipelines in the book use a Java package, called JSmallRNA, which is available for non-commercial use (registration required). Please write to Dr. Yun Zheng (zhengyun5488@gmail.com) for the JSmallRNA package.

Yun Zheng
Kunming, China
March 6, 2018