

# Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis

Francesca Finotello and Barbara Di Camillo

Advance Access publication date 18 September 2014

## Abstract

RNA-seq is a methodology for RNA profiling based on next-generation sequencing that enables to measure and compare gene expression patterns at unprecedented resolution. Although the appealing features of this technique have promoted its application to a wide panel of transcriptomics studies, the fast-evolving nature of experimental protocols and computational tools challenges the definition of a unified RNA-seq analysis pipeline. In this review, focused on the study of differential gene expression with RNA-seq, we go through the main steps of data processing and discuss open challenges and possible solutions.

**Keywords:** RNA-seq; differential gene expression; NGS; next-generation sequencing; transcriptomics

## INTRODUCTION

In every living organism, DNA encodes the whole information needed to determine all the properties and functions of each single cell. From this blueprint, cells can dynamically access and translate specific instructions through ‘gene expression’, namely, by selectively switching on and off a particular set of genes. The information encoded in the selected genes is transcribed into RNA molecules, which in turn can be translated into proteins or can be directly used to finely control gene expression. Thus, the set of RNAs transcribed in a certain condition and time reflects the current state of a cell and can reveal pathological mechanisms underlying diseases. More interestingly, the study of differential gene expression enables the comparison of gene expression profiles from different tissues and conditions to identify genes that play a major role in the determination of the phenotype. For instance, the comparison of healthy versus diseased tissues can provide new insights over the genetic variables involved in pathology.

In recent years, RNA-seq [1], a methodology for RNA profiling based on next-generation sequencing (NGS) [2], is replacing microarrays for the study of gene expression. The sequencing framework of RNA-seq enables to investigate at high resolution all the RNAs present in a sample, characterizing their sequences and quantifying their abundances at the same time. In practice, millions of short strings, called ‘reads’, are sequenced from random positions of the input RNAs. These reads can then be computationally mapped on a reference genome to reveal a ‘transcriptional map’, where the number of reads aligned to each gene gives a measure of its level of expression.

The powerful features of RNA-seq, such as high resolution and broad dynamic range, have boosted an unprecedented progress of transcriptomics research, producing an impressive amount of data worldwide. To support this exponential growth and to deal with the different steps of data analysis, several computational tools have been developed and updated at a fast pace. Nevertheless, the analysis

Corresponding author. Francesca Finotello, Department of Information Engineering, University of Padova, Padova 35131, Italy. Tel.: +39 049 8277964; Fax: +39 049 8277699; E-mail: francesca.finotello@dei.unipd.it

**Francesca Finotello** is a postdoctoral fellow at the Department of Information Engineering, University of Padova (Padova, Italy). Her research is focused on the analysis of Next-Generation Sequencing data for the study of genetic and complex diseases and on the development of bioinformatics methods.

**Barbara Di Camillo** is an assistant professor of Bioengineering at the Department of Information Engineering, University of Padova (Padova, Italy). She works on models of biological systems and on statistical methods for high-throughput data analysis.

scheme depicted above, which might seem simple at first glance, is in fact far more complex and consists in several processing steps.

In this review, we describe the current RNA-seq analysis framework, focusing on each computational step from read preprocessing to differential expression (DE) analysis. We review some of the most promising methodologies available, along with their underlying algorithmic strategies and mathematical models, and identify the research topics that require further investigation. We believe this work can provide a broad overview of RNA-seq analysis and can guide users to define and implement their own processing pipeline. Moreover, the dissection of the most challenging aspects of data analysis can help users to select the best methods depending on the characteristics of the specific study, considering both the driving biological question and the actual data features.

## INVESTIGATING GENE EXPRESSION WITH RNA-SEQ

The transcriptome is the whole set of RNAs transcribed from the genes of a cell. As discussed above, their relative abundances reflect the level of expression of the corresponding genes, for a specific developmental stage or physiological condition. Although RNAs are not the final products of the transcription–translation process, the study of gene expression and differential gene expression can unveil important aspects about the cell states under investigation.

In past years, hybridization-based approaches such as microarrays, were the most used solutions for gene expression profiling and DE analysis, thanks to their high throughput and relatively low costs [3]. These technologies consist in an array of probes, whose sequences represent particular regions of the genes to be monitored. The sample under investigation is washed over the array, and RNAs are free to hybridize to the probes with a complementary sequence. A fluorescent is used to label the RNAs, so that image acquisition of the whole array enables the quantification of the expressed genes. Although widely used in quantitative transcriptomics, these techniques have several limitations [3, 4]:

- reliance on prior knowledge about the genome for probe design;

- possibility to monitor only some portions of the known genes and not the actual sequences of all transcribed RNAs;
- high background levels due to cross-hybridization, i.e. imperfect hybridization between quasi-complementary sequences;
- limited dynamic range due to background noise and signal saturation;
- need for normalization to compare data from different arrays.

The advent of NGS has revolutionized transcriptomics and quickly established RNA-seq as the preferred methodology for the study of gene expression [3, 5]. The standard workflow of an RNA-seq experiment is described in the following. The RNAs in the sample of interest are initially fragmented and reverse-transcribed into complementary DNAs (cDNAs). The obtained cDNAs are then amplified and subjected to NGS. In principle, all NGS technologies can be used for RNA-seq, even though the Illumina sequencer (<http://www.illumina.com>) is now the most commonly used solution [6]. The millions of short reads generated can then be mapped on a reference genome and the number of reads aligned to each gene, called ‘counts’, gives a digital measure of gene expression levels in the sample under investigation.

Although RNA-Seq is still under active development, it is now widely used in place of microarrays to measure and compare gene transcription levels because it offers several key advantages over hybridization-based technologies [3–5, 7–9], such as:

- reconstruction of known and novel transcripts at single-base level;
- broad dynamic range, not limited by signal saturation;
- high levels of reproducibility.

The flexibility enabled by single-base resolution probably represents the most powerful feature, as it allows the quantification and sequencing of all the transcripts present in a sample. Compared with microarrays, that can only assay portions of transcripts corresponding to probes, RNA-seq leverages on the sequencing framework to overcome the pure quantification task, enabling new applications, such as transcriptome profiling of non-model organisms [10, 11], novel transcripts discovery [12],

investigation of RNA editing [13, 14] and quantification of allele-specific gene expression [15].

Despite all these newsworthy features and apparently easy scheme of data analysis, RNA-seq studies produce large and complex data sets, whose interpretation is not straightforward [16, 17]. Data analysis is further challenged by technical issues inherent to the specific NGS technology, such as sequencing errors in the output reads due to miscalled bases [2], or to biases introduced by the different steps of the RNA-seq protocol, such as amplification, fragmentation and reverse-transcription [18–20]. In particular, protocol-specific bias may under- or over-represent specific loci leading to biased results, thus necessitating careful data quality control and normalization. The latter issue is described in details in the ‘Count bias and normalization’ section. Nevertheless, if a well-annotated reference genome or transcriptome is available and if the aim of an RNA-seq study is the detection of DE genes, a basic data processing pipeline consists in the following steps: (i) read mapping, (ii) counts computation, (iii) counts normalization and (iv) detection of differentially expressed genes (Figure 1). More sophisticated pipelines can be tailored on the specific need by considering the addition of pre- and post-processing modules to be used before and after read mapping.

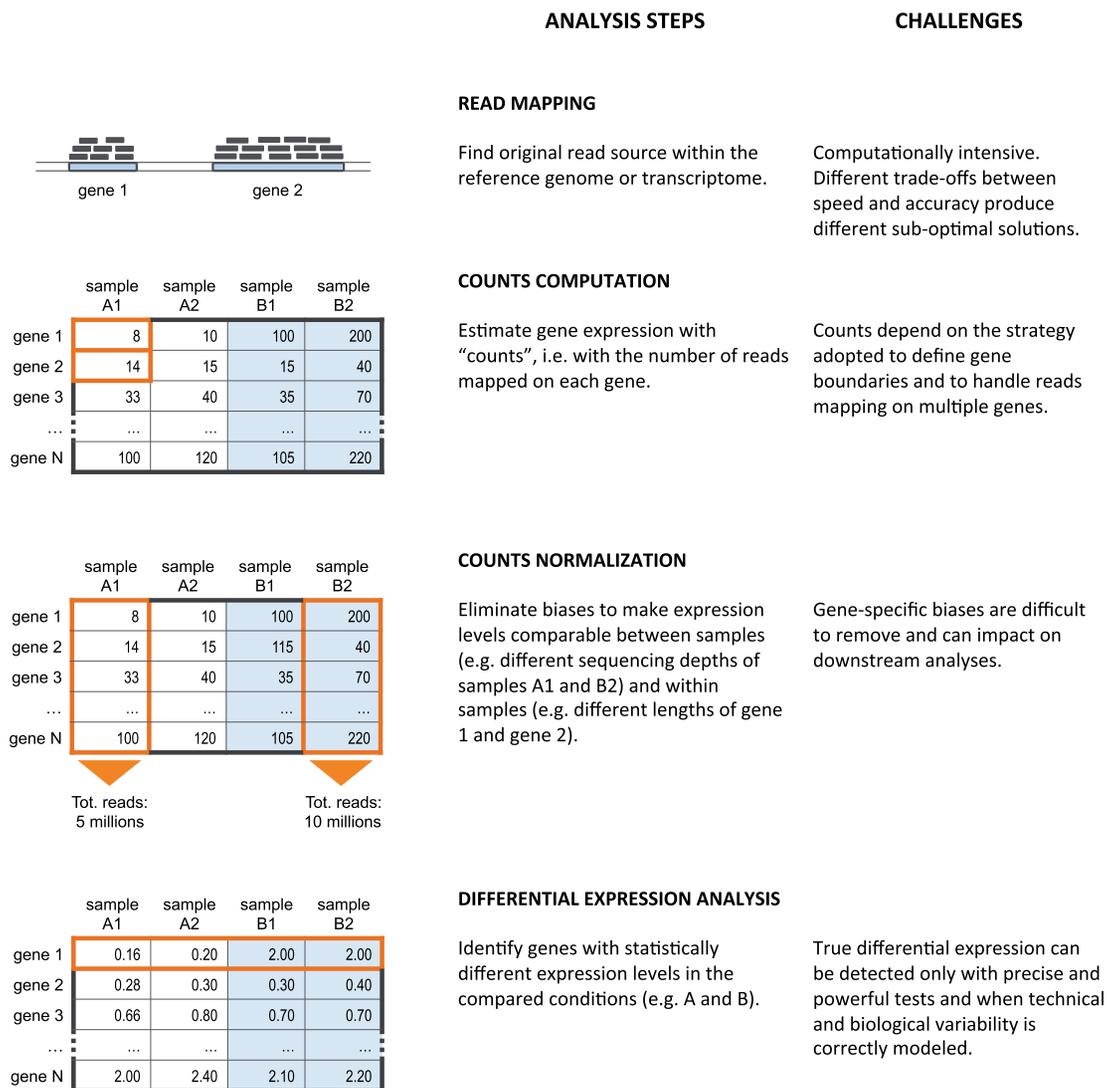
## ALGORITHMS FOR READ MAPPING

The first computational step of the RNA-seq data analysis pipeline is read mapping: reads are aligned to a reference genome or transcriptome by identifying gene regions that match read sequences. So far, many alignment tools have been proposed [21, 22]. In all cases, the mapping process starts by building an index of either the reference genome or the reads, which is then used to quickly retrieve the set of positions in the reference sequence where the reads are more likely to align. Once this subset of possible mapping locations has been identified, alignment is performed in these candidate regions with slower and more sensitive algorithms [21, 23]. The available mapping tools can be divided into two main categories based on the methodology used to build the index: hash tables or Burrows–Wheeler transform (BWT) (reviewed in [24]).

The hash table is a common data structure for indexing complex data sets so to facilitate rapid string searching. Mapping tools can build hash

tables either on the set of input reads or on the reference, considering all subsequences of a certain length  $k$  ( $k$ -mers) contained in the considered sequences. In the hash table, the key of each entry is a  $k$ -mer, while the value is the list of all positions in the reference where the  $k$ -mer was found. The two solutions have different advantages and drawbacks [21, 23]. For instance, building hash tables of the reference requires constant memory, for a given reference and parameter set, regardless of the size of the input read data. Conversely, building hash tables of reads typically requires variable but smaller memory footprint, depending on the number and complexity of the read set. However, this latter solution may require longer processing time to scan the entire reference sequence when searching for hits, even if the input read set is small, and is not suited for parallelization [21].

BWT [25] is a reversible string rearrangement that encodes the genome into a more compact representation, leveraging on redundancy of repeated subsequences. Methods based on BWT create an index of the BWT, called ‘FM-index’, that can be used to perform fast string searching in a reduced domain of available subsequences, without scanning the whole genome [26]. The combination of BWT and FM-index ensures both limited memory and space occupancy, but requires longer computational time for index construction than hash-based methods. However, since the index has to be constructed only once for a given reference and precomputed indexes for several model genomes are already available, this aspect has minimum impact on the total computational time. Conversely, the strategy used to extend the first partial high-quality hits identified thanks to hash- or BWT-based indexes into full-read alignments has a major impact on algorithm performance. Usually, hash-based algorithms implement a ‘seed-and-extend’ approach [24] leveraging on a bounded version of the Smith–Waterman (SW) algorithm [27]. BWT-based solutions sample substrings of the reference using the FM-index and then accommodate inexact matches by tolerating some mismatches, up to a certain threshold [27]. BWT implementations, which were developed for short (<50 nt) read alignment, impose very stringent constraints on inexact matches, which make them much faster than hash-based approaches, but less sensitive [21, 28]. As NGS technologies are producing increasingly longer reads (>100 nt), mapping tools are implementing hybrid solutions, which exploit the



**Figure 1:** DE analysis from RNA-seq data: computational steps and main methodological challenges.

efficiency of BWT and ‘FM-index’ for seeding and then perform alignment extension with SW-like algorithms [29–33].

Besides the specific indexing and string-matching approach implemented, differences in mapping solutions can be due to algorithmic strategies and heuristics specifically implemented to map reads:

- having no perfect matches with the reference;
- obtained with paired-end sequencing;
- generated from exon–exon junctions.

Owing to the presence of sequencing errors in NGS data, mapping algorithms must allow imperfect alignments. By tolerating a certain number of

mismatches, they are able to increase the percentage of mapped reads [21]. The available tools implement very different mismatch policies but, in general, allow the user specifying customized parameter settings. However, the mismatch policy strongly impacts on both mapping accuracy and computational performance, and the definition of its best configuration is not trivial [21, 22].

Besides systematic errors, the sequenced organism can present true single-nucleotide polymorphisms (SNPs), which result in nucleotidic differences between the reads and the reference. The flexibility/stringency given by the mismatch policy adopted is thus important to correctly map these reads, as reads having one or more SNPs have a lower probability of being mapped [28].

In addition to SNPs, reads can contain small insertions or deletions (indels). Algorithms that do not perform gapped alignment often fail to align reads containing indels [34]. Early NGS read mapping tools avoided or limited gaps in the alignment because of the computational complexity of choosing an indel location, but more recent software versions accommodate gapped alignment. Algorithms that do not perform gapped alignment have lower mapping accuracy, resulting in a significant reduction of the number of correctly mapped reads in correspondence of regions surrounding indels [35].

Another difference comes from the ability to map paired-end reads. Unlike conventional single-end sequencing, which can read only one end of each DNA fragment, paired-end protocols enable to sequence both ends, generating two reads per fragment. The information about the expected distance of the reads sequenced from these two ends, estimated from the distribution of fragment lengths, can be exploited to increase mapping or assembly accuracy. Paired-end reads are particularly useful to solve repeats, as they can cover long genomic regions (up to 20 kb), possibly extending into univocally determined sequences flanking the repeated ones. Moreover, they can be particularly useful for the identification of alternatively spliced isoforms and for the detection of fusion transcripts in cancer samples [36]. However, if an RNA-seq study is more focused on the quantification of (differential) gene expression than on the reconstruction of the exact transcript sequences, particular attention must be paid to ensure that the strategy for paired-end reads mapping is not too stringent, as it may result in a reduced number of mapped reads [21] and possibly in biased expression estimates.

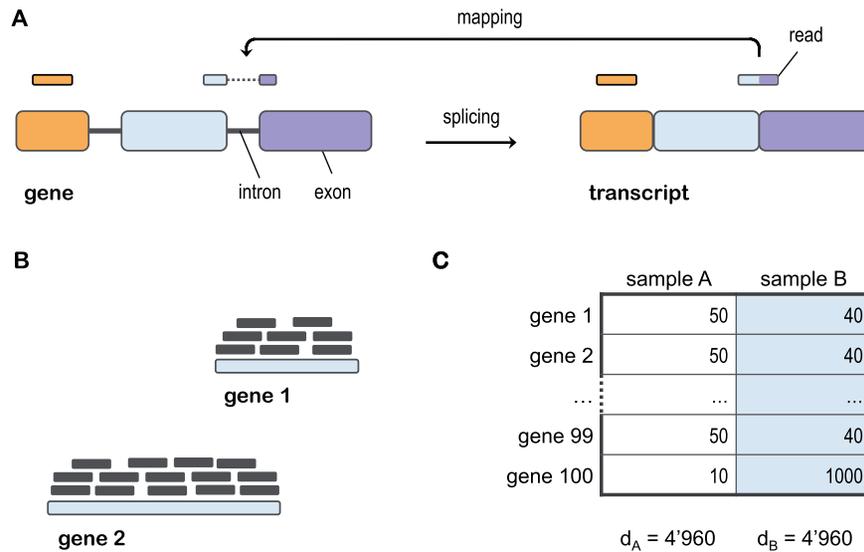
Unlike tools for genome-sequencing data mapping, algorithms developed for RNA-seq may have to handle ‘spliced reads’. Splicing is a post-transcriptional modification underwent by most of RNAs transcribed in eukaryotic organisms. During splicing, non-coding regions (introns) are removed and coding sequences (exons) are concatenated together. Although the order of exons is always preserved, some exons can be removed along with introns, giving rise to different RNAs. This process, called ‘alternative splicing’, enables to produce different protein isoforms starting from the same gene. Thus, RNAs in eukaryotes can give rise to spliced reads that span exon–exon junctions and

that cannot be directly mapped onto the genome, where exons are separated by introns. To map these spliced reads back to the genome, algorithms for RNA-seq data analysis must handle spliced alignment (Figure 2A). Generally, simple gapped alignment is not sufficient to account for introns because they can span a wide range of lengths [34]. To align spliced reads, many tools implement a two-step procedure: first, reads are mapped to the genome and used to identify putative exons; then, candidate exons are used to build all possible exon–exon junctions, which are considered for mapping the spliced reads that failed to map in the first step (e.g. [37, 38]).

Despite attempted in several works, the assessment and comparison of mapping algorithms, especially for RNA-seq reads, is not straightforward [17, 22, 39]. Ideally, the perfect algorithm would find, for each read, its true genomic source. However, the presence of sequencing errors, repeats and genetic variants, greatly increases uncertainty in read mapping and even challenges the definition of ‘correct mapping’ [21]. Moreover, the different features of the input data and the possibility to greatly change the parameter settings add further variability to the results [21]. In this scenario, it is impossible to identify the best tool, but the top performers have to be selected with respect to the specific application and input data, depending on the biological question under consideration [16, 21]. For instance, aligners that are suited for transcript quantification, might not be precise enough to study SNPs or RNA editing events.

## COUNTS: THE DIGITAL MEASURE OF GENE EXPRESSION

After mapping, the reads aligned to each coding unit, such as exon, transcript or gene, are used to compute counts, so to give an estimate of its expression level. The most used approach for computing counts considers the total number of reads overlapping the exons of a gene. However, even in well-annotated organisms, a fraction of reads map outside the boundaries of known exons [40]. Thus, an alternative strategy considers the whole length of a gene, also counting reads from introns. Moreover, if correctly handled in the mapping step, spliced reads can be used to model the abundance of different splicing isoforms of a gene [41, 42]. Particular attention should be paid to genes with overlapping



**Figure 2:** Examples of RNA-seq issues. **(A)** Spliced-reads mapping. Sequenced transcripts can produce spliced reads, generated in correspondence of exon–exon junctions. To be correctly mapped on the genome, where exons are separated by introns, spliced reads must be broken into shorter strings. **(B)** Length bias. Longer genes are more likely to generate more reads than shorter ones with similar expression levels. **(C)** Differences in library size composition. Example of a count data set where samples A and B have the same number of reads ( $d_B = d_A$ ), but different library compositions. The first 99 genes have the same counts in each sample (40 and 50 in sample A and B, respectively). In sample B, the reads available for most of the genes are ‘consumed’ by gene 100, which has very high expression. Library sizes can be computed excluding gene 100, to reflect the real sequencing state available for most of the genes in sample B ( $d_B = 0.8d_A$ ).

sequence. The ‘Union-Intersection gene’ model considers, for each gene, the union of the exonic bases that do not overlap with the exons of other genes [9]. Htseq-count implements instead a more flexible approach, which lets the user selecting the desired model for read counting in the presence of overlapping features [43]. Unlike the methods described so far, ‘maxcounts’ approach does not compute the sum of aligned reads, but estimates the expression level of each exon or single-isoform transcript as the maximum read coverage reached along its sequence [44]. This approach can be easily used for RNA-seq studies on prokaryotes, where transcripts are not subjected to splicing, while requires further research to define transcription models in eukaryotes that can be used for combining ‘maxcounts’, computed at exon level, into a measure of gene or transcript expression. Although the final strategy has the potential to significantly change expression estimates, limited research has been carried out to assess and compare the available approaches [16].

As explained above, quantification of gene expression from RNA-seq data is typically implemented in the analysis pipeline through two

computational steps: alignment of reads to a reference genome or transcriptome, and subsequent estimation of gene and isoform abundances based on aligned reads. Unfortunately, the reads generated by the most used RNA-Seq technologies are generally much shorter than the transcripts from which they are sampled. As a consequence, in the presence of transcripts with similar sequences, it is not always possible to uniquely assign short reads to a specific gene. In particular, the human genome contains duplicated and paralogous genes with high sequence similarity, and interspersed or tandem repeats that are likely to produce similar or identical short reads [38, 45, 46]. Thus, NGS data arising from repeated regions have to be handled properly in order not to bias the results [46–48]. RNA splicing makes transcriptome reconstruction even more challenging, as it generates alternatively spliced isoforms of the same gene that share a large part of their sequence and can be hardly assigned to one specific isoform. As a consequence, a non-negligible fraction of RNA-seq reads are ‘multireads’: reads that map with comparable fidelity on multiple positions of the reference. The fraction of multireads over total mapped reads depends on transcriptome complexity

and read length, varying from 10 to >50% [38, 45]. When considering reads mapping on multiple isoforms of the same gene, this percentage exceeds 70% [45].

One of the first strategies proposed for handling gene multireads was that of simply discarding them, so to estimate gene expression considering only uniquely mapping reads [1, 49]. Owing to the likelihood of assigning multireads to the wrong genomic location, introducing further bias in the results, multireads filtering is a commonly used approach in the analysis of RNA-seq and NGS data in general [46]. However, in RNA-seq studies, where the aim is both the reconstruction of transcripts sequences and the quantification of their relative abundances, discarding multireads causes an information loss and a systematic underestimation of expression levels in correspondence of repetitive regions. An alternative strategy reduces data loss by allocating multireads considering the coverage given by uniquely mapping reads and obtains expression estimates that are in better agreement with microarrays [8]. Ji *et al.* propose a more sophisticated approach that takes also into account the mismatch profiles between the unique reads and the sequence of the genomic locations they are aligned to [50]. Their method called BM-Map, calculates the posterior probability of mapping each multiread to a genomic location considering three sources of information: the sequencing error profile, the likelihood of true polymorphisms and the expression level of competing genomic locations. Conversely, the ‘proportional’ method described above only considers the latter information. The mismatch profile is also taken into consideration by MMSEQ [45], which estimates both isoform expression and allelic imbalance, namely expression differences between different alleles of the same gene or isoform. A two-step alignment procedure is used to reduce the uncertainty in read mapping. First, mismatch profiles are used to build a sample-specific transcriptome whose genotype can be different from that of the reference sequence. Then, once the reference transcriptome is updated considering the genotype, reads are realigned to estimate isoform expressions and allelic imbalance. More recent methods, such as RSEM, define a probabilistic model of RNA-Seq data and calculate maximum likelihood estimates of isoform expression levels using the Expectation-Maximization algorithm [38, 51]. True mappings are identified leveraging on the

information provided by the distribution of sequencing errors, fragment lengths and read coverage across transcripts, modeled as random variables and estimated from the data.

## COUNT BIAS AND NORMALIZATION

After the first optimistic expectation of a relative ease of analysis of RNA-seq data [3], many works have highlighted the need for a careful normalization of count data before assessing differential gene expression [9, 52–56] to correct for different sources of bias.

The first bias to be taken into account is the ‘sequencing depth’ of a sample, defined as the total number of sequenced or mapped reads. Let  $A$  and  $B$  being two RNA-seq experiments with no differentially expressed genes. If experiment  $A$  generates twice as much reads as experiment  $B$ , it is likely that the counts from experiment  $A$  will be doubled too. Hence, a common practice is that of scaling counts in each experiment  $j$  by the sequencing depth  $d_j$  estimated for that sample. In early works  $d_j$  was computed by counting the total number of reads sequenced or mapped in sample  $j$  (global scaling) [8, 49]. More recent approaches consider counts depending on the whole RNA population of the sequenced sample [57–59]. For instance, if there is a set of highly expressed genes in a sample, it will inevitably ‘consume’ the available reads, so that the expression level of the remaining genes will be underestimated [58]. A similar issue may result from the presence of contaminants. When a restricted set of highly-expressed genes accounts for the largest part of total counts, as happens in most of RNA-seq assays, global scaling techniques only capture and correct for differences related to these high-count genes (Figure 2B) [9, 59]. Bullard *et al.* propose a quantile normalization similar to that used for microarray preprocessing [60] and an alternative global scaling that adjusts counts distributions with respect to their third quartile, so to reduce the effect of high-counts genes [9]. More generally, slightly different normalizations can be defined by selecting different count quantiles [61]. Robinson and Oshlack *et al.* [58] propose the ‘Trimmed Mean of M-values’ (TMM) normalization to account for differences in library composition between samples. To reduce bias due to high-count genes, TMM is computed removing the

30% of genes that are characterized by the most extreme ‘M-values’ (i.e. log-fold-changes) for the compared samples. This normalization factor is then used to correct for differences in library sizes. Li *et al.* [59] propose a novel normalization method that assumes a Poisson model of counts and estimates the sequencing depth on a set of genes that are not differentially expressed. A Poisson goodness-of-fit statistic is used to determine which genes belong to this restricted set. In the R package ‘DESeq’ [62], the ratios between gene-wise counts in each sample  $j$  and the geometric mean of gene-wise counts across all samples are calculated, and the library size is computed as the median of these ratios across genes. Different studies (e.g., [44, 61]) indicated TMM and ‘DESeq’ methods as the most effective approaches for library size normalization. However, if the common assumption that the compared samples contain similar amount of RNA does not hold, count normalization methods are ineffective, and calibration techniques leveraging on spike-in RNA measurements can be used [63].

RNA-seq counts also show a gene length bias: the expected number of reads mapped on a gene is proportional to both the abundance and length of the isoforms transcribed from that gene. Indeed, longer genes produce more reads than shorter ones (Figure 2C), resulting in higher power for DE detection [9, 16, 64]. To reduce this bias, Mortazavi *et al.* [8] propose to summarize mapped reads as ‘Reads Per Kilobase of exon model per Million mapped reads’ (RPKM), computed dividing the number of reads aligned to gene exons by the total number of mapped reads and by the sum of exonic bases. An analogous measure is given by ‘Fragments Per Kilobase of exon per Million fragments mapped’ (FPKM) [41], which account also for paired-end data and estimate transcript abundances in terms of expected number of fragments, from which single-end and paired-end reads arise in a RNA-seq experiment. RPKM and FPKM are defined so to reduce both differences in library size and length bias. Other methods estimate and correct the dependence of counts on gene length and other sequence-specific covariates, such as GC-content and dinucleotide composition, using quantile regression [52, 65] and generalized linear models [66]. In DE analysis, as methods that correct counts for length bias can introduce additional biases [9, 44, 56, 61], normalizations that apply to DE test statistics while leaving gene counts unchanged have been

proposed [9, 64]. Differently from all the above described methods, ‘maxcounts’, which do not count the reads along exons or transcripts but select the best represented regions in terms of coverage, strongly reduce length bias before normalization [44].

Gene-specific covariates do not suffice to explain counts variability [44], which has been shown to vary greatly along gene sequences. The uneven distribution of reads along gene and transcript sequences are due to mapping errors and, primarily, to experimental biases. For instance, fragmentation methods based on restriction enzymes present sequence-specific efficiency [19]. Moreover, reverse-transcription can either over- or under-represent 3’ end of transcripts if performed with poly-dT oligomers or random hexamers, respectively [1, 3, 19]. More generally, RNAs and cDNAs can form secondary structures that depend on their primary sequences and that can either hamper or facilitate the binding of reverse-transcription primers and sequencing adapters [55]. Since the first RNA-seq experiment [1], several changes in library preparations and sequencing protocols have been introduced to reduce bias (e.g. postponing reverse transcription after fragmentation), but non-uniformity of read coverage remains an issue of state-of-the-art sequencing technologies [20]. Li *et al.* [55] call this sequence-specific bias ‘sequencing preference’: different regions of the same transcript can generate different amount of reads depending on their local nucleotidic sequence, which determines their ‘sequenceability’. They model read counts as Poisson variables with variable rates along transcripts and perform an iterative Poisson linear regression to fit the data. They also use multiple additive regression trees (MART) to capture non-linear relationships between counts and local sequences. The models are fitted using the top 100 genes with the highest expression levels and used to predict the sequencing preference of the remaining genes. This approach allows explaining up to ~50% of data variance due to coverage non-uniformity and predicting sequencing preferences that can be used in quantitative analysis of RNA-seq data to improve gene expression estimates.

## DE ANALYSIS

In recent years, a fervent research has characterized the RNA-seq field and many different tools for DE

detection have been developed [67–69]. At its simplest, methods for DE detection rely on a test statistic, used to identify which genes are characterized by a statistical significant change in gene expression in the compared conditions. In principle, non-parametric methods can be used (e.g. [70, 71]). However, because of the small number of replicates typically available in RNA-Seq experiments, non-parametric methods usually do not offer enough detection power, and parametric methods are preferred [67, 72]. Each parametric method assumes a specific model to describe the underlying distribution of count data, and seeks to identify the genes whose differences between the tested conditions exceed the variability predicted by the model. The models considered and implemented in most of the analysis tools are based on the Poisson and Negative Binomial (NB) distributions. In the following, we present a statistical description of the parameterization of RNA-seq count data and a more general summary of state-of-the-art approaches for DE analysis in RNA-seq studies. However, because of the high number of tools available, here we specifically focus on few interesting and well-characterized data modeling approaches implemented in recently developed methods.

### Models of RNA-seq count data

Let  $f = 1, \dots, F$  be the set of transcripts in the sample of interest  $j$ . For each transcript  $f$  in sample  $j$ , let  $l_f$  be its length and  $\theta_{fj}$  its expression level. All the positions within  $f$  that can give rise to a read, i.e. all possible read starts, are given by  $\theta_{fj}l_f$ . Therefore, the probability that a read comes from some transcript  $f$  in sample  $j$ , can be computed, similarly to [73], as

$$\pi_{fj} = \frac{\theta_{fj}l_f}{\sum_{f=1}^F \theta_{fj}l_f} \quad (1)$$

According to [73], the sequencing process can be modeled as a simple random sampling, in which every read is sampled independently and uniformly from sample  $j$ . Under this hypothesis, the number of reads arising from transcript  $f$ , namely counts, can be modeled as a random variable  $N_{fj}$  following a binomial distribution. Indeed, read sampling can be viewed as a Bernoulli's process, a random experiment with only two possible outcomes: 'success', when the read is sequenced from transcript  $f$ , and 'failure', when the read is sequenced from another transcript. If  $R_j$  is the number of reads sequenced in sample  $j$ ,

the random variable giving the number of successful events in  $R_j$  independent trials is given by the binomial distribution where the 'success' event has probability  $\pi_{fj}$  and the 'failure' event has probability  $1 - \pi_{fj}$ , that is

$$N_{fj} \sim \mathcal{B}(R_j, \pi_{fj}) \quad (2)$$

As  $R_j \geq 10^6$  and  $\pi_{fj} \ll 1$ , this distribution can be approximated by a Poisson distribution with parameter  $\lambda_{fj} = R_j \cdot \pi_{fj}$ :

$$N_{fj} \sim \mathcal{P}(\lambda_{fj}) \quad (3)$$

The  $\lambda_{fj}$  parameter of the Poisson model corresponds to both the mean  $\mu_f$  and the variance of the distribution. It has been demonstrated that the Poisson distribution captures the variability between RNA-Seq technical replicates sequenced in different lanes or flow-cells [9, 49, 74, 75]. In this case, we can assume  $\lambda_{fj} = \lambda_f$  for all  $j$ , considering that  $j = 1, \dots, J$  are technical replicates of the same sample. However, in the presence of biological replicates, i.e. when  $j = 1, \dots, J$  represents different biological samples belonging to the same experimental condition (e.g. different cell cultures), the expression of transcript  $f$  is not the same across different biological replicates and the resulting  $\lambda_{fj}$  is a random variable, with mean  $\mu_f$  and variance  $var(\lambda_{fj})$ . Thus, RNA-seq counts are affected by two sources of variation:

- 'Technical variation', due to the measurement error due to the adopted technology.
- 'Biological variation', representing the variability among samples belonging to the same treatment group or condition.

As a consequence, for biological replicates the variance is larger than the mean, and count data are said to be 'over-dispersed' [74–76]. In this case, the Poisson distribution cannot handle this additional variability, and models based on the NB distribution are preferred [62, 74–76]. If  $\lambda_{fj}$  is modeled with a Gamma distribution, the marginal probability distribution of counts is Negative Binomial, with mean  $\mu_f$  and variance that depends on the chosen parametrization of  $var(\lambda_{fj})$  [76].

If  $var(\lambda_{fj}) = \phi \mu_f^2$ , then

$$var(N_{fj}) = \mu_f(1 + \phi \mu_f) \quad (4)$$

More generally, if  $var(\lambda_{fj}) = \phi \mu_f^\alpha$ , then

$$var(N_{fj}) = \mu_f(1 + \phi \mu_f^{\alpha-1}). \quad (5)$$

The most used NB-based model of RNA-seq counts is that of Equation (4), with two parameters  $\phi$  and  $\mu_f$ :

$$N_{fj} \sim \mathcal{NB}(\mu_f, \phi). \quad (6)$$

The ‘overdispersion’ parameter  $\phi$  of Equations (4) and (5) accounts for the variance that is not explained by the Poisson model. When  $\phi = 0$ , the NB model reduces to the Poisson distribution. In summary, the NB distribution can be motivated as a Gamma mixture of Poisson distributions: the technical variability is Poisson, but the Poisson means differ between biological replicates according to a Gamma distribution.

### Tools for DE analysis of RNA-seq data

Given a specific statistical model of RNA-seq count data, all parametric tools for DE analysis consist in two main steps: estimation of model parameters from data and detection of DE genes with a test statistics. Library normalization can also be considered part of DE analysis [68], as it is implemented within all DE tools, despite with different approaches. So far, several studies have been focused on DE methods comparison [67–69, 72, 76–79], but a consensus on methods performance is challenged by the lack of gold-standard measures and by frequent tools updates, with several versions released each year [67, 72]. However, some findings are widely confirmed across different studies, such as the superior performance of NB-based methods over their Poisson-based counterparts [68, 74, 76–79].

The higher performance of NB-based tools is mainly because of their ability to capture biological variability. As discussed above, this variability is due to the stochastic nature of gene expression (i.e. some genes have more variable levels of expression than others), and is thus gene-specific and independent from the adopted technology [80]. Owing to the small sample-size that generally characterizes RNA-seq data sets, gene-wise estimation of  $\phi$  cannot be performed and different strategies are used to fit the data. edgeR [81] and DESeq [82], which are among the best performers in most of the comparative studies cited above, are both based on the NB model of Equation (4), but implement different strategies for dispersion estimation. The default strategy implemented in edgeR shrinks gene-wise dispersion estimates toward a common value. Alternatively, edgeR can compute a ‘trend’ estimate across genes in place

of a single value. DESeq considers the variance being a smooth function of the mean  $\mu_f$  and uses non-parametric regression to fit the variance as a function of the mean. Another approach, implemented in NBPseq [76], considers the model with three parameters described by Equation (5);  $\phi$  and  $\alpha$  are considered constant across genes and estimated jointly. Nevertheless, this approach does not outperform DESeq and edgeR [72].

More recently, Law *et al.* proposed to apply ‘limma’ [83], a method developed for microarrays and based on the normal distribution, to analyze RNA-seq data [78]. The underlying idea is that correctly modeling data mean–variance relationship is more important than exactly specifying the probabilistic count distribution. In their approach, called ‘limma voom’, the mean–variance relationship is estimated from data through lowess fit and used to estimate gene-wise variances. For each gene, the inverse of the variance is then used as weight in the ‘limma’ framework. Applied to RNA-seq data, ‘limma voom’ performs comparably with top-ranking NB-based approaches [67, 78]. Even though further assessments are needed to finally select the best approach for DE analysis from RNA-seq data, the promising results obtained with this strategy may enable to exploit a wide panel of methods developed for microarrays.

### CONCLUSIONS

RNA-seq has rapidly become the method of choice for the study of differential gene expression, as it enables the investigation and comparison of gene expression levels at unprecedented resolution. However, turning huge and complex RNA-seq data sets into biologically meaningful findings is not trivial. The interpretation of RNA-seq data requires the definition of a computational pipeline that comprises several steps: read mapping, count computation, normalization and testing for differential gene expression. Here, we reviewed some of the most used methodologies and models implementing these processing steps and discussed the main challenges of data analysis. We believe this review can guide users to define an accurate analysis pipeline.

RNA sequencing is evolving at a fast pace and emerging ‘Third-Generation’ technologies now enable single-molecule sequencing [84]; computational tools themselves are chasing this development to accommodate changes in the data features, and the

assessment and comparison of state-of-the-art methods must be constantly performed to implement an updated RNA-seq computational pipeline. However, some issues raised in this work, such as the impact of read mapping heuristics and count normalization, can be considered of broad interest, and should be carefully taken into consideration for all RNA-seq data analyses, independently from data features.

### Key points

- RNA-seq is a novel methodology based on NGS that enables to investigate differential gene expression at high resolution. However, data interpretation is not straightforward and requires several analysis steps: read mapping, counts computation, counts normalization and DE testing.
- Tools for read mapping provide different solutions depending on the specific algorithm and heuristics implemented. Particular care must be taken to handle reads mapping on multiple genomic locations to estimate correct gene expression levels even in the presence of high-similarity sequences.
- In RNA-seq studies, gene expression levels are measured by counts, i.e. by the number of reads mapped on each gene.
- Counts often depend on gene- and sample-specific covariates, such as gene length and library size, respectively. Between-sample differences in library size must be necessarily corrected before comparing samples to detect differentially expressed genes. Conversely, correction of gene-specific covariates is not mandatory and must be performed carefully to avoid information loss.
- DE analysis can be tested with parametric methods based on the Poisson or NB distribution. NB models are preferred, as they capture both technical and biological variability.

### FUNDING

This work was supported by Fondazione CARIPARO ["RNA sequencing for quantitative transcriptomics" PhD program]; and PRAT 2010 [CPDA101217].

### References

1. Nagalakshmi U, Wang Z, Waern K, *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.
2. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
4. Roy NC, Altermann E, Park ZA, *et al.* A comparison of analog and next-generation transcriptomic tools for mammalian studies. *Brief Funct Genomic* 2011;**10**:135–50.
5. Shendure J. The beginning of the end for microarrays? *Nat Methods* 2008;**5**:585–7.
6. Van Verk MC, Hickman R, Pieterse CM, *et al.* RNA-Seq: revelation of the messengers. *Trends Plant Sci* 2013;**18**:175–9.
7. Cloonan N, Forrest AR, Kolle G, *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 2008;**5**:613–19.
8. Mortazavi A, Williams BA, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
9. Bullard J, Purdom E, Hansen K, *et al.* Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010;**11**:94.
10. Crawford JE, Guelbeogo WM, Sanou A, *et al.* De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. *PLoS One* 2010;**5**:e14202.
11. Vera JC, Wheat CW, Fescemyer HW, *et al.* Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* 2008;**17**:1636–47.
12. Roberts A, Pimentel H, Trapnell C, *et al.* Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 2011;**27**:2325–9.
13. Peng Z, Cheng Y, Tan BC, *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* 2012;**30**:253–60.
14. Bahn JH, Lee J, Li G, *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* 2012;**22**:142–50.
15. Rozowsky J, Abyzov A, Wang J, *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 2011;**7**:522.
16. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol* 2010;**11**:220.
17. Korf I. Genomics: the state of the art in RNA-seq analysis. *Nat Methods* 2013;**10**:1165–6.
18. Aird D, Ross MG, Chen W, *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011;**12**:R18.
19. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 2010;**38**:e131.
20. Griebel T, Zacher B, Ribeca P, *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res* 2012;**40**:10073–83.
21. Hatem A, Bozda D, Toland AE, *et al.* Benchmarking short sequence mapping tools. *BMC Bioinformatics* 2013;**14**:184.
22. Engström PG, Steijger T, Sipos B, *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 2013;**10**:1185–91.
23. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods* 2009;**6**(Suppl. 11):S6–S12.
24. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinformatics* 2010;**11**:473–83.
25. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical report, 1994.
26. Ferragina P, Manzini G. Opportunistic data structures with applications. In: *Proceedings of 41st Annual*

- Symposium on Foundations of Computer Science*, IEEE Computer Society, 2000. pp. 390–8.
27. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
  28. Garber M, Grabherr MG, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* 2011;**8**:469–77.
  29. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 2010;**26**:589–95.
  30. Kielbasa SM, Wan R, Sato K, et al. Adaptive seeds tame genomic sequence comparison. *Genome Res* 2011;**21**:487–93.
  31. Lam TW, Sung W, Tam S, et al. Compressed indexing and local alignment of DNA. *Bioinformatics* 2008;**24**:791–7.
  32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
  33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint* 2013. arXiv:1303.3997v2.
  34. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**:R36.
  35. Grant GR, Farkas MH, Pizarro AD, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* 2011;**27**:2518–28.
  36. Mardis ER, Wilson RK. Cancer genome sequencing: a review. *Hum Mol Genet* 2009;**18**:R163–8.
  37. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**:1105–11.
  38. Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323.
  39. Marx V. The author file: Paul Bertone. *Nat Methods* 2013;**10**:1137.
  40. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**:768–72.
  41. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**:511–5.
  42. Gatto A, Torroja-Fungairiño C, Mazarotto F, et al. Fine-Splice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucleic Acids Res* 2014;**42**:e71.
  43. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *bioRxiv preprint* 2014, doi:10.1101/002824. <http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html>.
  44. Finotello F, Lavezzo E, Bianco L, et al. Reducing bias in RNA sequencing data: a novel approach to compute counts. *BMC Bioinformatics* 2014;**15**(Suppl. 1):S7.
  45. Turro E, Su S, Goncalves A, et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* 2011;**12**:R13.
  46. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2011;**13**:36–46.
  47. Chu H, Hsiao WW, Tsao TT, et al. SeqEntropy: genome-wide assessment of repeats for short read sequencing. *PLoS One* 2013;**8**:e59484.
  48. Finotello F, Lavezzo E, Fontana P, et al. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Brief Bioinformatics* 2012;**13**:269–80.
  49. Marioni JC, Mason CE, Mane SM, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**:1509–17.
  50. Ji Y, Xu Y, Zhang Q, et al. BM-Map: bayesian mapping of multireads for next-generation sequencing data. *Biometrics* 2011;**67**:1215–24.
  51. Nicolae M, Mangul S, Mandoiu II, et al. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol Biol* 2011;**6**:9.
  52. Risso D, Schwartz K, Sherlock G, et al. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 2011;**12**:480.
  53. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;**40**:e72.
  54. Dohm JC, Lottaz C, Borodina T, et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008;**36**:e105.
  55. Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* 2010;**11**:R25.
  56. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009;**4**:14.
  57. Lin Y, Li J, Shen H, et al. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 2011;**27**:2031–7.
  58. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**:R25.
  59. Li J, Witten DM, Johnstone IM, et al. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 2012;**13**:523–38.
  60. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**:249–64.
  61. Dillies M, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformatics* 2013;**14**:671–83.
  62. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
  63. Lovén J, Orlando DA, Sigova AA, et al. Revisiting global gene expression analysis. *Cell* 2012;**151**:476.
  64. Gao L, Fang Z, Zhang K, et al. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* 2011;**27**:662–9.
  65. Hansen KD, Irizarry RA, Zhijin W. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 2012;**13**:204–16.
  66. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-sequencing data. *BMC Bioinformatics* 2011;**12**:290.
  67. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinformatics* 2015;**16**:59–70.

68. Rapaport F, Khanin R, Liang Y, *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013;**14**:R95.
69. Guo Y, Li C, Ye F, *et al.* Evaluation of read count based RNAseq analysis methods. *BMC Genomics* 2013;**14**(Suppl. 8):S2.
70. Tarazona S, García-Alcalde F, Dopazo J, *et al.* Differential expression in RNA-seq: a matter of depth. *Genome Res* 2011;**21**:2213–23.
71. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* 2013;**22**:519–36.
72. Robles JA, Qureshi SE, Stephen SJ, *et al.* Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 2012;**13**:484.
73. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009;**25**:1026–32.
74. Oberg AL, Bot BM, Grill DE, *et al.* Technical and biological variance structure in mRNA-Seq data: life in the real world. *BMC Genomics* 2012;**13**:304.
75. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 2013;**14**:232–43.
76. Di Y, Schafer DW, Cumbie JS, *et al.* The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Stat Appl Genet Mol Biol* 2011;**10**:24.
77. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot* 2012;**99**:248–56.
78. Law C, Chen Y, Shi W, *et al.* Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.
79. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 2010;**11**:422.
80. Hansen KD, Wu Z, Irizarry RA, *et al.* Sequencing technology does not eliminate biological variability. *Nat Biotechnol* 2011;**29**:572–3.
81. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
82. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res* 2012;**22**:2008–17.
83. Smyth GK. Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer, 2005:397–420.
84. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**:R227–40.