Translational Bioinformatics 9
Series Editor: Xiangdong Wang, MD, Ph.D.

Jiaqian Wu *Editor*

Transcriptomics and Gene Regulation

Springer

Translational Bioinformatics

Volume 9

Series editor

Xiangdong Wang, MD, Ph.D.
Professor of Medicine, Zhongshan Hospital, Fudan University Medical School, China
Director of Shanghai Institute of Clinical Bioinformatics, (www.fuccb.org)
Professor of Clinical Bioinformatics, Lund University, Sweden

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Recently Published and Forthcoming Volumes

Single Cell Sequencing and Systems Immunology
Editors: Xiangdong Wang, Xiaoming Chen, Zhihong Sun, Jinglin Xia
Volume 5

Computational and Statistical Epigenomics
Editor: Andrew E. Teschendorff
Volume 7

Genomics and Proteomics for Clinical Discovery and Development
Editor: György Marko-Varga
Volume 6

Allergy Bioinformatics
Editors: Ailin Tao, Eyal Raz
Volume 8

More information about this series at http://www.springer.com/series/11057

Jiaqian Wu
Editor

Transcriptomics and Gene Regulation

Honor editor
Dong Kim

Springer

Editor
Jiaqian Wu
The Vivian L. Smith Department of Neurosurgery
Center for Stem Cell and Regenerative Medicine
The University of Texas Medical School at Houston
Houston, TX
USA

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-94-017-7448-2 ISBN 978-94-017-7450-5 (eBook)
DOI 10.1007/978-94-017-7450-5

Library of Congress Control Number: 2015952061

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Preface

This volume focuses on the modern computational and statistical tools for gene expression and regulation research to improve the understanding prognosis, diagnostics, prediction of severity, and therapies for human diseases. The recent advancements of microarray and next-generation sequencing technologies made it possible to detect gene expression at a genome-wide scale, which has greatly facilitated the identification of pathophysiological changes in various diseases. How are the global gene expression profiles regulated and what are the mechanisms underlying these changes? How are the signaling pathways altered under pathological conditions? How do the various regulatory molecules interact in a network to control disease states? How does the different genetic makeup of individuals affect the disease perceptibility and treatment outcome? These are fundamental questions for finding cures for complex human diseases and developing personalized medicine that is the future of health care. In this volume, we introduce the readers to some of state-of-the-art technologies as well as computational and statistical tools for translational bioinformatics in the areas of gene transcription and regulation, including the tools for next-generation sequencing analyses, alternative spicing, the modeling of signaling pathways, network analyses in predicting disease genes, as well as protein and gene expression data integration in complex human diseases. This volume is particularly suitable for researchers, physicians, or students in the field of molecular, clinical biology and bioinformatics. This exciting volume would not be possible without the expertise and dedication of all the contributing authors. Finally, I would like to dedicate this volume to my family for their unconditional love and support.

Houston, TX — Jiaqian Wu

Contents

About the Editors

Prof. Jiaqian Wu An assistant professor in the Vivian L. Smith Department of Neurosurgery and Center for Stem Cell and Regenerative Medicine at the University of Texas Medical School at Houston, Dr. Wu earned her doctorate in molecular and human genetics at Baylor College of Medicine and did her postdoctoral work at Yale and Stanford University. The Wu laboratory combines stem cell biology and systems-based approaches involving functional genomics, bioinformatics, and next-generation sequencing technologies to unravel gene transcription and regulatory mechanisms governing neural and blood development and differentiation. Dr. Wu's work has been recognized with prestigious honors and awards, including the National Institute of Health Pathway to Independence (PI) Award (K99/R00), R01 and the Senator Lloyd and B.A. Bentsen Investigator Award which she currently holds; the National Institutes of Health Ruth L. Kirschstein National Research Service Award for Individual Postdoctoral Fellows; and the International Society for Stem Cell Research (ISSCR) Annual Meeting Travel Award. A reviewer for NIH, MRC, the journals Nucleic Acids Research, Genome Research, and Genome Biology, Dr. Wu has presented invited talks and lectures on stem cell biology, functional genomics, and proteomics at international conferences, the Multiple Sclerosis Research Center of New York, Lawrence Livermore National Laboratory, and the University of Florida, etc. She has developed a patent, authored a book, and wrote many articles that have appeared in PNAS, the Journal of Neuroscience, Genome Biology, Plos Genetics, Genome Research, and Nature, among others.

Prof. Dong H. Kim Dr. Kim is Professor and Chair of the Department of Neurosurgery at the University of Texas in Houston. As director of the Mischer Neuroscience Institute (MNI), he also leads the clinical neuroscience efforts for the Memorial Hermann Healthcare System. Currently, his group includes over 100 faculty and residents/fellows. A graduate of Stanford and the University of California, San Francisco (UCSF) School of Medicine, he completed general surgery training at Harvard, then neurosurgery at UCSF. Prior to coming to Texas, he held positions at Harvard Medical School, Brigham and Women's Hospital, the Dana-Farber Cancer Institute, Cornell University Medical College, The New York Hospital and Memorial Sloan Kettering Cancer Center. Dr. Kim's research has focused on the origin, development, and treatment of brain aneurysms. His group recently identified the first gene defect proven to cause intracranial aneurysms in familial patients. His research effort is also on stem cell therapy for treating spinal cord and brain injury. Dr. Kim was mentioned in the US News and World Report's Top 1% Doctors, and America's Top Surgeons. He is the recipient of grants from the National Institutes of Health and the American Stroke Association.

Chapter 1
The Analyses of Global Gene Expression and Transcription Factor Regulation

Raquel Cuevas Diaz Duran, Sudheer Menon and Jiaqian Wu

Abstract A major challenge in molecular cell biology lies in understanding how the same genome can give rise to different cell types and how gene expression is regulated. Gene expression and regulation studies focus on the abundance and structure of transcripts as well as how RNA production is controlled. High-throughput sequencing technologies such as RNA sequencing have allowed more accurate profiling of the transcriptome and the rapid identification of differentially expressed genes among samples. The regulation of gene expression is orchestrated by transcription factors. The development of ChIP sequencing assay has made it possible to comprehensively identify transcription factor-binding sites in vivo, allowing rapid unraveling of signaling pathways. The following chapter described the common methods used in studying global gene expression and transcription factor regulation with a special emphasis on bioinformatic analyses. The final section illustrates an example of an integrated gene expression and regulation study for identifying key factors regulating self-renewal and differentiation in hematopoietic precursor cells.

Keywords RNA sequencing · ChIP sequencing · Transcription factors · Transcriptome

R. Cuevas Diaz Duran · S. Menon · J. Wu (✉)
The Vivian L. Smith Department of Neurosurgery, University of Texas Medical School at Houston, Houston, TX 77030, USA
e-mail: jiaqian.wu@uth.tmc.edu

R. Cuevas Diaz Duran · S. Menon · J. Wu
Center for Stem Cell and Regenerative Medicine, UT Brown Institution of Molecular Medicine, Houston, TX 77030, USA

J. Wu, *Transcriptomics and Gene Regulation*,
Translational Bioinformatics 9, DOI 10.1007/978-94-017-7450-5_1

1.1 Introduction

Gene transcription and regulation are important areas of study because they underlie many biological processes and phenotypic variations in living organisms. Aberrant gene expression and regulation lead to diseases. The transcriptome consists of all transcripts synthesized in an organism including protein-coding, noncoding, alternatively spliced, polymorphic, sense, antisense, and edited RNAs. Transcriptome data analyses, namely the analyses of gene expression levels and structures, are essential for interpreting the functional elements of the genome and understanding the molecular constituents of cells and tissues. The regulation of gene expression is a basic mechanism through which RNA production is coordinated, and it controls important events such as development, homeostasis, and responses to environmental stimuli. Transcription factors, a type of DNA-binding proteins which recognize specific sequences, and other proteins work together through a variety of mechanisms to regulate gene transcription.

In this volume, different aspects regarding the analyses of gene transcription and regulation are described in individual chapters. In this chapter, we focus on gene expression level analyses by RNA sequencing (RNA-Seq) and transcription factor regulation by chromatin immunoprecipitation coupled with sequencing (ChIP-Seq). First, we review some useful methods developed in the past for characterizing global gene transcription.

1.2 Methods for Characterizing Global Gene Transcription

1.2.1 EST Sequencing

It is widely recognized that expressed sequence tag (EST) sequencing has provided an invaluable resource for identification of novel human genes [1, 2]. EST clustering methods allow EST to be systematically mapped, so that the information is readily integrated into the positional cloning project UniGene database (http://www.ncbi.nlm.nih.gov/UniGene). Because ESTs are from single-pass sequencing, they have to be carefully analyzed to remove genomic DNA and other contaminating sequences, such as mitochondrial, ribosomal, vector, and bacterial sequences. However, EST databases still contain a significant portion of (estimated 5–10 %) artifact sequences such as intronic or intergenic DNA [3, 4]. This is likely due to the presence of heterogeneous nuclear RNA (hnRNA) in RNA samples where EST libraries were generated. Moreover, it is difficult to understand the relationships among short EST sequences. EST clustering may confuse genes sharing similarities and alternatively spliced transcripts. Additionally, because of their short length and generally low quality, ESTs only provide limited information about gene structure and function [1]. Since EST sequencing is biased toward genes

with high expression levels, the transcripts which are tissue-specific or low-abundant are less likely to be disclosed by EST sequencing. Therefore, methods that are less biased, more accurate, and sensitive are needed [5].

1.2.2 SAGE

The serial analysis of gene expression, or SAGE [6], is another technique used to quantify gene expression levels. SAGE method is designed to add a 9–14 bp tag adjacent to an *Nla*III restriction site at the gene's 3′ end. It measures transcript levels by automatically sequencing and counting each SAGE tag. The expression level of SAGE tags is analyzed and accessioned through the GEO repository. Additionally, an anatomic viewer "SAGE Genie" makes it easy to search and visualize the transcription level in different tissues and cell types of the human body (http://cgap.nci.nih.gov/SAGE). However, SAGE is a high-throughput technology which measures not the expression level of a gene, but a "tag" that represents a transcript. Due to alternative transcription, sequencing errors, and other potential effectors, sometimes two or more genes share the same tag or one gene may have more than one tag. Thus, the potential loss of fidelity should be taken into consideration.

Long serial analysis of gene expression (LongSAGE) is an adaptation of the original SAGE approach that can be used to rapidly identify novel genes and exons [7]. Instead of using an *Nla*III restriction site, LongSAGE uses a modification of longer tags (21 bp) added to a different restriction site (*Mme*I). The 21 bp tags include a constant 4 bp restriction site sequence where the transcript was cleaved and a unique 17 bp adjacent sequence of each transcript. The advantage of LongSAGE is the uniqueness of each tag in the human genome, which is not guaranteed by 14 bp tags. Sequencing tag concatamers and searching for the localization of tags in the genome help to verify predicted genes and to identity novel transcripts. LongSAGE has been reported to be at least an order of magnitude more efficient than EST sequencing [8].

1.2.3 Full-Length CDNA Sequencing Projects

In order to better access the biological information of genes including location of open reading frames, 5′ and 3′ untranslated regions, and splicing patterns, full-length and high-quality sequences of cDNAs are needed. cDNA sequencing is a valuable resource not only for characterizing the structure and function of known genes, but also for discovering novel genes. Especially with the completion of the human genome, the comparisons of the full-length and high-quality cDNA sequences with genomes are especially useful in identifying alternative gene structure and better understanding transcriptome composition during physiological and disease processes. Moreover, full-length cDNA sequencing projects paved the

way for proteomic study by identifying new enzymes and proteins, generating physical clones for expression profiling, testing protein interactions, and generating hypotheses for biochemical studies.

There have been multiple efforts that aim at capturing the sequence of full-length clones which can be directly obtained from cDNA libraries generated from mammals and other selected organisms, such as *zebra fish, drosophila,* and *Caenorhabditis elegans*, mouse, and human [9–15]. In particular, NIH Mammalian Gene Collection project, utilizing large-scale RT-PCR-based cloning methods, provided thousands of clones of full-length human and mouse open reading frames [16–18].

1.2.4 Microarrays

DNA microarray is a hybridization-based technology which enables researchers to analyze the expression of large number of genes in a single reaction. DNA microarray technology was developed in the early 1990s. The technical advancement of this methodology is to manufacture slides or chips with thousands of nucleic acid probes immobilized on a small surface area. DNA probes are conformed of several specific DNA sequences of genes to which cDNA samples are hybridized. Samples, also referred to as targets, may be obtained from cells in different biological or experimental conditions, tissues, organisms, or developmental stages. Probe–target hybridization is quantified by fluorescent labeling. The signal intensities captured as images after scanning are converted into a data matrix and processed using software specific to the application of the array to determine relative abundance of specific cDNA sequences from the samples. The DNA microarray is an effective tool to investigate the structure and activity of genes at a genome-wide scale, and it helps to elucidate the molecular mechanisms underlying normal and dysfunctional biological processes [19–24].

1.2.5 RNA-Seq

Even though microarray technology has provided valuable insights into gene function throughout the last decade, it suffers from limitations in resolution, dynamic range, and accuracy. The recently developed RNA-Seq methodology uses next-generation sequencing (NGS) to sequence cDNAs generated from RNA samples producing millions of short reads. The number of reads mapped within a genomic feature of interest (such as a gene or an exon) can be used as a measurement of the feature's abundance in the analyzed sample. Typical RNA-Seq procedure is depicted in Fig. 1.1. Briefly, RNAs are converted to a library of cDNA fragments with adaptors attached to one or both ends. The molecules, with or without amplification, are then sequenced with high-throughput technology, and

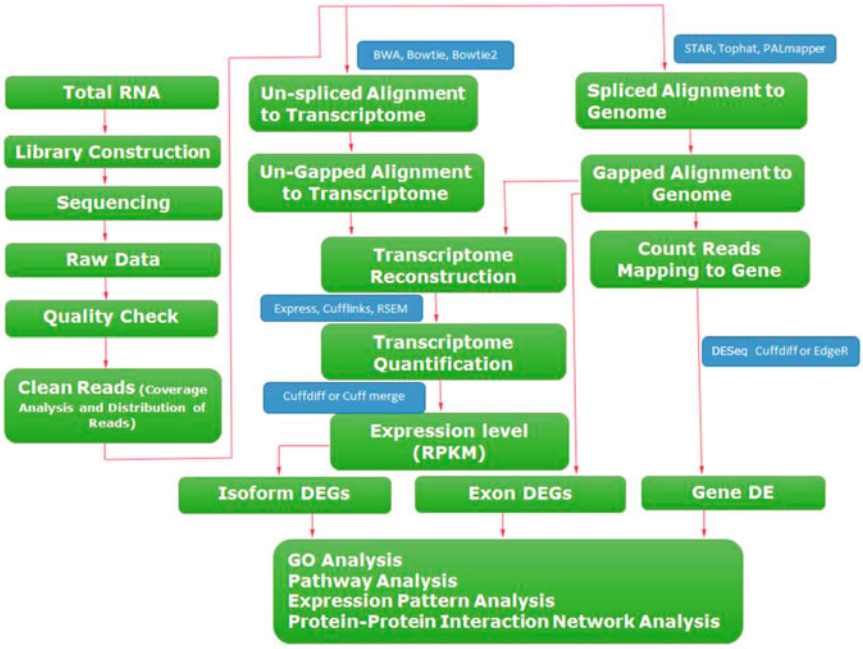

Fig. 1.1 RNA-Seq workflow. RNA-Seq begins with isolation of high-quality total RNA followed by conversion into cDNA, fragmentation, and adaptor ligation. Fragmented cDNA is used to construct a library for sequencing. Raw data, consisting of reads of a defined length, are preprocessed according to a set of quality control metrics, such as base quality, minimum read length, untrimmed adaptors, and sequence contamination. After filtering and trimming, reads are aligned to a reference genome or transcriptome depending on the objective of the experiment and the nature of the samples. Subsequently aligned reads are assembled. RNA-Seq assembly involves merging reads into larger contiguous sequences based on similarity. After assembly, reads are quantified in order to measure transcriptional activity. Read counts are generally computed in RPKM of FPKM in order to perform further downstream analysis, such as differential expression, pathway and gene set overrepresentation analysis, and interaction networks

short sequences from one end or both ends are obtained. The reads are typically 30–400 bp, depending on the DNA sequencing technology used. There are various high-throughput sequencing platforms available for RNA-Seq such as Illumina, Applied Biosystems SOLiD, and Roche 454 Life Science systems. The total reads obtained after sequencing are either aligned to a reference genome or transcriptome, or assembled de novo without genomic sequence guidance to create a genome-scale transcription map providing both transcriptional structure and expression level for each gene.

RNA-Seq has several advantages over microarrays [25–27]. First, sequencing technology is much more sensitive and quantitative than microarrays and it can provide a large dynamic range of detection (>9000-fold) [28]. Additionally, sequencing data are more specific and have less background. Moreover, sequencing experiments do not depend on the limited features of tiled microarrays and can

therefore be used to interrogate any location in the genome and to detect and quantify the expression of previously unknown transcripts and splicing isoforms. Finally, sequencing is not limited by array hybridization chemistry, such as melting temperature, cross-hybridization, and secondary structure concerns.

1.2.5.1 Data Analysis General Workflow

RNA-Seq experiments result in millions of short sequence reads which require computational methods for comprehensive transcriptome characterization and quantification. Steps for data analysis vary according to the desired biological problem to be assessed and to the availability of reference genome or transcriptome data. A generic overview of the routine analyses performed is included in Fig. 1.1. The main tasks of data analysis are read mapping also referred to as alignment, transcriptome assembly, expression quantification, and downstream applications. Steps for data analysis, although it is sequential, may be performed with different computational tools and algorithms which require specific data formats and external files. It is desirable to automate the multiple data analysis steps in a pipeline. A pipeline is a reusable script with defined inputs, outputs, and parameters for each processing step. Several software platforms which collect different RNA-Seq analysis tools for each task have been developed such as PRADA [29], Tuxedo [30], MAP-R-SEQ [31], and GENE-Counter [32]. However, pipelines may be custom-built by users, selecting the most appropriate tools according to experimental data and desired downstream analysis [33]. The following subsections describe the different RNA-Seq data analysis tasks.

1.2.5.2 Quality Control and Preprocessing

The first step in data analysis is quality control. Accuracy in library preparation and sequencing steps contribute to the quality of reads, which if overseen may lead to erroneous mappings, misassembles, and false expression estimates. The quality control should include the assessment of read length, GC content, sequence complexity, sequence duplication, polymerase chain reaction artifacts, untrimmed adapters, low-confidence bases, 3′/5′ positional bias, sequence contamination, and fragment biases [34]. Quality control metrics are obtained directly from raw reads. Raw reads are typically in FASTQ format, text-based files which contain a sequence identifier, a nucleotide sequence, and its corresponding quality score [35]. A brief overview of the most important quality control processes to be performed is described next.

(a) Base Quality: Since RNA-Seq technologies rely on complex interactions between chemistry, hardware, and optical sensors, sequencing platforms provide metrics for assessing error probabilities. Base quality is measured by computing the confidence on base calling, the process by which the sequencer

analyzes colorimetric sensor signals to predict individual bases. Base calling quality values are expressed in *Phred* scale, an error probability log transformation which has the advantage of converting low error probabilities to high-quality values and vice versa [36]. Quality values q are calculated for each base as:

$$q = -10 \log_{10}(p) \tag{4.1}$$

where p is the estimated error probability of a base call. Base quality values are encoded with ASCII characters with the sequence data in FASTQ format [35]. Typically, good reads should have base qualities greater than 20; however, this threshold depends on the platform used. It is important to inspect the reads' base quality distribution to detect regions of poor base quality, which may be filtered or trimmed preserving the order of reads, thus increasing mapping efficiency. This process is also referred to as quality trimming and will be discussed next.

(b) Filtering and Trimming: Reads should be inspected for the presence of sequencing adapters, tags, and contaminating sequences, which should be removed before quality control processing. Adapter sequences and tags used during library construction should be removed prior to mapping. Contaminating sequences such as DNA, rRNA, or sequences from other organisms or vectors should be filtered out.
Additionally, reads should be filtered according to mean base quality or to the proportion of bases whose quality is below a user-defined threshold. There is no consensus on the optimal base quality threshold for trimming, and it is rather a trade-off between mapping efficiency and coverage. Software for filtering generally outputs synchronized filtered reads. When low-quality bases are located at the ends of reads, trimming is a better option than filtering. The basic principle of read trimming is to assess base quality keeping the longest possible high-quality read segments. Trimming with respect to base quality may be performed using running sum algorithms or sliding window-based algorithms. Running sum algorithms basically find the summation of the differences between all base quality values against a quality threshold, and sequences are trimmed at the base that makes the running sum minimum. When analyzing reads with a sliding window, the user defines a window size and a mean base quality threshold. Depending on the tool used, the window may slide from the 5′ or the 3′ ends of reads. Sliding the window from the 5′ end will trim the read until the window's quality lies below a threshold, maintaining the beginning of the sequence, whereas sliding from 3′ end will trim until a passing quality window is encountered. An excellent evaluation on the performance of several trimming tools was published by Del Fabbro et al. [37]. Some common tools used for trimming are Trimmomatic [38], Cutadapt [39], PRINSEQ [34], and ConDeTri [40].

Reads with a high frequency of ambiguous bases, bases not identified during sequencing, should be filtered out since they can lead to erroneous mapping and misassemblies. Low-complexity sequences (homopolymers, di-trinucleotides) will also result in mapping errors and should therefore be trimmed.

(c) GC Content Determination: Another important metric that should be considered is reads' mean GC content, which if plotted should follow a normal distribution centered on the organism's normal content. Variations on the GC content might be due to the PCR amplification process and therefore may be sample specific. An approach to reduce GC content systematic bias is conditional quantile normalization, a technique in which the distribution of read counts is modified by estimating quantiles obtained with a median regression on a subset of genes [41].

(d) Minimum Read Length: The distribution of read lengths should be verified after trimming since reads may have ended up as very short fragments, which become difficult for mapping. The minimum read length is a user-defined variable, and its value depends on desired downstream applications.

(e) Fragment Biases: RNA fragmentation creates segments whose starting points were assumed to be located uniformly at random within a transcript. However, it has been demonstrated that there are both positional [42] and sequence-specific [43, 44] biases derived from fragmentation and reverse transcription. Positional bias describes the fact that reads' starting positions are non-uniformly distributed since they are preferentially located toward the transcripts' boundaries. Sequence-specific bias refers to the phenomenon by which the sequences at the reads' boundaries, such as the random hexamer primers used for reverse transcription priming, introduce biases in nucleotide composition and influence the likelihood of being sequenced. Furthermore, fragment length also generates bias since long transcripts result in more reads mapping to them than smaller transcripts. Thereby, for genes with equal levels of expression, the long genes will be overrepresented, distorting the relative expression among genes [45]. Since RNA-Seq read counts are proportional to transcript abundances, expression estimates should be made after fragment bias correction. An effective approach for fragment bias correction has been implemented in the Cufflinks [46] transcriptome assembly and differential expression RNA-Seq analysis tool. The fragment bias correction was based on an algorithm which learns the read sequences and models them as a likelihood function involving abundance and bias parameters such as the probability of finding a fragment with a specific length in a given position [47]. In this manner, bias and expression estimation are performed simultaneously.

1.2.5.3 Read Alignment

In order to determine transcript abundance from reads, it is necessary to align or map reads to a previously assembled reference genome or transcriptome to determine the read's origin. Mapping to a reference genome is more common since it

increases the potential information which may be obtained (e.g., identification of novel transcripts and genes) and because many transcriptomes are incomplete. Mapping is a challenging task since RNA-Seq reads are relatively short and they may match non-contiguous regions of the genome due to splice junctions. Furthermore, alignment tools must cope with mismatches and indels caused by genomic variation and sequencing errors. Many alignment tools have been developed. A comprehensive list of alignment tools and their properties was initially published by Fonseca et al. [48] and is kept updated on the Web [49]. A list of some common aligners and their main properties is included in Table 1.1.

The main consideration to be addressed when selecting an aligner is whether RNA-Seq reads span splice junctions. As depicted in Fig. 1.1, unspliced or contiguous alignment tools such as BWA [50], Bowtie [51], and Bowtie2 [52] are useful when mapping reads to a transcriptome, when sequencing microRNAs, or when the organism under study has no intronic regions. Unspliced aligners are thus limited to identifying known exons and do not allow for new splicing event identification. Spliced alignment tools are used when mapping to reference genomes without relying on previously known splice sites. Some of the most commonly used tools for spliced alignment are TopHat [53], TopHat2 [54], Palmapper [55], and STAR [56].

Alignment to a reference genome starts with indexing, the process with which auxiliary structures called indices are created for either the reference sequence or the sequenced reads to allow for faster queries. Indexing the reference genome is more time efficient and thus is used by most alignment tools. Alignment algorithms used for sequencing data analysis are mainly classified into hash tables and suffix trees according to the property of the index used. Hash table indexing was first introduced as an alignment tool by BLAST [67], using a seed and extend approach. In hash table indexing, reads are divided into short *k-mer* subsequences called "seeds" and stored in a hash table. The algorithm assumes that at least one "seed" in a read will match the reference. Once a "seed" is aligned, it is extended using more sensitive algorithms such as Smith–Waterman [68] or Needleman–Wunsch [69]. Modifications to hash table indexing algorithms have been performed, and they have been implemented in Novoalign [59], MAQ [65], SHRiMP2 [70], and BFAST [57], among other alignment tools. Suffix trees, on the other hand, are based on the premise that an inexact matching problem may be converted into an exact matching task by constructing a tree (an ordered tree data structure) with all the possible substrings that make up a sequence. The suffix tree data structure enables fast substring searches regardless of sequence size [71]. Among different suffix tree algorithms, one of the most efficient is the FM-index [72] which is based on the Burrows–Wheeler transform (BWT) [73]. BWT is a reversible permutation of characters in a string, and FM-indexing addresses permutations (nodes in a tree) constantly using a backward search. FM-index and BWT, both originally designed for data compression, have been successfully implemented for storing reference genomes and performing rapid queries.

Table 1.1 Overview of common alignment tools

ALIGNERS	Operating system	Language	Alignment algorithm	Input	Output	Paired-end mapping	Splice junction	Read length range	Ref.
BOWTIE	Unix-based, windows	C++	FM-index based on BWT	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–1 k	[51]
BOWTIE2	Unix-based, windows	C++	FM-index based on BWT, dynamic programming	FAST(A/Q)	SAM, TSV	Yes	No	4 bp–5000 k	[52]
PALMapper	Unix-based, web interface	C++	Reference indexing	FAST(A/Q)	SAM, BED (x), SHORE	Yes	Yes	12 bp–12 k	[55]
STAR	Unix-based	C++	Reference indexing	FAST(A/Q)	SAM	Yes	Yes	15 bp–10 k	[56]
BFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, TSV	Yes	No	25–100 bp	[57]
GENOME-MAPPER	Unix-based	C	Reference indexing	FAST (A/Q), SHORE	BED, SHORE	No	No	12 bp–2 k	[58]
NOVAALIGN	Unix-based	C++	Reference indexing	FAST (A/Q), CSFASTA	SAM	Yes	Yes	1–250 bp	[59]
SHRiMP2	Unix-based	Python	Reference indexing	FAST(A/Q)	SAM	Yes	No	30 bp–1 k	[60]
SOAP2	Unix-based	C++	BWT + reference indexing	FAST(A/Q)	SAM/BAM	Yes	No	27 bp–1 k	[61]
MrFAST	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, DIVET	Yes	No	25 bp–1 k	[62]
GNUMAP	Unix-based	C	Reference indexing	FAST(A/Q)	SAM, TSV	No	No	16 bp–1 k	[63]
RMAP	Unix-based	C++	Read indexing	FAST(A/Q)	BED	Yes	No	11 bp–10 k	[64]

(continued)