Preface

A PRACTICAL BOOK FOR VARIOUS AUDIENCES

RNA-seq offers unprecedented information about the transcriptome, but harnessing this information with bioinformatic tools is typically a bottleneck. The goal of this book is to enable readers to analyze RNA sequencing (RNA-seq) data. Several topics are discussed in detail, covering the whole data analysis workflow from quality control, mapping, and assembly to statistical testing and pathway analysis. Instead of minimizing overlap with existing textbooks, the aim is for a more comprehensive and practical presentation.

This book enables researchers to examine differential expression at gene, exon, and transcript levels and to discover novel genes, transcripts, and whole transcriptomes. In keeping with the important regulatory role of small noncoding RNAs, a whole section is devoted to their discovery and functional analysis.

This book is intended for students and advanced researchers alike. Practical examples are chosen in such a way that not only bioinformaticians, but also nonprogramming wet lab scientists can follow them, making the book suitable for researchers from a wide variety of backgrounds including biology, medicine, genetics, and computer science. This book can be used as a textbook for a graduate course, for an advanced undergraduate course, or for a summer school. It can serve as a self-contained handbook and detailed overview of the central RNA-seq data analysis methods and how to use these in practice.

The book balances theory with practice, so that each chapter starts with theoretical background, followed by descriptions of relevant analysis tools, and examples of their usage. In line with our desire for accessibility, the book is a self-contained guide to RNA-seq data analysis. Importantly, it caters also to noncomputer-savvy wet lab biologists, as examples are given

using the graphical Chipster software in addition to command line tools. All software used in examples are open source and freely available.

OUTLINE OF THE CONTENTS

The "Introduction" section in Chapters 1 and 2 discusses the different applications of RNA-seq, ranging from discovery of genes and transcripts to differential expression analysis and discovery of mutations and fusion genes. It gives an overview of RNA-seq data analysis and discusses important aspects of experimental planning.

The first part is devoted to mapping reads to references and *de novo* assembly. As the quality of reads strongly influences both, a chapter on quality control and preprocessing is also included. Chapter 3 discusses several quality issues characteristic to high-throughput sequencing data and tools for detecting and solving them. Chapter 4 describes the challenges in mapping RNA-seq reads to a reference and introduces some commonly used aligners with practical examples. Tools for manipulating alignment files and genome browsers for visualizing reads in genomic context are introduced. Chapter 5 describes the elements of transcriptome assembly. The relevance of data-processing steps such as filtering, trimming, and error correction in RNA-seq assembly is discussed. Fundamental concepts such as splicing graph, de Bruijn graph, and path traversal in an assembly graph are explained. Differences between genome and transcriptome assembly are discussed. Two approaches to reconstruct full-length transcripts are explained: mapping-based and *de novo* assembly. Both approaches are also demonstrated with practical examples.

The second part is devoted to statistical analyses predominantly carried out with the R software that is supplemented with the tools produced by the Bioconductor project. Chapter 6 discusses the different quantitation approaches and tools, as well as annotation-based quality metrics. Chapter 7 describes the R- and Bioconductor-based framework for the analysis of RNA-seq data and how to import the data in R. The main differences between the statistical and bioinformatic tools in R are also discussed. Chapters 8 and 9 discuss different options for analyzing differential expression of genes, transcripts, and exons and show how to perform the analysis using R/Bioconductor tools and some standalone tools. Chapter 10 offers solutions for annotating results, and Chapter 11 describes different ways of producing informative visualizations to display the central results.

The last part of the book is focused on the analysis of small noncoding RNAs using web-based or free downloadable tools. Chapter 12 describes the different classes of small noncoding RNAs and characterizes their function, abundance, and sequence attributes. Chapter 13 describes different algorithms that are used to discover small noncoding RNAs from next-generation sequencing data sets and provides a practical approach with workflows and examples of how small noncoding RNAs are discovered and annotated. In addition, the chapter describes downstream tools that can be used to elucidate functions of small noncoding RNAs.