Quantitative Biology 2016, 4(1): 22–35
DOI 10.1007/s40484-016-0060-7

Review

Mapping and differential expression analysis from short-read RNA-Seq data in model organisms

Qiong-Yi Zhao[1,*], Jacob Gratten[1], Restuadi Restuadi[1] and Xuan Li[2,*]

[1] The University of Queensland, Queensland Brain Institute, St Lucia, Qld 4072, Australia
[2] Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
* Correspondence: q.zhao@uq.edu.au, lixuan@sibs.ac.cn

Received November 16, 2015; Revised December 21, 2015; Accepted December 24, 2015

Recent advances in next-generation sequencing technology allow high-throughput RNA sequencing (RNA-Seq) to be widely applied in transcriptomic studies. For model organisms with a reference genome, the first step in analysis of RNA-Seq data involves mapping of short-read sequences to the reference genome. Reference-guided transcriptome assembly is an optional step, which is recommended if the aim is to identify novel transcripts. Following read mapping, the primary interest of biologists in many RNA-Seq studies is the investigation of differential expression between experimental groups. In this review, we discuss recent developments in RNA-Seq data analysis applied to model organisms, including methods and algorithms for direct mapping, reference-guided transcriptome assembly and differential expression analysis, and provide insights for the future direction of RNA-Seq.

Keywords: RNA-Seq; mapping; reference-guided transcriptome assembly; differential expression analysis

INTRODUCTION

RNA-Seq is an application of next-generation sequencing (NGS) technologies to perform transcriptome-wide profiling. As one of the most cost-effective approaches, RNA-Seq has been widely applied in humans, model organisms and non-model species and has provided unprecedented insights into the transcriptomic landscape [1–8]. The versatile applications of RNA-Seq include (i) whole transcriptome reconstruction based on *de novo* transcriptome assembly [9], (ii) identification of novel transcripts [4], (iii) detection of differentially expressed genes [10] or transcripts [11] between experimental groups, (iv) detection of alternatively spliced isoforms [12], (v) detection of allele-specific expression [13], (vi) construction of co-expression networks [14], (vii) identification of RNA editing sites [15], and (viii) identification of DNA variations in gene regions [16]. Many of these applications have been the subject of recent reviews [2,17–20], but the field is rapidly evolving, particularly with respect to methods for mapping of short-read RNA-Seq data in model organisms, which is a fundamental step for all forms of RNA-Seq data analysis. In this review we focus on recent progress in read-mapping algorithms for RNA-Seq data and reference-guided transcriptome assembly,which is recommended if the aim is to detect novel transcripts. Additionally, we discuss the latest developments in differential expression analysis from RNA-Seq data, which is the primary interest of biologists in many RNA-Seq studies. We conclude with a perspective on future directions for RNA-Seq.

STRATEGIES FOR TRANSCRIPTOMIC ANALYSIS WITH A REFERENCE GENOME

For organisms with a reference genome, direct mapping to the reference and/or reference-guided transcriptome assembly are more computationally efficient than *de*

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

novo assembly and are the most commonly used strategies.

Direct mapping is a straightforward option for transcriptomic analysis in model organisms with a well-annotated reference genome or transcriptome. Using this strategy, RNA-Seq reads are directly aligned to the reference genome or to transcript sequences using mapping tools such as Tophat [21], Tophat2 [22], HISAT [23], HISAT2 [23], MapSplice [24], SOAPSplice [25] or STAR [26] for splice-junction mapping, or Bowtie [27], Bowtie2 [28], BWA [29], BWA-SW [30], BWA-MEM [31], SOAP [32] or SOAP2 [33] for non-splice-junction mapping. Based on the annotation, each feature (i.e., gene, transcript or exon) is assigned a count value or a normalized count value by counting the number of RNA-Seq reads covering the feature, with these count values representing the relative abundance of features in the transcriptome. Comprehensive annotation is advantageous for this approach, but a simulation study has shown that the method is robust to the presence of incomplete annotation and any incorrect transcripts present in a curated set do not absorb much signal [34]. In summary, direct mapping to the reference is a popular approach for analysis of RNA-Seq data, both because the analysis workflow is straightforward and due to the availability of many well-developed downstream software tools (e.g., edgeR [35], DESeq [36], DESeq2 [37], SAMseq [38], baySeq [39], NOIseq [40], limma [41], NBPSeq [42], TSPM [43] and EBSeq [44] for differential expression analysis).

Reference-guided transcriptome assembly is a more ambitious approach for transcriptomic analysis. This method involves aligning reads to a reference genome and uses both the alignment outcomes and curated annotations to infer the transcript structures. This strategy is attractive because it can leverage a reference genome and existing annotations for the discovery of novel transcripts. In theory, this strategy is superior to the direct mapping approach because it offers the possibility of obtaining a more complete set of gene/transcript sequences, as has now been shown in many RNA-Seq studies [11,45,46], whereas direct mapping relies on current annotations for model organisms that are often incomplete. However, a potential caveat is that due to typical limitations in RNA-Seq data, such as short read length, sequencing errors and biases, and/or errors introduced during alignment and assembly, reference-guided transcriptome assembly may generate massive partial transcripts and even assembled artefacts that can confound transcriptomic analyses. Recent studies have shown that these assembled artefacts can account for a substantial proportion of the signal when performing expression analyses [34]. In addition, compared to the direct mapping approach, there are fewer downstream tools supporting this methodology. Cufflinks [11,47] and Scripture [45] were the first software tools to implement reference-guided transcriptome assembly. Trinity was initially designed for *de novo* transcriptome assembly [48], but now it also offers reference-guided transcriptome assembly in recently released versions. More recently a related method called StringTie has been released that claims to have improved performance compared to Cufflinks [49]. Reference-guided transcriptome assembly is unquestionably the best option if the objective is to identify novel transcripts. Alternatively, direct mapping to the reference is arguably the best choice for analysis of RNA-Seq data in well-annotated model organisms.

MAPPING ALGORITHMS FOR SHORT-READ DATA

Ideally, the first step in analysis of RNA-Seq data would involve mapping of short-read sequences to a reference transcriptome. However, because the complexity of the transcriptome is incompletely annotated, even for well-studied species, mapping RNA-Seq reads to a reference genome is preferable for organisms whose reference genomes are available.

A wide variety of mapping algorithms and software tools have been developed over the past few years. For example, more than 60 aligners are listed in Fonseca et al. (2012)'s study [50] and the number continues to increase (e.g., 84 aligners were listed in http://www.ebi.ac.uk/~nf/hts_mappers/ as at 09/2015). The growing number of aligners is indicative of the importance of sequence alignment to the research community and is evidence of the active development of mapping tools. However, it also present challenges to researchers in terms of selecting a suitable aligner for their studies.

Mapping tools for short-read sequencing data can be divided into two major groups: (i) unspliced aligners that are designed to align continuous reads to a reference without consideration for splicing junctions, and (ii) spliced aligners that are capable of splitting reads at intron-exon boundaries. For RNA-Seq studies, unspliced aligners are mainly applied when (i) organisms do not contain introns in their genomes (e.g., most bacteria and some eukaryotic microorganisms), or (ii) sequence reads are mapped to a library of known transcript sequences (i.e., a reference transcriptome) rather than a reference genome sequence. On the other hand, spliced aligners are capable of mapping RNA-Seq data to a reference genome. Below we briefly discuss the mapping algorithms and tools that are commonly applied to RNA-Seq data. Note that most unspliced aligners discussed below are designed for short-read NGS data rather than specifically for RNA-Seq data. However, these unspliced aligners can be

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

applied in RNA-Seq studies under the aforementioned scenarios.

Three broad categories of mapping algorithms are commonly used in analysis of short-read data (reviewed in Ref. [51]): algorithms based on hash tables, algorithms based on suffix trees and algorithms based on merge sorting (Figure 1).

Algorithms and tools based on hash tables

A hash table is a key-value data structure used to implement an associative array. The most important feature of this type of data structure is that it can map keys to values very efficiently. The idea of hash table based algorithms, which essentially follow the seed-and-extend paradigm by matching a short section (i.e., a seed) of each read to the reference and extending these seed matches to the full length of the read, can be traced back to when the BLAST algorithm was first developed [52,53].

Arguably, Eland was the first successful aligner integrated into the Illumina data processing package that utilized the seed-and-extend paradigm in short-read

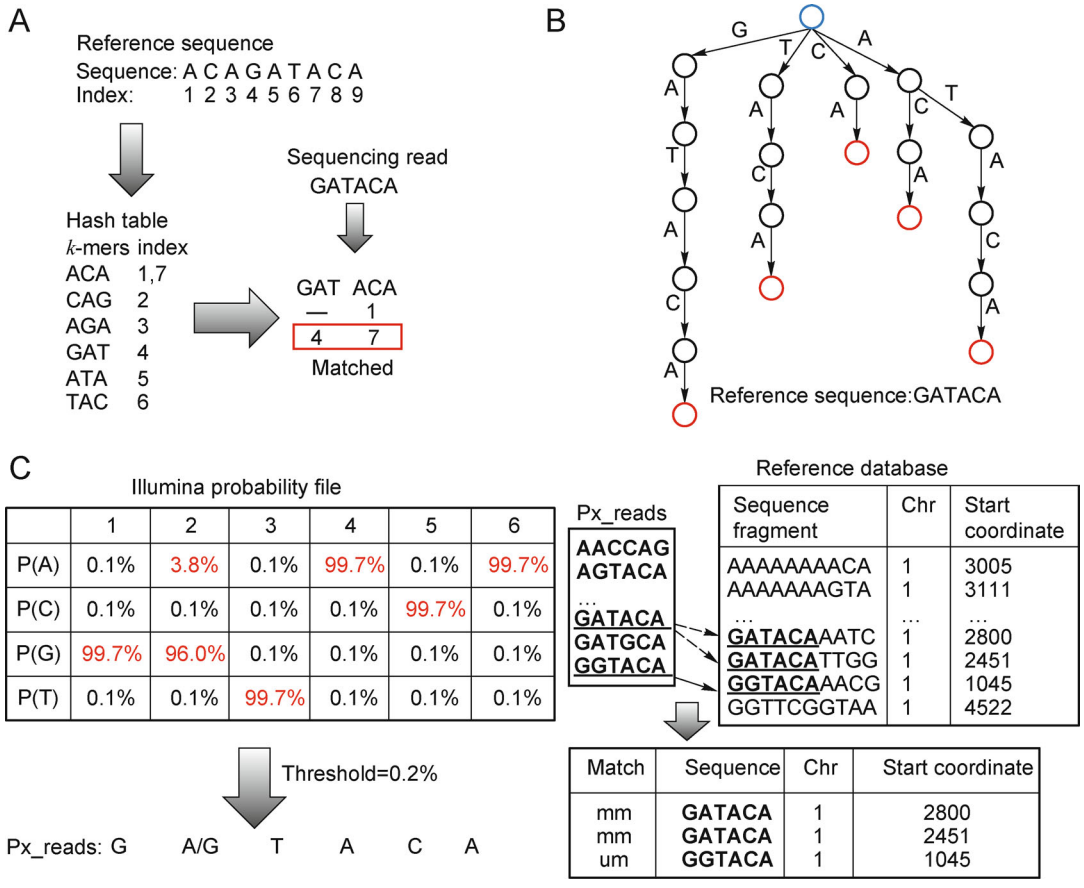

Figure 1. The general concept and data structures for three broad categories of mapping algorithms. (A) Algorithms based on hash tables. A set of *k*-mers and their positional indexes are extracted from each reference sequenceand stored in a hash table. The alignment is performed by searching *k*-mers in sequencing reads against keys in the hash table with matched index(es). In this example, 3 bp *k*-mers and their indexes are generated and stored in a hash table. (B) Algorithms based on suffix trees. A reference sequence "GATACA" is shown as an example for the data structure based on suffix trees. The five paths from the root (blue circle) to the leaves (red circles) corresponding to the five suffixes (from left to right: "GATACA", "TACA", "CA", "ACA" and "ATACA"). (C) Algorithms based on merge sorting. Illumina probability files generated from the Illumina Genome Analyzer platform are used as input files. Shown is an example where we assume a read is 6 bp (left side). All possible sequences are generated as Px_reads considering every possible nucleotide with a probability higher than a certain threshold (a 0.2% value is used by default). The algorithms scans both the lexographically sorted reference database and the lexographically sorted Px_reads to generate all unique matches (um) and multiple matches (mm). In this example, the set of imput sequences is 6 bp, which is aligned to a reference database of 10 bp oligos created by a sliding window across the reference sequences. Reads that match are indicated in bold and underlined with an example of a unique match (solid line) and a multiple match (dashed line).

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

alignment (A. J. Cox, unpublished). The concept of Eland is to split a read into segments, creating a memory-resident hash table for all read segments and scanning inexact matches using combinations of segments as exact hash-keys. This seed strategy is also known as space seed. This approach inspired the development of many other short-read aligners based on space seed, such as SOAP [32], MAQ [54], RMAP [55,56], and ZOOM [57], among others. The downside of the space seed approach is that gaps are not permitted within the seed. More recent methods have sought to overcome this limitation by use of dynamic programming to detect gaps during the extension step or by attempting small gaps at each read position [32,58]. Ultimately, the problem was overcome by the *q*-gram filter and multiple seed hits approaches. The *q*-gram filter is based on the observation that the substrings of an approximate match must have a certain number of *q*-grams (i.e., strings of length *q*) in common [59]. In general, methods based on space seed and *q*-gram are similar insomuch as they both rely on a hash table for fast and exact matching. Space seed initiates seed extension from one long-seed match while *q*-gram initiates extension usually with multiple relatively short-seed matches. SHRiMP [60] and RazerS [61] are two successful examples implementing the *q*-gram filter that provides a solution to building an index that allows gaps. Later, RazerS 3 [62] was developed as a successor to RazerS with a superior running time and the capability to mapping reads of various lengths with many insertion and deletion errors. In addition to using the *q*-gram filter, RazerS 3 makes use of Open Multi-Processing to provide a share-memory parallelization with dynamic load balancing, a pigeonhole-based filter with controllable sensitivity, and an implementation of a banded version of Myers' bit-vector algorithm for verification to improve the performance on both running time and sensitivity [62].

Major improvements on seed extension were also achieved by accelerating the standard Smith-Waterman algorithm with vectorization (i.e., multiple query sequences can be processed in one CPU cycle) and by constraining dynamic programming around seeds. These improvements enabled significant acceleration of the alignment process. For example, the striped Smith-Waterman algorithm (i.e., the Smith-Waterman implementation where the Single-Instruction Multiple-Data (SIMD) registers are parallel to the query sequence, but are accessed in a striped pattern) achieved a 2–8 fold performance improvement over other SIMD based Smith-Waterman implementations [63]. Novoalign (http://www.novocraft.com/products/novoalign/), CLC Genomics workbench (http://www.clcbio.com/products/clc-genomics-workbench/), SHRiMP [60] and SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) are examples of software that utilize the accelerated Smith-Waterman algorithm in the alignment. BWA-MEM [31]also recently joined this category, and introduced several innovations including seeding and re-seeding, improved seed extension, and chaining (i.e., linking a group of seeds that are collinear and close to each other) and chain filtering (i.e., filtering overlapping short chains by some criteria), all designed for optimal alignment of 70 bp or longer reads.

Algorithms and tools based on suffix trees

A suffix tree is a compressed trie containing all the suffixes (i.e., substrings) of the given sequence (e.g., a genome sequence) by pre-processing the sequence data into a space-efficient data structure. After the construction of suffix trees, fast query searches can be performed easily, for instance by locating a substring with specific mismatches. Algorithms based on suffix trees essentially reduce the inexact matching problem to the exact matching problem. This is achieved by first identifying exact matches and then building inexact matches supported by these exact matches [51].

Use of a trie greatly enhances alignment efficiency because multiple loci that share an identical substring in a reference need only be aligned once (since identical alignments collapse on a single path in the trie), whereas alignment needs to be performed independently for each locus using the hash table approach.The suffix tree is undoubtedly one of the most important and widely used data structures in string processing. However, algorithms based on suffix trees are memory intensive because even the most space efficient implementation [64] requires at least 12.5 bytes per bp, which equates to $>$ 37 G bytes for the human genome (~3 Gbp). Continuous improvements have been made to overcome this obstacle, culminating in the enhanced suffix array [65] and FM-index [66]. An enhanced suffix array uses a basic suffix array enhanced with several auxiliary arrays, leading to a reduction in space consumption to 6.25 bytes per bp. An FM-index (Full-text index in Minute space) is a compressed full-text substring index based on Burrows-Wheeler transform [67], which allows compression of the input text while still supporting fast substring queries.

A number of publicly available aligners have been developed based on suffix tree algorithms. For example, Segemehl [68] use an enhanced suffix array, Bowtie [27], BWA [29], SOAP2 [33], and BWA-SW [30] are based on the FM-index. Bowtie2 combines ultrafast FM-index-based seeding with efficient extension by dynamic programming in order to obtain gapped alignments [28]. RSEM is a software package for quantifying gene and isoform abundances from short-read RNA-Seq data [69].

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

It uses the Bowtie/Bowtie2 alignment program to align reads against transcript sequences rather than a genome reference, with parameters specifically chosen for transcript quantification from RNA-Seq data (e.g., the "--estimate-rspd" option enables RSEM to use the data to learn how RNA-Seq reads are distributed across a transcript). TopHat is one of the few tools that supports splice junction mapping for RNA-Seq reads. It first maps RNA-Seq reads to a genome reference using Bowtie, and then analyses the mapping results to identify splice junctions between exons [21]. TopHat2 [22] is the descendant of TopHat. By using Bowtie or Bowtie2 as the underlying mapping engine and adopting a two-step approach — these being (i) detection of potential splice sites for introns and (ii) use of these candidate splice sites in a subsequent step to correctly align multiexon-spanning reads — TopHat2 is able to align reads spanning insertions and deletions on the same chromosome, even if these are very large, and reads spanning translocations involving different chromosomes [22]. MapSplice [24] and SOAPSplice [25] use a similar two-step approach for splice junction mapping. Spliced Transcripts Alignment to a Reference (STAR) is another popular splice junction mapper based on an algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and a stitching procedure [26]. Another recent and highly promising method is HISAT [23] and its upgraded version HISAT2. In addition to using one global FM index that represents a whole genome, HISAT and HISAT2 use massive local FM indexes that collectively cover the whole genome for the effective alignment of RNA-Seq reads [23].

Algorithms and tools based on merge sorting

The alignment algorithm based on merge sorting uses not only the most probable base, but also all possible bases with a probability above a certain base probability threshold provided by the Illumina probability file. It then generates all possible reads with probability above a certain read probability threshold. For the core alignment, it sorts all these generated reads in lexicographical order and then crosses it sequentially with a pre-sorted table of windows of reference sequences and their reverse complements. This approach eliminates the need for an indexed structure by replacing random I/O with sequential I/O. Currently, the only software using this approach are Slider [70] and its descendant SliderII [71].

Summary for mapping algorithms and tools

In general, the strengths of hash table based algorithms are that they can tolerate high levels of genomic variation and easily perform partial alignment (such as for exon-exon junction reads), but this comes at the cost of high memory requirements for hashing and poor sensitivity for alignment of reads in repetitive regions. In comparison, the strengths of algorithms based on suffix trees are that they are able to perform fast alignment, especially for exact matches, and these algorithms offer alignments with high sensitivity in repetitive regions. However, suffix tree based algorithms are generally less tolerant of high genomic variation than hash table based algorithms. Merge sorting based tools, since Slider and SliderII, are becoming less popular, primarily because they use Illumina probability files as input rather than more standard file format (such as fastq), and recent sequencing platforms (such as HiSeq 2000) do not provide Illumina probability files. Table 1 lists some popular aligners that have been widely applied in short-read sequence alignment. Note that this is not a complete list. Readers are referred to Fonseca *et al*. (2012)'s study [50] and the well-maintained high-throughput sequencing mappers website (http://www.ebi.ac.uk/~nf/hts_mappers/) for a more comprehensive list of aligners.

DIFFERENTIAL EXPRESSION ANALYSIS

The primary goal of most RNA-Seq studies is to identify differentially expressed genes (DEGs) and/or differentially expressed transcripts (DETs) between experimental groups. Prior to data analysis, quality control is usually performed to assess the quality of sequencing reads, including sequence quality scores, GC content, sequence duplication levels. There are a number of tools that are designed for this purpose, such as FastQC and FASTX-Toolkit. To quantify gene expression, RNA-Seq reads need to be aligned to a reference genome for model organisms (e.g., using HISAT2 [23]) or to a library of transcriptome sequences reconstructed using *de novo* assembly strategies for organisms without reference sequences (e.g., using Trinity [48] for *de novo* assembly, and RSEM [69] for mapping and detection of DETs). If detecting novel isoforms is of interest in a study, then reference-guided assembly needs to be performed (e.g., using StringTie [49]), followed by a merge step to generate a non-redundant set of transcripts (e.g., using Cufflinks-Cuffmerge [72]) for downstream analyses. Following alignment, the expression level of genes/transcripts is quantified by counting the number of reads aligned to each feature (e.g., using HTSeq or StringTie [49] for generating gene-level or transcript-level count tables, respectively). Subsequently, a range of statistical methods can be applied to assess the significance of differences in expression level observed between experimental groups (e.g., using edgeR [35] or Ballgown [73] for detection DEGs or DETs, respectively). A general

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

Table 1. Popular short-read aligners.

Aligner	Spliced (Y/N)	Supported NGS data
Aligners based on hash tables		
BWA-MEM	N	Illumina (>70 bp), 454 and long-read data (e.g., PacBio). The developer of BWA recommends BWA-MEM over BWA-SW as it is faster and more accurate than BWA-SW
CLC genomics	N	Almost all NGS data (commercial software package)
Eland	N	Illumina (implemented by Illumina)
MAQ	N	Illumina. MAQ has not been maintained since 2008. The developer of MAQ recommends people to use other tools (such as BWA) rather than MAQ
Novoalign	N	Illumina (commercial software package)
RazerS	N	Illumina
RazerS 3	N	Illumina, 454 and long read platforms (e.g., PacBio). RazerS 3 is a successor to RazerS
RMAP	N	Illumina and bisulfite-treaded Illumina reads
SHRiMP	N	Illumina and SOLiD. SHRiMP has not been maintained since 2012
SMALT	N	Illumina and 454
SOAP	N	Illumina
ZOOM	N	Illumina and SOLiD
Aligners based on suffix trees		
Bowtie	N	Illumina, 454, SOLiD. Works best when aligning short reads to large genomes
Bowtie2	N	Illumina, 454 and long-read data. For reads >50 bp, Bowtie2 is generally faster, more sensitive, and uses less memory than Bowtie
BWA	N	Illumina ($\leqslant 100$ bp)
BWA-SW	N	Illumina (>70 bp), 454. BWA-SW has better sensitivity when alignment gaps are frequent
HISAT	Y	Illumina, 454. HISAT is >50 times faster than TopHat2 with better alignment quality
HISAT2	Y	Illumina, 454. HISAT2 is a successor to both HISAT and TopHat2
Segemehl	N	Illumina and bisulfite-treaded Illumina data, 454 and long-read data
SOAP2	N	Illumina
TopHat	Y	Illumina, 454, SOLiD. It uses Bowtie or Bowtie2 as the underlying mapping engine
TopHat2	Y	Illumina, 454, SOLiD. TopHat2 is a successor to TopHat
Aligners based on merge sorting		
Slider	N	Data from Illumina Genome Analyzer
SliderII	N	Data from Illumina Genome Analyzer

workflow for the differential expression analysis is illustrated in Figure 2.

Tools and methods for RNA-Seq differential expression analysis

Accurate quantification of gene expression and detection of DEGs and DETs is non-trivial [74,75] due to (i) biases and errors inherent in NGS technology [76–78], (ii) biases of abundance measures due to the effects of nucleotide composition and the varying length of genes or transcripts [79,80], (iii) undetermined effects of both sequencing depth and the number of replicates, (iv) the mixture of technical and biological variation, and (v) the existence of alternative gene isoforms and overlapping sense-antisense transcripts [72]. A lot of efforts have been made to address these difficulties [72,81,82]. In early RNA-Seq studies lacking biological replicates, the distribution of feature counts across technical replicates was reported to fit well to a Poisson distribution where the variance is equal to the mean [76,83]. However, when biological replicates are included in RNA-Seq studies, the Poisson distribution underestimates the variation seen in many studies [84,85], a problem known as overdispersion. Several methods have been proposed to account for overdispersion in RNA-Seq differential expression analysis, including Auer *et al.*'s (2011) [43] two-stage Poisson model based on quasi-likelihood, the negative binomial (NB) distribution [35,36], and non-parametric methods such as NOISeq [40] and SAMseq [38]. Among all these methods, NB has achieved a dominant position in the methodologies to model feature counts for RNA-Seq data [35,36,80] due to the capability of accounting for both technical and biological variance. A number of software tools were

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

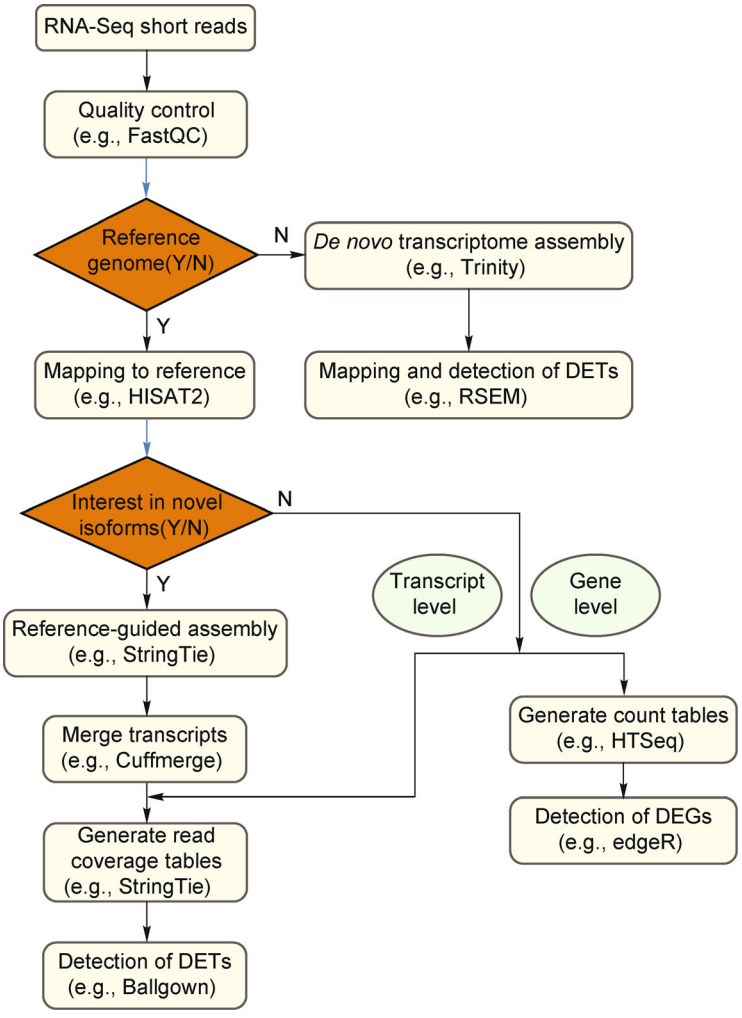

Figure 2. The workflow for RNA-Seq differential expression analysis.

developed based on NB, including DESeq [36], DESeq2 [37], edgeR [35], and baySeq [39], among others.

Although most existing tools were developed for differential expression analysis at the gene level, it is worth noting that Cufflinks-Cuffdiff [11] and its upgraded version Cuffdiff2 [72] implemented a more ambitious method for transcript-level differential expression analysis. Cuffdiff2 estimates count variances for each transcript among biological replicates under a beta negative binomial model of fragment count variability [72]. Another software package, RSEM, computes maximum likelihood abundance estimates at transcript-level resolution using the Expectation-Maximization algorithm for its directed graphical model [69]. A key feature of RSEM is that it only requires the user to provide a set of reference transcript sequences, such as one produced by a *de novo* transcriptome assembler, which allows for RNA-Seq analysis of species for which only transcript sequences are available [69]. Ballgown is another recently developed software tool that performs linear model-based differential expression analysis at transcript-level resolution [73]. It also offers functionality for visualization of the transcript assembly on a gene-by-gene basis and extraction of abundance estimates for exons, introns, transcripts or genes [73].

Differential expression analysis at transcript-level resolution is unquestionably an ideal approach as all RNA-Seq reads originate from transcripts whereas gene-based analyses represent a combination of all isoforms in the same gene locus. One simple scenario that illustrates this point is when two isoforms are differentially expressed in different directions (i.e., one isoform is up-regulated and the other isoform is down-regulated), in which case one may not detect any gene-level differtial expression. However, a key challenge in transcript-level quantification from RNA-Seq data is that lists of transcripts are incomplete, even for well-studied model organisms. As a consequence, if a gene has novel

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

isoforms, then RNA-Seq reads originated from these isoforms may be assigned to other known isoforms, leading to incorrect quantification of those known isoforms.

The field of differential expression analysis, although maturing, is still growing quickly and new software tools are continuously being developed. A few comparison studies have been reported to evaluate the performance of different RNA-Seq software tools. Soneson and Delorenzi [86] evaluated 11 software packages (DESeq [36], edgeR [35], NBPSeq [42], TSPM [43], baySeq [39], EBSeq [44], NOISeq [40], SAMseq [38], ShrinkSeq [87] and two versions of limma [36,41]) mainly based on simulated data sets and concluded that the method of choice in a particular situation depends on the experimental conditions. Rapaport *et al*. [88] evaluated six of the most commonly used differential expression software packages (Cuffdiff [89], edgeR [35], DESeq [36], PoissonSeq [90], baySeq [39], and limma [41]) by considering a number of key features, including normalization, accuracy of differential expression detection and differential expression analysis when one condition has no detectable expression. They found significant differences among the methods, but comparable performance was found between array-based methods (e.g., limma) adapted to RNA-Seq data and methods specially designed for RNA-Seq (e.g., edgeR). Seyednasrollah *et al*. [91] performed a systematic comparison of eight widely used software packages (edgeR [35], DESeq [36], baySeq [39], NOIseq [40], SAMseq [38], limma [41], Cuffdiff2 [72] and EBSeq [44]) for detecting differential expression between sample groups, focusing on measures that are of practical interest to researchers when analysing RNA-Seq data sets, including the number of DEGs identified using different numbers of replicates, their consistency within and between pipelines, the estimated proportion of false discoveries and the runtimes. They found marked differences among software packages, and the number of replicates and the heterogeneity of the samples should be taken into account when selecting the analysis pipeline [91]. Zhang *et al*. [75] have recently demonstrated that edgeR outperforms DESeq and Cuffdiff2 using both real and simulated RNA-Seq data sets by consideration of number of replicates, sequencing depth, and balanced vs. unbalanced sequencing depth within and between groups. These comparison studies have provided useful guidelines for a proper study design and a suitable software tool for RNA-Seq differential expression analysis. However, new software tools such as DESeq2 [37] and Ballgown [73] have since been developed and most existing tools have been upgraded (typically resulting in improved performance). The fast growing number of new tools and active development of existing tools also makes it difficult to choose the best (or the most suitable) software tool for differential expression analysis in a given RNA-Seq study, though edgeR and limma were previously reported to perform well under many circumstances compared with others [75,91].

Key factors in study design: sequencing depth and sample size

Sequencing depth and sample size are two key factors that affect differential expression analysis. Zhang *et al*. [75] have shown that the performance of Cuffdiff2 is sensitive to sequencing depth, whereas DESeq and edgeR appear relatively stable and thus are a better choice for differential expression analysis when sequencing depth is low (i.e., number of reads < 10 M). There is evidence that the number of DEGs discovered in RNA-Seq studies is positively correlated with sequencing depth [40,75], suggesting a strong effect of sequencing depth on differential expression analysis. Unbalanced sequencing depth between groups can also have negative effects on the performance of differential expression analysis for some software tools [75].

Another key factor for RNA-Seq differential expression analysis is the sample size (i.e., the number of biological replicates in each group). In theory, one would expect an increase in statistical power for the identification of DEGs with an increasing number of biological replicates, and indeed a positive correlation between DEGs and the number of biological replicates has been reported by Seyednasrollah *et al*. [91] and Zhang *et al*. [75]. However, different versions of software tools may have opposite effects on correlations between DEGs and the number of biological replicates, as discovered by Seyednasrollah *et al*. that with a different version of Cuffdiff2 the number of detected DEGs decreased when the number of samples increased [91].

Since budgetary constraints are common with RNA sequencing, an optimal experimental design needs to balance the sequencing depth for each sample with the number of replicates for each group. The consensus position of many studies [75,88,92] is that the overall impact of the sequencing depth is not as critical as sample size, and thus including sufficient biological replicates should be the prime consideration for RNA-Seq study designs. The required number of biological replicates depends on a number of factors including the amount of biological variation in the samples to be sequenced. Several studies have suggested that 4–6 biological replicates from inbred mice cell populations, and at least 14 biological replicates from human cell lines (unrelated individuals in the same ethnic group) are required for RNA-Seq differential expression analysis [75,91]. Larger sample sizes are likely to be required in animal/human tissue samples compared to cell lines or

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

cells from inbred lab strains. However, to determine the optimal number, more gold standard datasets and comprehensive evaluations based on these datasets are required to guide future RNA-Seq study designs.

FUTURE DIRECTIONS

Long-read RNA-Seq

Long reads have greater potential than short reads at many levels. For transcriptomic analysis with a genome reference, long-read RNA-Seq data has greater power than short-read data to (i) unambiguously map to the reference genome [51,93], (ii) detect indels and structural variations, especially for variants in repeat regions [94], (iii) produce full-length transcripts without assembly [95,96], (iv) resolve transcriptional complexity for gene loci with a massive number of isoforms and/or antisense transcripts [96,97], and (v) detect allele-specific expression and allele-specific AS patterns [93].

Roche 454 was the first high-throughput sequencing platform offering long-read sequencing using the pyrosequencing technology and sequencing-by-synthesis approach [98]. It can generate relatively long reads of up to 1 kp (average read length of 450 bp with the Roche 454 FLX Titanium sequencer [78]). The use of 454 sequencing has led to a better understanding of the structure of the human genome [99] since its launch in 2005, enabling the first non-Sanger sequence of an individual human [100] and opening up new approaches for transcriptomic studies [101].

The PacBio SMRT (single molecule real-time) sequencing platform, also known as one of the third-generation sequencing platforms, has been pioneered by Pacific Biosciences [102,103]. PacBio SMRT sequencing is built upon several key innovations (i.e., zero-mode waveguides and phospholinked nucleotides) that harness the natural process of DNA replication and enable real-time observation of DNA synthesis [102]. It offers long-read sequencing with an average read length > 10 kb, and a proportion of reads longer than 60 kb. Despite the relatively high error rate associated with the PacBio SMRT techonology, since the SMRT sequencing platform was commercially launched in early 2011, it has achieved many successful applications in the RNA-Seq field, including but not limited to obtaining comprehensive gene sets for non-model eukaryotes [95], characterization of full-length alleles in complex gene loci [96], and resolving the transcriptomic complexity [104]. Another important innovation based on the PacBio SMRT platform is Iso-Seq (the isoform sequencing: http://www.pacb.com/applications/rna-sequencing/), a method for the production of complete and unbiased full-length complementary DNA (cDNA) sequences without transcriptome reconstruction. This approach provides accurate information about alternatively spliced exons, transcriptional start sites and alternative polyadenylation sites directly from sequencing.

Oxford Nanopore technologies MinION offers a new approach for long-read sequencing. MinION uses the nanopore sequencing technology that can discriminate individual nucleotides by measuring the change in electrical conductivity as DNA molecules pass through the nanopore [105,106]. As the first commercially available sequencer that uses nanopores, MinION offers read lengths of tens of kilobases, with theoretically no instrument-imposed limitation on the size of sequenced reads [107]. An important feature of nanopore sequencingis that the sequencing process does not rely on DNA replication. It has the advantage of reading full-length molecules in real-time and has the potential for sequencing RNA without conversion to cDNA, which is extremely attractive because it has the potential of recognizing the modified RNA bases during real-time sequencing and therefore may shed light on the types and putative functions of RNA modifications. Currently, direct sequencing of RNA using the nanopore technology is yet to be developed but this development is expected in the near future.

The broad application of long-read sequencing is currently constrained by relatively high error rates of sequenced nucleotides and relatively high sequencing and computational costs [108] (e.g., it was estimated that 32.86 years CPU time would be required to process the PacBio raw reads for error-correction-overlap at ~ 44X sequencing coverage in Pendleton *et al*.'s study [108]), compared to short-read sequencing. Nonetheless, it is foreseeable that long-read sequencing will play a more important role in future RNA-Seq studies.

Single-cell RNA-Seq

Cells are the basic units of biological structure and function. Each tissue is a mixture of different cell types, and these subpopulations, or indeed individual cells in a single subpopulation, may have temporal and spatial variation in gene expression. There is a growing demand for single-cell profiling that is driven by the need for (i) direct analysis of rare cell types or cells with insufficient material for conventional RNA-Seq protocols, (ii) identification of cell subpopulations in tissues [109] and (iii) profiling interesting subpopulations of cells from a heterogeneous population [110]. To fully understand how complex tissues work in development and physiology, it will be important and essential to study transcriptional programs at single-cell resolution.

With the application of RNA imaging techniques such as RNA-FISH (fluorescent *in situ* hybridization targeting

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

ribonucleic acid molecules), single-cell measurements of gene expression are now possible. Previous studies have provided important insights into the dynamics of transcription and cell-to-cell variation in gene expression [111–113]. However, such approaches can only examine the expression of a small number of genes in each experiment, thus restricting our ability to perform transcriptome-wide examinations of gene expression and co-expression patterns.

Recent technological advances have enabled RNA-Seq whole-transcriptome analysis of a single cell [114]. Several such methods for profiling single cells have emerged, such as CEL-Seq [115], Smart-seq2 [116] and MARS-Seq [7]. Typically, these methods first separate the cells by fluorescence-activated cell sorting (FACS) [6] or microfluidics [8], and then amplify each cell's transcriptome separately for RNA-Seq, typically profiling hundreds to a few thousand cells in one experiment. To overcome the low-throughput issue, two droplet-based RNA-Seq approaches: inDrop RNA-Seq [117] and Drop-Seq [118] have recently been developed to enable fast profiling of the transcriptome for thousands of individual cells. Both approaches encapsulate cells into droplets and use novel barcoding strategies to match each mRNA to its cell-of-origin; inDrop RNA-Seq uses a microfluidic platform for droplet barcoding whereas Drop-Seq uses a split-pool synthesis approach to generate large numbers of distinctly barcoded beads into individual droplets [117,118]. Klein *et al*. also claimed that the inDrop RNA-Seq method has a theoretical capacity to barcode tens of thousands of cells per run [117], which will be important for its future application for profiling large populations of cells when sequencing throughput is high enough to afford multiplexing tens of thousands of cell samples in a single run. Meanwhile, G&T-Seq offers a powerful method for simultaneously sequencing a single cell's genome and transcriptome, thereby enabling direct identification of genetic variations and their effect on gene expression at single-cell resolution [104]. Macaulay *et al*. demonstrate the power of G&T-Seq by sequencing the genome and transcriptome of a single cell in parallel by discovery of many cellular properties that could not be inferred from DNA or RNA sequencing alone.

Methods and tools for single cell RNA-Seq analysis are only now beginning to emerge. Pollen *et al*. reported an analysis strategy for unbiased analysis and comparison of cell populations from heterogeneous tissue by microfluidic single-cell capture and low-coverage sequencing of many cells using existing tools [119]. Meanwhile, Trapnell *et al*. reported a tool kit called Monocle which is an unsupervised algorithm that increases the temporal resolution of transcriptome dynamics using single-cell RNA-Seq data collected at multiple time points[120]. Another recent method called scLVM (single-cell Latent Variable Model) has been developed to tease apart different sources of gene expression heterogeneity in single-cell transcriptomes, in particular that due to cell cycle-induced variation [109]. Interest in single-cell RNA-Seq is growing rapidly. It is foreseeable that single-cell RNA-Seq will significantly accelerate biological discovery by enabling routine transcriptional profiling at single-cell resolution, revolutionizing our view of the transcriptome. In addition, the integrated analysis of a cell's transcriptome, genome and eventually epigenome will enable a more complete understanding of the molecular machinery of cells and how this relates to higher order phenotypic variation.

ABBREVIATIONS

RNA-Seq, high-throughput cDNA sequencing; NGS, next-generation sequencing; SIMD, single-instruction multiple-data; FM-index, full-text index in minute space; DEG, differentially expressed gene; DET, differentially expressed transcript; NB, negative binomial; SMRT, single molecule real-time; cDNA, complementary DNA; Iso-Seq, the isoform sequencing; RNA-FISH, fluorescent *In Situ* hybridization targeting ribonucleic acid molecules; FACS, fluorescence-activated cell sorting.

ACKNOWLEDGEMENTS

The authors would like to thank Ms. Jennifer Whitehead for critical reading of the manuscript. This work is supported by the National Health and Medical Research Council project grants (Nos. APP1067795 and APP1087889), National Basic Research Program of China (No. 2012CB316501) and National Natural Science Foundation of China (No. 31571310).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Qiong-Yi Zhao, Jacob Gratten, Restuadi Restuadi and Xuan Li declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Wang, E. T., Sandberg, R., LuoS., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature, 456, 470–476
2. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet., 10, 57–63
3. Nilsen, T. W. and Graveley, B. R. (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature, 463, 457–463
4. Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., *et al.* (2011) The developmental transcriptome of *Drosophila melanogaster*. Nature, 471, 473–479
5. Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

species. Science, 338, 1587–1593
6. Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature, 498, 236–240
7. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science, 343, 776–779
8. Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature, 510, 363–369
9. Wang, X. C., Zhao, Q. Y., Ma, C. L., Zhang, Z. H., Cao, H. L., Kong, Y. M., Yue, C., Hao, X. Y., Chen, L., Ma, J. Q., *et al.* (2013) Global transcriptome profiles of *Camellia sinensis* during cold acclimation. BMC Genomics, 14, 415
10. Jhaveri, D. J., O'Keeffe, I., Robinson, G. J., Zhao, Q. Y., Zhang, Z. H., Nink, V., Narayanan, R. K., Osborne, G. W., Wray, N. R. and Bartlett, P. F. (2015) Purification of neural precursor cells reveals the presence of distinct, stimulus-specific subpopulations of quiescent precursors in the adult mouse hippocampus. J. Neurosci., 35, 8132–8144
11. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515
12. Shao, W., Zhao, Q. Y., Wang, X. Y., Xu, X. Y., Tang, Q., Li, M., Li, X. and Xu, Y. Z. (2012) Alternative splicing and trans-splicing events revealed by analysis of the *Bombyx mori* transcriptome. RNA, 18, 1395–1407
13. Muzzey, D., Sherlock, G. and Weissman, J. S. (2014) Extensive and coordinated control of allele-specific expression by both transcription and translation in *Candida albicans*. Genome Res., 24, 963–973
14. Hong, S., Chen, X., Jin, L. and Xiong, M. (2013) Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Res., 41, e95
15. Blanc, V., Park, E., Schaefer, S., Miller, M., Lin, Y., Kennedy, S., BillingA. M., Hamidane, H. B., Graumann, J., MortazaviA., *et al.* (2014) Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. Genome Biol., 15, R79
16. Piskol, R., Ramaswami, G. and Li, J. B. (2013) Reliable identification of genomic variants from RNA-seq data. Am. J. Hum. Genet., 93, 641–651
17. Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C. (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods, 8, 469–477
18. Martin, J. A. and Wang, Z. (2011) Next-generation transcriptome assembly. Nat. Rev. Genet., 12, 671–682
19. Ozsolak, F. and Milos, P. M. (2011) RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet., 12, 87–98
20. Han, L., Vickers, K. C., Samuels, D. C. and Guo, Y. (2015) Alternative applications for distinct RNA sequencing strategies. Brief. Bioinform., 16, 629–639
21. Trapnell, C., Pachter, L. and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111
22. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., 14, R36
23. Kim, D., Langmead, B. and Salzberg, S. L. (2015) HISAT: a fast spliced aligner with low memory requirements. Nat. Methods, 12, 357–360
24. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res., 38, e178
25. Huang, S., Zhang, J., Li, R., Zhang, W., He, Z., Lam, T., Peng, Z., Yiu, S. (2011) SOAPsplice: genome-wide *ab initio* detection of splice junctions from RNA-Seq Data. Front. Genet., 2,46
26. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21
27. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25
28. Langmead, B. and Salzberg, S. L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359
29. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760
30. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589–595
31. Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997
32. Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. Bioinformatics, 24, 713–714
33. Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25, 1966–1967
34. Jnes, J., Hu, F., Lewin, A. and Turro, E. (2015) A comparative study of RNA-seq analysis strategies. Brief. Bioinform., 16, 932–940
35. Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140
36. Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, R106
37. Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 15, 550
38. Li, J. and Tibshirani, R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res., 22, 519–536
39. Hardcastle, T. J. and Kelly, K. A. (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 11, 422
40. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. Genome Res., 21, 2213–2223
41. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 3, 1–25
42. Di, Y. M., Schafer, D. W., Cumbie, J. S. and Chang, J. H. (2011) The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat. Appl. Genet. Mol. Biol., 10

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

43. Auer, P. L. and Doerge, R. W. (2011) A two-stage Poisson model for testing RNA-Seq data. Stat. Appl. Genet. Mol. Biol., 10, 1–26
44. Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. and Kendziorski, C. (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics, 29, 1035–1043
45. Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., *et al.* (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol., 28, 503–510
46. Chen, G., Wang, C., Shi, L., Tong, W., Qu, X., Chen, J., Yang, J., Shi, C., Chen, L., Zhou, P., *et al.* (2013) Comprehensively identifying and characterizing the missing gene sequences in human reference genome with integrated analytic approaches. Hum. Genet., 132, 899–911
47. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 27, 2325–2329
48. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol., 29, 644–652
49. Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T. and Salzberg, S. L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol., 33, 290–295
50. Fonseca, N. A., Rung, J., Brazma, A. and Marioni, J. C. (2012) Tools for mapping high-throughput sequencing data. Bioinformatics, 28, 3169–3177
51. Li, H. and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform., 11, 473–483
52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410
53. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402
54. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851–1858
55. Smith, A. D., Xuan, Z. and Zhang, M. Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics, 9, 128
56. Smith, A. D., Chung, W. Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z. and Zhang, M. Q. (2009) Updates to the RMAP short-read mapping software. Bioinformatics, 25, 2841–2842
57. Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. and Li, M. (2008) ZOOM! Zillions of oligos mapped. Bioinformatics, 24, 2431–2437
58. Jiang, H. and Wong, W. H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics, 24, 2395–2396
59. Jokinen, P. and Ukkonen, E. (1991) Two algorithms for approxmate string matching in static texts. Mathematical Foundations of Computer Science 1991. Lect. Notes Comput. Sci., 520, 240–248
60. Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol., 5, e1000386
61. Weese, D., Emde, A. K., Rausch, T., Döring, A. and Reinert, K. (2009) RazerS—fast read mapping with sensitivity control. Genome Res., 19, 1646–1654
62. Weese, D., Holtgrewe, M. and Reinert, K. (2012) RazerS 3: faster, fully sensitive read mapping. Bioinformatics, 28, 2592–2599
63. Farrar, M. (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics, 23, 156–161
64. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S. L. (2004) Versatile and open software for comparing large genomes. Genome Biol., 5, R12
65. Abouelhoda, M. I., Kurtz, S. and Ohlebusch, E. (2004) Replacing suffix trees with enhanced suffix arrays. J. Discrete Algorithms, 2, 53–86
66. Ferragina, P. andManzini, G., (2000) Opportunistic data structures with applications.In Proceedings, 41st Annual Symposium, 390–398
67. Burrows, M. and Wheeler, D. J. (1994) A block-sorting lossless data compression algorithm. Systems Research Center, 124
68. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. and Hackermüller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput. Biol., 5, e1000502
69. Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323
70. Malhis, N., Butterfield, Y. S., Ester, M. and Jones, S. J. (2009) Slider—maximum use of probability information for alignment of short sequence reads and SNP detection. Bioinformatics, 25, 6–13
71. Malhis, N. and Jones, S. J. M. (2010) High quality SNP calling using Illumina data at shallow coverage. Bioinformatics, 26, 1029–1035
72. Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol., 31, 46–53
73. Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L. and Leek, J. T. (2015) Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat. Biotechnol., 33, 243–246
74. Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J. and Taylor, J. M. (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 13, 484
75. Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R., *et al.* (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One, 9, e103207
76. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res., 18, 1509–1517
77. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. and Hackermüller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput. Biol., 5, e1000502
78. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. and Konstantinidis, K. T. (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One, 7, e30087

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

79. Mamedov, T. G., Pienaar, E., Whitney, S. E., TerMaat, J. R., Carvill, G., Goliath, R., Subramanian, A. and Viljoen, H. J. (2008) A fundamental study of the PCR amplification of GC-rich DNA templates. Comput. Biol. Chem., 32, 452–457
80. Oshlack, A., Robinson, M. D. and Young, M. D. (2010) From RNA-seq reads to differential expression results. Genome Biol., 11, 220
81. Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res., 38, e131
82. McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J. and Nuzhdin, S. V. (2011) RNA-seq: technical variability and sampling. BMC Genomics, 12, 293
83. Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics, 11, 94
84. Robinson, M. D. and Smyth, G. K. (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23, 2881–2887
85. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 320, 1344–1349
86. Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91
87. Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W. and Van Wieringen, W. N. (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics, 14, 113–128
88. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol., 14, R95
89. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc., 7, 562–578
90. Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics, 13, 523–538
91. Seyednasrollah, F., Laiho, A. and Elo, L. L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. Brief. Bioinform., 16, 59–70
92. Liu, Y., Zhou, J. and White, K. P. (2014) RNA-seq differential expression studies: more sequence or more replication? Bioinformatics, 30, 301–304
93. Cho, H., Davis, J., Li, X., Smith, K. S., Battle, A. and Montgomery, S. B. (2014) High-resolution transcriptome analysis with long-read RNA sequencing. PLoS One, 9, e108095
94. Zavodna, M., Bagshaw, A., Brauning, R. and Gemmell, N. J. (2014) The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. PLoS One, 9, e113862
95. Minoche, A. E., Dohm, J. C., Schneider, J., Holtgräwe, D., Viehöver, P., Montfort, M., Sörensen, T. R., Weisshaar, B. and Himmelbauer, H. (2015) Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol., 16, 184
96. Westbrook, C. J., Karl, J. A., Wiseman, R. W., Mate, S., Koroleva, G., Garcia, K., Sanchez-Lockhart, M., O'Connor, D. H. and Palacios, G. (2015) No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. Hum. Immunol., 76, 891–896
97. Gao, Q., Sun, W., Ballegeer, M., Libert, C. and Chen, W. (2015) Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing. Mol. Syst. Biol., 11, 816
98. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376–380
99. Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science, 318, 420–426
100. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature, 452, 872–876
101. Droege, M. and Hill, B. (2008) The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol., 136, 3–10
102. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. Science, 323, 133–138
103. Uemura, S., Aitken, C. E., Korlach, J., Flusberg, B. A., Turner, S. W. and Puglisi, J. D. (2010) Real-time tRNA transit on single translating ribosomes at codon resolution. Nature, 464, 1012–1017
104. Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., *et al.* (2015) G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods, 12, 519–522
105. Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G. and Bayley, H. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc. Natl. Acad. Sci. USA, 106, 7702–7707
106. Olasagasti, F., Lieberman, K. R., Benner, S., Cherf, G. M., Dahl, J. M., Deamer, D. W. and Akeson, M. (2010) Replication of individual DNA molecules under electronic control using a protein nanopore. Nat. Nanotechnol., 5, 798–806
107. Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K. and Studholme, D. J. (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif., 3, 1–8
108. Pendleton, M., Sebra, R., Pang, A. W., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods, 12, 780–786
109. Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. and Stegle, O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol., 33, 155–160
110. Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., *et al.* (2011) Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol., 29, 1120–1127
111. Levsky, J. M., Shenoy, S. M., Pezo, R. C. and Singer, R. H. (2002)

 © Higher Education Press and Springer-Verlag Berlin Heidelberg 2016

Single-cell gene expression profiling. Science, 297, 836–840

112. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. and Tyagi, S. (2008) Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods, 5, 877–879
113. Taniguchi, Y., Choi, P. J., Li, G. W., Chen, H., Babu, M., Hearn, J., Emili, A. and Xie, X. S. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science, 329, 533–538
114. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods, 6, 377–382
115. Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012) CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports, 2, 666–673
116. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G. and Sandberg, R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods, 10, 1096–1098
117. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell, 161, 1187–1201
118. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161, 1202–1214
119. Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol., 32, 1053–1058
120. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel P., Li, S. Morse, M., Lennon, N. J., Livak K. J., Mikkelsen, T. S., Rinn, J. L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol, 32, 381–386

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2016